Applied Soft Computing Journal 104 (2021) 107182

Contents lists available at ScienceDirect

Applied Soft Computing Journal

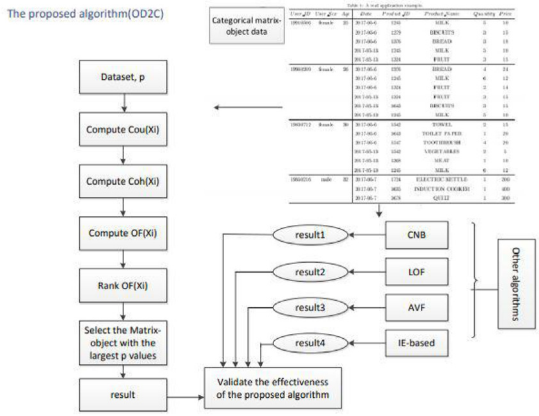
journal homepage: www.elsevier.com/locate/asoc

An outlier detection algorithm for categorical matrix-object data

Fuyuan Cao [*], Xiaolin Wu, Liqin Yu, Jiye Liang

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 14 June 2019
Received in revised form 21 December 2020
Accepted 10 February 2021
Available online xxxx

Keywords:
Outlier detection algorithms
Categorical matrix-object data
Data mining
Machine learning

ABSTRACT

Outlier detection is a significant problem in data mining and machine learning which aims to discover objects in a data set that do not conform to well-defined notions of expected behavior. Generally, the input of the existing outlier detection algorithms is a collection of n objects and each object is described by a feature vector. However, in many real world applications, an object is not only described by one feature vector, but a number of feature vectors. In this paper, we define an object described by more than one feature vector as a matrix-object. Inspired by the concepts of cohesion and coupling in software engineering, we define the coupling of a matrix-object based on the average distance between it and other matrix-objects, and define its cohesion based on information entropy and mutual information. On this basis, the outlier factor of a matrix-object is given, and an outlier detection algorithm for categorical matrix-object data is proposed. The experimental results on real and synthetic data sets have shown that the proposed outlier detection algorithm can effectively detect outliers for the matrix-object data set compared with other algorithms.

1. Introduction

Outlier detection is an active research area in data mining and machine learning [1–3], which aims to find useful exceptional and novel or rare events. Hawkins [4] first gave the essential definition of outliers: an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Outlier detection can be formally described: given a set of n data points or objects, and the expected number of outliers p, then finds the significantly different, abnormal or inconsistent first p objects. The objects detected are also called anomalies, surprises, etc. Outlier

[*] Corresponding author.
E-mail addresses: cfy@sxu.edu.cn (F. Cao), 1144191845@qq.com (X. Wu), 1367545521@qq.com (L. Yu), ljy@sxu.edu.cn (J. Liang).

https://doi.org/10.1016/j.asoc.2021.107182
1568-4946/© 2021 Elsevier B.V. All rights reserved.

detection has been applied into credit card fraud, financial audit, network monitoring, e-commerce, fault detection, inclement weather forecasting and health system monitoring, etc [5].

At present, many outlier detection algorithms have been proposed [1]. There are two basic tasks in outlier detection: one is in a given data set to define what kind of data can be considered inconsistent, the other is to find an effective way to mine outliers. The existing outlier detection algorithms are mainly classified as statistical-based algorithm [6], distance-based algorithm [7], density-based algorithm [8], deviation-based algorithm [9], clustering-based algorithm [10]. In recent years, many researchers have also proposed corresponding outlier detection algorithms for different applications such as multi-view field [11, 12].

However the input of the existing outlier detection algorithms is typically a data set containing n objects, each of which consists of one vector. In many real world applications, an object is not only described by one feature vector, but a number of feature vectors. The object described by more than one feature vector is called as a matrix-object. There is no outlier detection algorithm for matrix-object data at this stage. If we use the existing algorithms, the data needs to be preprocessed in advance. In order to better find outliers in a matrix-object data set, an outlier detection algorithm for categorical matrix-object data is proposed in this paper.

The rest of this paper is organized as follows. In Section 2, the related work is described. In Section 3, the problem description is discussed. In Section 4, an outlier detection algorithm is proposed. In Section 5, experimental results on the data sets are reported. The conclusions are given in Section 6.

2. The related work

The related work is introduced in this section, including statistical-based algorithms, distance-based algorithms, density-based algorithms, deviation-based algorithms, clustering-based algorithms and other recent algorithms.

2.1. Statistical-based algorithms

The earliest outlier detection algorithms are mostly used in statistics [6]. The basic idea of a statistical-based algorithm is to assume a probability model of data distribution according to the characteristics of the data set, and then determine the anomaly based on the inconsistency of the model. But the problem is that in many real applications the distribution of data is unknown, and the actual data often do not conform any kind of ideal state of the mathematical distribution, so it is difficult for us to detect the outliers in the later stage. Moreover, the statistical-based methods are more suitable for low-dimensional data sets, but the actual data are mostly high-dimensional data sets. Therefore, it is very difficult to estimate the distribution of data in advance.

2.2. Distance-based algorithms

The distance-based method was proposed by Knorr and Ng [7]. An outlier is an object that has a higher distance relative to all other objects. Distance-based outlier $DB(p,d)$ can be defined as follows: if the data set X has at least p parts, and the distance between a given object and the objects of p parts is greater than the distance d, the object is a distance-based outlier with parameters p and d, i.e. $DB(p,d)$ [13]. The distance-based outlier detection methods can analyze multidimensional data in an unknown data distribution state without too much calculations, thus, it can be applied to any feature space that can be measured by some distance mechanism. For categorical data sets, *ORCS* [14] uses the Hamming distance and *CNB* uses neighborhoods based distance to measure the distance between classified objects. The *CNB* algorithm consists of two steps, generating the neighbors and mining the outliers. In the first step, we obtain the k nearest neighbors of all objects with the same threshold. In the second step, we calculate the outlier factor of each object by their neighbors. Several objects with the maximal outlier factor are selected as outliers.

2.3. Density-based algorithms

The density-based approach focuses only on the density of neighbors around an object (the number of nearest neighbor). The points with larger number of neighbors are not outliers, and the points with smaller number of neighbors may be outliers. The goal of the density-based outlier detection method is to detect the local density by different density estimation strategy. Therefore, a local outlier detection algorithm based on density model is proposed. Breunig et al. [8] uses the local outlier factor (LOF) to express the isolation degree of a point, and uses the given minimum number of neighbors k and the minimal distance between the point and its neighbors to determine the neighborhood. Given a point, the LOF is represented by the ratio of the average reachable density of its neighborhood and the reachable density of itself after calculating its k nearest neighbor distance, reachable distance and reachable density. If the density of a point is different from the density of other points in the neighborhood (LOF value is larger), then the point is determined as a local isolated point.

2.4. Deviation-based algorithms

The deviation-based approach aims at determining the anomaly by examining the main characteristics of objects. For a given object, if its characteristics deviate too much from the given description, it can be taken as an outlier. The existed deviation-based methods mainly include sequential anomaly technique and *OLAP* data cube method [15]. The former describes the basic characteristics of the sample set by taking its variance as the dissimilarity function. All the samples that deviate from these characteristics are abnormal samples. This method is too idealistic for the assumption of anomalies, and the effects on complex data are not very good. The latter identifies anomalies in large-scale multidimensional data. If a cell value of a cube is significantly different from the value obtained from the statistical model, the cell is considered to be an outlier. This method is a form of discovering driven exploration, but it is very difficult to detect artificially because of the large search space.

2.5. Clustering-based algorithms

The clustering-based approaches aim to transform the process of outlier detection into a clustering process and identify outliers as clusters of small sizes [10]. Su et al. [16] proposed a clustering-based outlier detection algorithm, and the smaller clusters are considered as outliers. However, the method ignores the distance between small clusters and large clusters. When the smaller clusters are very closer to the large clusters, these points in the small clusters are more likely to be the boundary points of the large clusters rather than outliers. For categorical matrix-object data sets, Cao [17] proposed a novel clustering method. However, the proposed algorithm is a typical k-type partition method that is applied into the spherical-shape distribution data. It is not suitable for finding outliers.

2.6. Other recent algorithms

Many researchers have also proposed corresponding outlier detection algorithms for different applications. For example, most multi-view outlier detection algorithms express the complex distribution between different views by learning new latent features with pairwise constraints on different view data, but it is expensive to extend from two-view data to three-view (or more) data. Zhao et al. [11] propose a novel multi-view outlier detection method, which performs consensus regularization on the latent representation. Specifically, each outlier is clearly described through inherent cluster assignment labels and sample-specific errors. Detecting outliers in the spatio-temporal trajectory data is critical to improving data quality and the accuracy of subsequent trajectory data mining tasks. Han et al. [12] propose a trajectory outlier detection algorithm based on a Bidirectional Long Short-Term Memory (Bi-LSTM) model. First, a six-dimensional motion feature vector is extracted for each trajectory point, and then we construct a Bi-LSTM model. The model input is the trajectory data feature vector of a certain sequence length, and its output is the class type of the current track point. The BiLSTM model can automatically learn the difference between the normal points and adjacent abnormal points in the motion characteristics by combining the LSTM unit and the bidirectional network.

3. Problem description

A given matrix-object data set is formulated as follows. Suppose that $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is a set of n objects described by m attributes $\{A_1, A_2, \ldots, A_m\}$, where $X_i = \begin{pmatrix} x_{i11} & \ldots & x_{i1m} \\ \vdots & \ddots & \vdots \\ x_{ir_i1} & \ldots & x_{ir_im} \end{pmatrix}$. x_{ijs} denotes the sth attribute value of X_i on the jth record. We call X_i as a matrix-object and $\mathbf{X}$ as a matrix-object data set. Suppose that V^s represents the domain values of the attribute A_s in $\mathbf{X}$ and $V_{X_i}^{A_s}$ denotes a set of values on the attribute A_s for X_i. Obviously, $\bigcup_{i=1}^{n} V_{X_i}^{A_s} = V^s$. In traditional data representation, an object is only described by a feature vector or a record while a matrix-object is usually represented by multiple feature vectors or records. Therefore, a matrix-object is a general representation of a traditional object. Table 1 is an example of matrix-object data in real applications.

There are two parts in Table 1. The left half is the user information. The right half is the customer's purchase information and each customer is described by numerical and categorical attributes together. Clearly, there is a typical one to many relationship between the two parts in Table 1. The data in Table 1 have the following features.

- **Correlation**: The two parts of the data may have some correlation. Users with different sex or age maybe have different preferences. For example, the female user of 25 years old and the female user of 26 years old from Table 1 show a similar tendency to buy, they only buy foods. However, the male user of 32 years old buys household appliances and daily necessities.
- **One-many**: Each user corresponds with more than one record in Table 1. For example, the user 19910506 has 5 records while the user 19830712 has 6 records.
- **Mixed**: In the majority of cases, an object is described by categorical and numerical attributes together. The *Quantity* is a numeric attribute and the *Product_Name* is a categorical attribute.
- **Evolution**: Some attribute values will change over time. For example, the user 19830712 buys supplies in a week, but he only buys foods for the next week, so we should focus on the user's buying dynamics.

In the right half part, we can see clearly that every user buys more than one item and one item is bought by many users. Besides, a commodity may also be purchased many times by one person. Through the transaction records we can find the behavior characteristics of different users. For example, the first three users mainly purchase food and daily necessities, but the last person buy household appliances, not buy food and daily necessities. We can see that the last person bought something that is not the same as that of the first three people. Therefore, we can say that his behavior is different from it of the other three. In data mining and machine learning, we call objects that are different from the general behavior as outliers.

The data shown in Table 1 are widely available in banks, insurance, telecommunications and medical databases. Finding outliers in a matrix-object data set can be formalized as: for each $X_i \in \mathbf{X}$, calculate the outlier factor and rank X_i by the outlier factor, then select the p matrix-objects with the largest value as the outliers in the matrix-object data set. Since there is no outlier detection algorithm for categorical matrix-object data at this stage, if we use the existing algorithms, the data need to be preprocessed in advance [18–21]. Given a matrix-object, we can use the mean to represent it for numeric attributes, and use the mode to represent it for categorical attributes. Thus, a lot of information of the data may be lost, resulting in the user's behavior characteristics cannot be reflected enough. In order to better find outliers in a matrix-object data set, we need to develop some new outlier detection algorithms. In this paper, we only investigate the outlier detection algorithm for the right half part whose each record is described by categorical attributes. For numeric attribute, the discretization can be used to change the continuous values to the categorical values, thus the proposed algorithm can be used to solve the problem of outlier detection. The mixed attributes consists of categorical attributes and numeric attributes, so the problem of outlier detection on the mixed attributes also can be solved.

4. The proposed method

High cohesion and low coupling are the concepts in software engineering. The coupling degree and cohesion degree are usually used as the standard to measure the independence degree of modules. Coupling is also called block correlation and is a measure of the closeness between modules in a software system structure. The closer the connection between modules is, the stronger the coupling is and the worse the independence of the modules is. Cohesion is also called intra-block connection and is a measure of the functional strength of a module. If the elements in a module are more closely related, the cohesion of a module is higher. Inspired by high cohesion and low coupling, we give an innovative two-phase strategy to solve the problem of outlier detection when the input data set is a matrix-object data set. In this paper, we assume that a matrix-object is a module. By defining the coupling degree between a matrix-object and others and the cohesion degree of itself, a new outlier factor of the matrix-object is given.

4.1. Coupling degree

Coupling is a measure of the closeness between objects. Using the concept of clustering to understand is objects in the same cluster are closer than objects in different clusters. Therefore, we can use the distance formula [17] to consider the coupling degree. The distance between two matrix-objects is defined as follows.

Table 1
A real application example.

User_ID	*User_Sex*	*Age*	*Date*	*Product_ID*	*Product_Name*	*Quantity*	*Price*
19910506	Female	25	2017-05-6	1245	MILK	5	10
			2017-05-6	1279	BISCUITS	3	15
			2017-05-6	1376	BREAD	3	18
			2017-05-13	1245	MILK	5	10
			2017-05-13	1324	FRUIT	3	15
19900209	Female	26	2017-05-6	1376	BREAD	4	24
			2017-05-6	1245	MILK	6	12
			2017-05-6	1324	FRUIT	2	14
			2017-05-13	1324	FRUIT	3	15
			2017-05-13	1643	BISCUITS	3	15
			2017-05-13	1245	MILK	5	10
19830712	Female	30	2017-05-6	1542	TOWEL	2	15
			2017-05-6	1643	TOILET PAPER	1	20
			2017-05-6	1547	TOOTHBRUSH	4	20
			2017-05-13	1542	VEGETABLES	2	5
			2017-05-13	1368	MEAT	1	10
			2017-05-13	1245	MILK	6	12
19850216	Male	32	2017-05-7	1724	ELECTRIC KETTLE	1	200
			2017-05-7	1635	INDUCTION COOKER	1	400
			2017-05-13	1678	QUILT	1	300

Definition 1. Given two matrix-objects X_i and X_j, which are described by m attributes $\{A_1, A_2, \dots, A_m\}$, the dissimilarity measure between X_i and X_j is defined as

$$d(X_i, X_j) = \frac{1}{m}\sum_{s=1}^{m}\delta(X_{is}, X_{js}), \tag{1}$$

where

$$\delta(X_{is}, X_{js}) = \frac{1}{2}\sum_{v\in V_{X_i}^{A_s}\bigcup V_{X_j}^{A_s}}\left|\frac{\sum_{p=1}^{r_i} f(v, x_{ips})}{r_i} - \frac{\sum_{q=1}^{r_j} f(v, x_{jqs})}{r_j}\right| \tag{2}$$

and

$$f(x, y) = \begin{cases} 1, & if \quad x == y. \\ 0, & otherwise. \end{cases} \tag{3}$$

Here, $f(\cdot, \cdot)$ is a function and its value is 1 if two parameter values are equal, otherwise its value is 0. $|\cdot|$ represents the absolute value of a value.

In addition, add a normalization factor $\frac{1}{2}$ in Eq. (2) is to ensure $0 \le \delta(X_{is}, X_{js}) \le 1$. We have $\delta(X_{is}, X_{js}) = 1$ when $V_{X_i}^{A_s}\bigcap V_{X_j}^{A_s} = \emptyset$.

We have proved that the dissimilarity measure $d(X_i, X_j)$ is a distance metric satisfying three properties as follows [17].

(1) Nonnegativity: $d(X_i, X_j) \ge 0$ and $d(X_i, X_i) = 0$;
(2) Symmetry: $d(X_i, X_j) = d(X_j, X_i)$;
(3) Triangle inequality: $d(X_i, X_j) + d(X_j, X_k) \ge d(X_i, X_k)$.

For any $X_i \in \mathbf{X}$, the coupling degree is defined as

$$Cou(X_i) = \frac{1}{\frac{1}{n}\sum_{j=1}^{n}\ d(X_i, X_j)}. \tag{4}$$

In general, a matrix-object is more isolated if it is farther away from other matrix-objects in the data set. If the average distance between a matrix-object and other matrix-objects is larger, the coupling degree of the matrix-object and other matrix-objects is smaller and the greater the degree of isolation is. Thus, the coupling degree of a matrix-object can reflect its isolation.

4.2. Cohesion degree

Entropy can be used to measure the confusion degree of values on an attribute, while mutual information can be used to measure the dependency between two attributes. We use entropy and mutual information to consider the cohesion degree of a matrix-object.

4.2.1. Entropy

The concept of entropy is derived from statistical thermodynamics and is used to measure the degree of chaos in the system. C. E. Shannon used the method of probability and statistics to give the definition of entropy, the entropy is the measure of information and uncertainty of a random variable [6]. Given a data set X containing n objects $\{x_1, x_2, \dots, x_n\}$, each object is described by attributes $\{A_1, A_2, \dots, A_m\}$, where m is the number of attributes, $x_i = (x_{i1}; x_{i2}; \dots; x_{im})$. $p(x_i)$ is the probability distribution, the entropy of X can be written as follows:

$$E(X) = -\sum_{i=1}^{n} p(x_i)log_2 p(x_i). \tag{5}$$

Entropy can be used as a global measure of outlier detection [22]. The greater the entropy of a random variable is, the more uncertainty it has and the confusion is higher. If the value of an attribute is unknown, the entropy of this attribute indicates how much information we need to predict the correct value. A subset of objects is good outlier candidates if their removal from the data set causes significant decrease of the entropy of the data set.

4.2.2. Object-based entropy

For a matrix-object data set, each object has multiple records, so we can regard each object as a small data set. We define the entropy of a matrix-object as follows.

Definition 2. Given a data set $\mathbf{X}$ including n objects $\{X_1, X_2, \dots, X_n\}$, each object described by m attributes $\{A_1, A_2, \dots, A_m\}$. Based on the definition of entropy, the entropy of the object X_i on attribute A_s can be written as follows

$$E_{A_s}(X_i) = -\sum_{v_s\in V_{X_i}^{A_s}} p(v_s)log_2 p(v_s). \tag{6}$$

Owing to $0 \le p(v_s) \le 1$, the value of entropy is nonnegative. When $p(v_s) = 1$, the entropy gets the minimum 0. When $p(v_s) = 1/|V_{X_i}^{A_s}| = 1/r_i$, the entropy obtain the maximum value $log_2 r_i$. The entropy of the object X_i on the attribute A_s is denoted as

$$H_{A_s}(X_i) = E_{A_s}(X_i)/log_2 r_i. \tag{7}$$

4.2.3. Mutual information

In the previous consideration, the entropy only can be used to measure the uncertainty of distributions of attribute values on each attribute. However, in fact, some attributes are highly dependent each other. That is, the relationship between any pair of attributes values should be calculated. On this basis, the definition of mutual information [23] was proposed. The mutual information is a useful information measure in information theory, which can be seen as the amount of information contained in a random variable about another random variable, or the uncertainty that a random variable reduces by knowing another random variable. Without loss of generality, given a data set X containing n objects $\{x_1, x_2, \ldots, x_n\}$, and with m attributes, the dependence degree between each pair of attributes A_j and A_k $(j, k \in 1, 2, \ldots, m)$ can be calculated based on the mutual information [23], which is defined as

$$I(A_j; A_k) = \sum_{v_j \in V^j} \sum_{v_k \in V^k} p(v_j, v_k) log_2 \frac{p(v_j, v_k)}{p(v_j)p(v_k)}. \tag{8}$$

Here, V^j represents the domain values of the attribute A_j in X. The items $p(v_j)$ and $p(v_k)$ represent the probability of the two attribute values in the data set and $p(v_j, v_k)$ is the joint probability distribution function. Intuitively, the mutual information measures the information shared by A_j and A_k. If the attributes A_j and A_k are independent of each other, we know that A_j does not provide any information about A_k, so their mutual information is zero, and vice versa. If A_j is a deterministic function of A_k, and A_k is also a deterministic function of A_j, all the information passed is shared by A_j and A_k, so their mutual information is maximal.

4.2.4. Object-based mutual information

We generally use the mutual information to compute the relationship of any pair of attributes in a data set. When the input data set is a matrix-object data set, we can regard each matrix-object as a small data set, so the mutual information can be used to compute the relationship of any pair of attributes in a matrix-object, the definition is given as follows.

Definition 3. Given a matrix-object data set **X** with n objects, each matrix-object is described by m attributes $\{A_1, A_2, \ldots, A_m\}$, the mutual information $I_{X_i}(A_j; A_k)$ of each pair of attributes for the matrix-object X_i is defined as

$$I_{X_i}(A_j; A_k) = \sum_{v_j \in V_{X_i}^{A_j}} \sum_{v_k \in V_{X_i}^{A_k}} p(v_j, v_k) log_2 \frac{p(v_j, v_k)}{p(v_j)p(v_k)}. \tag{9}$$

The minimum value of the mutual information is 0, which means that a random variable does not make any effect to the determination of another random variable. The maximum value is the entropy of the random variable, which means that a random variable can completely eliminate the uncertainty of another random variable. According to [24], we normalize the mutual information with a joint entropy, which is denoted as

$$R_{X_i}(A_j; A_k) = \frac{I_{X_i}(A_j; A_k)}{E_{X_i}(A_j, A_k)}, \tag{10}$$

where the joint entropy $E_{X_i}(A_j, A_k)$ is calculated by

$$E_{X_i}(A_j, A_k) = -\sum_{v_j \in V_{X_i}^{A_j}} \sum_{v_k \in V_{X_i}^{A_k}} p(v_j, v_k) log_2 [p(v_j, v_k)]. \tag{11}$$

Table 2
An example of a categorical matrix-object data.

Object	A_1	A_2
X_1	2	4
	2	5
	3	4
X_2	2	4
	3	5
	3	3
	2	4
X_3	2	4
	1	5
	2	3
	3	4
X_4	2	4
	2	5
	3	4
	1	5
X_5	2	4
	6	2
	7	1

4.2.5. Cohesion degree

To measure the compactness degree of a matrix-object, we need to consider the based-object entropy and the based-object mutual information together. The cohesion degree of a matrix-object is defined as

$$Coh(X_i) = \frac{\sum_{j=1}^{m} H_{A_j}(X_i)}{m} \times \frac{\sum_{j=1}^{m-1} \sum_{k=j+1}^{m} R_{X_i}(A_j; A_k)}{m(m-1)/2}. \tag{12}$$

When a given data set is only described by one attribute, the cohesion degree $Coh(X_i)$ is set to $\frac{\sum_{j=1}^{m} H_{A_j}(X_i)}{m}$.

4.3. Outlier factor

We measure the isolation of a matrix-object by calculating its cohesion degree and coupling degree and define the outlier factor of a matrix-object as follows.

$$OF(X_i) = \frac{1 + Coh(X_i)}{Cou(X_i)}. \tag{13}$$

To compute the outlier factor of a given matrix-object, we not only consider the distance between it and other objects, but also consider the distribution of the object itself. Generally, compared with the cohesion degree, the coupling degree can better reflect the isolation degree of an object. So we added 1 to the cohesion degree to make the coupling degree be the main factor in the computation of outlier factor. The greater the outlier factor is, the more likely it is to be an outlier.

Next, we give an example to calculate the outlier factor of the matrix-objects.

Example 1. Given five matrix-objects described by two categorical attributes whose values are represented by integers. The details are illustrated in Table 2.

Firstly we calculate the coupling degree of the objects. According to Eq. (1), the distance between X_1 and X_2 on the attribute A_1 can be calculated as $|\frac{2}{3} - \frac{2}{4}| + |\frac{1}{3} - \frac{2}{4}| = \frac{4}{12}$. Similarly, the dissimilarity measure is $\frac{6}{12}$ on the attribute A_2. Therefore we have $d(X_1, X_2) = (\frac{4}{12} + \frac{6}{12})/(2 \times 2) = \frac{5}{24}$.

Similarly, we can get the distance matrix between the matrix-objects in Table 3.

We can see from Table 3, the sum of the distance between the fifth object and other objects is significantly larger than the other four objects, so it has the minimal coupling degree; the sum of

Table 3
Distance matrix for the five matrix-objects.

Object	X_1	X_2	X_3	X_4	X_5
X_1	0	$\frac{5}{24}$	$\frac{1}{4}$	$\frac{5}{24}$	$\frac{2}{3}$
X_2	$\frac{5}{24}$	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{2}{3}$
X_3	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{1}{8}$	$\frac{2}{3}$
X_4	$\frac{5}{24}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{2}{3}$
X_5	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	0
$Cou(X_i)$	$\frac{15}{4}$	4	$\frac{30}{7}$	4	$\frac{15}{8}$

Table 4
The newly generated data set by Table 2.

Object	A_1	A_2
X_1	2	4
X_2	2	4
X_3	2	4
X_4	2	4
X_5	2	4

the distance between the third object and others is minimal, so the coupling degree is largest. From Table 2, we can see that the values of the matrix-object X_5 are obviously different from other objects. In the attribute A_1, we see that the values 6,7 are only in the object X_5; the values 1,2 in the attribute A_2 appear none of the rest of objects except X_5. So we consider the object X_5 has the biggest isolation.

Then we calculate the degree of cohesion of each object. According to Eq. (7), we respectively calculate the object-based entropy of the five objects on attribute A_1, i.e $H_{A_1}(X_1) = -\frac{-\frac{2}{3}log_2\frac{2}{3}-\frac{1}{3}log_2\frac{1}{3}}{log_2 2}$ =0.57938. Similarly, $H_{A_1}(X_2)$= 0.5, $H_{A_1}(X_3)$= 0.75, $H_{A_1}(X_4)$= 0.75 and $H_{A_1}(X_5)$= 1. We also calculate the object-based mutual information of two attributes of each object. According to Eq. (10), we have $R_{X_1}(A_1;A_2)$= 0.15876, $R_{X_2}(A_1;A_2)$= 0.6667, $R_{X_3}(A_1;A_2)$= 0.5, $R_{X_4}(A_1;A_2)$= 0.25 and $R_{X_5}(A_1;A_2)$=1. Finally, the cohesion degree of each object can be calculated by Eq. (12), $Coh(X_1)$=0.092, $Coh(X_2)$=0.4166, $Coh(X_3)$= 0.375, $Coh(X_4)$= 0.1562 and $Coh(X_5)$=1.0.

In summary, we can obtain the outlier factor of each object according to Eq. (13), $OF(X_1)$=0.2912, $OF(X_2)$=0.3542, $OF(X_3)$= 0.3208, $OF(X_4)$=0.2891 and $OF(X_5)$=1.0667. We can clearly see that the outlier factor of the object X_5 is the largest, so we think it is an outlier.

If using the existing method to process a matrix-object data set, we have to preprocess the data set according to the frequency of the attribute values. We simplify a matrix-object data by selecting attribute values whose frequency are highest, if the frequency of attribute values are equal, the first value is selected. The newly generated data set are illustrated in Table 4.

The values of each object after simplifying are same. Many information in the original data have been lost and we cannot detect which object is an outlier.

4.4. OD2C algorithm

We propose a new algorithm, called a new outlier detection algorithm based on the coupling degree and the cohesion degree (abbr. OD2C). The details are described in *Algorithm* 1.

Algorithm 1 *OD2C Algorithm*

```
Input: : A matrix-object set X and the number of outliers requested $p$;
Output: : $p$ outliers;
Method:
for $i = 1$ to $n$ do
    Compute $Cou(X_i)$ by (4);
    Compute $Coh(X_i)$ by (12);
    Compute $OF(X_i)$ by (13), get the outlier factor for each object;
end for
Rank $OF(X_i)(1 \le i \le n)$ in ascending order, then select the first $p$ matrix-objects as outliers;
Return $p$ outliers;
```

4.5. Time complexity

In the algorithm, we calculate the outlier factor of each matrix-object in two steps, the distance between objects $Cou(X_i)$ for all objects are computed, the time complexity of computing $d(X_i, X_j)$ is $\mathcal{O}(m \times |V'|)$, where $|V'| = max\{|V^s|, 1 \le s \le m\}$, $|V^s|$ is the number of different values for each attribute, so the time complexity of computing $Cou(X_i)$ is $\mathcal{O}(n^2 \times m \times |V'|)$. For each object, the time complexity of computing $Coh(X_i)$ is $\mathcal{O}(m^2)$, the final time complexity of the algorithm can be written as $\mathcal{O}(m^2 + n^2 \times m \times |V'|)$, when the number of attributes is small, the m^2 is much smaller, can be omitted.

5. Experiments

To investigate the effectiveness of the proposed algorithm, some experiments are conducted on three real data sets and five synthetic data sets. Firstly, the experimental data sets are given. Then, the evaluation indexes are introduced. Finally, the compared results between the proposed algorithm and existing four outlier detection algorithms are showed.

5.1. The real data sets

To our best knowledge, there are no public matrix-object data sets with outlier objects. To solve this problem, we conduct data cleaning on some real matrix-object data sets to find their outlier objects. As the outlier objects deviate so much from other objects in a given matrix-object data set, we need to obtain the distribution of the data set in order to find outlier objects. The multidimensional scaling technique [25] can be used to find the distribution of a data set by visualizing the data sets. The main objective of the technique is to obtain a configuration of n points (rows) in P dimensions (cols) by passing the n-by-n distance matrix obtained by Eq. (1) to the function *mdscale* of the *MATLAB*. To visualize the data, we set P = 2. Thus, n points can be visualized to reflect the distribution of data. From the visualized data, we can find some isolated points and get a data set with outlier objects. Then we can run the proposed algorithm on the new data set and evaluate the performance of the algorithm. The following is the data cleaning process of the data sets.

5.1.1. Market basket data

Market Basket data set downloaded from Data website[1] records 1001 customers transactions, each customer is described by four attributes, Customer_Id, Time, Product_Name and Product_Id. Here, we just need to consider Customer_Id and Product_Id, regardless of attributes Time and Product_Name, because all customers have the same values on the attribute Time and the values of Product_Name correspond to Product_Id one by one. In addition, the outstanding characteristic of the Market Basket data is that each customer in it has 7 transactional records. Therefore,

[1] http://www.datatang.com/datares/go.aspx?dataid=613168.

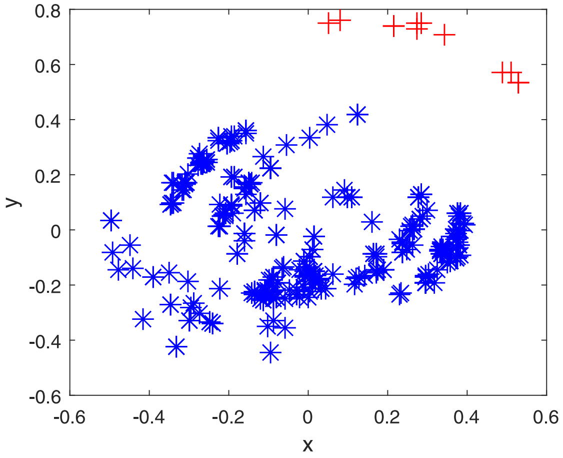


Fig. 1. The distribution of Market Basket data after cleaning.

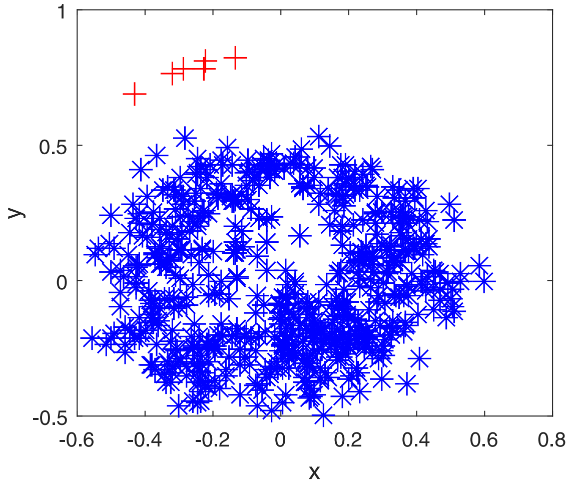


Fig. 2. The distribution of Microsoft Web data after cleaning.

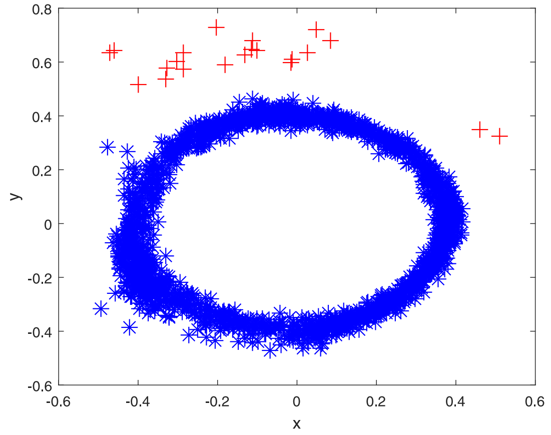


Fig. 3. The distribution of MovieLens data after cleaning.

Table 5
Data sets after cleaning.

Data set	Matrix-objects	Attributes	Records	p
Basket	319	2	2233	12
Web	589	2	5610	6
Movie	1929	3	338778	21

each customer is a typical matrix-object and the data set is a matrix-object data set.

The process of data cleaning for Market Basket data is shown as follows. We firstly visualize the data by the multidimensional scaling technique, then select some objects in the coordinate system which locate the position of $x < -0.2, y < 0.2$ or $x > 0.5$ to form a new data set of 319 objects. The distribution of the new data set is shown in Fig. 1. It is clearly from the newly generated data set that the objects depicted by the symbol $'+'$ can be taken as the outliers of the new data set.

5.1.2. Microsoft web data

Microsoft Web data set which is downloaded from UCI is created by sampling and processing the www.microsoft.com logs. It recorded 32711 anonymous, randomly selected users who visited the site in the week of February 1998. Each user is represented by two attributes, User_Id and Web_Id, and visited more than one web site. Therefore, each user is a matrix-object.

The data cleaning process is described as follows. After visualizing the initial data, we select the points whose abscissa values are in the position of $x < -0$ or $x > 0.68$ to form a new data set of 589 objects, then visualize the new data set in Fig. 2. We can see clearly that the data set consists of two parts. The objects depicted by the symbol '+' can be taken as the outliers of the new data set.

5.1.3. Movielens data

MovieLens data are gathered and made from the MovieLens website[2] and they contain three parts of different sizes, MovieLens 100k, MovieLens 1M and MovieLens 10M. We conducted some experiments on the MovieLens 1M data, which include movies data, ratings data and users data. We do not take movies data and users data into consideration, because they only describe the basic information of movies and users.

In this experiment we only use the ratings data set, which randomly selected 6040 users to record 1000209 ratings of the 3900 movies. And it is represented by four attributes, UserID, MovieID, Rating and Timestamp. The property Timestamp has different values for each record so we delete the attribute.

The data cleaning process is described as follows. We firstly select objects that are in the range of $-0.5 < x < 0.5, y < 0.2$ or $y > 1$ in the coordinate system as a new set after the visualization of the initial data set, then visualize the new data in Fig. 3. It is clearly that the data set consists of a large class and some scattered points. These scattered points depicted by the symbol '+' can be treated as outliers to the new data set.

5.1.4. The data sets after data cleaning

The final data sets after data cleaning are listed in Table 5, where p is the number of outliers. These data sets are used to evaluate the proposed algorithm.

5.2. The synthetic data sets

As there is no public high-dimensional matrix-object data sets, we generate some synthetic data sets and conduct some experiments on them to validate the effectiveness of the proposed algorithm on high-dimensional data sets. The generating process is shown as follows.

Suppose the generated data set contains 200 matrix-objects and each matrix-object contains 2–5 records randomly, the values

[2] http://grouplens.org/datasets/movielens/.

Table 6
The comparison of five algorithms on Precision for different data sets.

Data set	CNB	LOF	AVF	$IE-based$	$OD2C$
Basket	0.1562	0.2000	0.1667	0.2500	1.0000
Web	0.0000	0.1667	0.1667	0.1667	1.0000
Movie	0.0000	0.0476	0.6667	0.6667	0.9524
$syn10$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn20$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn40$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn60$	0.4000	0.6000	0.8000	0.8000	1.0000

Table 7
The comparison of five algorithms on Recall for different data sets.

Data set	CNB	LOF	AVF	$IE-based$	$OD2C$
Basket	0.1526	0.2000	0.1667	0.2500	1.0000
Web	0.0000	0.1667	0.1667	0.1667	1.0000
Movie	0.0000	0.0476	0.6667	0.6667	0.9524
$syn10$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn20$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn40$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn60$	0.4000	0.6000	0.8000	0.8000	1.0000

Table 8
The comparison of five algorithms on F-measure for different data sets.

Data set	CNB	LOF	AVF	$IE-based$	$OD2C$
Basket	0.1562	0.2000	0.1667	0.2500	1.0000
Web	0.0000	0.1667	0.1667	0.1667	1.0000
Movie	0.0000	0.0476	0.6667	0.6667	0.9524
$syn10$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn20$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn40$	0.4000	0.6000	0.8000	0.8000	1.0000
$syn60$	0.4000	0.6000	0.8000	0.8000	1.0000

Table 9
The correlation coefficient between outlier factor and coupling degree, cohesion degree, respectively.

Data set	correlation coefficient with outlier factor	
	Coupling degree	Cohesion degree
Basket	−0.9517	−0.0634
Web	−0.9755	NaN
Movie	−0.9208	0.6654
$syn10$	−0.9815	0.1301
$syn20$	−0.9983	0.0992
$syn40$	−0.9993	0.1444
$syn60$	−0.9993	0.0316

of the first 195 matrix-objects are random numbers in 1–30, and the values in the last five matrix-objects are random numbers in 31–40. The dimensions of the generated matrix-object data sets are set to 10, 20, 40, 60 to generate five different dimensional data sets, which are named as syn10, syn20, syn40, syn60. Then we visualize the four data sets in Fig. 4. It can be clearly seen that the data sets consist of a large class and some scattered points, these scattered points depicted by the symbol $'+'$ can be treated as outliers to the generated high-dimensional data sets.

5.3. Evaluation index

To demonstrate the performance of the outlier detection algorithm, we take into account three evaluation metrics which are Precision, Recall, F-measure [26]. The formulas for them are written as follows.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}, \tag{14}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}, \tag{15}$$

$$F-measure = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{16}$$

Thus, the Precision is the proportion of the target outcomes in the evaluation of the results obtained; Recall, as the name suggests, is the proportion of target categories recalled from the areas of interest; the F-measure is the comprehensive evaluation index of these two indicators, which is used to reflect the overall indicators. The values of Precision and Recall are between 0 and 1, and if the values are closer to 1, the better the experimental results are.

5.4. Experimental results

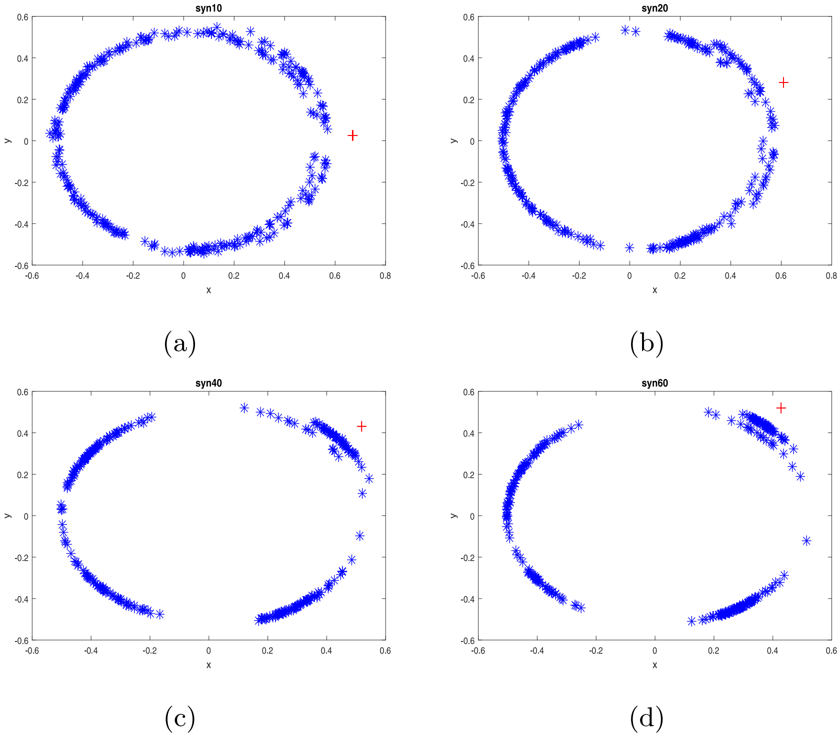
In order to verify the effectiveness of the proposed outlier detection algorithm, we compare it with some existing approaches, including CNB [18] from the proximity-based approach, LOF [19] from the density-based method, AVF [20] and $IE - Based$ [21]. The first three compared algorithms are classic algorithms among different types, the fourth algorithm uses information entropy, so we choose these four algorithms as compared algorithms. Since the four algorithms cannot directly apply into the matrix-object data, we first need to preprocess them to meet the input of existing algorithms. That is, a matrix-object is only described by a record. The experimental results are shown in Tables 6–8.

It can be seen from Tables 6–8 that the proposed outlier detection algorithm outperforms other algorithms. All outliers can be found completely correctly in the Basket, Web data sets and the four synthetic data sets. But the comparison methods are not ideal. For example, the CNB cannot find an outlier in the Web and Movie data sets. In addition, for each data set, we can see that the values of the three evaluation indexes in Tables 6–8 are same for the same algorithm. This is because the number of outliers p in the data sets is determined, and the values of $FalsePositive$ and $FalseNegative$ are equal. Before we use the comparison algorithms, the data sets are simplified. We usually use the mode variable to represent the categorical data. In the new data, an object has only one record, thus many relevant information is lost for a given matrix-object. Therefore, the results of the comparison algorithms are not ideal.

5.5. The effect of a single outlier value

To test the effect of the coupling and cohesion degree on the outlier factor, we compute the values of three variables for each matrix-object of each data set, shown in Figs. 5–6. Thus, for each data set, we can compute the correlation coefficient between the coupling degree, the cohesion degree and the outlier factor, respectively. Table 9 shows the correlation coefficient.

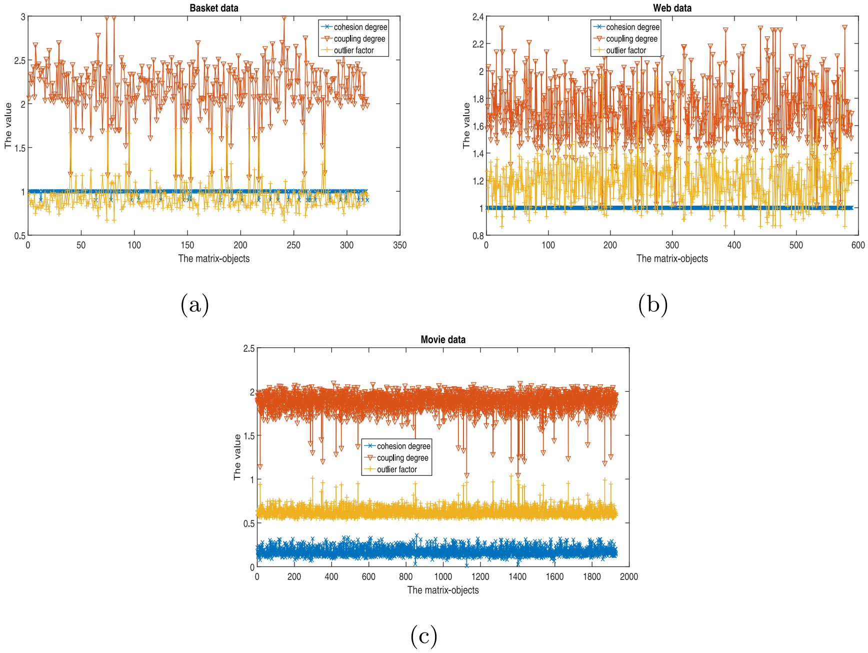
From Figs. 5–6, we can see that the curve fluctuation of the outlier factor is more similar to it of the coupling degree, compared with it of cohesion degree. From Table 9, we can see that the absolute value of correlation coefficient between outlier factor and coupling degree is in range of 92%–99%. In summary, the coupling degree is the main factor for the computation of the outlier factor. Furthermore, we can see from Table 9 that the outlier factor is positively correlated with the cohesion degree while it is negative correlation with the coupling degree for most data sets. It verifies that a higher cohesion degree and a lower coupling degree will result in a bigger outlier factor. For the Web data set, as they are only described by one attribute and all attribute values on each object are different, the entropy of each object is computed as 1, resulting in an abnormal value



(a) (b)

(c) (d)

Fig. 4. The distribution of four synthetic data sets.



(a) (b)

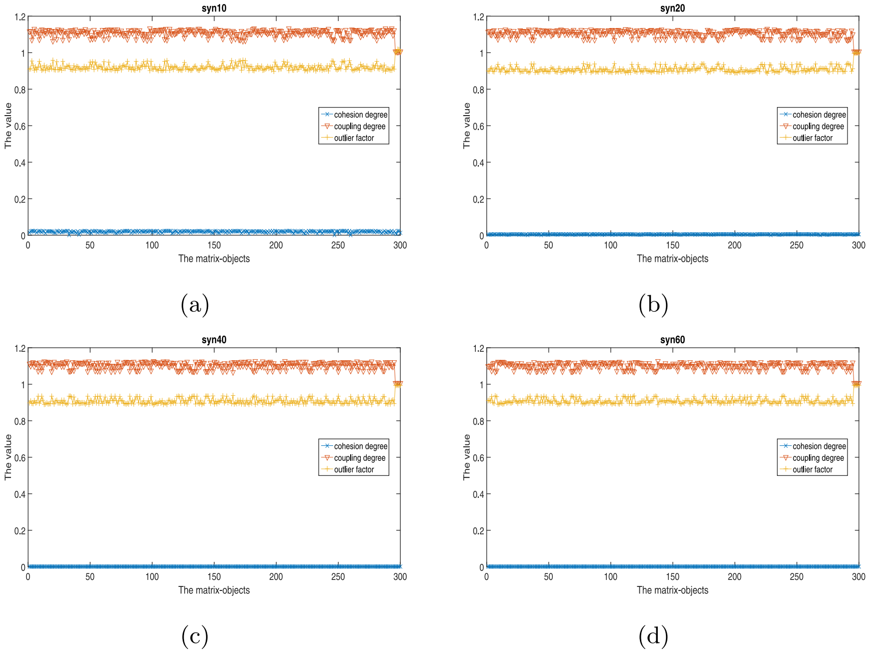
(c)

Fig. 5. The value of outlier factor, coupling degree and cohesion degree of each matrix-object on real data.

of correlation coefficient between cohesion degree and outlier factor.

6. Conclusions

The matrix-object data sets have been widespread in many practical applications. In this paper, in order to solve the problem of outlier detection for categorical matrix-object data set, we propose a new outlier detection algorithm. The algorithm uses an innovative two-phase strategy to find outliers in matrix-object data set. We firstly calculate coupling degree of an object and other objects, then calculate cohesion degree of the object itself, finally we combine these two parts to find outliers. The effectiveness of the method has been shown by experiments in real and synthetic data sets. In the future work, we will focus on the outlier detection problem for numerical matrix-object data or local outlier detection problem for categorical matrix-object data.



(a) (b)

(c) (d)

Fig. 6. The value of outlier factor, coupling degree and cohesion degree of each matrix-object on synthetic data.

CRediT authorship contribution statement

Fuyuan Cao: Conceptualization, Methodology, Writing - original draft, Software, Resources, Funding acquisition. **Xiaolin Wu:** Writing - original draft, Investigation, Project administration, Visualization. **Liqin Yu:** Writing - review & editing. **Jiye Liang:** Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (under grants 61976128, 61773247), the Key Research and Development Projects of Shanxi Province, China (under grant 201803D31022), the Natural Science Foundation of Shanxi Province, China (under grant 201901D111035).

References

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (3) (2009) 1–58.
[2] J. Takeeuchi, K. Yamanishi, A unifying framework for detecting outliers and change points from time ssries, IEEE Trans. Knowl. Data Eng. 18 (4) (2006) 482–492.
[3] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection for discrete sequences: A survey, IEEE Trans. Knowl. Data Eng. 24 (5) (2012) 823–839.
[4] D.M. Hawkins, Identification of Outliers, Chapman and Hall, London, 1980.
[5] V.J. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2) (2004) 85–126.
[6] C.E. Shannon, W. Weaver, A Mathematical Theory of Communication, University of Illinois Press, Urbana, 1949.
[7] H.P. Kriegel, P. Kroger, E. Schubert, A. Zimek, LoOP: local outlier probabilities, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 1649–1652.
[8] M.M. Breunig, H.P. Kriegel, R. Ng, J. Sander, LOF: identifying density-based local outliers, ACM Sigmod Record 29 (2) (2000) 93–104.
[9] A. Arning, R. Agrawal, P. Raghavan, A linear method for deviation detection in large databases, KDD 1141 (50) (1996) 972–981.
[10] W. Jin, A. Tung, J. Han, W. Wang, Ranking outliers using symmetric neighborhood relationship, Adv. Knoml. Discov. Data Min. (2006) 577–593.
[11] H.D. Zhao, H.F. Liu, et al., Consensus regularized multi-view outlier detection, IEEE Trans. Image Process. 27 (1) (2018) 236–248.
[12] Z.R. Han, T.L. Huang, W.J. Ren, et al., Trajectory outlier detection algorithm based on Bi-LSTM model, J. Radars 8 (1) (2019) 36–43.
[13] E.M. Knorr, R.T. Ng, Algorithms for mining distance-based outliers in large datasets, in: Proceeding of the 24th VLDB Conference, 1998, pp. 427–438.
[14] S.D. Bay, M. Schwabacher, Mining distance-based outliers in neat linear time with randomizition and a simple pruning rule, in: Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge and Data Mining, 2003.
[15] J. Han, M. Kamber, J. Pei, Data mining: concepts and techniques, in: Data Mining Concepts Models Methods and Algorithms, vol. 5, no. 4, second ed., 2011, p. 1C18.
[16] D. Ren, I. Rahal, W. Perrizo, A vertical outlier detection algorithm with clusters as by-product, in: IEEE International Conference on TOOLS with Artificial Intelligence, 2004, pp. 22—29.
[17] F. Cao, L. Yu, Z. Huang, J. Liang, k-mw-modes: An algorithm for clustering categorical matrix-object data, Appl. Soft Comput. 57 (2017) 605–614.
[18] S. Li, R. Lee, S. Lang, Mining distance-based outliers from categorical data, in: Proc. IEEE Seventh Int'l Conf. Data Mining Workshops, 2007.
[19] M. Breunig, H.P. Kriegel, R. Ng, J. Sander, LOF: Identifying density-based local outliers, in: Proc.ACM SIGMOD Int'l Conf. Management of Data, 2000.
[20] K. Anna, E.G. Ortiz, et al., A scalable and efficient outlier detection strategy for categorical data, in: IEEE International Conference on Tools with Artificial Intelligence, IEEE, 2007.
[21] F. Jiang, Y.F. Sui, C.G. Cao, An information entropy-based approach to outlier detection in rough sets, Expert Syst. Appl. 37 (9) (2010) 6338–6344.
[22] S. Wu, S. Wang, Information-Theoretic Outlier Derection for Large-Scale Categorical Data, IEEE Educational Activities Department, 2013.
[23] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge Univ. Press, Cambridge, U.K., 2003.
[24] W.H. Au, K.C.C. Chan, A.K. Wong, W.Y. Wang, Attribute clustering for grouping, selection, and classification of gene expression data, IEEE/ACM Trans. Comput. Biol. Biointormat. 2 (2005) 83–101.
[25] S. Schiffman, L. Reynolds, F. Young, Introduction to Multidimensional Scaling: Theory, Methods, and Applications, Academic Press, 1981.
[26] M. Markou, S. Singh, A neural network-based novelty detector for image sequence analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1664–1677.