IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 17, 2022, accepted March 11, 2022, date of publication March 22, 2022, date of current version March 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161481

Unsupervised Outlier Detection for Mixed-Valued Dataset Based on the Adaptive k-Nearest Neighbor Global Network

YU WANG, XUEJING CAO, AND YUPENG LI
Department of Industrial Engineering, School of Mines, China University of Mining and Technology, Xuzhou 221116, China

Corresponding author: Yupeng Li (ypeng_li@163.com)

This work was supported in part by the National Natural Science Foundation, China, under Grant 51505480 and Grant 72001203; and in part by the Graduate Research and Innovation Projects of Jiangsu Province, China, under Grant KYCX21_2477.

ABSTRACT Outlier detection aims to reveal data patterns different from existing data. Benefit from its good robustness and interpretability, the outlier detection method for numerical dataset based on k-Nearest Neighbor (k-NN) network has attracted much attention in recent years. However, the datasets produced in many practical contexts tend to contain both numerical and categorical attributes, that are, the datasets with mixed-valued attributes (DMAs). And, the selection of k is also an issue that is worthy of attention for unlabeled datasets. Therefore, an unsupervised outlier detection method for DMA based on an adaptive k-NN global network is proposed. First, an adaptive search algorithm for the appropriate value of k considering the distribution characteristics of datasets is introduced. Next, the distance between mixed-valued data objects is measured based on the Heterogeneous Euclidean-Overlap Metric, and the k-NN of a data object is obtained. Then, an adaptive k-NN global network is constructed based on the neighborhood relationships between data objects, and a customized random walk process is executed on it to detect outliers by using the transition probability to limit behaviors of the random walker. Finally, the effectiveness, accuracy, and applicability of the proposed method are demonstrated by a detailed experiment.

INDEX TERMS Unsupervised outlier detection, k-nearest neighbor, mixed-valued dataset; network model, random walk process.

I. INTRODUCTION

As an important task in data mining, the purpose of outlier detection is to reveal data patterns different from existing data [1]. An outlier can be defined as "an observation which deviates so much from other observations as to arouse suspicions that is generated by a different mechanism [2]". Another definition is given by Barnett and Lewis, "an outlier is an observation (or a set of observations) which appears to be inconsistence with the remainder of the given dataset [3]". Outliers can be anomalies, novelties, noise, deviations, and exceptions [4]. And an outlier usually represents a new perspective or a specific mechanism which attracts higher interest than the normal instances. Therefore, outlier detection has been widely used in different domains, e.g., human activity recognition [5]; credit card fraud detection [6]; medical diagnosis [7]; video detection [8] and fault diagnosis [9].

Generally, according to whether the dataset is labelled or not, outlier detection methods can be roughly divided into three categories, namely supervised methods [10], semi-supervised methods [11] and unsupervised methods [12]. Most of the datasets collected in real engineering contexts are unlabeled, and the labeling are problematic or cost unacceptable. Therefore, unsupervised outlier detection method is very popular because it does not require a labelled training dataset. In recent decades, various unsupervised outlier detection technologies have been proposed, most of which are based on the nearest neighbor of data objects [13]. There exist two kinds of the nearest neighbor concepts, i.e., ε-Nearest Neighbor and k-Nearest Neighbor (k-NN) [14], among which, the k-NN is more widely adopted. The core idea of the k-NN is to select a specific k for a dataset and find k data objects with the greatest similarity or the shortest distance from each data object in the dataset. When the

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

value of k is too large, the neighbors of each data object will contain useless information or even lead to errors in subsequent algorithms. Conversely, if the value of k is too small, the data object will have fewer neighbors and contain limited useful information, which will reduce the accuracy of the algorithm. For different datasets, the value of k is often different in order to ensure the optimal performance of the outlier detection results. Therefore, a search algorithm for the appropriate value of k is proposed to automatically determine the value of k for different distributed datasets.

Besides, the datasets often include both numerical and categorical attributes simultaneously, that are, datasets with mixed-valued attributes (DMAs) for many real-world problems. For example, some unpredictable outliers often appear in the actual operation of the warehousing system, especially in the emerging warehousing system, which increase the liabilities of the warehousing industry [15]. The abnormal operation status in a warehousing system is described as an object with mixed-valued attributes so as to realize the health diagnosis of the warehouse system [16]. More and more scholars focus on the outlier detection methods for the DMA to solve practical problems. However, most outlier detection methods based on the k-NN are developed to handle the datasets with only numerical attributes [17]–[19]. Therefore, the analysis and research of outlier detection methods for the DMA based on the k-NN is theoretically and practically significant.

Furthermore, the random walk process has been widely used for a variety of information retrieval tasks, including web search [20], keyword extraction [21], and text summarization [22]. These methods usually build a network model for data objects and perform a random walk process on it to evaluate the centrality or importance of each data object. Moonesinghe and Tan [23] applied the random walk process to outlier detection, and verified the accuracy of the detection results through real datasets and synthetic datasets. The k-NN focuses on the local information of each data object and ignores the possible internal connections in the whole dataset. The random walk process on the network model just makes up for this drawback and emphasizes the overall or partial structure of the dataset. The combination of the k-NN and a random walk process can relevantly characterize the relationships between data objects and the hermit connections in the dataset [24]. Motivated and inspired by the above observations, an unsupervised outlier detection method based on an adaptive k-NN network to handle the DMA is proposed in this study. First, considering the influence of different attributes on data objects, the heterogeneous distance function is combined to measure the spatial distance between mixed-valued data objects in a DMA. Second, an adaptive search algorithm for the appropriate value of k is proposed according to the distribution characteristics of the dataset, then the k-NN for each mixed-valued data object is obtained. Then, an adaptive k-NN global weighted and directed network model is constructed based on the neighborhood relationships between data objects, and a customized random walk process is implemented on it. After the random walk process converges to an equilibrium state, the element in the stationary distribution vector is used to construct the outlier score of each data object. Finally, the proposed method is compared with other existing related outlier detection methods on three different types of datasets from University of California Irvine (UCI) Machine Learning Repository, and the results show that the proposed method is more effective, accurate, and applicable.

The main contributions of this paper are threefold.

(1) We propose an adaptive search algorithm for the value of k based on the k-NN. It can automatically search the appropriate value of k according to the distribution characteristics of data objects in different datasets. This algorithm enables the unsupervised mechanism in the proposed outlier detection method for the DMA.

(2) We combine the k-NN with a random walk process to construct the outlier score for mixed-valued data objects. The k-NN obtains the information around each data object, and the random walk process on the network model explores the relationships between data objects from the perspective of the whole dataset. In this way, we can not only make full use of the local information of each mixed-valued data object, but also consider the outlier degree of each data object from the global perspective.

(3) We validate the effectiveness, accuracy, and applicability of the proposed methodology with three other related outlier detection methods by using three UCI datasets with different data types, which contain a dataset with numerical attributes, a dataset with categorical attributes, and a DMA. Several evaluation metrics that include precision, recall, rank power, and time consumption are employed to evaluate the performance of the methods in the experiments.

The remainder of this paper is organized as follows. The related works are presented in Section 2. The proposed methodology is detailed in Section 3. The detailed experiment on three types of UCI datasets is implemented in Section 4. The conclusions and future work are provided in Section 5.

II. RELATED WORKS

In this section, two kinds of outlier detection methods related to this study are investigated: (1) outlier detection methods based on the k-NN and (2) outlier detection methods based on random walk process, which are listed in Table 1.

A. OUTLIER DETECTION BASED ON THE k-NN

For decades, scholars have contributed a lot on outlier detection based on the k-NN. Dong and Yan [25] propose a multivariate outlier detection method based on the k-NN. Wang *et al.*. [19] give each data object a local outlier score and a global outlier score based on the k-NN medoid to measure whether a data object is an outlier. Muthukrishnan [26] introduces the Reverse Nearest Neighbor (RNN) which lays a foundation for the RNN-based outlier detection method. Uttarkabat *et al.* [13] use the statistical information of the RNN and the k-NN, define the distance factor to measure the

TABLE 1. Descriptions of the related works.

Method	Strategy	Contribution
[25]	*k*-NN, mahalanobis distance	The location and scatter of the high-dimensional data are calculated precisely and the outliers can be effectively detected and eliminated.
[19]	*k*-NN, *k*-means method	This method combines *k*-medoids clustering algorithm and *k*NN-based outlier detection.
[26]	RNN	It presents a general approach for solving RNN queries and an efficient R-tree based method for large data sets.
[13]	*k*-NN, RNN	It utilizes *k*-NN and RNN efficiently to identify outliers between sparse and dense clusters.
[27]	R*k*-NN	The update of insertion or deletion only needs one scan of the current window, which improves efficiency.
[28]	improved monarch butterfly optimization, mutual nearest neighbor	An enhanced algorithm is offered for enhanced search precision and run time efficiency by a fresh adaptation provider. The class markers of unknown data objects are defined by mutually next to one another, and pseudo near neighbors can be identified and taken not into account in the prediction process.
[29]	nearest neighbors, information entropy	It proposes a method based on the entropy to measure the observability factor of each iteration, and optimize the value of parameter *k*.
[30]	*k*-NN, mutual neighbor graph	The number of cliques (complete graphs) in the mutual neighbor graph is used to search the stable state. When it reaches the stable state, the appropriate value of *k* can be found,
[23]	random walk, Markov chain	It investigates the effectiveness of random walk approach for outlier detection.
[31]	random walk	It proposes a method to measure the outlier degree of nodes in complex networks by simultaneously considering both local and global information of each node based on distance measure of the random walk process of a Brownian particle and the dissimilarity index.
[32]	random walk	It proposes a method based on the random walk process to identify the spatial outliers. Two weighted graphs are established according to the spatial and non-spatial attributes of spatial objects, respectively, and the outlier score is defined by the random walk process.
[24]	random walk	It designs a local information graph aiming to capture the differences and interdependencies of various types of data objects. The weighted directed graph can be later utilized by a random walk process to effectively distinguish potential outlier and inlier objects.
[33]	*k*-NN, random walk	Combining the local information with the implicit connections in the graph representation of the original dataset, proposing a new outlier detection model named Virtual Outlier Score.
[34]	random walk	It proposes a novel method for outlier detection via kernel preserving embedding and random walks.

outlier degree of data objects. The generalized form of the RNN is reverse *k*-NN (R*k*-NN). Cao *et al.*. [27] propose a novel stream outlier detection method based on the R*k*-NN to avoid multi-scan of the dataset and to capture concept drift. Batchanaboyina and Devarakonda [28] propose an efficient outlier detection approach using improved monarch butterfly optimization and mutual nearest neighbors.

However, the value of k is still an important factor affecting the accuracy of outlier detection results in this kind of methods. In order to deal with the problem of parameter sensitivity, some novel strategies have been proposed. Inspired by the concept that outlying objects are less easily selected than inlying objects in blind random sampling, Ha *et al.* [29] solve the problem of the effect of parameter k. However, the algorithm still has parameters, such as iteration times and sampling size, which does not fundamentally eliminate the dependence on parameters. Ning *et al.*. [30] propose a parameter selection method based on a mutual neighbor graph, but this is at the expense of the accuracy and efficiency of identification.

In conclusion, the k-NN which can effectively express the local information around the data objects is widely used in outlier detection methods and its effectiveness is proved by a large number of studies. However, the selection of the value of parameter k is still an issue that needs to be tackled.

B. OUTLIER DETECTION BASED ON RANDOM WALK PROCESS

In order to broaden the application fields of the random walk process, Moonesinghe and Tan [23] apply it to outlier detection and propose two strategies for constructing networks, using appropriate similarity measures and the number of shared neighbors, respectively. The accuracy of the detection results is verified by real datasets and synthetic datasets. Afterwards, people have further explored the outlier detection methods based on the random walk process, e.g., [31], [32], [24], [33], and [34]. In these methods, each data object is modelled as a node in a network, and the relationship between objects is defined as an edge that connects the nodes. The nodes and edges in a network are analyzed by deeply mining the characteristics of topological structure to define the outlier score of each data object [35].

From the above, the keys of this kind of methods are how to build the network model and how to determine the index to measure the outlier degree of data objects. Scholars mostly build the network model based on the neighbourhood system of each data object, and the combination of the k-NN and the random walk process is only used to deal with the dataset with numerical attributes, and the outlier detection results are affected by the value of k. Therefore, this paper proposes an unsupervised outlier detection method which can deal with the DMA combining the k-NN and the random walk process.

III. METHODOLOGY

The proposed methodology will be discussed in detail in this section, which includes three parts: a) data preprocessing, b) constructing an adaptive k-NN global weighted and directed network model, and c) performing a random walk process to identify outliers.

A. DATA PREPROCESSING

Let $X=\{x_1, x_2, \ldots, x_n\}$ be a set of n mixed-valued data objects, and $A=\{a^1, a^2, \ldots, a^d\}$ be a set of attributes. Each data object is described by d_1 numerical attributes and d_2 categorical attributes, and $d_1 + d_2 = d$. A data object x_i $(i = 1, 2, \ldots, n)$ is represented as $x_i = [x_i^{j^{N'}}, x_i^{j^C}]$, where $x_i^{j^{N'}}$ $(j^N = 1, 2, \ldots, d_1)$ and $x_i^{j^C}$ $(j^C = 1, 2, \ldots, d_2)$ represent the value of the i^{th} data object on the j^{th} numerical attribute and the j^{th} categorical attribute, respectively.

The distribution of the original data objects on each attribute is different. The attribute has great impact on the overall deviation degree of data objects when the discrete degree of the distribution of data objects on it is large. Therefore, the entropy weight method [36] is applied to objectively weight d attributes according to the discreteness of the distribution of data objects on the attributes.

$$w^{j^N} = \frac{1 - E^{j^N}}{d - \sum_{j^N=1}^{d_1} E^{j^N} - \sum_{j^C=1}^{d_2} E^{j^C}}, \tag{1}$$

$$w^{j^C} = \frac{1 - E^{j^C}}{d - \sum_{j^N=1}^{d_1} E^{j^N} - \sum_{j^C=1}^{d_2} E^{j^C}}, \tag{2}$$

where, E^{j^N} and E^{j^C} are the information entropy of data objects on numerical attributes and categorical attributes, respectively.

$$E^{j^N} = -\frac{1}{\ln n} \sum_{i=1}^{n} \frac{x_i^{j^{N'}}}{\sum_{i=1}^{n} x_i^{j^{N'}}} \ln \frac{x_i^{j^{N'}}}{\sum_{i=1}^{n} x_i^{j^{N'}}}, \tag{3}$$

here, if $x_i^{j^{N'}} = 0$, $\frac{x_i^{j^{N'}}}{\sum_{i=1}^{n} x_i^{j^{N'}}} \ln \frac{x_i^{j^{N'}}}{\sum_{i=1}^{n} x_i^{j^{N'}}} = 0$.

The values of data objects on the categorical attributes are expressed by

$$X^{j^C} = \{x_1^{j^C}, \quad x_2^{j^C}, \ldots, x_c^{j^C}, \tag{4}$$

where, the elements in the set X^{j^C} are mutually exclusive, $x_l^{j^C}$ $(l = 1, 2, \ldots, c)$ indicates the l^{th} value of the data object on the j^{th} categorical attribute, and c indicates the number of data objects taking different values on the j^{th} categorical attribute. The appearance frequency of $x_l^{j^C}$ in X is $X/X^{j^C} = \{F_1^{j^C}, F_2^{j^C}, \ldots, F_c^{j^C}$. Information entropy of a group of data objects in the categorical attribute is calculated by:

$$E^{j^C} = -\frac{1}{\ln n} \sum_{l=1}^{c} \frac{F_l^{j^C}}{n} \ln \frac{F_l^{j^C}}{n}. \tag{5}$$

In order to eliminate the differences in dimensions and orders of magnitude of the original data objects on numerical attributes, the min-max normalization method [37] which linearly transforms the variables is used to process the numerical data objects, and the standardized ones are recorded as,

$$x_i^{j^N} = \frac{x_i^{j^{N'}} - \left\{x_i^{j^{N'}}\right\}}{\left\{x_i^{j^{N'}}\right\} - \left\{x_i^{j^{N'}}\right\}}, \tag{6}$$

where, $\left\{x_i^{j^{N'}}\right\}$ and $\left\{x_i^{j^{N'}}\right\}$ represent the maximum and minimum value of data objects regarding numerical attributes, respectively.

B. CONSTRUCTING AN ADAPTIVE k-NN GLOBAL WEIGHTED AND DIRECTED NETWORK MODEL

The k-NN of the data object x_i is a collection of data objects, that is, the distance from x_i is less than or equal to the distance from x_i to its k^{th} neighbor (represented by D_i^*). It is defined as follows:

$$k - NN(x_i) = \{x_m | (x_m \in X) \cap D_{im} \le D_i^*\} \quad (i, m = 1, 2, \ldots, n, m \neq i), \tag{7}$$

where, D_{im} is the distance between x_i and x_m. The Heterogeneous Euclidean-Overlap Metric [38] is used to calculate the distance among data objects on the mixed-valued attributes, that is

$$D_{im} = \sqrt{\sum_{j^N=1}^{d_1} w^{j^N} \left(d_{im}^{j^N}\right)^2 + \sum_{j^C=1}^{d_2} w^{j^C} \left(d_{im}^{j^C}\right)^2}, \tag{8}$$

where,

$$d_{im}^{j^N} = \left| x_i^{j^N} - x_m^{j^N} \right|, \tag{9}$$

$$d_{im}^{j^C} = \begin{cases} 0, & x_i^{j^C} = x_m^{j^C}; \\ 1, & x_i^{j^C} \neq x_m^{j^C}. \end{cases} \tag{10}$$

The value of k has a great impact on the accuracy of outlier detection results. Scholars have proposed plenty of methods to obtain it, however, most of them determine the value of k by combining parameter optimization algorithms and evaluation indexes on the premise that the outlier objects in the dataset are known. For unlabelled datasets, this kind of methods are difficult to give an appropriate k. Specially, the distribution characteristics of data objects are various for different datasets. In order to ensure the accuracy of detection results, the corresponding k needs to be determined according to the distribution characteristics of datasets. Therefore, based on the principle of the k-NN, an adaptive search algorithm is given for the appropriate value of k according to the distribution characteristics of datasets, shown as Algorithm 1. where, $\Omega_r(x_i)$ is a collection of data objects, in which the element is the data object that regards x_i as the r-NN. $N(\Omega_r(x_i))$ represents the number of the elements in $\Omega_r(x_i)$.

Then, an adaptive k-NN weighted and directed network model (shown as Fig. 1 (b)) $M = (X, \mathbf{E})$ is constructed based on the k-NN of each data object, where the elements of $X = (x_1, x_2, \ldots, x_n)$ represent the nodes (data objects) in the network model, $\mathbf{E}$ is a binary matrix and the element E_{im} in it represents that exists an edge from x_i to its neighbor x_m, that is:

$$E_{im} = \begin{cases} 1, & \text{if } x_m \in k - NN(x_i); \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The adaptive k-NN weighted and directed network of the dataset can be represented by an adjacent matrix $\mathbf{A}$ under the

Algorithm 1 Automatic Search Algorithm for the Value of k

```
Input: Dataset $X$, $r$=1, $\Omega_r(x_i) = \emptyset$, $N(\Omega_r(x_i))$ =0, $P = 0$
Output: the adaptive $k$
for $i \leftarrow 1$ to $n$
 for $m \leftarrow 1$ to $n$
  $D_{im} \leftarrow \sqrt{\sum_{jN=1}^{d_1} w^{jN}\left(d_{im}^{jN}\right)^2 + \sum_{jC=1}^{d_2} w^{jC}\left(d_{im}^{jC}\right)^2}$
  sort $x_m$ based on $D_{im}$ in ascending order for $x_i$
 end for
end for
while $P = 0$ do
for $i \leftarrow 1$ to $n$
  $\Omega_r(x_i) \leftarrow \{x_m|(m \neq i)(x_i \in r\text{-NN}(x_m))\}$
end for
foreach $x_i$ in $X$
 if exists $N(\Omega_r(x_i)) = 0$
   $P \leftarrow 0$
 else
   $P \leftarrow 1$
 end if
end
$r = r + 1$
end while
$k = r - 1$
return $k$
```

rule that: if $A_{im} > 0$, then there will be a directed edge from x_i to x_m, and the weight on this edge is A_{im}. The elements in matrix $\mathbf{A}$ are defined as follows:

$$A_{im} = \begin{cases} 1 - D_{im}, & \textit{if } E_{im} = 1; \\ 0, & \textit{if } E_{im} = 0. \end{cases} \tag{12}$$

In this network, it is assumed that there is an edge between one node and its k-NN, which is from the node to its k-NN, and the weight of the connecting edge is defined by Eq. (12). As shown in Fig. 1 (b), it is found that there are two sub networks. With the different distribution of datasets, there may be multiple sub networks, or isolated nodes, etc. Therefore, a node x_G is introduced, and it is assumed that there are bidirectional edges between x_G and x_i $(i = 1, 2, \ldots, n)$, so as to form an adaptive k-NN global weighted and directed network, as shown in Fig. 1 (c).

The edge weights of the bidirectional edges between x_G and x_i $(i = 1, 2, \ldots, n)$ are expressed as,

$$\omega_{iG} = \{A_{im}, (A_{im} \neq 0)\}, \tag{13}$$

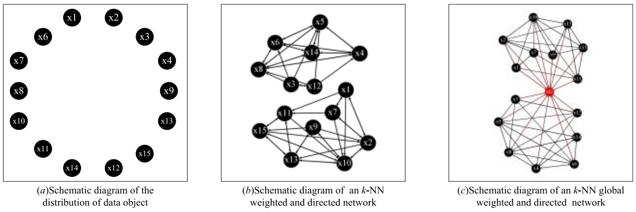
$$\omega_{Gi} = \sum\nolimits_{m=1}^{n} A_{im}. \tag{14}$$

The adjacency matrix $\mathbf{A}_G$ of the adaptive k-NN global weighted and directed network is denoted as,

$$\boldsymbol{A}_G = \begin{bmatrix} 0 & A_{12} & \cdots & A_{1n} & \omega_{1G} \\ A_{21} & 0 & \cdots & A_{2n} & \omega_{2G} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{n1} & A_{n2} & \cdots & 0 & \omega_{nG} \\ \omega_{G1} & \omega_{G2} & \cdots & \omega_{Gn} & 0 \end{bmatrix}. \tag{15}$$

C. PERFORMING A RANDOM WALK PROCESS TO IDENTIFY OUTLIERS

In a specific network, a random walk process is defined as a stochastic process that a random walker moves from x_i to x_m $(i,\ m = 1, 2, \ldots, n,\ m \neq i)$ in the next random step with specific probability. And the transition probability of the random walker moving from one node to another only depends on the current state and remains unchanged throughout the process, expressed as:



(a)Schematic diagram of the distribution of data object

(b)Schematic diagram of an k-NN weighted and directed network

(c)Schematic diagram of an k-NN global weighted and directed network

FIGURE 1. Schematic diagram of the network model.

$$p_{im}^{t} = p_{im}^{t-1} = p\left(x^t = x_m | x^{t-1} = x_i\right)\left(\forall i : \sum\nolimits_m p_{im} = 1\right). \tag{16}$$

where x^t, x^{t-1} represent the position of the random walker at time step t and $t-1$, respectively. p_{im}^t represents the transition probability of the random walker moving from x_i to x_m at time step t.

The adaptive k-NN global weighted and directed network represented by an adjacent matrix $\mathbf{A}_G$ uniquely defines a random walk process. The transition probability matrix is obtained from $\mathbf{A}_G$:

$$\mathbf{P}_G = \mathbf{A}_G \times \mathbf{D}^{-1}, \tag{17}$$

where $\mathbf{D}$ is a diagonal matrix, and each element in the matrix is equal to the sum of the corresponding rows of $\mathbf{A}_G$.

The random walker is expected to jump to a neighbor with a greater similarity to the current node at the next time step. Using the edge weights between nodes to customize the transition probability of the random walk process can achieve the above effect and effectively complete the transition process between nodes. Starting from any state at any time, the random walk process will converge to an equilibrium state after a certain number of iterations, and the probability of the random walker at each node will not change. Using an iterative method, the stationary distribution vector of the random walk process on the adaptive k-NN global weighted and directed network can be estimated. The iterative process is formalized as follows:

$$\boldsymbol{\pi}^t = \boldsymbol{\pi}^{t-1} \times \mathbf{P}_G, \tag{18}$$

where, the element in $\boldsymbol{\pi}^t = (\pi_1^t, \pi_2^t, \ldots, \pi_n^t, \pi_G^t)$ represents the probability of a random walker remaining at the node at time step t. And $\boldsymbol{\pi}^0 = (0)$ is an initialized random probability vector.

After the random walk process reaches equilibrium, each element in the stationary distribution vector π^t can be explained as the visited probability for the corresponding node in the adaptive k-NN global weighted and directed network. Based on the constraints on the behavior of the random walkers, it can be inferred that the potential outlier notes will have less chances to be visited by the random walker, therefore they will be assigned relative smaller scores in the stationary distribution vector. Considering the influence of x_G, an index (outlier score) is developed by assigning the visited probability of x_G to each real node to measure the outlier degree of data objects, which is denoted as,

$$\Psi_i = \frac{1}{\pi_i^t + \frac{\omega_{Gi}}{\sum_{i=1}^{n} \omega_{Gi}} \pi_G^t}. \tag{19}$$

In order to ensure the convergence of the random walk process, x_G is added in the construction process to ensure the connectivity of the network model, and non-existent edge pointing to itself is restricted on the network model to avoid self- reinforcement. The transition probability matrix of the random walk process is defined by the similarity between the data object and its neighbors. When the random walker is located on a real node, it always jumps to its neighbor with the greatest similarity at the next time step. The setting of Eqs. (11)-(14) ensures the completion of the following process, that is, the random walker always jumps with a greater probability to the neighbor with the greatest similarity, x_G, and the node which has the greatest similarity with its all neighbors at the next time step, when it is located at the real node, the isolated node, and x_G at the current time, respectively. The value in the stationary distribution vector is used to express the probability that a node is accessed when the random walk process reaches a stable state. Based on the transition mechanism of the random walk process, the potential outliers should get a smaller access probability. In order to eliminate the influence of x_G on the detection results, the visited probability of x_G to real nodes based on Eq. (13) and Eq. (14) is assigned to define outlier score (Ψ_i). It can be seen from Eq. (19) that the larger Ψ_i is, the more outlying x_i tends to be.

The proposed outlier detection methodology is presented in Algorithm 2.

The proposed method first searches the value of k, which has an $O(n^2)$ complexity. Next to calculate the neighborhood relations in universe X with $O(n^2)$ complexity. Then, to compute the adjacent matrix of the k-NN weighted and directed network with an $O(n^2)$ complexity. The iterative method to compute the stationary distribution vector has an $O(n^2)$ complexity, and the last step to distribute the abnormal score has an $O(n)$ complexity. To sum these up, the total time complexity for the algorithm is $O(n^2)$.

IV. EXPERIMENTS

The proposed method is experimentally verified on three UCI datasets with three related outlier detection methods: the neighborhood information entropy-based outlier detection method (NIEOD) [39], the outlier detection using random walks (OutRank) [23], and the virtual outlier score model (VOS) [33], and the experimental environment is shown in Table 2.

Algorithm 2 Calculate the Outlier Score for Each Data Object

```
Input: Dataset $X$, $k$
Output: outlier score of each data object
for $i \leftarrow 1$ to $n$
    $k$-NN$(x_i) \leftarrow \{x_m | (m \neq i)(x_m \in X) \cap D_{im} \leq D_i^*\}$
end for
for $i \leftarrow 1$ to $n$
    for $m \leftarrow 1$ to $n$
        $E_{im} \leftarrow \begin{cases} 1, & \text{if } x_m \in k - \text{NN}(x_i); \\ 0, & \text{otherwise.} \end{cases}$
        $A_{im} \leftarrow \begin{cases} 1 - D_{im}, & \text{if } E_{im} = 1; \\ 0, & \text{if } E_{im} = 0. \end{cases}$
    end for
end for
for $i \leftarrow 1$ to $n$
    $\omega_{iG} \leftarrow \{A_{im}, (A_{im} \neq 0)\}$
    $\omega_{Gi} \leftarrow \sum_{m=1}^{n} A_{im}$
end for
$\mathbf{P}_G \leftarrow A_G \times \mathbf{D}^{-1}$
$\pi^t \leftarrow \pi^{t-1} \times \mathbf{P}_G$
for $i \leftarrow 1$ to $n$
    $\Psi_i \leftarrow \frac{1}{\pi_i^t + \frac{\omega_{Gi}}{\sum_{i=1}^{n} \omega_{Gi}} \pi_G^t}$
end for
return $\Psi_i$
```

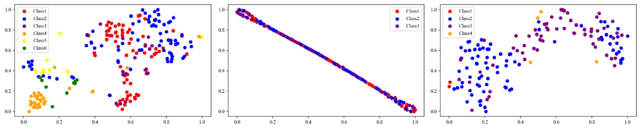
TABLE 2. Experimental environment.

Software and Hardware Environment	Parameter
CPU	2.40 GHz Intel Core i7-55000U
Operating System	Windows 10
Internal Storage	8.00 GB
Development Tool	Visual Studio 2015
Compiling Environment	C#

The NIEOD is proposed to detect outliers in the dataset with numerical, categorical and mixed-valued data by using the neighborhood information system and information entropy. The heterogeneous distance and self-adapting radius are applied to determine the neighborhood information system of the dataset, the neighborhood information entropy, relative neighborhood entropy, deviation degree, and outlier factor are further constructed based on the neighborhood information around each data object to measure the outlier degree of each data object. It has extended the outlier detection methods which are based on the traditional distance and rough set, and more applies to the datasets with some uncertainty mechanisms. There are two parameters in this method, namely, the neighborhood radius adjustment parameter and the judgement threshold.

The OutRank explores the application of the random walk process in outlier detection for the dataset with numerical data, the random walk process can effectively capture not only the uniformly dispersed outliers but also small clusters of outliers. It builds a weighted and undirected neighborhood network model and uses the cosine similarity between objects and the shared-nearest neighbor density to define similarity metric, respectively. It has yield higher detection rates with lower false alarm rates than the outlier detection methods based on distance and density on both real and synthetic datasets.

The VOS improves the outlier detection methods based on the random walk process, combines the k-NN with the network model to identify outliers in the dataset with



(*a*) Glass Identification dataset (*b*) Hayes-Roth dataset (*c*) Lymphography dataset

FIGURE 2. Distribution of data objects on the above three datasets.

numerical attributes. It uses the top-k similar neighbors of each data object to construct the network model, and implements outlier detection by executing a tailored random walk process. By making full use of the local information of each data object and considering the outlier degree of each data object from a global perspective, the effectiveness of this method is demonstrated theoretically.

The proposed method in this paper can identify the outliers in a dataset with numerical, categorical or mixed-valued data integrating the advantages of the above methods, which overcomes the defects of these methods to a certain extent. And, the predetermined parameter such as k is not necessary in the proposed method.

A. UCI DATASETS

Glass Identification dataset, Hayes-Roth dataset, and Lymphography dataset in the UCI machine learning library [40] are applied to evaluate the performance of the proposed method. These datasets are marked for classification and the rare classes are known, the data objects in the rare classes are considered as outliers.

Glass Identification dataset contains 214 data objects, which described by 1 name attribute, 9 numerical attributes, and 1 class attribute. The 214 data objects are divided into 6 categories with 70, 76, 17, 13, 9, and 29 data objects in each category, respectively. The class 5 has the fewest objects and can be treated as the rare class, in which the data objects are outliers.

Hayes-Roth dataset contains 132 data objects with 1 name attribute, 4 categorical attributes, and 1 class attribute. The 132 data objects are divided into 3 categories and each category has 51, 51, and 30 data objects, respectively. The class 3 has the fewest objects and can be treated as the rare class, in which the data objects are outliers.

Lymphography dataset contains 148 data objects, which described by 3 numerical attributes, 15 categorical attributes, and 1 class attribute. The 148 data objects are divided into 4 categories with 2, 81, 61, and 4 data objects in each category, respectively. The class 1 and class 4 have the fewest objects and can be treated as the rare classes, in which the data objects are outliers.

The data distribution of the above three datasets is shown in Fig. 2 (*a*), (*b*), and (*c*) respectively.

B. EVALUATION METRICS FOR THE PERFORMANCE OF METHODS

In order to quantitatively analyze the experimental results of different outlier detection methods, three traditional information system quality metrics "precision"(*Pre*), "recall"(*Rec*) and "rank power"(*RP*) are used in this paper [41].

Pre calculates the proportion of the real outliers in the dataset identified as outliers in the first z data objects:

$$Pre = \frac{N_{identified}}{z}, \tag{20}$$

where, $N_{identified}$ represents the number of the real outliers identified as outliers in the first z data objects.

Rec measures the percentage of the real outliers identified in the first z data objects and all real outliers in the dataset:

$$Rec = \frac{N_{identified}}{N_{real}}, \tag{21}$$

where, N_{real} is the number of the real outliers in the dataset.

Pre and *Rec* estimate the accuracy of detection results of an outlier detection method, but neither of them can accurately compare the quality of detection results of different methods. For example, *Pre* and *Rec* of two real outliers identified by the method at the first two positions are the same as those at any two positions in the first z data objects. The location, where the real outliers are identified, is usually an important factor in comparing different outlier detection methods. Therefore, *RP* is introduced to measure simultaneously the position and number of the real outliers:

$$RP = N_{identified} \frac{N_{identified} + 1}{2\sum_{L=1}^{N_{identified}} O_L}, \tag{22}$$

where $RP \in [0, 1]$, O_L represents the position of the L^{th} real outlier. In particular, $RP = 1$ if and only if all real outliers are at the top of the objects selected by the outlier detection method. Obviously, larger *RP* means better performance of the outlier detection method.

Pre and *Rec* are positively correlated with the effectiveness of outlier detection methods. When *Pre* and *Rec* are the same, the larger the *RP* is, the more effective the outlier detection method is.

C. EXPERIMENT RESULTS

Here, the detection results of the proposed method with those of other three methods on three different types of UCI datasets are compared by the above indexes. The above four methods no longer classify the data objects into normal objects and outliers. Instead, each data object is assigned a score to measure the outlier degree, and then the data objects are sorted (in the ascending or descending order) based on the judgment mechanism of the method. The first z data objects which may include the real outlier and/or the identified (but not real) outlier are chosen to verify the performance of the methods.

The scores of data objects in the dataset based on the proposed method, the NIEOD, and the VOS are sorted in a descending order, while based on the OutRank is sorted in an ascending order considering the different mechanism of each method to construct the index to measure the outlier

TABLE 3. Detection results of the four outlier detection methods on the Glass Identification dataset.

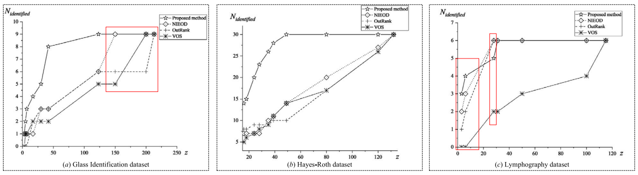
Top ratio (z)	$N_{identified}$ (Coverage (%))			
	Proposed method	NIEOD	OutRank	VOS
1%(3)	1(11.1)	1(11.1)	0(0)	0(0)
2%(5)	2(22.2)	1(11.1)	0(0)	1(11.1)
3%(7)	3(33.3)	1(11.1)	0(0)	1(11.1)
8%(17)	4(44.4)	1(11.1)	2(22.2)	2(22.2)
14%(30)	5(55.6)	3(33.3)	3(33.3)	2(22.2)
20%(42)	8(88.9)	3(33.3)	3(33.3)	2(22.2)
57%(123)	9(100)	6(66.7)	6(66.7)	5(55.6)
70%(150)	9(100)	9(100)	6(66.7)	5(55.6)
93%(200)	9(100)	9(100)	6(66.7)	9(100)
99%(213)	9(100)	9(100)	9(100)	9(100)
Mean	5.9	4.3	3.5	3.6

TABLE 4. Detection results of the four outlier detection methods on the Hayes-Roth dataset.

Top ratio (z)	$N_{identified}$ (Coverage (%))			
	Proposed method	NIEOD	OutRank	VOS
12%(16)	14(46.7)	7(23.3)	8(26.7)	5(16.7)
14%(18)	15(50)	7(23.3)	8(26.9)	6(20)
18%(24)	20(66.7)	7(23.3)	9(30)	7(23.3)
21%(28)	23(76.7)	7(23.3)	9(30)	8(26.7)
27%(35)	26(86.7)	10(33.3)	9(30)	9(30)
30%(39)	28(93.3)	11(36.7)	10(33.3)	11(36.7)
37%(49)	30(100)	14(46.7)	10(33.3)	14(46.7)
60%(80)	30(100)	20(66.7)	17(56.7)	17(56.7)
90%(120)	30(100)	27(90)	26(86.7)	26(86.7)
100%(132)	30(100)	30(100)	30(100)	30(100)
Mean	24.6	14	13.6	13.3

TABLE 5. Detection results of the four outlier detection methods on the Lymphography dataset.

Top ratio (z)	$N_{identified}$ (Coverage (%))			
	Proposed method	NIEOD	OutRank	VOS
2%(3)	3(50)	2(33.3)	1(16.7)	0(0)
4%(6)	4(66.7)	3(50)	2(33.3)	0(0)
7%(28)	5(83.3)	6(100)	6(100)	2(33.3)
19%(31)	6(100)	6(100)	6(100)	2(33.3)
34%(50)	6(100)	6(100)	6(100)	3(50)
68%(100)	6(100)	6(100)	6(100)	4(66.7)
78%(115)	6(100)	6(100)	6(100)	6(100)
Mean	5.1	5	4.7	2.4



(*a*) Glass Identification dataset

(*b*) Hayes-Roth dataset

(*c*) Lymphography dataset

FIGURE 3. Detection results on three datasets based on the four methods.

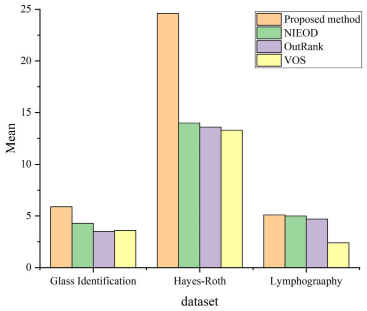


FIGURE 4. Mean of detection results on datasets for the four outlier detection methods.

TABLE 6. Analysis of detection results on the Glass Identification dataset.

z	Proposed method			NIEOD			OutRank			VOS		
	Pre	*Rec*	*RP*	*Pre*	*Rec*	*RP*	*Pre*	*Rec*	*RP*	*Pre*	*Rec*	*RP*
3	0.33	0.11	0.33	0.33	0.11	0.33	0	0	0	0	0	0
5	0.4	0.22	0.38	0.2	0.11	0.33	0	0	0	0.2	0.11	0.2
7	0.43	0.33	0.4	0.14	0.11	0.33	0	0	0	0.14	0.11	0.2
17	0.24	0.44	0.31	0.06	0.11	0.33	0.12	0.22	0.11	0.12	0.22	0.18
30	0.17	0.56	0.24	0.1	0.33	0.21	0.1	0.33	0.12	0.07	0.22	0.18
42	0.19	0.89	0.20	0.07	0.33	0.21	0.07	0.33	0.12	0.05	0.22	0.18
123	0.07	1	0.15	0.05	0.67	0.07	0.05	0.67	0.08	0.04	0.56	0.06
150	0.06	1	0.15	0.06	1	0.06	0.04	0.67	0.08	0.03	0.56	0.06
200	0.05	1	0.15	0.05	1	0.06	0.03	0.67	0.08	0.05	1	0.04
213	0.04	1	0.15	0.04	1	0.06	0.04	1	0.05	0.04	1	0.04

TABLE 7. Analysis of detection results on the Hayes-Roth dataset.

z	Proposed method			NIEOD			OutRank			VOS		
	Pre	*Rec*	*RP*	*Pre*	*Rec*	*RP*	*Pre*	*Rec*	*RP*	*Pre*	*Rec*	*RP*
16	0.88	0.47	0.79	0.44	0.23	0.56	0.5	0.27	0.51	0.31	0.17	0.56
18	0.83	0.5	0.79	0.39	0.23	0.56	0.44	0.27	0.51	0.33	0.2	0.48
24	0.83	0.67	0.80	0.29	0.23	0.56	0.38	0.3	0.5	0.29	0.23	0.42
28	0.82	0.77	0.81	0.25	0.23	0.56	0.32	0.3	0.5	0.29	0.27	0.40
35	0.74	0.87	0.79	0.29	0.33	0.37	0.26	0.3	0.5	0.26	0.3	0.36
39	0.74	0.93	0.78	0.28	0.37	0.35	0.26	0.33	0.44	0.28	0.37	0.33
49	0.61	1	0.76	0.29	0.47	0.32	0.20	0.33	0.44	0.29	0.47	0.31
80	0.38	1	0.76	0.25	0.67	0.30	0.21	0.57	0.25	0.21	0.57	0.28
120	0.25	1	0.76	0.23	0.9	0.28	0.22	0.87	0.23	0.22	0.87	0.24
132	0.23	1	0.76	0.23	1	0.27	0.23	1	0.23	0.23	1	0.23

TABLE 8. Analysis of detection results on the Lymphography dataset.

z	Proposed method			NIEOD			OutRank			VOS		
	Pre	*Rec*	*RP*	*Pre*	*Rec*	*RP*	*Pre*	*Rec*	*RP*	*Pre*	*Rec*	*RP*
3	1	0.5	1	0.67	0.33	1	0.33	0.17	1	0	0	0
6	0.67	0.67	0.83	0.5	0.5	0.86	0.33	0.33	0.5	0	0	0
28	0.18	0.83	0.375	0.21	1	0.54	0.21	1	0.34	0.07	0.33	0.12
31	0.19	1	0.30	0.19	1	0.54	0.19	1	0.34	0.06	0.33	0.12
50	0.12	1	0.30	0.12	1	0.54	0.12	1	0.34	0.06	0.5	0.09
100	0.06	1	0.30	0.06	1	0.54	0.06	1	0.34	0.04	0.67	0.08
115	0.05	1	0.30	0.05	1	0.54	0.05	1	0.34	0.05	1	0.06

degree of data objects. With the increase of the first z objects selected, the number of the real outliers identified by the above methods is different, and the detection results on the three datasets are shown in Table 3, Table 4 and Table 5, respectively.

There are some supplementary explanations of detection results in Table 3, Table 4 and Table 5. ‘‘Top ratio’’ indicates the proportion of the first z data objects selected in the whole dataset, and ‘‘Coverage’’ represents the proportion of the real outliers identified in the first z data objects in all real outliers in the dataset. ‘‘Mean’’ measures the average number of the real outliers identified by each method when the four methods can identify all real outliers in the dataset in the first z data objects selected.

For better visualization, the detection results on different datasets in Table 3, Table 4 and Table 5 are illustrated in Fig. 3 (*a*), (*b*), and (*c*), respectively, and the Mean of outlier detection methods on UCI datasets is shown as Fig. 4.

The analysis of detection results based on evaluation indexes for each method on the three datasets are shown in Table 6, Table 7 and Table 8, respectively, and the corresponding results are shown in Fig. 5, Fig. 6, and Fig. 7, respectively.

The running time of the four outlier detection methods on the three datasets are exhibited in Table 9.

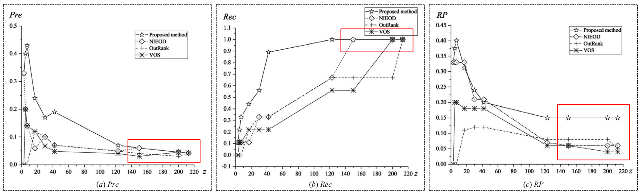


FIGURE 5. Analysis results of the four methods on the Glass Identification dataset.

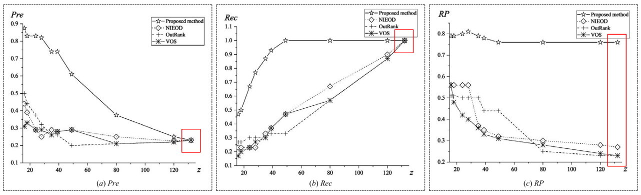


FIGURE 6. Analysis results of the four methods on the Hayes-Roth dataset.

TABLE 9. Experiments results of running time (*s*).

Dataset	Proposed method	NIEOD	OutRank	VOS
Glass Identification	4.2	207.5	15.1	8858.4
Hayes-Roth	0.87	1.7	1.5	306.4
Lymphography	1.1	4.7	5.1	302.0

D. DISCUSSION AND ANALYSIS

The acquisition details of the experiment results in Section 4.3 are as follows:

1) GLASS IDENTIFICATION DATASET

As shown in Table 3, Fig. 3 (*a*), and Fig. 4, the number of the real outliers identified by the proposed method is the largest compared with the other three methods with the increase of the first z data objects. When the first 123 (57%) data objects are selected, the proposed method identifies all real outliers in the dataset, and the other three methods identify 6, 6 and 5 real outliers, respectively. And they identify all real outliers when the first 150 (70%), 213 (99%), and 200 (93%) data objects are selected, respectively. When the first 213 data objects are selected, the four methods identify all real outliers in the dataset, and the average number of the real outliers identified by each method is 5.9, 4.3, 3.5, and 3.6, respectively. From Fig. 5, it can be found that when the selected data objects are less than or equal to 150, the *Pre* and *Rec* of the proposed method are larger than the other three methods. When the value of z is larger than 150, the four methods have the same values of *Pre* and *Rec*, but the proposed method has a larger *RP* value than the other three methods. Besides, the running time shown as in Table 9 of the proposed method on this dataset is only 4.2 s, while the other three running times are 207.5 s, 15.1 s, and 8858.4 s, respectively. In conclusion, the proposed method in this paper has better detection performance than the NIEOD, the OutRank, and the VOS on the Glass Identification dataset, which is a dataset with only numerical attributes.

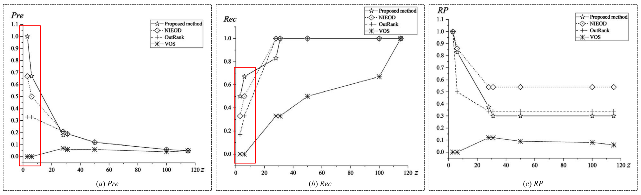


FIGURE 7. Analysis results of the four methods on the Lymphography dataset.

2) HAYES-ROTH DATASET

The proposed method detects all real outliers in the dataset when the first 49 (37%) data objects are selected, while the other three methods detect 14, 10, and 14 data objects as shown in Table 4 and Fig. 3 (*b*), respectively. When the value of z is 132 (100%), the NIEOD, the OutRank, and the VOS detect all real outliers. In Fig. 4, the Mean of the proposed method is 24.6 which significantly higher than that of the comparative methods. With the increase of the first z data objects selected, the *Pre* and *Rec* of the proposed method are significantly larger than the other three methods. When z is 132, their *Pre* and *Rec* reach the same. At this time, the *RP* of the proposed method is the largest among them as shown in Fig. 6 (*c*). Moreover, the proposed method has the shortest running time of 0.87 s, and the other methods are 1.7 s, 1.5 s, and 306.4 s, respectively. Compared with the other three methods, the proposed method has obvious advantages in identifying the number of the real outliers, *Pre*, *Rec*, and running time on the Hayes-Roth dataset which only contains categorical attributes.

3) LYMPHGRAPHY DATASET

From Table 5, Fig. 3 (*a*), and Fig. 4, the proposed method identifies 3 real outliers in the dataset when the first 3 (2%) data objects are selected, while the NIEOD, the OutRank, and the VOS identify 2, 1, and 0, respectively. The four methods detect 4, 3, 2, and 0 real outliers respectively when the value of z is 6 (4%). The NIEOD and the OutRank identify all real outliers first when the selected data objects are 28, the proposed method and the VOS identify 5 and 2, respectively. And the proposed method detects all real outliers in the dataset when the value of z is 31. However, the Mean of the proposed method is slightly higher than the other three methods. In Table 8 and Fig. 7, the *Pre* and *Rec* of the proposed method are the highest when the value of z is less than 28, consistent with the NIEOD and the OutRank when the selected data objects are greater than 28, and slightly lower than the NIEOD and the OutRank only when the first 28 data objects are selected. Additionally, the running time of the proposed method is 1.1 s, and the other methods are 4.7 s, 5.1 s, and 302.0 s, respectively. Therefore, the proposed method can be applied to identify the real outliers in the Lymphgraphy dataset, which has simultaneously numerical and categorical attributes.

Based on the above observations, the following conclusions can be drawn:

(1) The proposed method can be used to detect outliers in different types of datasets, including the dataset with numerical attributes, categorical attributes, and mixed-valued attributes.

(2) The adaptive k is obtained automatically according to the distribution characteristics of the dataset, which ensures higher quality of outlier mining and reduces the cost in the process of parameter adjustment.

(3) In the proposed method, the k-NN is used to mine the local information of data objects, and a customized random walk process is used to explore the long-term correlation between related data objects from a global perspective. The combination of the k-NN and the random walk process improves the accuracy of detection results.

V. CONCLUSION

In this paper, an unsupervised outlier detection method based on the adaptive k-NN global network is proposed to identify the outliers in a DMA. First, the process of determining the weight of numerical and categorical attributes is given respectively to reduce the influence of different attributes on data objects, and the spatial distance between mixed-valued data objects is measured based on the Heterogeneous Euclidean-Overlap Metric. Second, an adaptive search algorithm for the appropriate value of k is introduced, which can automatically obtain the k according to the distribution of data objects in different datasets. And the k-NN of each data objects is obtained. Next, a network model is constructed based on the neighborhood relationship between data objects, and a customized similarity measurement is applied to calculate the edge weight of the network, in which the edge weight is directly proportional to the similarity between the data object and its neighbor. Then, a special random walk process is performed on the network model by defining the transition probability matrix using the edge weight. After the random walk process reaching the equilibrium state, outlier score (Ψ_i) is constructed to measure the outlier degree of each data object. Finally, a detailed empirical study is devised to illustrate the effectiveness, accuracy, and applicability of our method in detecting outliers using three typical UCI datasets. The proposed method has higher *Pre* and *Rec* in the detection results compared with the other three methods. It can be employed in the dataset with numerical attributes, categorical attributes or mixed-valued attributes.

However, calculating the stationary distribution vector is a time-consuming process, which limits the application of the outlier detection methods based on the random walk process to the large-scale and stream dataset. The main work in the next stage is how to apply this method to practical application scenarios with large datasets.

REFERENCES

[1] H. O. Marques, R. J. G. B. Campello, J. Sander, and A. Zimek, "Internal evaluation of unsupervised outlier detection," *ACM Trans. Knowl. Discovery from Data*, vol. 14, no. 4, pp. 1–42, Aug. 2020, doi: 10.1145/3394053.

[2] D. Ghosh and A. Vogt, "Outliers: An evaluation of methodologies," in *Joint Statistical Meetings*, San Diego, CA, USA: American Statistical Association, 2012, pp. 3455–3460.

[3] W. R. Buckland, "Outliers in statistical data," *J. Oper. Res. Soc.*, vol. 30, no. 7, pp. 674–675, Jul. 1979, doi: 10.1057/jors.1979.165.

[4] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, Jun. 2017, doi: 10.1016/j.neucom.2017.02.039.

[5] M. Munoz-Organero, "Outlier detection in wearable sensor data for human activity recognition (HAR) based on DRNNs," *IEEE Access*, vol. 7, pp. 74422–74436, 2019, doi: 10.1109/ACCESS.2019.2921096.

[6] P. C. Cynthia and S. T. George, "An outlier detection approach on credit card fraud detection using machine learning: A comparative analysis on supervised and unsupervised learning," in *Intelligence in Big Data Technologies-Beyond the Hype*, vol. 1167. Singapore: Springer, Jul. 2020, pp. 125–135, doi: 10.1007/978-981-15-5285-4_12.

[7] C.-M. Kim, E. J. Hong, and R. C. Park, "Chest X-ray outlier detection model using dimension reduction and edge detection," *IEEE Access*, vol. 9, pp. 86096–86106, 2021, doi: 10.1109/ACCESS.2021.3086103.

[8] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, Jul. 2021, doi: 10.1016/j.neucom.2019.12.148.

[9] C. Xu, S. Zhao, and F. Liu, "Sensor fault detection and diagnosis in the presence of outliers," *Neurocomputing*, vol. 349, pp. 156–163, Jul. 2019, doi: 10.1016/j.neucom.2019.01.025.

[10] Y. Yi, W. Zhou, Y. Shi, and J. Dai, "Speedup two-class supervised outlier detection," *IEEE Access*, vol. 6, pp. 63923–63933, 2018, doi: 10.1109/ACCESS.2018.2877701.

[11] S. Liu, Z. Qin, X. Gan, and Z. Wang, "SCOD: A novel semi-supervised outlier detection framework," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2019, pp. 316–321.

[12] Z. Cheng, E. Zhu, S. Wang, P. Zhang, and W. Li, "Unsupervised outlier detection via transformation invariant autoencoder," *IEEE Access*, vol. 9, pp. 43991–44002, 2021, doi: 10.1109/ACCESS.2021.3065838.

[13] S. Uttarkabat, N. D. Sunkara, and B. K. Patra, "RSOD: Efficient technique for outlier detection using reverse nearest neighbors statistics," in *Proc. 4th Int. Conf. Comput. Intell. Netw. (CINE)*, Kolkata, India, Feb. 2020, pp. 1–6.

[14] S. S. Stevens, "Mathematics, measurement, and psychophysics," in *Handbook of Experimental Psychology*. Hoboken, NJ, USA: Wiley, Jan. 1951, pp. 1–49.

[15] N. Li, C. Zhang, W. Xie, and Y. Li, "Exceptional events classification in warehousing based on an integrated clustering method for a dataset with mixed-valued attributes," *Int. J. Comput. Integr. Manuf.*, vol. 31, no. 11, pp. 1078–1096, Aug. 2018, doi: 10.1080/0951192x.2018.1509129.

[16] Y. Li, Y. Wang, and N. Li, "Abnormal operation status identification in warehousing based on neighborhood information entropy considering mixed-valued attributes," *Int. J. Prod. Res.*, vol. 59, no. 18, pp. 5647–5660, Jul. 2020, doi: 10.1080/00207543.2020.1788736.

[17] S. Xu, C. Hu, L. Wang, and G. Zhang, "Support vector machines based on K nearest neighbor algorithm for outlier detection in WSNs," in *Proc. 8th Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Shanghai, China, Sep. 2012, pp. 1–4.

[18] S. K. Sahu, S. K. Jena, and M. Verma, "K-NN based outlier detection technique on intrusion dataset," *Int. J. Knowl. Discovery Bioinf.*, vol. 7, no. 1, pp. 58–70, Jan. 2017, doi: 10.4018/IJKDB.2017010105.

[19] X. Wang, H. Jiang, and B. Yang, "A K-nearest neighbor medoid-based outlier detection algorithm," in *Proc. Int. Conf. Commun., Inf. Syst. Comput. Eng. (CISCE)*, Beijing, China, May 2021, pp. 601–605.

[20] L. Li, G. Xu, Y. Zhang, and M. Kitsuregawa, "Random walk based rank aggregation to improving web search," *Knowl.-Based Syst.*, vol. 24, no. 7, pp. 943–951, Oct. 2011, doi: 10.1016/j.knosys.2011.04.001.

[21] M. R. Islam and M. R. Islam, "An improved keyword extraction method using graph based random walk model," in *Proc. 11th Int. Conf. Comput. Inf. Technol.*, Khulna, Bangladesh, Dec. 2008, pp. 225–229.

[22] X. Liu, Z. Li, X. Zhao, and Z. Zhou, "Using concept-level random walk model and global inference algorithm for answer summarization," in *Information Retrieval Technology*, vol. 7097. Berlin, Germany: Springer, Dec. 2011, pp. 434–445, doi: 10.1007/978-3-642-25631-8_39.

[23] H. D. K. Moonesinghe and P.-N. Tan, "Outrank: A graph-based outlier detection framework using random walk," *Int. J. Artif. Intell. Tools*, vol. 17, no. 1, pp. 19–36, 2008, doi: 10.1142/S0218213008003753.

[24] C. Wang, H. Gao, Z. Liu, and Y. Fu, "A new outlier detection model using random walk on local information graph," *IEEE Access*, vol. 6, pp. 75531–75544, 2018, doi: 10.1109/ACCESS.2018.2883681.

[25] Y. Dong and X. Yan, "Multivariate outlier detection approach based on K-nearest neighbors and its application for chemical process data," *J. Chem. Eng. Jpn.*, vol. 47, no. 12, pp. 876–886, 2014, doi: 10.1252/jcej.13we346.

[26] F. Korn and S. Muthukrishnan, "Influence sets based on reverse nearest neighbor queries," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 201–212, Jun. 2000, doi: 10.1145/335191.335415.

[27] L. Cao, X. Liu, T. Zhou, Z. Zhang, and A. Liu, "A data stream outlier delection algorithm based on reverse K nearest neighbors," in *Proc. Int. Symp. Comput. Intell. Design*, Hangzhou, China, Oct. 2010, pp. 236–239.

[28] M. Batchanaboyina and N. Devarakonda, "Efficient outlier detection for high dimensional data using improved monarch butterfly optimization and mutual nearest neighbors algorithm: IMBO-MNN," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 2, pp. 63–73, Apr. 2020, doi: 10.22266/ijies2020.0430.07.

[29] J. Ha, S. Seok, and J.-S. Lee, "A precise ranking method for outlier detection," *Inf. Sci.*, vol. 324, pp. 88–107, Dec. 2015, doi: 10.1016/j.ins.2015.06.030.

[30] J. Ning, L. Chen, C. Zhou, and Y. Wen, "Parameter K search strategy in outlier detection," *Pattern Recognit. Lett.*, vol. 112, pp. 56–62, Sep. 2018, doi: 10.1016/j.patrec.2018.06.007.

[31] L. Berton, J. Huertas, B. Araujo, and L. Zhao, "Identifying abnormal nodes in complex networks by using random walk measure," in *Proc. IEEE Congr. Evol. Comput.*, Barcelona, Spain, Jul. 2010, pp. 1–6.

[32] X. Liu, C. T. Lu, and C. Feng, "Spatial outlier detection: Random walk based approaches," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, New York, NY, USA, 2010, pp. 370–379.

[33] C. Wang, Z. Liu, H. Gao, and Y. Fu, "VOS: A new outlier detection model using virtual graph," *Knowl.-Based Syst.*, vol. 185, Dec. 2019, Art. no. 104907, doi: 10.1016/j.knosys.2019.104907.

[34] E. Li, H. Liu, K. Su, and S. Zhang, "Outlier detection via kernel preserving embedding and random walk," in *Proc. IEEE Int. Conf. Knowl. Graph (ICKG)*, Nanjing, China, Aug. 2020, pp. 20–25.

[35] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Adv. Phys.*, vol. 56, no. 1, pp. 167–242, 2007, doi: 10.1080/00018730601170527.

[36] Y. He, H. Guo, M. Jin, and P. Ren, "A linguistic entropy weight method and its application in linguistic multi-attribute group decision making," *Nonlinear Dyn.*, vol. 84, no. 1, pp. 399–404, Apr. 2016, doi: 10.1007/s11071-015-2595-y.

[37] S. Kappal, "Data normalization using median & median absolute deviation MMAD based Z-score for robust predictions vs min-max normalization," *London J. Res. Sci., Natural Formal*, vol. 19, no. 4, pp. 39–44, Jun. 2019, doi: 10.13140/RG.2.2.32799.82088.

[38] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *J. Artif. Intell. Res.*, vol. 6, no. 1, pp. 502–516, 2017, doi: 10.1109/ISIE.1997.648935.

[39] Z. Yuan, X. Zhang, and S. Feng, "Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures," *Expert Syst. Appl.*, vol. 112, pp. 243–257, Dec. 2018, doi: 10.1016/j.eswa.2018.06.013.

[40] S. D. Bay. (1999). *The UCI KDD Repository*. [Online]. Available: http://kdd.ics.uci.edu

[41] H. Cao, G. Si, Y. Zhang, and L. Jia, "Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8090–8101, Dec. 2010, doi: 10.1016/j.eswa.2010.05.079.

YU WANG received the B.S. degree in traffic engineering from the China University of Mining and Technology, in 2019, where she is currently pursuing the master's degree in management science and engineering. Her current research interests include data mining, pattern recognition, and outlier detection.

XUEJING CAO received the B.S. degree in engineering from the Zhengzhou University of Aeronautics, Zhengzhou, China, in 2018. She is currently pursuing the M.S. degree with the China University of Mining and Technology, Xuzhou, China. Her research interests include outlier detection, pattern recognition, and outlier detection.

YUPENG LI received the B.S. degree in industrial engineering and the M.S. degree in resource development and planning from the China University of Mining and Technology, Xuzhou, in 2006 and 2009, respectively, and the Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University, in 2015. He is currently an Associate Professor with the School of Mines, China University of Mining and Technology. His research interests include product evolution control and operation management of manufacturing and service systems.