ICIC Express Letters
Volume **12**, Number **11**, November **2018**

ICIC International ©2018 ISSN 1881-803X
pp. **1079–1088**

MICRO-CLUSTER INSTABILITY PRIVACY PRESERVING OUTLIER DETECTION

Zhaoyu Shou[1], Akang Liu[1] and Simin Li[2]

[1]School of Information and Communication
[2]Key Laboratory of Cognitive Radio and Information Processing Ministry of Education
Guilin University of Electronic Technology
No. 1, Jinji Road, Guilin 541004, P. R. China
{ guilinshou; siminl }@guet.edu.cn; kangluu@foxmail.com

Received April 2018; accepted July 2018

Abstract. *Aiming at the problems of the difficulty of measuring global outliers for uneven distribution data, and the outlier information being easily disturbed by malicious third-party in the information sharing or transmission process, an algorithm of micro-cluster instability privacy preserving outlier detection (MIPPOD) is proposed. Firstly, the original data is divided into micro-clusters based on the density and neighbors. Secondly, outlier detection based on instability factor is performed for each cluster. Finally, the directional transformation is performed according to the micro-cluster instability factors to achieve the privacy preserving of outlier information. The receiver can identify anomaly objects effectively and the hidden data can be restored completely. The verification experiments are performed on simulated and real datasets respectively. By comparison and security analysis, this algorithm has obvious advantages on outlier detection and a good privacy preserving performance.*

Keywords: Outlier detection, Privacy preserving, Micro-cluster instability factor, Directional transformation

1. **Introduction.** Outlier detection [1] is a very important research issue in data mining and data management. The main goal is to quickly and accurately detect the outliers from the complex data environment, which deviate so much from the normal (expected) objects. At present, outlier detection algorithm [2-5] can be roughly divided into distribution-based, depth-based, distance-based, clustering-based and density-based approaches. In recent years, outlier detection has been widely applied in credit card fraud detection, network intrusion, medical health and other fields [6,7].

With the expansion of application requirements, outlier detection must also consider privacy preserving issues. The so-called privacy preserving [8-11] is to hide information that the user considers private by means of some techniques, so as to improve the security of sensitive information. Combining privacy preserving techniques with outlier detection can protect sensitive outlier information. The outlier information is easily disturbed by malicious third-party in the information sharing or transmission process. It is necessary to study the information privacy preserving methods of anomaly data so that it becomes very similar to normal data. While studying how to detect anomaly data efficiently and accurately, it can ensure the security and integrity of the anomaly data in the sharing or transmission process, and reduce the amount of information and parameters transmitted.

In this paper, we present an algorithm of micro-cluster instability privacy preserving outlier detection. The technical contributions of this paper are summarized as follows.

1) Outlier detection based on micro-cluster instability factor is proposed to overcome the difficulty of measuring global outliers for uneven distribution data significantly;

DOI: 10.24507/icicel.12.11.1079

2) Directional transformation is proposed based on the micro-cluster instability factors to achieve the privacy preserving of outlier information, and the corresponding restoration of hidden data is presented to recover the original data completely;

3) The experimental evaluations conducted demonstrate the promising efficiency and effectiveness of the proposed algorithm for privacy preserving outlier detection.

The rest of the paper is organized as follows. Some statements and definitions are introduced in Section 2. Section 3 presents our algorithms for micro-cluster instability privacy preserving outlier detection. The restoration of hidden data is presented in Section 4. Section 5 presents the experiment evaluations. The final section concludes the whole paper and points out the future research directions.

2. Problem Statement and Formal Definitions.

[6] proposes a new robust outlier detection using the instability factor (INS). INS improves the problem that is sensitive to parameter. The changes of accuracy by INS are minor when the parameter is changed. However, INS hardly finds a proper parameter to detect the local outliers and global outliers simultaneously. Huang et al. [7] present a non-parameter outlier detection algorithm based on natural neighbor. As the KD-tree is introduced into the natural neighbor searching, it has high complexity. And it does not consider privacy preserving.

In order to solve the above problems, the datasets are divided into micro-clusters, and then the outlier detection algorithm is carried out for each cluster. Due to the uneven distribution data, the ideas of density and nearest neighbor are used to divide the micro-clusters. Meanwhile, the perturbation technique, namely directional transformation, is explored based on the micro-cluster instability factor. An algorithm of micro-cluster instability privacy preserving outlier detection is proposed to effectively overcome the difficulty of measuring global outliers for uneven distribution data. The definitions and calculations involved are as follows.

Definition 2.1. The k center of gravity. *Let $\sigma_k(p)$ denote a set of k nearest neighbors of an object p, namely $\sigma_k(p) = \{q|d(p,q) \leq d(p,p_k)\}$, where $d(p,q)$ is the distance between objects p and q, and p_k is the k-th nearest neighbor of object p. The center of gravity at a given $\sigma_k(p)$ is then defined as a centroid of the objects in $\sigma_k(p)$, which is given by*

$$m_k(p) = \frac{1}{k+1} \sum_{q \in \sigma_k(p)} X_q \tag{1}$$

where $X_q = (x_{q1}, x_{q2}, \ldots, x_{qd})$ is the coordinates of the object q observed in a d-dimensional space (under the assumption that the space is Euclidean).

Definition 2.2. Absolute difference. *Let $\theta_i(p)$ denote the distance between $m_i(p)$ and $m_{i+1}(p)$, which is defined by the following equation:*

$$\theta_i(p) = d\left(m_i(p), m_{i+1}(p)\right), \quad i = 1, 2, \ldots, k-1 \tag{2}$$

The absolute difference between $\theta_i(p)$ and $\theta_{i+1}(p)$, denoted as $\Delta\theta_i(p)$, is defined as:

$$\Delta\theta_i(p) = \left|\theta_i(p) - \theta_{i+1}(p)\right|, \quad i = 1, 2, \ldots, k-2 \tag{3}$$

Definition 2.3. Instability factor. *The instability factor [6], and $INS(p,k)$ are defined by the following equation:*

$$INS(p,k) = \sum_{i=1}^{k-2} \Delta\theta_i(p) \tag{4}$$

A high value of $INS(p,k)$ indicates that p is a good candidate for an outlier.

Definition 2.4. Directional transformation. *Directional transformation belongs to a new method of perturbation techniques. Before the directional transformation operation, the instability factor is normalized, and its calculation is as below:*

$$INS_p = \frac{INS(p)}{10 * INS_{\max}} \tag{5}$$

where INS_p is the instability factor after the object p is normalized, and $INS_{\max}$ represents the maximum instability factor in the micro-cluster. INS_p is constructed by 10 times, which is normalized to $[0, 0.1]$.

The directional transformation illustrated in Figure 1, is constructed and calculated by the micro-cluster instability factor. A, B, C and D are anomaly objects and they are transformed into A′, B′, C′ and D′. Due to perturbation direction being considered in the directional transformation, the formula is as follows:

$$d(p, p') = (1 - INS_p)d(p, p_{\min}) \tag{6}$$

where p' represents the object after the directional transformation, and $p_{\min}$ is the minimum instability factor in the micro-cluster. $d(p, p')$ represents the distance between the objects. In the restoration of hidden data, the directional transformation can be inversely operated according to Formula (6).

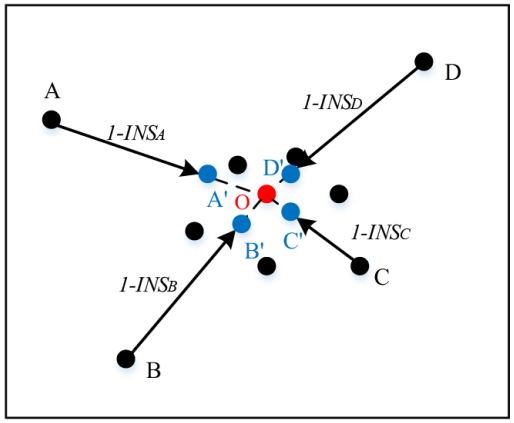

FIGURE 1. Directional transformation

3. **Privacy Preserving Outlier Detection.** Micro-cluster instability privacy preserving outlier detection (MIPPOD) involves the micro-cluster partitioning and the instability factor. Firstly, in order to effectively distinguish different clusters, the dataset is divided into micro-clusters based on the density and neighbors. Secondly, the instability factors of data objects are calculated for each micro-cluster, and outliers are detected by the *Top-N*. Finally, the directional transformation is performed according to the micro-cluster instability factors of outliers. Therefore, these anomaly objects are completely transformed into the data similar to the normal data and hidden in the normal dataset, to achieve the privacy preserving of outlier information.

3.1. **Micro-cluster partitioning.** Clustering is a way of partitioning data objects based on data relationships (distances). Its main goal is to make the objects as similar as possible in the cluster, while the objects in different clusters are as different as possible. DBSCAN [12] is commonly used for clustering. The micro-cluster partitioning in this proposed algorithm involves the density of the DBSCAN algorithm. After the DBSCAN being performed to identify the core points and boundary points, the non-core objects are divided by those queried neighbor cluster identifiers. So all data objects are divided into clusters. The algorithm of micro-cluster partitioning is as follows.

Algorithm 1: Micro-cluster Partitioning

Input: Original Dataset D; Neighborhood Density Threshold *MinPts*
Output: Micro-cluster Dataset

1. Repeat
2. Determine whether the initial point is the core object;
3. Find all direct density points in the r neighborhood of the core object;
4. Until all objects are judged;
5. Repeat
6. For the direct density reachable objects, the maximum densities of connected object clusters are found, which involve cluster merging;
7. Until the r neighborhood of all core objects are traversed;
8. For those non-core object, query which identified cluster those nearest neighbors belong to and assign them to the neighbor clusters.

This Algorithm 1 involves the neighborhood radius r and the neighborhood density threshold *MinPts*. The neighborhood radius can be adaptive to the mathematical model and only the threshold *MinPts* needs to be determined.

3.2. **MIPPOD.** The MIPPOD is operated based on the micro-cluster partitioning. The k nearest neighbor and *Top-N* parameters are involved [6]. The k nearest neighbor is used to calculate the center of gravity, and the *Top-N* is used to determine the number of anomaly objects. Due to micro-cluster partitioning, the k parameter here is selected by the relative neighbors in the micro-clusters, namely $n_1 - 1$, where n_1 is the number of objects in the micro-cluster. The number of anomaly objects in each micro-cluster is different. There are multiple values for the *Top-N* parameters [5,11], which need to be presupposed in each cluster.

The privacy dataset includes cluster identifiers and instability factors, and the cluster identifiers need a one-to-one correspondence with the original data. The instability factors of these data objects participating in the directional transformation are reserved in the micro-cluster, and the instability factors of other data objects are set to 0. Because of the relatively fewer outliers, the amount of information transmitted in the privacy dataset is less, and the data transfer rate is improved. The instability factor in this algorithm is the key parameter, which is used for outlier detection, directional transformation and the restoration of hidden data. The MIPPOD algorithm is as follows.

Algorithm 2: MIPPOD

Input: Original Dataset D; Neighborhood Density Threshold *MinPts*; *Top-N*
Output: The Hidden Dataset; Privacy Dataset

1. Algorithm 1 is used to obtain the micro-cluster dataset;
2. For each cluster:
3. Find the k nearest neighbor objects for each data object;
4. Calculate the center of gravity, absolute difference and instability factor;
5. Each object is normalized by the maximum instability factor and sort them in descending order;
6. Outliers are detected and marked by the presupposed *Top-N*;
7. According to the formula and the data object with the minimum instability factor, directional transformations of these marked outliers are carried out;
8. Return a privacy dataset of cluster identifiers and instability factors;
9. Return the hidden dataset.

3.3. **The complexity analysis.** Assuming that n is the number of objects in the dataset, the complexity of MIPPOD is $T(n) = O(n^2) + O(n_1^2) = O(n^2)$, where n_1 is the number of objects in the micro-cluster. In terms of complexity, the proposed MIPPOD is the same as the DBSCAN, LOF (local outlier factor) [13] and INS. This algorithm protects the outlier information after outlier detection, which is different for the purpose of other algorithms. In addition, the number of objects in the micro-clusters is less than those in the original dataset. Therefore, the space complexity of this algorithm is lower than that of the traditional algorithms. So the complexity of the whole MIPPOD algorithm has a certain advantage.

4. **The Restoration of Hidden Dataset.** After receiving the hidden dataset and privacy dataset, the receiver needs to recover the data so as to further explore the overall data. In the process of data transmission, the sender performs directional transformation of the original data, making the sensitive outlier information hidden and protected.

The receiver receives the hidden dataset and privacy dataset. According to the marks of privacy dataset, clusters are partitioned and data objects with the minimum instability factor are found in each cluster. Then the directional transformation is inversely operated to restore the hidden data. The data objects with the instability factor of 0 represent the original data, while the remaining data objects are involved in the directional transformation, in which the data information corresponds to each other. The hidden data can be restored accurately and correctly.

The receiver can directly obtain the anomaly objects according to the received dataset, without secondary outlier detection, and the overall data information can be gained by the restoration. Only the receiver knows the outliers information, and the third-party cannot obtain or tamper with sensitive outlier information so as to ensure the whole dataset security and privacy.

5. **Experimental Results and Analysis.** The verification experiments are performed to verify the feasibility and effectiveness of the algorithm by simulated datasets and real datasets. The software and hardware environment of this experiment are as below:

Operating system: Windows7 x64; **Software platform:** MATLAB R2014a; **Processor:** Intel Pentium-B960; **Master frequency:** 2.20GHz;**Running memory:** 4GB.

5.1. **Experiment of simulated datasets.** In order to show the visualization effects of MIPPOD, two simulated datasets are selected. One is a regularly distributed dataset, including 1000 data objects, and three clusters randomly generated by Gaussian distribution. The other is an irregularly distributed dataset that contains 832 data objects and is divided into two clusters. Figure 2 and Figure 3 show the detection results of MIPPOD on these two simulated datasets respectively. According to the detection results of the regularly distributed dataset and irregularly distributed dataset, the MIPPOD can effectively detect and hide anomaly objects. The effect on the irregularly distributed datasets is more obvious, ensuring the security of data information and achieving a good result.

Figure 4 and Figure 5 show the restoration of hidden dataset with regular distribution and irregular distribution respectively. It can be seen that the anomaly objects can be restored completely and the effects are ideal.

5.2. **Experiment of real datasets.** The contrast experiments are conducted on real datasets taken from UCI machine learning repository [14] to further validate the priority over other algorithms. Seven real datasets are chosen to carry out the performance comparison and analysis, with different scales and dimensions. The brief information of chosen datasets is described in Table 1.

A. Detection Rate (DR). DR [2,11] is defined as the ratio between the number of outliers detected by the system to the total number of outliers presented in the dataset,

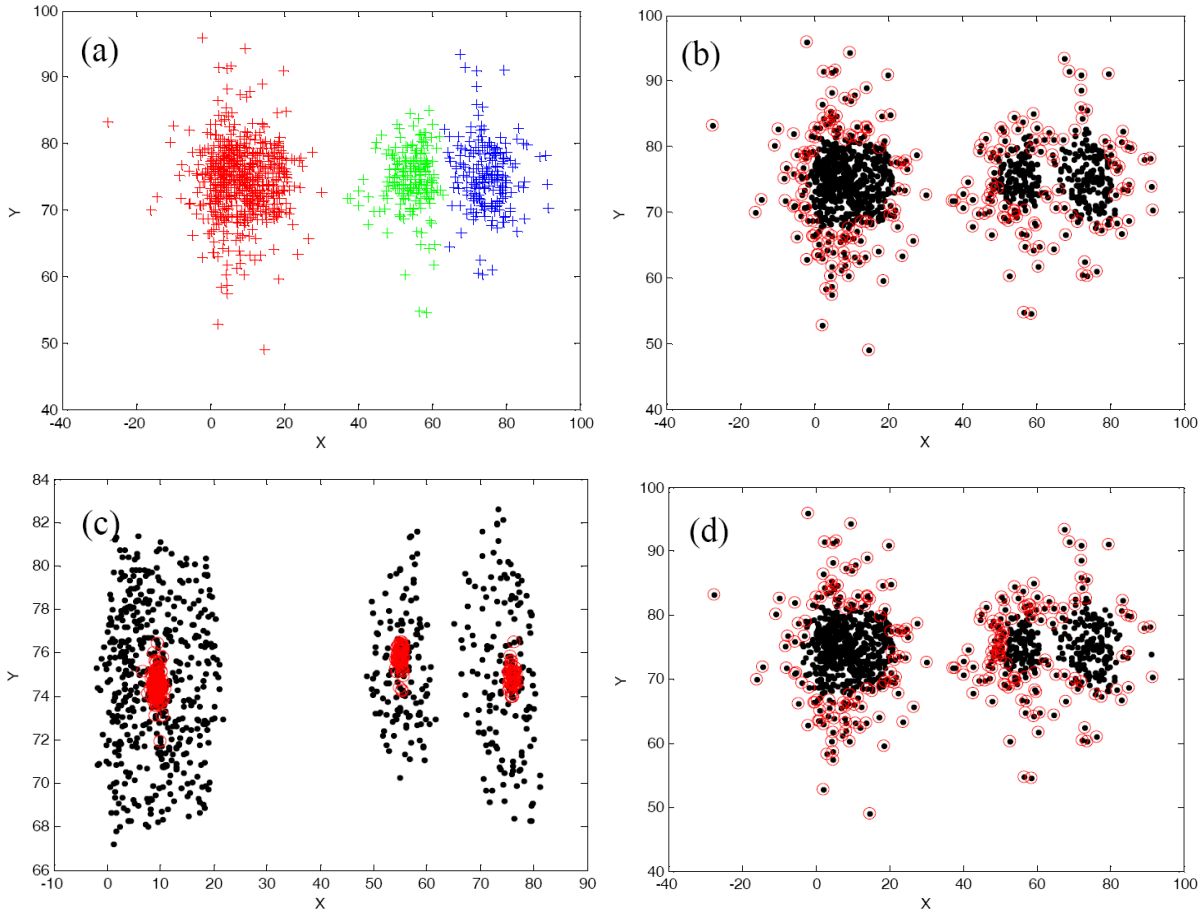

FIGURE 2. Detection results of the regularly distributed dataset: (a) original dataset, micro-cluster partitioning; (b) the results of MIPPOD, cluster 1 *Top-N* = 125, cluster 2 *Top-N* = 50, cluster 3 *Top-N* = 60; (c) the hidden dataset; (d) the results of INS on the original dataset, *Top-N* = 235

i.e., DR = (No. of detected outliers)/(No. of outliers). The greater the value is, the more the number of outliers is detected. The seven datasets are tested by the proposed algorithm DBSCAN, LOF and INS. Due to the inconsistent dimensions of the selected datasets, the datasets are analyzed and preprocessed first. The DR results are shown in Table 2, and a line chart is drawn as Figure 6.

As shown in Figure 6, the DR of MIPPOD is the same as that of LOF and INS in Pima and Optdigits, but they are prior to the DBSCAN. The DR of MIPPOD is higher than that of DBSCAN, LOF and INS in other datasets, which are related to the data attributes. In comparison, the effect of MIPPOD is better.

B. Repetition Rate (RR). In view of the outlier objects detected in the original dataset, whether these objects are detected as anomaly objects after the perturbation processing is measured by repetition rate [11]. It is shown as RR = (No. of outliers again)/(No. of detected outliers). And its value is smaller, explaining that the outlier information hiding is better. The detection results are shown in Table 3. It can be seen that the DR of the hidden datasets is lower, and the number of outliers detected is smaller in the hidden dataset. Meanwhile, using the MIPPOD to detect the anomaly objects, there are no outliers that have been detected before. The anomaly objects information obtained before and after the directional transformation is completely different, and the RR of the outliers are 0.

C. Hiding Failure. Hiding failure is the portion of sensitive information that is not hidden by the application of a privacy preservation technique. The MIPPOD can detect

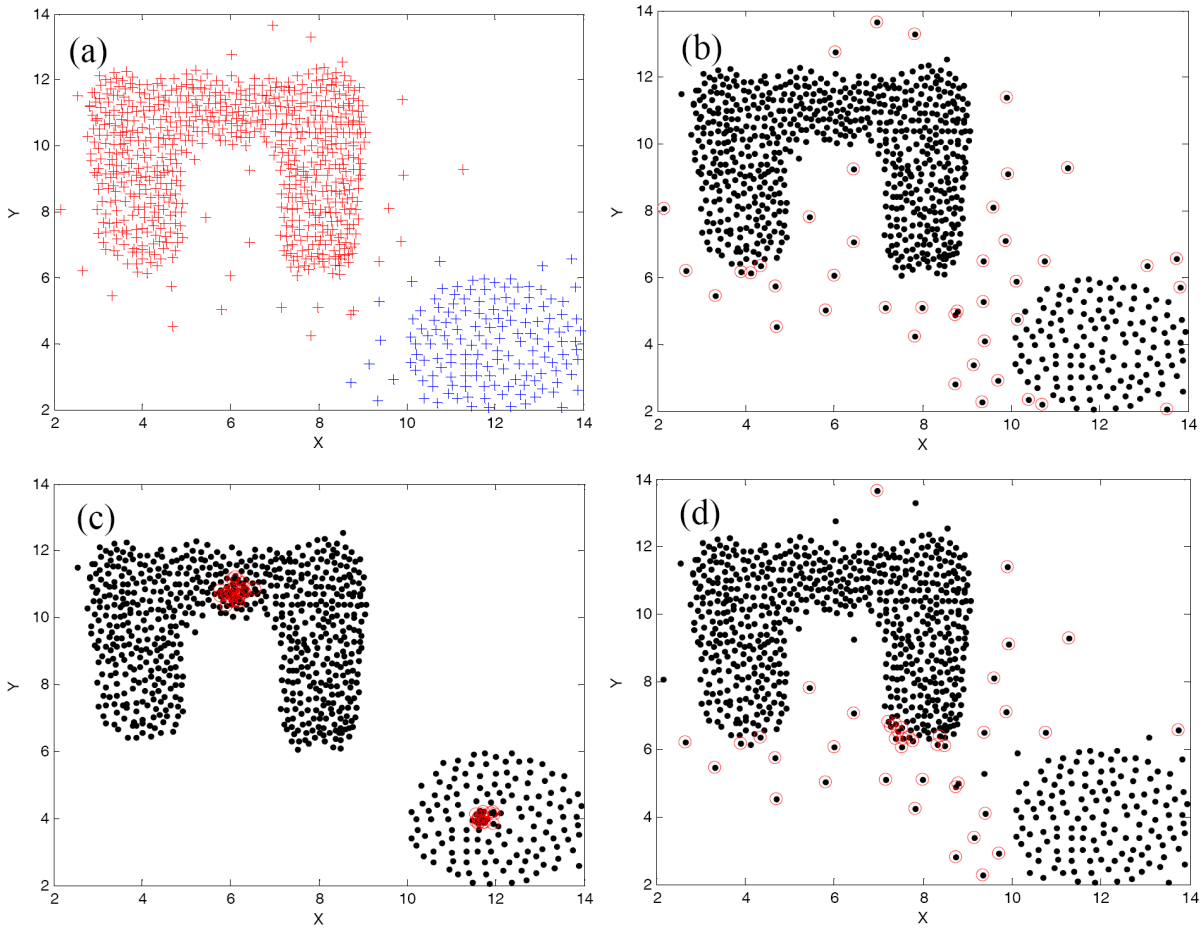

FIGURE 3. Detection results of the irregularly distributed dataset: (a) original dataset, micro-cluster partitioning; (b) the results of MIPPOD, cluster 1 *Top-N* = 27, cluster 2 *Top-N* = 15; (c) the hidden dataset; (d) the results of INS on the original dataset, *Top-N* = 42

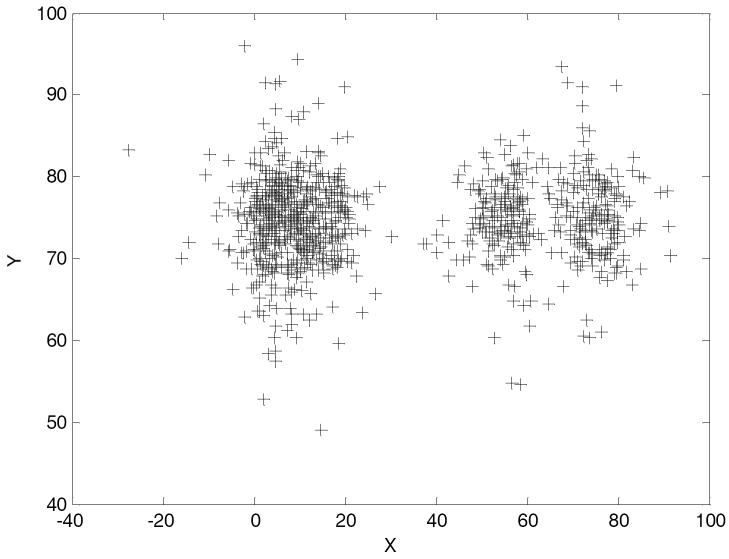

FIGURE 4. The restoration of the regularly distributed dataset

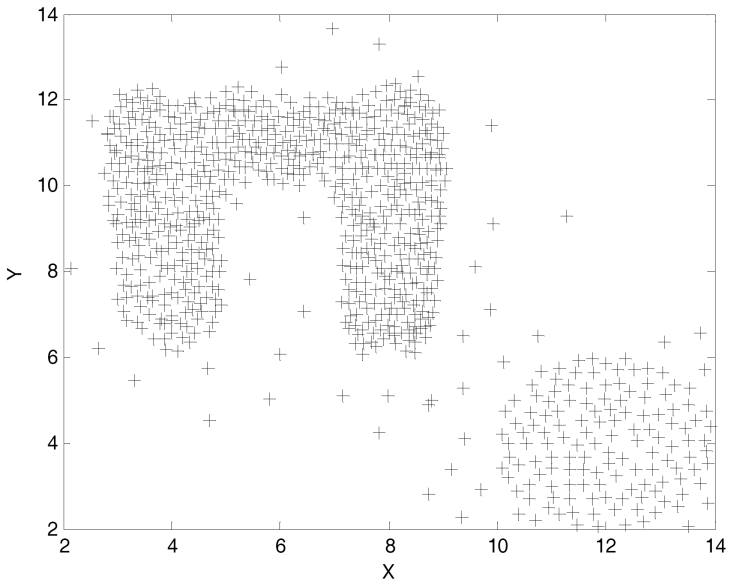

FIGURE 5. The restoration of the irregularly distributed dataset

TABLE 1. Properties of datasets

Dataset	No. of records	No. of attributes	No. of outliers
Ionosphere	351	34	126
Wdbc	569	30	212
Pima	768	8	268
Spambase	4601	57	1813
Optdigits	1797	64	449
Statlog	6435	36	626
Covtype	286048	10	2747

TABLE 2. The DR results with DBSCAN, LOF, INS and MIPPOD

Dataset	DBSCAN/%	LOF/%	INS/%	MIPPOD/%
Ionosphere	84.7	86.4	87.8	92.1
Wdbc	84.5	88.1	89.7	92.0
Pima	85.8	89.4	89.2	89.2
Spambase	83.5	87.2	87.9	89.9
Optdigits	83.9	85.8	85.5	86.1
Statlog	83.7	90.9	91.1	93.1
Covtype	80.6	88.7	89.9	90.9

anomaly objects, which are hidden by the directional transformation into normal data, and their data information completely changes. Some normal data objects are also selected in the process of data perturbation. They can be completely hidden for the outliers that need to be protected, so the hidden failure rate is 0.

5.3. **Adversary attack on data privacy.** The micro-cluster partitioning is the key first step. Different micro-clusters affect the calculation of the instability factor and the perturbation scale. Selecting different *Top-N* in each cluster will lead to different anomaly objects. In order to improve the security level, the *Top-N* is selected a slightly larger value, including a part of relatively normal data objects. Different levels of perturbation processing are performed on each cluster to achieve the differentiation protection.

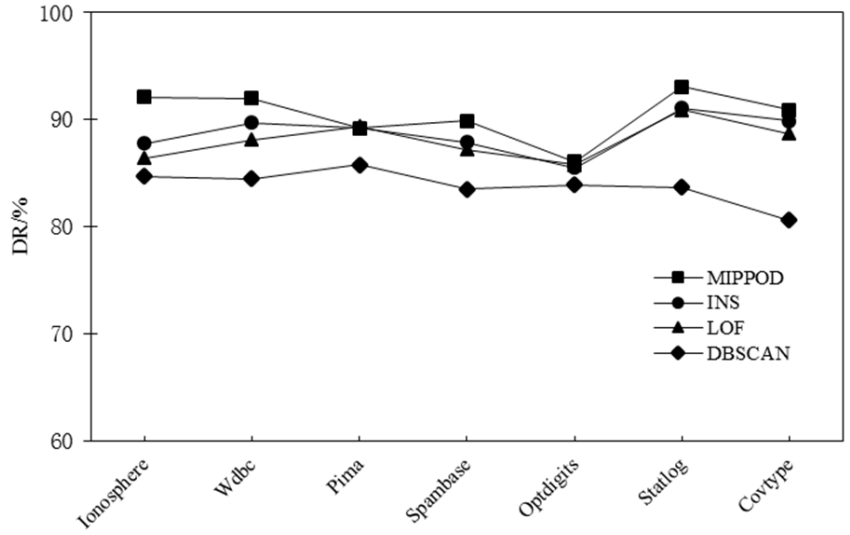

FIGURE 6. The DR results with DBSCAN, LOF, INS and MIPPOD

TABLE 3. The detection results of original dataset and hidden dataset

Dataset	DR of original dataset/%	DR of hidden dataset/%
Ionosphere	92.1	6.3
Wdbc	92.0	7.1
Pima	89.2	12.3
Spambase	89.9	7.9
Optdigits	86.1	13.7
Statlog	93.1	5.6
Covtype	90.9	10.5

The hidden dataset and the privacy dataset are transmitted separately. Even if the third-party obtains the hidden dataset without the privacy dataset, the third-party cannot obtain anomaly objects, which are completely hidden into the normal data. If only the privacy dataset is obtained, it is even more impossible to know the original data. The receiver can directly obtain the anomaly objects and complete the restoration to obtain the whole data information. In this process, only the receiver knows the privacy dataset, and the third-party cannot obtain or tamper with the sensitive outlier information.

6. **Conclusions.** In this paper, the MIPPOD algorithm is introduced in detail. The micro-cluster partitioning algorithm is proposed based on the density and neighbors. Outlier detection based on micro-cluster instability factor is proposed to overcome the difficulty of measuring global outliers for uneven distribution data. The directional transformation is performed based on the micro-cluster instability factors to achieve the privacy preserving of outlier information. The data receiver can identify anomaly objects effectively and the hidden data can be restored completely. This proposed algorithm guarantees the security and integrity of the anomaly data in the information sharing or transmission process. It is a differentiation scheme of anomaly data privacy preserving, which can achieve different privacy protection according to the instability factors. In the MIPPOD, the time complexity of the micro-cluster partitioning is relatively high, which affects the computational complexity of the whole algorithm. In the future work, other micro-cluster partitioning algorithms are explored and make this algorithm more efficient.

Acknowledgment. This work is partially supported by the following foundations: the National Natural Science Foundation of China (61662013, 61362021, U1501252); Natural

Science Foundation of Guangxi Province (2016GXNSFAA380149); Guangxi Innovation-Driven Development Project (Science and Technology Major Project) (AA17202024); the Key Laboratory of Cognitive Radio and Information Processing Ministry of Education (2011KF11); Innovation Project of GUET Graduate Education (2017YJCX34, 2018YJCX37); the Guilin Scientific Research and Technological Development Project (2016010404-4).

REFERENCES

[1] K. Bhaduri, M. D. Stefanski and A. N. Srivastava, Privacy-preserving outlier detection through random nonlinear data distortion, *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol.41, no.1, pp.260-272, 2011.

[2] Z. Shou, M. Li and S. Li, Outlier detection based on multi-dimensional clustering and local density, *Journal of Central South University*, vol.24, no.6, pp.1299-1306, 2017.

[3] T. R. Bandaragoda, M. T. Kai, D. Albrecht et al., Isolation-based anomaly detection using nearest-neighbor ensembles: iNNE, *Computational Intelligence*, no.3, 2018.

[4] M. Bai, X. Wang, J. Xin et al., An efficient algorithm for distributed density-based outlier detection on big data, *Neurocomputing*, vol.181, pp.19-28, 2016.

[5] J. Huang, Q. Zhu, L. Yang et al., A novel outlier cluster detection algorithm without top-n parameter, *Knowledge-Based Systems*, vol.121, pp.32-40, 2017.

[6] J. Ha, S. Seok and J. S. Lee, Robust outlier detection using the instability factor, *Knowledge-Based Systems*, vol.63, no.2, pp.15-23, 2014.

[7] J. Huang, Q. Zhu, L. Yang et al., A non-parameter outlier detection algorithm based on natural neighbor, *Knowledge-Based Systems*, vol.92, pp.71-77, 2016.

[8] J. Liu and Y. Xu, Privacy preserving clustering by random response method of geometric transformation, *The 4th International Conference on Internet Computing for Science and Engineering*, pp.181-188, 2010.

[9] C. Gokulnath, M. K. Priyan, E. V. Balan, K. P. R. Prabha and R. Jeyanthi, Preservation of privacy in data mining by using PCA based perturbation technique, *International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials*, pp.202-206, 2015.

[10] K. A. Ahmed and H. A. Rauf, Privacy preserving data using fuzzy hybrid data transformation technique, *Indian Journal of Science & Technology*, vol.10, no.24, pp.1-6, 2017.

[11] Z. Shou, A. Liu, S. Li et al., Spatial outlier information hiding algorithm based on complex transformation, *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*, pp.241-255, 2017.

[12] M. Hahsler, *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*, 2017.

[13] M. M. Breunig, H. P. Kriegel and R. T. Ng, LOF: Identifying density-based local outliers, *ACM SIGMOD International Conference on Management of Data*, vol.29, no.2, pp.93-104, 2000.

[14] M. Lichman, *UCI Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA, http://archive.ics.uci.edu/ml.