Journal of Computers Vol. 33 No. 2, April 2022, pp. 201-209
doi: 10.53106/199115992022043302018

Density Space Clustering Algorithm Based on Users Behaviors

Boqing Feng[1*], Mohan Liu[2], Jiuqiang Jin[2], Xiuyu Sui[3]

[1] Institute of Electronic Computing, Technology of China Railway Academy Group Co., Ltd., Beijing, China
[2] China Railway Network Company Limited, Beijing, China
[3] China Railway Wuhan Group Co., Ltd., Wuhan, China
fengboqing@rails.cn, liumohan@crnet.com.cn, jinjiuqiang@crnet.com.cn, xy.sui@163.com

Received: 1 February 2022; Revised: 7 April 2022; Accepted: 15 April 2022

Abstract. At present, insider threat detection requires a series of complex projects, and has certain limitations in practical applications; in order to reduce the complexity of the model, most studies ignore the timing of user behavior and fail to identify internal attacks that last for a period of time. In addition, companies usually categorize the behavior data generated by all users and store them in different databases. How to collaboratively process large-scale heterogeneous log files and extract characteristic data that accurately reflects user behavior is a difficult point in current research. In order to optimize the parameter selection of the DBSCAN algorithm, this paper proposes a Psychometric Data & Attack Threat Density Based Spatial Clustering of Applications with Noise algorithm (PD&AT-DBSCAN). This algorithm can improve the accuracy of clustering results. The simulation results show that this algorithm is better than the traditional DBSCAN algorithm in terms of Rand index and normalized mutual information.

Keywords: user behavior analysis, cluster analysis, detection efficiency

1. Introduction

1.1 Research Background

In the digital age, the complexity of information and technology makes more and more organizations vulnerable to both external attacks and internal threats.

Internet criminals can break through network boundaries such as firewalls, steal the keys of other users, or launch internal attacks by pretending to be insiders of the organization. Therefore, comprehensive defense has become the focus of research in many universities and enterprises. On the premise of finding the common characteristics of internal attacks, by modeling internal user behavior, internal threats that may cause losses to individuals, organizations, or countries can be detected and prevented in a targeted manner. Internal threat detection is different from previous external intrusions. The focus is on how to identify "may" or "already" anomalies from users authorized by the organization [1]. In addition, in real scenarios, the number of insider threat behaviors is small, and they are scattered in a large number of normal behaviors generated by users. Looking for abnormal data from massive amounts of data is diffcult. The existing internal threat detection algorithms are mainly machine learning, but this method relies heavily on feature engineering, so it is difficult to accurately capture the behavior differences between malicious users and normal users.

The Carnegie Mellon University cyber security team conducted an in-depth study of insider threats, which showed that existing insider threats are more risky and concealed than traditional attack methods. In addition, insider attacks are not limited to a single domain, a single point in time, or a single threat scenario. It can be seen that insider threat detection faces new challenges across domains and time nodes. Therefore, the goal of this article is as follows: find a suitable deep learning algorithm, use the algorithm to analyze the relationship between user psychology and internal threats from a psychological perspective, and organizations or institutions can establish a dual internal threat mechanism of prevention and dynamic monitoring by analyzing its correlation.

After investigating related technologies and algorithms, this paper combines psychological characteristics and internal threat density spatial clustering algorithm to analyze the correlation between internal user psychological characteristics and internal threats. This paper proposes a PD&AT-DBSCAN algorithm. This algorithm is improved from DBSCAN. First, select the optimal parameters by constructing the K-nearest neighbor matrix and analyzing the distribution characteristics of the data, and then combine the nearest neighbor and the reverse near-

* Corresponding Author

est neighbor to determine the initial point with the highest density, and use this point as the core point to cluster to improve the accuracy of clustering results.Compared with other algorithms, we optimize the clustering effect by setting the optimal parameters through the difference quotient analysis, which makes our results better.

2. Related Work and Theories

2.1 Overview of Insider Threat Detection

There is no unified definition of insider threat, and people are doing intrusion detection [2]. R Garfinkel et al. [3] define internal threat users from the perspective of whether they have computer and network authorization and whether they have internal knowledge. This is the first time the concept of internal users has been defined. William R. et al. propose an insider threat indicator to potentially identify malicious activity at or prior to an attack [4]. At present, it seems that this traditional definition method is obviously one-sided. Althebyan et al. propose a mobile edge computing mitigation model to address insider threats in the cloud [5]. According to the definition of the above researchers, insider threats are malicious behaviors made by internal users who have a certain understanding of the organization [6].

Among the different definitions of insider threats, everyone generally agrees with the definition of insider threats in the "Guiding Documents on Insider Threats" issued by the United States Computer Emergency Response Team (CERT) in 2012: insiders refer to the employees of the organization. Outsourcing personnel, partners, and these people are authorized to access the system and network [7]. The constant changing of technologies have brought to critical infrastructure organisations numerous information security threats such as insider threat. Critical infrastructure organisations have difficulties to early detect and capture the possible vital signs of insider threats due sometimes to lack of effective methodologies or frameworks [8]. So we have to take precautions.

2.2 DBSCAN Clustering Method

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering method proposed by Martin Ester et al. [9]. DBSCAN regards the data node with sufficient density area as the distance center and then continuously expands the area until it cannot be expanded. The algorithm is powerful, and it can successfully distinguish irregular graphs without specifying the number of clusters.

DBSCAN does not need to specify the number of clusters in advance and can find clusters of any shape. At the same time, the algorithm is based on the neighborhood parameters to determine the clustering results. Its neighborhood contains two important parameters: Eps and MinPts. The former is to define the radius of the neighborhood of a data point, and the latter is a condition that needs to be met when defining the data point as a core point, that is, the minimum number of points required to form a cluster. For simplicity, Eps and MinPts are abbreviated as ϵ and M, respectively. Consider the data set $X = \{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$, and introduce concepts: ϵ neighborhood; density; core points; boundary points; noise points; The direct density can be reached; the density can be reached; the density is connected.

Fig. 1 intuitively explains the definition of core points, boundary points and noise points, and it satisfies $X=X_c \cup X_{bd} \cup X_{noi}$.

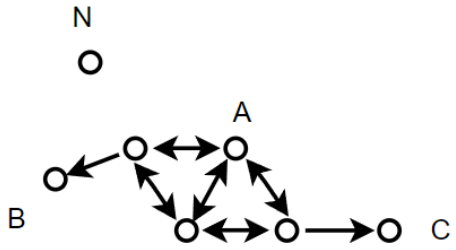

Fig. 1. Core point, boundary point and noise point

As shown in Fig. 1, the middle area has the highest density, and the density of point A is greater than 3, so the middle data is the core point. The density of the data points on both sides within their radius is less than the density corresponding to the red data points. Points B and C cannot continue to expand outward, so the data B and C are boundary points. N data points are not within the radius of any other data points, so N data points are noise points.

2.3 Principle of DBSCAN Algorithm

According to the relevant content described in the previous section, the core idea of the DBSCAN algorithm can be described as: judging whether each sample is a core point, and if it is a core point, classify the sample and the samples in its neighborhood.

From the above introduction, the input of the DBSCAN algorithm is divided into three parts. The first part is the data set D = $\{x_1, x_2, x_3, ...x_m\}$ that needs to be classified, and the second part is the neighborhood parameter (ϵ, MinPts), and finally One is the measurement method between data points. In practical applications, Euclidean distance is usually used. The output is divided cluster C=$\{ C_1, C_2, C_3,..., C_k \}$. The specific process of the algorithm is as follows:

Step 1: Set (ϵ, MinPts) neighborhood parameters

Step 2: Traverse the sample set and randomly select a point with this point as the center. If the distance between multiple data points and the center in other samples is less than ϵ, and the number of such samples is not less than MinPts, then this sample is the first A core point, stop traversing.

Step 3: The sample point selected in the previous step is the center, traverse the entire sample set other than that, find all the samples whose distance from the center is less than ϵ, and classify this into one category.

Step 4: Subtract the classified samples from the overall sample set, and then repeat steps 2 and 3 until the sample set to be classified is empty and the classification ends.

The DBSCAN algorithm has many advantages. It can cluster dense data sets of different shapes, and find noise points during clustering. At the same time, the traditional DBSCAN algorithm also has some shortcomings. Since boundary points can be connected by more than one cluster density, different data processing sequences may lead to different clustering results, so uncertainty is one of the problems of DBSCAN. The clustering effect of DBSCAN will be affected by the common problem of Euclidean distance dimensionality disaster. At the same time, it is very difficult to select the minimum number of samples MinPts for data with large differences in density. So how to choose the neighborhood distance ϵ and the minimum number of samples MinPts in the neighborhood is a very critical issue in the DBSCAN algorithm. For this reason, many scholars have designed various improved DBSCAN algorithms to optimize this problem [10-12].

3. Research on Internal User Psychology Based on PD&AT-DBSCAN Algorithm

3.1 PD&AT-DBSCAN Algorithm

In the traditional DBSCAN algorithm, the MinPts parameter is used to smooth the density estimation. In comparison, the setting of ϵ is difficult. J. Sander et al. [13] suggested using the parameter combination of (2*dim-1) nearest neighbor and MinPts=2*dim to determine the final clustering result. Ester et al. suggested using the distance value of the 4th nearest neighbor as ϵ in the two-dimensional data. In the DBSCAN algorithm, the neighborhood parameters (ϵ, MinPts) and the choice of initial points determine the clustering effect. Therefore, many experts and scholars at home and abroad have devoted themselves to studying the parameter optimization problem of DBSCAN. Aiming at the selection of neighborhood parameters, this paper improves on the original DBSCAN algorithm, and proposes a PD&AT-DBSCAN algorithm that dynamically selects neighborhood parameters.

The essence of the density-based clustering algorithm is to find the denser data set, and the distance between the data points in the set is relatively small [14]. Based on this, this paper constructs the K-nearest neighbors of the data points, and finds the optimal ϵ through the difference quotient analysis of the data set by constructing the similarity matrix between samples. [15] Based on the determination of ϵ, use the method of combining Mean matrix and statistical characteristics to determine MinPts. When selecting the initial point, this paper discards the method of randomly selecting a point as the initial point in the traditional DBSCAN clustering algorithm, but finds the best initial point through the nearest neighbor and the reverse nearest neighbor.

(1) K-nearest neighbor and reverse K-nearest neighbor [16]

Suppose X is a data set of size n, and $\forall x \epsilon X$, $x \epsilon R^d$. For any two observation points x, y that appear in the data

set, use the Euclidean distance formula $dist(x,y) = \left(\sum_{i=1}^{d}(y_i - x_i)^2\right)^{0.5}$ to return one of the two points The distance between. The K-nearest neighbor function is defined based on the distance formula of a given k, where 0≤k≤n-1.

The K nearest neighbor of point x is given by the function $N_k(x)$=N. The reverse K nearest neighbor formula for point x is $R_k(x)$=R, and needs to satisfy $R \subseteq X\backslash\{x\} \cap \forall y \in R, x \in N_k(y)$.

(2) Similarity matrix

According to the input data set D, calculate the similarity matrix $Matrix_{n*n}$, and the calculation formula is as follows:

$$Matrix_{n\times n} = \{dist(i,j) \mid 1 \le i \le n, 1 \le j \le n\} \quad \textbf{(1)}$$

In the above formula: n is the number of data points in the data set, $Matrix_{n\times n}$ and is a real symmetric matrix representing the Euclidean distance between data points.

(3) Hermite interpolation

No matter what kind of data set, there is a specific distribution function corresponding to it. In order to find the specific distribution, we often use the fitting function to approximate the distribution. When looking for a curve that can reflect the distribution of the data set to the greatest extent, using Hermite interpolation to fit is appropriate.

Suppose that the function f(x) has n+1 different points on the interval [a, b], and the function value and derivative value at $a \le x_0 \le x_1 \le x_2 \le \ldots < x_n \le b$ are known, then there is one The polynomial function whose degree does not exceed 2n+1 is called the Hermitian interpolation polynomial of f(x). The average difference of its n multiple nodes is $f[x_0, x_1, \ldots, x_n] = f^{(n)}(x_0)/n!$. Combine it with Taylor's formula $f(x) = \sum_{n=0}^{\infty} f^{(n)}(a)(x-a)^n/n!$ to get Elmi as shown below Special interpolation polynomial:

$$P_n(x) = f(x_0) + f(x_0)(x - x_0) + \cdots + f(n)(x_0)(x - x_0)^n / n! \quad \textbf{(2)}$$

3.2 PD&AT-DBSCAN Algorithm Design

This algorithm includes three parts: Eps, MinPts parameter setting and the selection of the initial point of clustering.

(1) Steps to determine Eps parameters:

Step 1: Construct a similarity matrix $Matrix_{n*n}$

Step 2: Calculate the value of each element in the similarity matrix through the data set, and then arrange them in ascending order row by row. $Matrix_{n*i}$ is the element in the nth row and the i-th column. The k_dist distribution can be obtained by also sorting the elements in the i-th column in ascending order.

Step 3: Select the kth nearest neighbor curve that best reflects the k_dist distribution, and take the k+1th column in $Matrix_{n*n}$ to analyze.

Step 4: Fit the curve by interpolation and find the inflection point from gentle to steep, and set the k_dist value at this point as the parameter Eps.

(2) Steps to determine MinPts parameters:

Step 1: From the perspective of the Eps neighborhood determined in the previous step, different data nodes have different Eps neighborhoods, so as to find the number of Eps neighborhoods P_i for each data point.

Step 2: Calculate the expected value of all points and use it as the neighborhood density threshold parameter of the data set:

$$MinPts = (\sum_{i=1}^{n} P_i)/n \quad \textbf{(3)}$$

(3) Steps to determine the optimal initial point:

Initialize k;

Step 1: Construct a similarity matrix $Matrix_{n*n}$

Step 2: Calculate the value of each element in the matrix through the data set, and then arrange them in ascending order row by row. The values in the first column are all 0. Take the values in the second column and calculate

the number c of the reverse nearest neighbors (k=1) of each point. The parameter c reflects the density value of the point.

Step 3: Calculate the distance from each sample to its k-th nearest neighbor by constructing the formula (4) shown below.

$$initialpt_x = c^{-1} + \sum_{i=1}^{k} dist_i \quad \textbf{(4)}$$

Step 4: Arrange the initialpt values of all points from small to large, and cluster the smallest point as the initial point. After clustering with this point as the center, delete the point and continue to select the point cluster with the minimum initialpt value.

The pseudo code of PD&AT-DBSCAN algorithm is shown in Table 1.

Table 1. Pseudo code of PD&AT CSDN

PD&AT-DBSCAN
Input: *D*: a data set containing several objects, ε: radius, MinPts: density neighborhood threshold
Output: a collection of clustersMark all objects in *D* as “unvisited”;
do
Select the point x with the largest initialpt_x as the initial point;
Mark x as “visited”;
Let N be the set of ε neighborhood objects of *x*;
for each object *q* in *N*
if q is “unvisited”
Then mark q as “visited”;
if *q* is the core object
Then add all the object sets in the ε neighborhood of *q* to *N*;
if *q* does not belong to other clusters
Then add *q* to cluster *C*;
End For;
Output *C*
until all objects in *D* are “visited”

(4) PD&AT-DBSCAN algorithm steps are as follows:

Input: Data set D containing n objects, neighborhood radius Eps, density threshold MinPts. Expected output: all generated clusters.

Step 1: Initialization parameters: set all objects in the database as noise and unvisited.

Step 2: Select the optimal initial point from the database as the core object, use this object as the starting object to create a new class, recursively find all objects whose density is reachable from the object, add it to the class, and mark it as Visited.

Step 3: Until all objects have been visited, output the clustering results, and the algorithm ends; otherwise, go to step 2.

4. Simulation and Result Analysis

4.1 Data Set Introduction

This article uses the psychometric.csv file in the CMU-CERTv4.2 data set. As shown in Fig. 2, the file mainly contains the five psychological personalities of 1,000 internal employees: O, C, E, A, N, that is, Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism.

employee_name	user_id	O	C	E	A	N
Calvin Edan Love	CEL0561	40	39	36	19	40
Christine Reagan Deleon	CRD0624	26	22	17	39	32
Jade Felicia Caldwell	JFC0557	22	16	23	40	33
Aquila Stewart Dejesus	ASD0577	40	48	36	14	37
Micah Abdul Rojas	MAR0955	36	44	23	44	25
Gail Rhiannon Mcconnell	GRM0868	21	25	20	13	28
April Alika Levy	AAL0706	37	14	28	13	25
Rama Vielka Clayton	RVC0232	34	20	47	38	25
Tasha Casey Dalton	TCD0009	44	28	44	38	23

Fig. 2. Information of psychometric.csv

In this experiment simulation, the description information of the experiment environment is as follows:
System environment: macOS operating system
Hardware configuration: 2.6 GHz Intel Core i7 16G memory 256G hard disk
Programming environment: python3.6
Compiler: Jupyter Notebook, pycharm

4.2 Evaluation Index

A good clustering effect should be that the distance between data points in the same class is small, and the distance between data points between different classes is large. Generally speaking, for the evaluation of clustering effect, external evaluation standards are commonly used. It is aimed at given a benchmark, according to this benchmark to evaluate the clustering results.

Before introducing the two external evaluation standards, first define the paired variables: a and b. a is the number of samples in the data set that belong to both the same cluster C and the same cluster K. b is the number of samples in the data set that do not belong to the same cluster C or the same cluster K.

(1) Rand Index (RI)

RI is used to measure the similarity between clusters, as shown in formula (5):

$$RI = (a+b) / C_n^2 \tag{5}$$

In order to solve the problem that the RI of the randomly assigned cluster vector is gradually increased with the increase of the number of clusters, the Adjusted Rand Index (ARI) is proposed. ARI calculation is shown in equation (6).

$$ARI = (RI - E[RI] / (max(RI) - E[RI]) \tag{6}$$

The range of ARI is [-1,1]. The closer the ARI is to 1, the better the clustering effect.

(2) Normalization Mutual Information (NMI)

The normalized mutual information measures the clustering effect by comparing the clustering result with the real label. The calculation is shown in formula (7).

$$NMI(U,V) = 2I(U,V) / H(U) + H(V) \tag{7}$$

4.3 Result Analysis

In order to evaluate the performance of the proposed algorithm, it is compared with the classic AF-DBSCAN and I-DBSCAN algorithms. In order to facilitate the visualization of the effect, we use a two-dimensional manual data set for effect comparison. The basic information of the data set is shown in Table 2.

Table 2. Manufacture dataset

Data set	Number of instances	Dimension	Number of categories
Compound	352	2	5
Aggregation	788	2	7

As shown in Fig. 3 and Fig. 4, from left to right are I-DBSCAN, AF-DBSCAN, PD&AT-DBSCAN, the clustering effects of the three clustering methods on the Compound and Aggregation data sets.

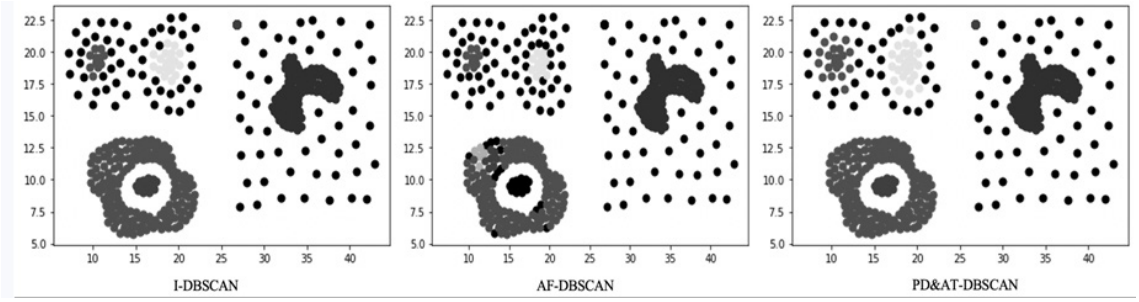

Fig. 3. Comparison effect on compound dataset

As can be seen from Fig. 3, for the Compound dataset, the PD&AT-DBSCAN algorithm performs better in clustering. A few data points in categories B and C of PD&AT-DBSCAN are classified as black dots, and the rest are correct. Classification. I-DBSCAN Second, AF-DBSCAN performs the worst. The data in the lower left part A is divided into other categories, and the clustering result is 8 categories, which is different from the number of categories in the original data set.

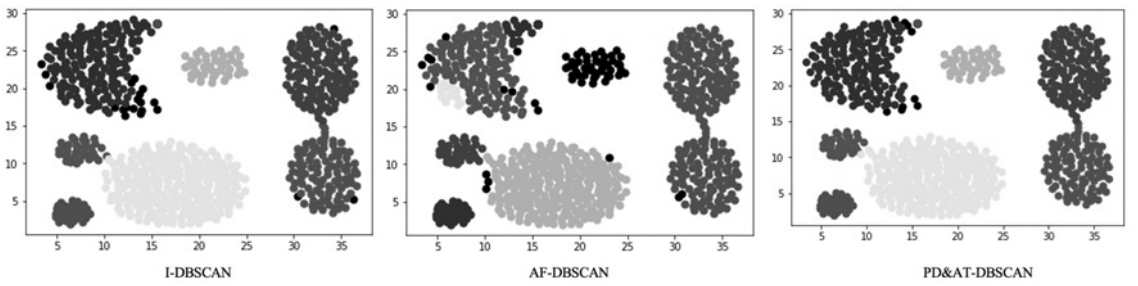

Fig. 4. Clustering effect on PD&AT-DBSCAN algorithm

It can be seen from Fig. 4 that for the Aggregation data set, the PD&AT-DBSCAN algorithm has the best clustering effect, followed by I-DBSCAN, and finally AF-DBSCAN.

The comparison of the three clustering algorithms on the artificial data set is shown in Table 3.

Table 3. Comparison of evaluation indicators of three algorithms on artificial datasets

Data set	Method	ARI	NMI	Cluster
Compound	PD&AT-DBSCAN	0.9039	0.8870	5
	I-DBSCAN	0.8696	0.8664	5
	AF-DBSCAN	0.7095	0.7065	8
Aggregation	PD&AT-DBSCAN	0.9795	0.9776	7
	I-DBSCAN	0.9611	0.9572	7
	AF-DBSCAN	0.8113	0.8604	9

It can be seen from Table 3 that compared to the traditional AF-DBSCAN algorithm and I-DBSCAN algorithm, the Rand index and normalized mutual information of the PD&AT-DBSCAN algorithm are both the highest.

The internal user psychological characteristic data set psychometric.csv is clustered using PD&AT-DBSCAN algorithm. By observing the data distribution, select the k=15 curve to fit the data distribution of the entire data set.

Use Hermite interpolation to fit the curve and select its inflection point as the corresponding function value as eps, and calculate minpts. After calculation, eps is 8.404 and minpts is 7.604.

The PD&AT-DBSCAN algorithm is used to cluster user psychological data in the internal threat data set, and the result is shown in Fig. 5.

Fig. 5. Clustering results of insider user's psychological situation

By analyzing the correspondence between cluster labels and user ids, it can be seen that among the 70 threatening users, 17 are located in the noise point of the "-1" label, and 53 are located in the category of the "0" label. Therefore, when a new employee joins, cluster analysis of his psychological situation. If the user's point belongs to the "-1" or "0" tag category, it is necessary to strengthen the supervision of the user's operation behavior to reduce the possibility of internal attacks. The simulation results show that this algorithm is better than the traditional DBSCAN algorithm in terms of Rand index and normalized mutual information.

5. Summary and Outlook

Internal attacks caused by improper operation of internal users can pose a serious threat to security. Considering its particularity, the related technologies of external threat detection cannot be directly applied to internal threat detection. Due to the lack of research on analyzing internal threats from the field of psychology, this paper proposes the PD&AT-DBSCAN clustering algorithm combined with psychological data of internal threats.

This paper combines psychological data and internal attacks, in order to solve the problem of unsatisfactory clustering effect caused by the fixed parameters in the traditional DBSCAN algorithm, proposes the PD&AT-DBSCAN algorithm. We set the optimal parameters to optimize the clustering effect through the difference quotient analysis. The simulation results show that the ARI and NMI indexes of the PD&AT-DBSCAN algorithm are higher than those of the traditional DBSCAN algorithm.

In the future, internal threats can be detected separately from the perspective of user behavior and psychological data, the fusion of D-S evidence theory gives the uncertain result that the user becomes a threat user.

Acknowledgement

This work has been supported by scientific research project "Research on Key Technology and Typical Application Scenarios of Railway Industry Internet Data Security Exchange and Sharing Service" (Project No.: 2021YJ203).

References

[1] Y. Bao, P. Song, G. Yu, Z. Zhang, Research on the technical architecture of personal information protection based on data security, China Security and Protection (Z1)(2022) 42-48.
[2] J. Peng, Q. Yu, M. He, WSN Intrusion Detection Technology Based on Traffic Prediction, Computer Applications and Software 33(2)(2016) 310-313.
[3] S.L. Garfinkel, A. Juels, R. Pappu, RFID privacy: an overview of problems and proposed solutions, IEEE Security & Privacy 3(3)(2005) 34-43.
[4] W.R. Claycomb, C.L. Huth, B. Phillips, L. Flynn, D. McIntire, Identifying indicators of insider threats: Insider IT sabotage, in: Proc. 2013 47th International Carnahan Conference on Security Technology (ICCST), 2013.
[5] Q. Althebyan, A Mobile Edge Mitigation Model for Insider Threats: A Knowledgebase Approach, in: Proc. 2019 International Arab Conference on Information Technology (ACIT), 2019.
[6] C. Probst, J. Hunker, D. Gollmann, M. Bishop (Eds), Insider threats in cyber security, Springer, Boston, MA, 2010 (pp. 115-137).
[7] D.M. Cappelli, A.P. Moore, R.F. Trzeciak, The CERT guide to insider threats: how to prevent, detect, and respond to information technology crimes (Theft, Sabotage, Fraud), Addison-Wesley, 2012.
[8] J. Ikany, H. Jazri, A Symptomatic Framework to Predict the Risk of Insider Threats, in: Proc. 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 2019.
[9] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proc. KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.
[10]K.M. Kumar, A.R.M. Reddy, A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method, Pattern Recognition 58(2016) 39-48.
[11]J.-H. Kim, J.-H. Choi, K.-H. Yoo, A. Nasridinov, AA-DBSCAN: an approximate adaptive DBSCAN for finding clusters with varying densities, The Journal of Supercomputing 75(1)(2019) 142-169.
[12]W. Lai, M. Zhou, F. Hu, K. Bian, Q. Song, A new DBSCAN parameters determination method based on improved MVO, IEEE Access 7(2019) 104085-104095.
[13]K. Xu, Research on Parallel Density-based Clustering Algorithm in Big Data, [dissertation], Jiangxi: Jiangxi University of Science and Technology, 2021.
[14]W. Li, S. Yan, Y. Jiang, S. Zhang, C. Wang, Research on Method of Self-Adaptive Determination of DBSCAN Algorithm Parameters, Computer Engineering and Applications 55(5)(2019) 1-7.
[15]Y. Zhao, H. Wang, Intelligent classification of railway communication equipment manufactures information based on cluster analysis, Railway Computer Application 27(7)(2018) 75-79.
[16]A. Bryant, K. Cios, RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates, IEEE Transactions on Knowledge and Data Engineering 30(6)(2018) 1109-1121.