**International Journal of Data Science and Analysis**
2022; 8(5): 163-168
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220805.17
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

# User and Entity Behavior Analytics Method Based on Adaptive Mixed-Attribute-Data Density Peaks Clustering

**Shihua Liu**[1, 2]

[1]School of Artificial Intelligence, Wenzhou Polytechnic, Wenzhou, China

[2]Wenzhou Network Security Detection and Protection Engineering Technology Research Center, Wenzhou, China

**Email address:**
Chaoshua@foxmail.com

**To cite this article:**
Shihua Liu. User and Entity Behavior Analytics Method Based on Adaptive Mixed-Attribute-Data Density Peaks Clustering. *International Journal of Data Science and Analysis*. Vol. 8, No. 5, 2022, pp. 163-168. doi: 10.11648/j.ijdsa.20220805.17

**Received**: October 6, 2022; **Accepted**: October 25, 2022; **Published**: October 29, 2022

**Abstract:** In the era of digital economy, new technologies emerge in an endless stream, and the network environment becomes increasingly complex. Traditional security products, technologies and solutions cannot meet the needs. In order to deal with the increasingly severe network security challenges, User and Entity Behavior Analytics (UEBA) technology provides a new solution. The application of new technologies such as statistical analysis, machine learning and deep learning also increases the adaptability and effectiveness of UEBA technology. User and entity behavior analysis technology based on machine learning has also become one of the research hotspots in current academia. In this paper, An User and Entity Behavior Analytics Method based on Adaptive Mixed-Attribute-Data Density Peaks Clustering is proposed. Firstly, the relevant access behavior data records of user entities are extracted from the access logs of the servers that need to be monitored. Since these records contain mixed attributes, adaptive mixed-attribute-data density peak clustering (AMDPC) can be used for clustering. Then, by constructing the user behavior baseline in each cluster, suspicious users and behaviors are analyzed. Combined with log backtracking and expert manual verification, the threat behavior is finally determined. This method has been applied in a company's network security situation awareness platform, and has achieved good practical results.

**Keywords:** User and Entity Behavior Analytics, UEBA, Density Peaks Clustering, AMDPC, Cybersecurity

## 1. Introduction

With the development of information technology, the era of the digital economy is coming at an accelerated pace. The application of new technologies such as cloud computing, big data analysis, industrial Internet technology, Internet of Things, intelligent algorithms etc. has made the network environment more diversified and the access mode diverse, and the network boundary gradually blurred or even disappeared. With the proliferation of enterprise data and complex personnel turnover. Digital transformation poses significant cybersecurity challenges [1]:

1) There are more and more external attacks, including advanced attacks driven by interests or states that are difficult to detect;
2) Various insider threats caused by malicious spies, negligent employees, lost accounts, and lost hosts;
3) The vulnerability and risk exposure of the digital infrastructure are increasing, and the changeable business needs continue to intensify;
4) Security teams are insufficiently staffed or have limited capabilities, and are deeply involved in asymmetric "security wars".

Traditional security products, technologies and schemes are basically based on known features for rule matching analysis and detection. Methods based on features, rules and manual analysis have security visibility blind spots, serious lag effect, inability to detect unknown attacks, easy to be bypassed, and difficult to adapt to the network reality of attack and defense confrontation, rapidly changing enterprise environment, external threats, and other problems.

Safety is a game of human aggression and defense. All intentions must be expressed through behavior. This is the most important and valuable piece of the puzzle in safe operation and it is also the most lacking in the traditional way. In view of the shortcomings of the traditional way, the security industry gradually strengthens the detection and analysis based

on big data drive, machine learning, probability analysis, pattern recognition, and other methods based on "behavior". In order to better detect potential threats and more accurately discover security problems, the technology of User and Entity Behavior Analytics (UEBA) is proposed. UEBA is developed on the basis of User and Behavior Analytics (UBA) and Security Information and Event Management (SIEM) [2], It is a threat detection method that analyzes internal and external network threats and comprehensively evaluates the risks faced by the system through multi-dimensional dimensions [3]. The concept of Entity added in it emphasizes the role of device behavior in network attacks and threat detection. Establishing a baseline according to user and entity behaviors and finding out abnormal behaviors of users and entities can not only realize internal behavior detection of enterprises, but also solve external network security problems [4]. UEBA technology has been applied in enterprise internal behavior analysis [5], host intrusion detection [6], user portrait research [7, 8], complex behavior modeling [9], recommendation system [10] and other fields. The application of machine learning algorithms to user and entity behavior analysis has also received more and more attention from academia and security industry.

In Section 2, this paper first highlights the concepts and techniques related to User and Entity Behavior Analytics (UEBA), then in the third section, the unsupervised learning-based user entity behavior analysis technology is introduced, the adaptive mixed-attribute data density peaks Clustering (AMDPC) algorithm proposed by the authors is introduced to the user and entity behavior analytics in Section 4. Finally, a brief summary is provided.

## 2. Introduction to UEBA

Gartner defined user entity behavior analytics in detail for the first time in its “Market Guide for User and Entity Analytics” in 2015 [11]: UEBA provide images and based on all kinds of analysis method of anomaly detection, usually is the basic analysis method (using the rules of the signature, pattern matching, simple statistics, threshold, etc.) and advanced analysis method (supervision and unsupervised machine learning, etc.), use the packaging analysis to evaluate the user and other entities (host, application, network, data base, etc.), Discover potential events related to activities that are abnormal from the standard portrait or behavior of the user or entity. These activities include abnormal access to the system by trusted internal or third-party personnel (user exceptions), or intrusions by external attackers that bypass security controls (abnormal users).

UEBA is a kind of model used to track and monitor abnormal behaviors of users, IP addresses, hosts, etc., which can analyze potential evil activities through context correlation of behaviors [3]. UEBA's analysis objects include user behavior and entity behavior. User behavior refers to user operations on terminal devices [12], such as using applications, interaction with data, click behavior, mouse movement, execution of command line statements, etc. Entity behaviors mainly refer to behaviors that cannot be directly correlated with real users [13], such as the operation logs of some apps themselves, the action records of virus Trojans, and the behavior tracks of some Advanced Persistent Threat (APT) [14].

UEBA is a detection method that focuses on behaviors and actions. Its core idea is to identify behaviors that deviate from the normal benchmark, and then deduce the possible results in reverse.

Data, Use Cases and Analytics are the three pillars of UEBA. As shown in Figure 1 below.

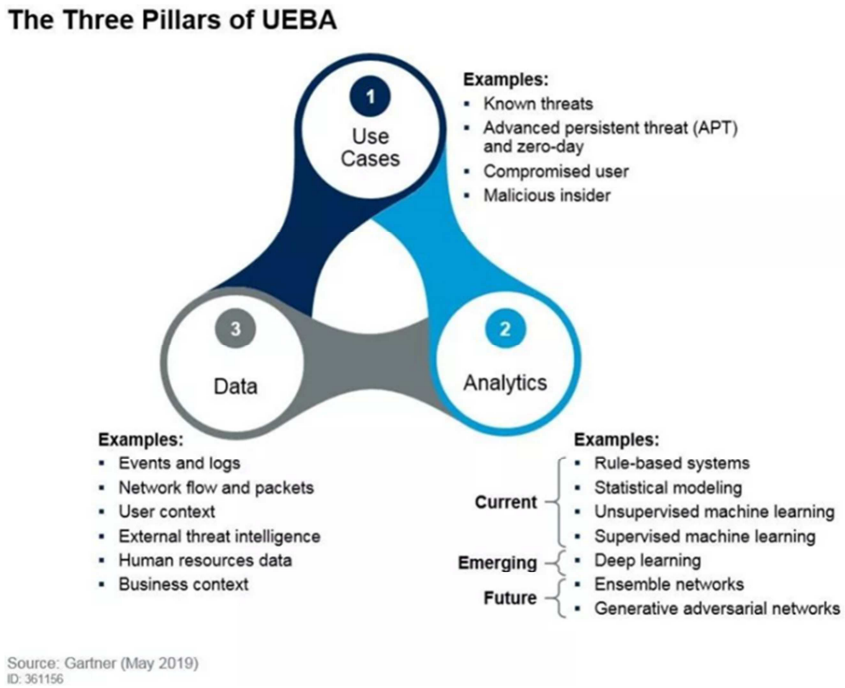


***Figure 1.*** *The Three Pillars of UEBA.*

1) Data

Data is the foundation of UEBA. Current data sources include events and logs generated by security devices, network traffic packets, external threat intelligence, and basic enterprise information. The quality of the collected data will directly affect the accuracy of the analysis results.

If you input a lot of garbage data, even if the model is good and the user case is rich, the output will be garbage or even misleading. In addition, the collection of data is insufficient to support the analyzed model, and it is difficult to produce satisfactory results. Of course, more data is not always better. More UEBA data will rely on the scene you expect to analyze for targeted selection and more accurate discovery of abnormal traces.

2) Use Cases

The application scenarios of UEBA are very rich, such as account anomaly detection [15], network security situation awareness [16], data leakage prevention analysis [17], data security internal control risk management [18] and other fields. Its focus is more on the field of internal security, such as abnormal login (time, place, frequency) of accounts, illegal operations of users on servers (upload, download, execute sensitive instructions), abnormal actions on terminals (initiate Intranet scanning, abnormal DNS query requests, large file transfer), etc.

3) Analytics

Most of the current UEBA analysis methods and models are still based on "rules" and statistical models, relying on historical events, expert experience and artificially set thresholds, and it is inevitable to adjust and optimize parameters according to different use environments. With the development of machine learning and deep learning technology, UEBA technology is mainly based on statistical learning, deep learning and reinforcement learning [19].

Statistical learning is composed of supervised learning, unsupervised learning, semi-supervised learning and other research categories. Unsupervised learning is the case where pointer pairs have data but no labels, which is mainly applied to cluster analysis, outlier detection, etc. In actual intrusion detection and behavior analysis, data is often generated without labels, and labeling data in some ways will cost certain resources. Unsupervised learning can learn certain rules from unlabeled data and use these rules to analyze new data. The representative algorithms for unsupervised learning are k-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Principal Component Analysis (PCA), and so on. When unsupervised learning methods such as clustering are used for threat identification in UEBA, the representative characteristics of most data samples are generally learned and classified according to the characteristics of the data itself. Finally, data samples that deviate from the group are analyzed to determine whether they are outliers. In recent years, various types of network devices and numbers show an increasing trend, and the massive data generated are difficult to label. Therefore, the research of unsupervised threat detection algorithm is one of the hot spots of current research.

# 3. Adaptive Mixed-Attribute-Data Density Peaks Clustering

## *3.1. Density Peaks Clustering Algorithm*

A clustering algorithm for fast search and discovery of density peaks (known as DPC algorithm) was proposed by A. Rodriguez and A. Laio [20]. This algorithm belongs to the density clustering algorithm in the above unsupervised learning. The algorithm has good clustering effect, high precision, high algorithm execution efficiency, and fewer parameters to be adjusted., can find the number of clusters, and can cluster data with different shapes. And can automatically identify outliers.

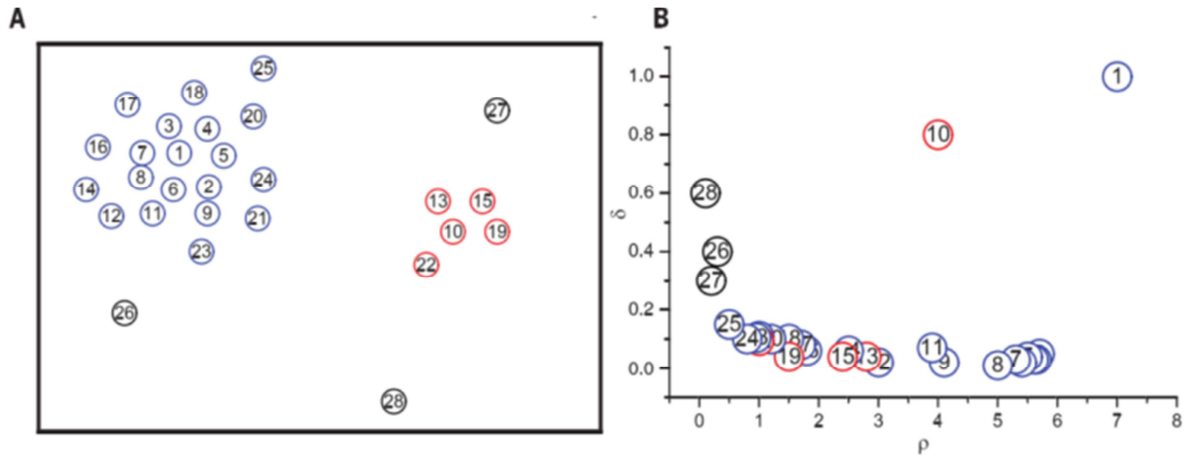


***Figure 2.*** *DPC algorithm sample dataset and its decision graph [20].*

The philosophical idea of DPC algorithm is that the cluster center point has a high local density, the local density of the points around it is lower than it, and the distance between the cluster center point and other cluster center points is relatively far. Therefore, the DPC algorithm constructs the decision graph by calculating the local density $\rho_i$ and the

relative distance $\delta_i$, so as to find the clustering center of the data set. The remaining data points in the dataset will be assigned to the nearest cluster whose local density is higher than its own. The sample data set of the DPC algorithm and its decision graph [20] are shown in Figure 2, points 1 and 10 are cluster centers for which the local density and the relative distance are both large.

$X= \{X_1, X_2, ..., X_n\}$ represents the data set to be clustered, which contains n data points. The distance between the data points $X_i$ and $X_j$ is defined as $d_{ij} = dist(X_i, X_j)$. For each data point $X_i$, its local density $\rho_i$ and its distance $\delta_i$ are defined as shown in Equations (1) and (2), where $\chi(d_{ij} - d_c) =$ 1when $d_{ij}-d_c<0$ and 0 otherwise, and $d_c$ is a cutoff distance.

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \tag{1}$$

$$\delta_i = min_{j:\rho_j>\rho_i}(d_{ij}). \tag{2}$$

$\delta_i$ is measured by computing the minimum distance between the point $X_i$ and any other point with higher density; For the data point with highest density, we take the relative distance is the maximum distance from this point to all other points. That is $\delta_i = \max_j(d_{ij})$.

When the data set has few data points, the local density is generally calculated by a Gaussian kernel, as shown in Equation (3).

$$\rho_i = \sum_{j\neq i} exp(-\left(\frac{d_{ij}}{d_c}\right)^2). \tag{3}$$

Given the local density $\rho_i$ and the relative distance $\delta_i$ for each data point, the user can draw a decision graph. Based on the decision graph, users can explicitly choose the cluster centers on the decision graph. After the cluster center is determined, other data points can be classified into the cluster of their nearest and denser neighbors.

### *3.2. Adaptive Mixed-Attribute-Data Density Peaks Clustering*

Since DPC algorithm requires users to manually select the cluster center point in the decision graph to determine the number of clusters, and cannot be directly applied to the clustering of mixed attribute data, Liu et al. studied its adaptive improvement and proposed an adaptive clustering method of mixed attribute data based on density peak AMDPC [21]. The algorithm uses the uniform distance metric of the mixed attribute data to construct the distance matrix, calculates the local density based on the K-nearest neighbor defined by DPC-KNN [22], and proposes an automatic clustering center determination method based on three inflection points of $\rho_i$, $\delta_i$, and their production $\gamma_i$, here $\gamma_i=\rho_i\times\delta_i$.

The local density calculation formula is expressed by Equation (4).

$$\rho_i = exp(-\frac{1}{K}\sum_{X_j\in KNN(X_i)} d^2(X_i, X_j)), \tag{4}$$

The specific process of the AMDPC is as follows, the mixed attribute dataset (DS) is feeded into the AMDPC, and the result is the cluster label vector (CL).:

Step 1. load the mixed-attribute dataset DS, and divide it into a numerical attribute subset *Dr* and a categorical attribute subset *Dc*, the real class label vector of the dataset is loaded at the same time.

Step 2. The distance between every two data points is calculated by using the unified distance metric formula.

Step 3. The local density $\rho_i$ and the relative distance $\delta_i$ of each data point according to Equation (4) and Equation (2) respectively.

Step 4. Calculate the product $\gamma$ of normalized $\rho$ and $\delta$: $\gamma_i=\rho_i\times\delta_i$.

Step 5. Select the cluster center points and set each point a different label using an adaptive method proposed in [21].

Step 6. Sort the original data points by local density descending order; then, the non-cluster center points are successively assigned the same class label as their nearest high-density points.

Step 7. Output the cluster label CL as the result.

The pseudo-code of the algorithm is shown in Algorithm 1 below:

```
Algorithm 1: AMDPC_clustering
Input: DS (the mixed-attribute data set).
Output: CL (cluster label vector)
    //Step 1. load the dataset DS and the ture label subset.
[Dr,Dc,truelabel]=loadseparate (DS);
    //Step 2: Calculate the distance.
distmatrix=distamdpc (Dr,Dc);
    //Step 3. Calculate ρi and δi according to Equation (2).
rho= kNNrho (distmatrix);
delta =calcdelta (distmatrix);
    //Step 4. find the cluster center.
Sc= findClusterCenter (rho,delta);
    //Step 5. Assign the class label.
    //Step 5.1. initialize the class label vector CL.
NCLUST=0;
for i=1 to sizeof(DS)
   CL(i)=-1;
End
    //Step 5.2. initialize the class label for center points;
for j=1 to sizeof(Sc)
    NCLUST=NCLUST+1;
    CL(Sc(j)) =NCLUST;
End
    //Step 6. complete the clustering
    //Step 6.1. sort the local density rho.
 Orho=sort(rho,descend);
    //Step 6.2. assign the class label for non-center points
for k=1 to sizeof(DS)
   if (CL(Orho(k))==-1)
      CL(Orho(k))=CL(nneigh(Orho(k)));
 End
    //Step 7. print the result CL.
 print(CL)
```

In algrithm 1, the function *findClusterCenter* (rho,delta) which is used to determine the cluster center point adaptively is described in [21].

# 4. User and Entity Behavior Analytics Method Based on AMDPC

In this paper, a user entity behavior analysis method based on density peak adaptive clustering is proposed, which is applied in a company's network security situation awareness platform, and has achieved good practical results. The results analyze several cases of abnormal user behavior, which ensures the company's network security.

The flow chart of the UEBA method based on the AMDPC is as follows:

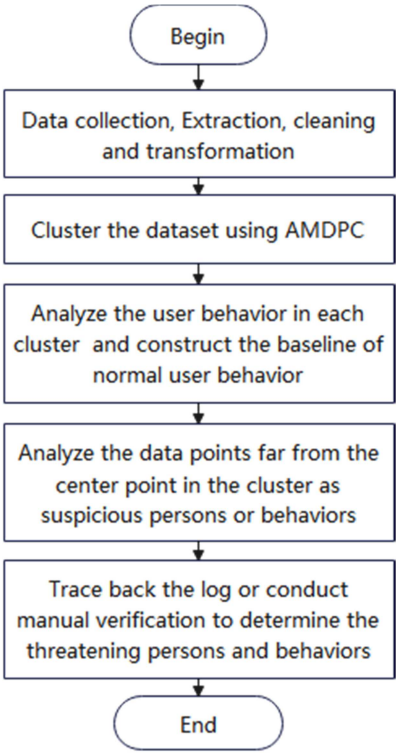


***Figure 3.*** *The flow chart of the UEBA method based on the AMDPC.*

As shown in the flowchart in Figure 3, the access behavior data of user entities need to be collected, cleaned, and transformed. The user access data of the servers in the analysis period is extracted from the log system for the servers that are monitored in focus, and the data set is constructed based on the access times and access behaviors. In the data set, each user or entity process forms a record at each visit, which serves as a data point to be analyzed.

Because the above collected data set may contain numerical attributes and categorical attributes, the AMDPC algorithm can be used to cluster the data sets to form clusters of similar user behaviors. Firstly, the local density of user data points and the distance between each other in the data set are calculated, the product of local density and distance is calculated, and the product is arranged from large to small in order to determine the inflection point. Then, the point before the inflection point is taken as the center point. Finally, the remaining data points are assigned to the proper cluster where the nearest cluster center point is located.

After the clustering is completed, the user behavior types in each data point cluster are analyzed and compared with the user information database. At the degree of same time, the matching between user rights and behavior types is analyzed to build the baseline user behavior.

Then, according to the user behavior baseline constructed above, the abnormal behavior information of the user is analyzed, and the user is judged as a suspicious person. The association analysis is performed on the suspicious person, and the expected associated person is analyzed.

Finally, according to the list of suspicious persons screened, log search is used to backtrack their operations and finally confirm their threatening behaviors.

Among them, the specific steps to build the user behavior baseline are as follows:

Step 1: Identify whether the user is a new user according to the user information data;

Step 2: if the user is a new user, extraction of the user's permissions, according to the behavior of the default permissions model as a baseline user behavior;

Steps: 3: if the user is old users, from the user information in the database data, draw on the history of the old user behavior analysis of historical data, get the behavior model as a baseline user behavior.

Through the above steps, it can effectively identify the encroachment caused by external intrusion or new intrusion, user rights appropriation, and old user change, and greatly increase the network security performance.

# 5. Conclusion

User and Entity Behavior Analytics (UEBA) integrates security analysis of network traffic logs and terminal logs, identifies abnormal behaviors, and associates alarm attack behaviors to make threat event analysis "understand", providing a new idea to help users discover hidden threats. Machine learning is the most core technology in UEBA. The research, improvement and application of density peak clustering algorithm in machine learning algorithm are also the research hotspots in the academic field. This paper proposes to apply the Adaptive Mixed-Attribute-Data Density Peaks Clustering (AMDPC) algorithm to user entity behavior analysis, which integrates adaptive clustering, baseline construction and evaluation, log backtracking and expert analysis, and realizes the screening of suspicious users and abnormal behaviors in enterprise network security analysis, and achieves good practical results.

Due to the complex and diverse application scenarios of security, it is not realistic to completely rely on machine learning algorithms to automatically complete user internship behavior analysis that meets the requirements of all scenarios. In order to achieve better results, the participation of domain experts is also needed. How to use technical means to strengthen the practicability and accuracy of user entity behavior analysis and reduce user investment is the direction of further research in the future.

## Acknowledgements

This research was supported by the Qingshan Lake Science and Technology City Joint Fund of Zhejiang Provincial Natural Science Foundation of China under Grant No. LQY19F020001 and the Wenzhou Polytechnic Major Scientific Research Projects [Grant No. WZYSDCY2018002].

## References

[1] Security Research Institute of China Academy of Information and Communications Technology, Hangzhou Anheng Information Technology Co., LTD., "User Entity Behavior Analytics Technology (UEBA) (2020).", 2020.

[2] SINGH K, SINGH P, KUMAR K. User Behavior Analytics-Based Classification of Application Layer HTTP-GET Flood Attacks [J]. Journal of Network and Computer Applications, 2018, 112 (15): 97-114.

[3] SHASHANKA M, SHEN M-Y, WANG J. User and entity behavior analytics for enterprise security [C] // IEEE International Conference on Big Data. Washington, DC, USA: IEEE, 2016: 1867-1874.

[4] GUPTA R, TANWAR S, TYAGI S, et al. Machine Learning Models for Secure Data Analytics: A taxonomy and threat model [J]. Computer Communications, 2020, 153: 406-440.

[5] WEN Yu, WANG W, MENG D. Mining User Cross-Domain Behavior Patterns for Insider Threat Detection [J]. Chinese Journal of Computers, 2016, 39 (8): 1555-1569.

[6] LI Z, SONG L. Research on internal threat detection based on user window behavior [J]. Computer Engineering, 2020, 46 (4): 135-142, 150.

[7] YANG A, ZHUANSUN Y, LIU C, et al. Design of Intrusion Detection System for Internet of Things Based on Improved BP Neural Network [J]. IEEE Access, 2019, 7: 106043-106052.

[8] AHMIM A, DERDOUR M, FERRAG M A. An intrusion detection system based on combining probability predictions of a tree of classifiers [J]. International Journal of Communication Systems, 2018, 31 (9): 1-17.

[9] BELOUCH M, EL S, IDHAMMAD M. A Two-Stage Classifier Approach using RepTree Algorithm for Network Intrusion Detection [J]. International Journal of Advanced Computer Science and Applications, 2017, 8 (6): 389-394.

[10] SHAILENDRA SINGH, ABDULSALAM YASSINE. Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting [J]. Energies, 2018, 11 (2): 1-26.

[11] Gartner, 'Market Guide for User and Entity Behavior Analytics', Gartner, 2019. https://www.gartner.com/en/documents/3917096.

[12] MIAH S J, VU H Q, GAMMACK J, et al. A Big Data Analytics Method for Tourist Behaviour Analysis [J]. Information & Management, 2017, 54 (6): 771-785.

[13] WANG K, ZHENG H, LOURI A. TSA-NoC: Learning-Based Threat Detection and Mitigation for Secure Network-on-Chip Architecture [J]. IEEE Micro, 2020, 40 (5): 56-63.

[14] XU S, QIAN Y, HU R Q. Edge Intelligence Assisted Gateway Defense in Cyber Security [J]. IEEE Network, 2020, 34 (4): 14-19.

[15] Mo Fan, He Shuai, Sun Jia, Fan Yuan, and Liu Bo, "Application of user entity behavior analysis technology based on Machine learning in account anomaly detection," Communications Technology, Vol. 53, No. 5, pp. 1262 -- 1267, 2020.

[16] Xu Fei, "Status and Development Analysis of network security Situation Awareness Technology Based on UEBA," Network Security Technology and Application, No. 10, pp. 10 -- 13, 2020.

[17] Shaoyong Hu, "Data Leakage Analysis Based on UEBA," Information Security and Communication Confidentiality, No. 8, PP. 26-28. 2018.

[18] Liu Jin, Li Jiangbo, and Ye Bing, "Research on the Internal Control Risk Management of UEBA Data Security," Cyberspace Security, Vol. 12, No. Z3, pp. 43-48 +55, 2021.

[19] Cui Jing-yang, Chen Zhen-guo, Tian Li-qin, and Zhang Guang-hua, "A survey of user and entity behavior analysis techniques based on Machine learning," Computer Engineering, pp. 1-20, 2021, doi: 10.19678/j.issn.1000-3428.0062623.

[20] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science, vol. 344, no. 6191, pp. 1492-1496, June 2014, doi: 10.1126/science.1242072.

[21] S. Liu, 'Adaptive Mixed-Attribute Data Clustering Method Based on Density Peaks', Complexity, p. 13, 2022.

[22] M. Du, S. Ding and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis", Knowledge-Based Systems, vol. 99, pp. 135-145, May 2016, doi: 10.1016/j.knosys.2016.02.001.