K-means Optimized Network Security Management in the Internet Context

Cuijuan Liu* and Liya Song

School of Information and Engineering, Hebei GEO University, Shijiazhuang, 050031, China
E-mail: liucuijuan@hgu.edu.cn
**Corresponding Author*

Received 11 May 2024; Accepted 10 October 2024

Abstract

The rapid development of the Internet has had a broad and profound impact on humanity, making information acquisition and dissemination more convenient. It has also brought significant opportunities and benefits to business and the economy. However, there are some issues, such as personal factors and data security concerns. In order to solve the above problems, the K-means algorithm is optimized from the perspectives of K-value validity index, feature weighting and three-branch decision making. First, the optimal clustering results are determined according to K-value validity index, and the influence of different dimensional features on clustering is considered for feature weighting, and the uncertain objects in the three-branch decision deadlock class are divided into the boundary domain. The delay decision of the boundary domain data is carried out, and the K-means clustering optimization algorithm is improved by combining the above three aspects, and the intelligent network security management system is developed on this basis. The results showed that the K-means optimization algorithm achieved the highest average accuracy rate, adjusted Morandi index, and adjusted mutual information across various datasets, with values of 96.01%, 0.866, and 0.869, respectively. In practical network attack scenarios, the K-means optimization

Journal of Cyber Security and Mobility, Vol. 13_6, 1467–1490.
doi: 10.13052/jcsm2245-1439.13611

algorithm attained an attack threat recognition accuracy of 94.38%. Under unknown network attack types, its detection rate and false alarm rate were 94.63% and 1.32%, respectively. Surveys conducted post-implementation of the intelligent network security management system indicated that over 90% of users were satisfied with their experience of the system. In summary, the proposed method accurately identifies potential network threats in network data, fulfilling performance requirements for network security management systems and ensuring the security of network resources.

Keywords: Internet, K-means optimization, Network security management system, Three-way decision.

1 Introduction

The Internet, serving as a carrier and transmission system for information, is the bond connecting global computers, integrating news, communication, entertainment, and resource sharing [1–3]. Enterprises can not only obtain massive business information from the Internet but also present themselves to the international community through it [4, 5]. However, it inevitably brings many issues, such as information security, leakage of confidential information, and network management problems. The K-means clustering algorithm (K-means) is advantageous in data processing and analysis in network security management due to its simple principles, fast clustering speed, etc., but it has drawbacks like the tendency to converge to local optima and the need for presetting relevant parameters [6, 7]. Many scholars have conducted in-depth analysis and discussion on this issue. Pullagura I. addressed network information security concerns by combining the K-means algorithm with an ordered weighted averaging method and ensemble models to create a new hybrid algorithm. The research demonstrated satisfactory results in parallel usage and provided an accurate solution for the imbalance learning of intrusion detection systems [8]. Jian Y. and team designed a hybrid method combining a sparse autoencoder with extreme learning machines to reduce the dimensionality of abnormal data features caused by network intrusions. The study showed that, compared to K-means algorithm and support vector machines, their proposed method exhibited superior performance in the effective detection and identification of abnormal data [9]. In order to ensure the efficiency and security of key application services of wireless sensor networks, Gulganwa P et al proposed an energy-saving and safety-weighted clustering algorithm based on data-driven and machine learning,

and designed a centralized intrusion system detection on this basis. The simulation results show that the detection accuracy rate is as high as 90%. In addition, in the real-time scenario, the research method can replicate about 75% of the performance, and the superiority of the research method is verified from the performance of network service quality such as packet transmission rate, throughput and energy consumption [10]. The improvement measures of the above research provide certain reference value for the use of feature weighted optimization K-means algorithm in this research.

Based on the above content, it can be found that the application of K-means and other clustering algorithms in the context of network security has achieved great results and greatly improved the accuracy and convenience of data mining. However, the performance of K-means algorithm currently applied in network security still has a large room for improvement. Aiming at the problems existing in the application of K-means algorithm in network security, the study first integrates three-way decision-making to make the K-means algorithm more aligned with human cognition. It then optimizes the algorithm further through feature weighting and K-value effectiveness indicators, resulting in the K-means Optimization Based on Three-way Decision-making (TDO-K-means). Finally, an Intelligent Network Security Management system (INSM) is designed based on this. The study aims to rapidly identify abnormal behaviors and potential threats in networks using data mining techniques, thereby enhancing network security defenses and providing new perspectives in the field of network security. The innovations of the study are twofold: firstly, the optimization of the K-means algorithm through the introduction of three-way decision-making, feature weighting, and K-value effectiveness indicators; secondly, the application of the TDO-K-means algorithm in network security defense to enhance the security defense capability of network data. The contribution of the research is to build a more intelligent, deep and powerful network security defense system to ensure the application of big data security, and further improve the convenience of human society to work, live and learn. The structure of the study is divided into four parts. The first part is a summary of related research results. The second part is the design of TDO-K-means algorithm and the development of INSM system. The third part is the validity and feasibility verification of the proposed method. The last part is the summary of the research. Through research, we hope to solve the problem of personal information and property security by strengthening network security protection, ensure the safe operation of commercial and financial services, and promote the sustainable development and innovation of information and communication technology.

2 Related Works

Due to the diverse and open nature of the Internet, it is susceptible to attacks such as hacking and viruses. To prevent and mitigate these attacks and ensure the security of network information, the Network Security Management System plays a significant role. Ferrag M. A. and others proposed a set of indicators for assessing network security threats, specifically for evaluating the performance of intrusion detection systems in the context of Agriculture 4.0. Their research demonstrated that the proposed evaluation indicators could effectively assess network security intrusion detection [11]. Shu L. and colleagues observed that wireless sensor and control systems become more intelligent after connecting to the Internet. To enhance data transmission efficiency, computational and storage functions are shifted to edge devices. However, this intelligent integration poses challenges to the reliability and security of sensor and control systems [12]. Elmorshidy A. aimed to evaluate factors influencing mobile applications used for accessing and controlling home internet security cameras. The study collected survey data from 397 mobile users in Southern California and employed a structural equation model for hypothesis testing. The results indicated that when users access home and office security cameras, an increase in security controls leads to improved convenience, flexibility, and privacy [13]. Wang P. and team addressed potential information security issues in IoT applications by designing a blockchain-based IoT data security storage model. Simulation results demonstrated the effectiveness, scalability, and improved data security protection of the proposed model for IoT applications [14].

The K-means algorithm is commonly applied in network anomaly detection. It clusters access requests into groups, distinguishing between normal and abnormal access patterns. Sirisha A. and others, acknowledging the exponential growth of network security threats with internet usage, experimented with K-means algorithm, random forests, naive Bayes, and other machine learning algorithms. The results suggested that, compared to other algorithms, K-means algorithm performed better [15]. Vedavathi N. proposed a butterfly weed optimization algorithm by combining intrusion weed optimization with butterfly optimization. They used a rough K-means algorithm for course grouping, and the results indicated that the proposed method could offer suitable course recommendations to learners [16]. Aiming at the problem of effective channel estimation in industrial Internet of Things communication, Wang H et al. proposed a block SBL algorithm that uses the block sparse structure of sparse multipath channel model to estimate

channel performance. Computer simulation results prove the robustness of the proposed method in the filter group multi-carrier biased orthogonal AM system. The research method can obtain lower mean square error and bit error rate [17]. Abolfathi M et al. aimed to better understand the privacy vulnerabilities of HTTPS traffic in order to cope with the evolving traffic analysis attacks, so the research designed an HTTPS website fingerprint attack model and HTTPS confusion defender for super learner attacks. The research results showed that the HTTPS traffic had a high accuracy of more than 97%. Superior to existing attack models, HTTPS obfuscation Defender significantly reduced the accuracy of website fingerprinting from 97.2% to 2.89% [18].

In summary, there is a substantial body of research on the application of the K-means algorithm in network intrusion detection. However, the corresponding performance has been found to be suboptimal. Therefore, the research proposes the introduction of three decision criteria, feature weighting, and the effectiveness index of the K value to optimize the K-means algorithm, leading to the construction of the INSM system.

3 Construction of INSM System Based on TDO-K-means

The advent of the Internet era has permeated various industries worldwide, significantly altering people's lifestyles. However, the accompanying challenges in network information security have become a major concern for the development of the Internet. To address these issues, this research introduces three decision branches and proposes a TDO-K-means algorithm. Based on this algorithm, an INSM system is constructed.

3.1 TDO-K-means Algorithm

With the rapid development of Internet technology, the global number of Internet users has surpassed 5 billion, constituting 63% of the world's population. The Internet has become an indispensable part of people's production and daily life. Serving as a crucial engine for global economic development, it provides limitless business opportunities across various industries and a broader platform for innovation, contributing to societal progress and development [19, 20]. However, the escalating prominence of network security issues and the increasing stealthiness of cyber attacks, coupled with prolonged attack durations, pose significant challenges. Traditional network security management primarily relies on passive defense, initiating detection

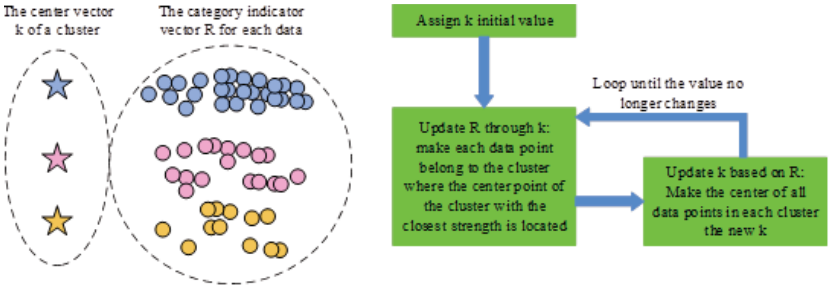


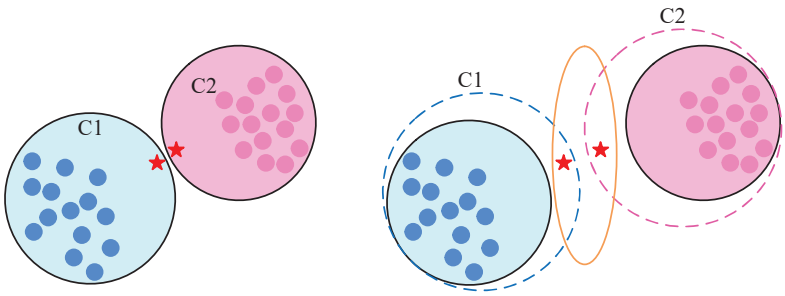
Figure 1 Process diagram of K-means algorithm.

and mitigation tasks only after an attack has commenced. Consequently, there is a need to augment existing network security management systems with proactive security defense mechanisms to enhance overall security capabilities. The K-means algorithm, categorized as an unsupervised clustering analysis algorithm, proves effective in establishing patterns for detecting network threats through learning. This facilitates the identification and detection of network security attack behaviors from extensive data packets. The schematic diagram of the K-means process is illustrated in Figure 1.

In Figure 1, the process begins by selecting K points from the dataset as initial cluster centers. Subsequently, the distance between each sample point and its corresponding cluster center is computed to determine the shortest distance and assign the sample point to the relevant cluster. The mean of all sample points in each cluster is then calculated to establish new cluster centers. This process iterates until the cluster centers stabilize, and the objective function converges, signaling the conclusion of the algorithm. The clustering analysis problem essentially involves optimizing the objective function, which is computed using the Error Sum of Squares (ESS), as shown in Equation (1).

$$ESS = \sum_{i=1}^{K} \sum_{x \in C_i} (x - \bar{x}_i)^2 \tag{1}$$

In Equation (1), $\bar{x}_i$ represents the mean of class C_i. While the K-means algorithm is efficient, simple, and exhibits good scalability when handling large datasets, its relationship between identified objects and classes is binary – either belonging or not. However, in practical applications, there may be uncertain objects within clusters. Forcing their assignment to a cluster can lead to high decision risks, and determining their values is challenging, often relying on empirical ranges. Additionally, selecting initial cluster centers is



(a) Two branch clustering effect (b) Three branch clustering effect

Figure 2 The clustering effect of two branch decision and three branch decision.

difficult and can result in local optimization problems if there is a significant difference between them and the final initial cluster centers. To address these issues, the research focuses on optimizing from three perspectives: uncertain object points, K values, and initial cluster centers. An improved approach for uncertain object points utilizes a three-decision method, where a two-decision model typically renders judgments in a black-and-white manner, while a three-decision model adds an additional judgment, reflecting human cognition more accurately. The clustering effects of two-decision and three-decision models are illustrated in Figure 2.

In Figure 2, there are two points with significant differences from their corresponding clusters. The two-decision approach leads to an unreasonable division, whereas the three-decision method places them in a separate region for processing. The optimization steps for the three-decision method are as follows: firstly, input the dataset into the K-means algorithm to obtain the two-decision clustering result set C. Next, set 1/10 of the mean of all class samples as the neighborhood scale, defining the q neighborhood. Then, iterate through C and data objects for partitioning, ultimately obtaining the result set C' of core and boundary regions. Subsequently, perform a second K-means clustering on the boundary region result set to obtain the three-decision clustering result set C'. For the improvement of K values and initial cluster centers, the study employs feature weighting and effectiveness metrics to allocate weights to the features of network data. Let X be a dataset with N data points in M dimensions, and the relationship between the variance σ_k^m and mean μ_k^m of class k in dimension k is described in Equation (2).

$$\sigma_k^m = \frac{1}{N_k} \sum_{i=1}^{N_k} (x_{im} - \mu_k^m)^2, x_{im} \in X^d \tag{2}$$

The total sum of means of X for all clusters in dimension m is calculated as shown in Equation (3).

$$S_{\mu}^{m} = \sum_{i=1}^{K} \sum_{j=1}^{K} (\mu_i^m - \mu_k^m)^2, i \neq j, m = 1, 2, \ldots, M \tag{3}$$

The total sum of variances for all clusters in dimension m is calculated as shown in Equation (4).

$$S_{\sigma}^{m} = \sum_{i=1}^{K} \sum_{j=1}^{K} (\sigma_i^{m^2} - \sigma_k^{m^2}), i \neq j, m = 1, 2, \ldots, M \tag{4}$$

To evaluate the clustering quality in dimension m, the research introduces the evaluation parameter $R(m)$, which represents the proportion of inter-cluster distance to intra-cluster dispersion, as shown in Equation (5).

$$R(m) = S_{\mu}^{m} / S_{\sigma}^{m}, m = 1, 2, \ldots, M \tag{5}$$

The calculation of the clustering weight for this feature is expressed in Equation (6).

$$\omega(m) = \frac{R(m)}{\sum_{1}^{M} R(m)} \tag{6}$$

The distance feature weighting expression for any x_i to the cluster center c_j is given by Equation (7).

$$d(x_i, c_j) = \sum_{m=1}^{M} \omega_m \bullet d(x_{im}, c_{jm}) \tag{7}$$

In Equation (7), $d(x_{im}, c_{jm})$ represents the distance in m dimension from x_i to cluster center c_j. In the optimization of K value effectiveness metrics, the study first defines a dataset A, where intra-cluster distance $d_{in}(C_i)$ is calculated as shown in Equation (8).

$$d_{in}(C_i) = \sum_{x,y \in C_i} \sum_{m=1}^{M} \omega_m \bullet |x - y|^2 \tag{8}$$

In Equation (8), x and y are both data objects of class C_i. The inter-cluster distance $d_{ic}(C_i, C_j)$ from C_i to C_j is described in Equation (9).

$$d_{ic}(C_i, C_j) = \sum_{i \neq j, j=1}^{k} \frac{1}{bg} \sum_{x \in C_i, y \in C_j} \sum_{m=1}^{M} \omega_m \bullet |x - y|^2 \tag{9}$$

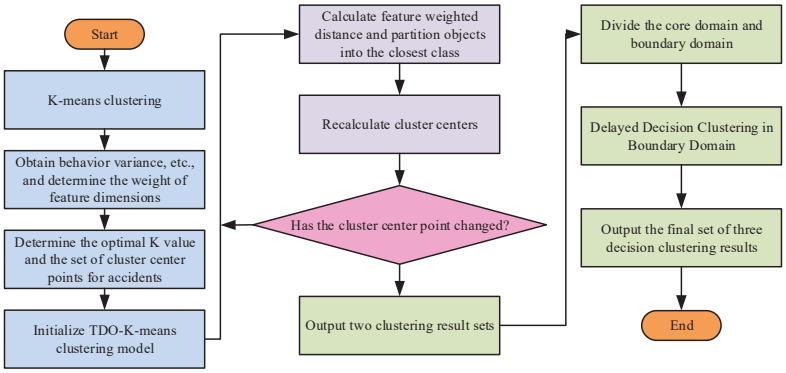


Figure 3 Process diagram of TDO-K-means algorithm.

In Equation (9), b and g are the respective quantities of data points corresponding to C_i add C_j. Based on this, the K value effectiveness metric S is obtained, as expressed in Equation (10).

$$S = \frac{\sum_{i=1}^{k} d_{in}(C_i)}{\sum_{i=1}^{k} \sum_{i \neq j, j=1}^{k} d_{ic}(C_i, C_j)} \tag{10}$$

The calculation for the optimal number of clusters is given in Equation (11).

$$k_{best} = \arg \min_{2 \leq k \leq \sqrt{n}} S \tag{11}$$

Combining the aforementioned optimization steps results in the TDO-K-means algorithm, as depicted in the schematic diagram in Figure 3.

In Figure 3, it is necessary to first use the K-means algorithm to obtain the mean of all data objects for different feature dimensions in X. Subsequently, through optimization methods, determine K values and initial cluster centers to initialize the K-means algorithm model. Then, compute $d(x_i, c_j)$ and classify, followed by updating cluster centers. If the cluster centers continue to change, recalculate $d(x_i, c_j)$ and classify. Otherwise, C can be obtained, and after performing a second K-means clustering on the boundary region, the final clustering result C' is obtained.

3.2 Design of the INSM System Based on the TDO-K-means

The Internet has profoundly changed the way people access and disseminate information. However, there are significant security vulnerabilities in current network security management, such as malicious attacks on corporate

networks, theft of confidential files and data, posing a challenging problem in the field of cybersecurity. Traditional network security management technologies include antivirus software, firewalls, and access control lists. Antivirus software scans critical areas such as files and memory in a computer, searching for virus characteristics that match known virus databases. Common antivirus software includes 360 Total Security, Huorong Security Defender, and Kaspersky. However, with the continuous mutation and upgrading of viruses, some may evade the detection mechanisms of antivirus software, leading to delayed discovery and removal. In severe cases, viruses may even attack the antivirus software itself, disrupting its operational environment. Firewalls primarily protect the security of internal networks by controlling network access. They are typically deployed at the boundary between internal and external networks, filtering and inspecting data packets entering the internal network, allowing only packets that comply with security rules to pass. Popular products include Sangfor, Topsec, and Huawei. Access control lists filter data packets on interfaces based on set conditions, controlling user access to the network. However, these traditional technologies currently cannot meet the growing security demands of the Internet. Therefore, there is a need to introduce big data technology and optimize the defense capabilities of network security systems through data mining techniques. The proposed TDO-K-means algorithm can be applied to the INSM system, thereby enhancing its security performance. With the integration of the TDO-K-means algorithm into the INSM system, analysis can be conducted based on actual requirements. This leads to the identification of functional requirements for the INSM supported by TDO-K-means, as illustrated in Figure 4.

In Figure 4, the INSM system mainly comprises three functions: acquiring, analyzing, and post-processing for data. Data acquisition is the most crucial and essential function, involving the collection of data from various sources such as log data, software data, and network data streams. It specifically includes Transmission Control Protocol (TCP), User Datagram Protocol (UDP), log data, device data, and other network data. In the data analysis section, the TDO-K-means algorithm is introduced to enhance the real-time security defense of the INSM system for identification, detection, and analysis. In the post-processing section, hash verification is constructed to provide human-machine interaction functionality for the data mining component. Based on the analysis of these functionalities and the design of the TDO-K-means algorithm, the structure schematic of the INSM supported by TDO-K-means is depicted in Figure 5.

Figure 4 Functional requirements of INSM supported by TDO-K-means.

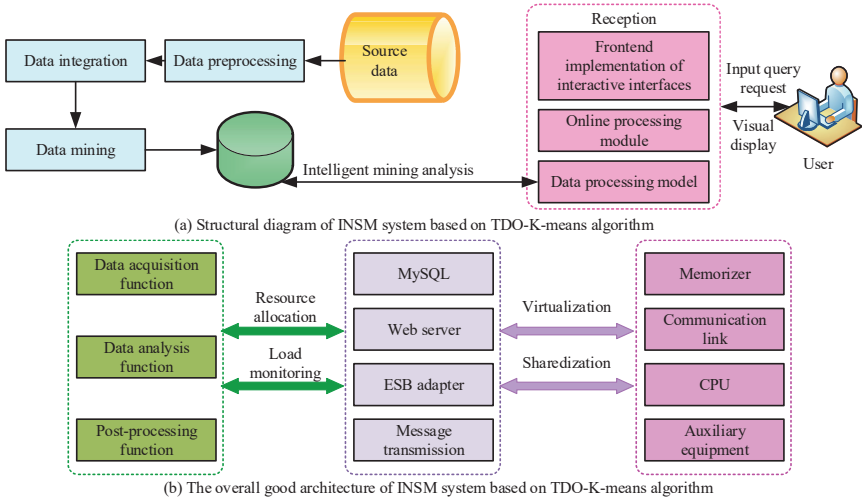


Figure 5 The structure and overall good architecture of INSM system based on TDO-K-means algorithm.

In Figure 5, the first step is to establish a data analysis engine using the TDO-K-means algorithm to explore potential viruses and other network security threats in massive data. Secondly, the backend includes two modules: offline and data processing. The offline module can collect, preprocess, and mine raw network data. If potential viruses are found during the data mining process, they can be stored in the virus pattern library. The real-time interaction in the front-end section consists of online modules, which establish a connection between users and rules to complete real-time interacting. In addition, the offline module in the INSM system can provide support for the online module, thereby completing the management. In the systematic operation, the front-end real-time interaction and back-end logical business processing are logically analyzed, with a focus on the back-end part mining and analysis of data, to ensure that the system can meet the current real-time transmitting requirements and accurate service. In the overall architecture of the INSM system based on the TDO-K-means algorithm, the INSM system will assign accounts to users, who can enter the system through the login terminal and connect to the network security data acquisition module to realize data mining and analysis. After mining results everywhere, the defense measures can be further configured and started. After the INSM system is loaded, the relevant business data can be stored in the database and processed by the Web logic business of the trap. Then it can integrate different types of data, such as network log data, device operation data, etc., by dividing the data analysis results into different levels to provide users. Human-machine interaction allows users to communicate directly with the INSM system, thus achieving powerful information processing and processing services. The composition structure of the final database section is shown in Figure 6.

In Figure 6, the INSM system's database primarily consists of four entities: security policies, security administrators, network management departments, and network operation logs. Combining the above content, the INSM supported by TDO-K-means is derived. To better evaluate the clustering performance of the TDO-K-means in the INSM system, common metrics such as entropy and purity are chosen for evaluation. For entropy calculation, the probability p_{ij} of a sample belonging to class j in each cluster i needs to be determined first, and then the entropy of the clustering result can be obtained, as shown in Equation (12).

$$EY_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij} \tag{12}$$

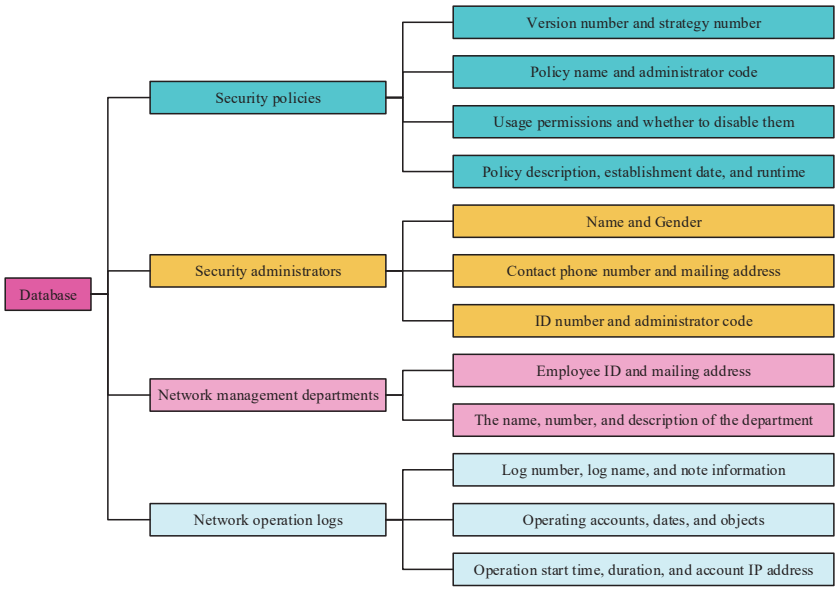


Figure 6 The composition structure of the INSM system database.

In Equation (12), L represents the number of clusters. The total entropy calculation for the entire sample set is given by Equation (13).

$$EY = \sum_{i=1}^{K} \frac{m_i}{m} EY_i \tag{13}$$

In Equation (13), m represents the total amount of data in the sample set, and m_i represents the total number of sample objects in cluster S. For purity PY_i of i, the purity calculation is shown in Equation (14).

$$PY_i = \max(PY_{ij}) \tag{14}$$

In Equation (14), PY_{ij} represents the possibility that each sample in i belongs to class j. The expression for the total purity PY of the sample set is shown in Equation (15).

$$PY = \sum_{i=1}^{K} \frac{m_i}{m} PY_i \tag{15}$$

According to the above designation, a scientific evaluation of the clustering effect of the clustering algorithm can be conducted.

4 Results Analysis of INSM System Based on TDO-K-means Algorithm

To validate the performance and application effectiveness of the proposed TDO-K-means algorithm, the study initially focuses on verifying the performance of the TDO-K-means algorithm. This serves as the foundation for assessing the subsequent application effectiveness of the INSM system. Finally, an analysis of the application scenarios of the INSM system is conducted.

4.1 Results Analysis Based on TDO-K-means Algorithm

In order to assess the performance of the TDO-K-means algorithm, the experimental platform employs a computer with Windows 10 as the operating system, an Intel i5 CPU, and 8GB of RAM. The software utilized is Weka. Additionally, the study employs seven commonly used datasets for evaluating clustering algorithms, namely Flame, Spiral, Iris, Seeds, Jain, Aggregation, and Wine. The corresponding numbers of instances are 240, 321, 150, 210, 373, 788, and 178, with category numbers of 2, 3, 3, 3, 2, 7, and 3, and dimensions of 2, 2, 4, 7, 2, 2, and 13, respectively. To scientifically validate the effectiveness of the proposed algorithm, the study introduces mainstream algorithms for comparative experiments. These algorithms include the Adaptive Multi Density Peak Subcluster Fusion Clustering (AMDPSFC) and the Optimization of Density Peak Fast Search Clustering for Cuckoo Birds (ODPFSCCB). The AMDPSFC algorithm introduces the idea of natural neighbors into density peak clustering, proposes an automatic cluster center selection strategy to determine the initial subcluster centers, uses a two-stage allocation strategy to reduce the probability of chain effect, and designs a metric criterion based on K-nearest neighbor similarity for fusion. The ODPFSCCB algorithm introduces the cosine similarity principle, combines the direction and the actual distance, and better distinguishes the belonging degree of data points in the middle region of the two kinds of clusters. The initial step involves verifying whether the S of TDO-K-means algorithm can help determine the optimal number of clusters. The study utilizes the ESS for comparing results before and after the algorithm's improvement.

Figure 7(a) illustrates the variation curve of S and the number of clusters for different datasets. It shows that the k values obtained when S is minimized are consistent with the actual number of clusters across different datasets. Figure 7(b) presents the ESS results before and after the improvement of the K-means algorithm for different datasets. Compared to the K-means

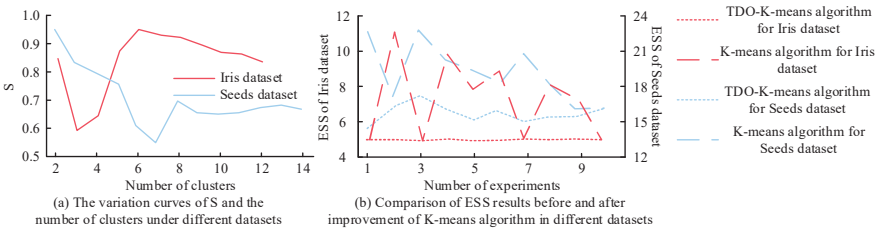


Figure 7 Validity test results of S on different datasets.

algorithm, the TDO-K-means algorithm achieves a smaller objective function value, and the overall curve exhibits no significant changes. These results indicate that the TDO-K-means algorithm's determination of k values is effective, and the optimized initial cluster centers can mitigate the adverse effects of random selection on clustering performance. Due to the challenges of visualizing clustering results in high-dimensional datasets, the study focuses its experiments on the Spiral, Jain, and Aggregation datasets.

Figures 8(a) and 8(b) respectively show the clustering results of different algorithms on the Spiral and Jain datasets. It can be observed that the K-means algorithm performs poorly in handling spiral-shaped sample data, dividing points into three uniform parts only, while other algorithms demonstrate better performance on the Spiral dataset. Due to the uneven distribution of samples in the Jain dataset, clustering is more challenging, resulting in comparatively poorer performance for all three algorithms, with TDO-K-means showing relatively better results. These findings indicate that the TDO-K-means algorithm outperforms other algorithms in clustering across different types of datasets. To further quantitatively analyze the clustering performance, accuracy, Adjusting the Morand Index (AMI), and Adjusting Mutual Information (AMIN) are employed for evaluation [21, 22]. Among them, AMI is the mainstream external evaluation index of clustering. The effectiveness of clustering is evaluated by calculating the number of sample pairs assigned to the same or different types of clusters in real labels and clustering results. AMIN is mainly used to solve the problem that mutual information is sensitive to cluster size, and the evaluation effect is more robust.

Figures 9(a)–9(c) compare the accuracy, AMI, and AMIN of different clustering algorithms on various datasets. It is evident that in more than half of the datasets, the TDO-K-means algorithm and the ODPFSCCB algorithm exhibit superior accuracy. The TDO-K-means algorithm, in particular, achieves the highest average accuracy at 96.01%, surpassing the ODPFSCCB

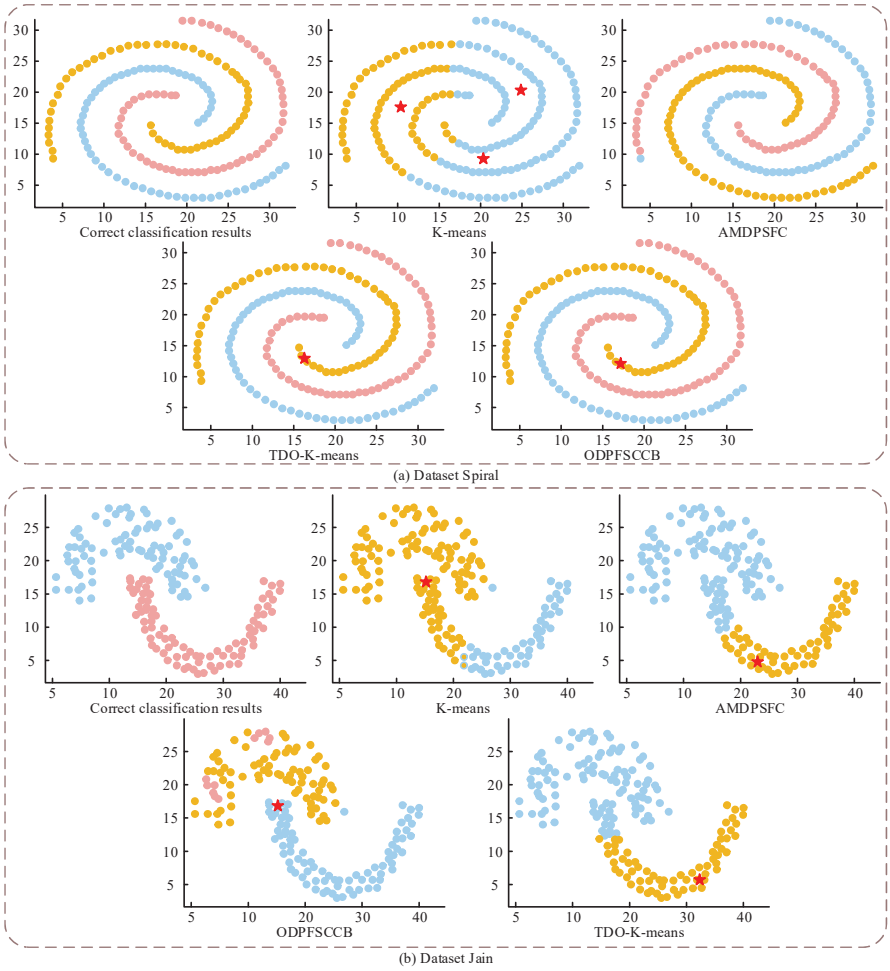


Figure 8 The clustering performance of different algorithms on different datasets.

algorithm by 3.36%. In terms of AMI and AMIN results, except for the dataset Seeds where the TDO-K-means algorithm's two evaluation metrics are lower than the AMDPSFC algorithm, the TDO-K-means algorithm outperforms in the remaining datasets, with average AMI and AMIN values of 0.866 and 0.869, respectively. These results suggest that the proposed algorithm effectively determines the value of K and initial cluster centers, demonstrating outstanding performance and establishing a solid foundation for subsequent applications. In order to further explore the application effect

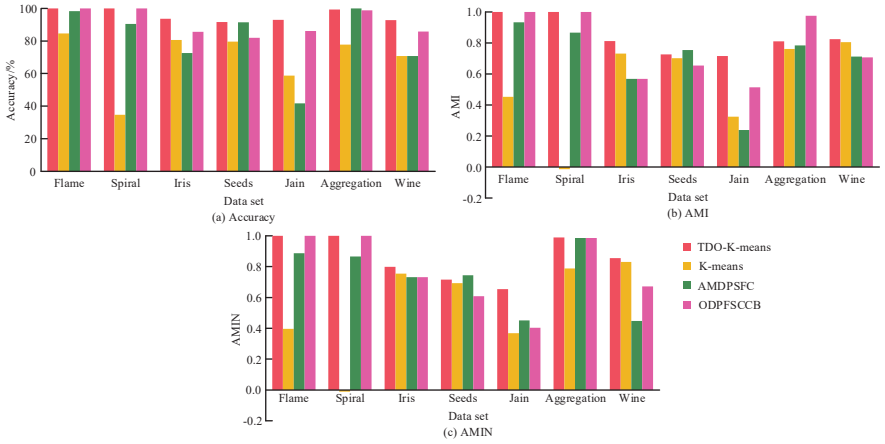


Figure 9 Comparison of quantitative clustering evaluation indicators for different clustering algorithms on different datasets.

Table 1 Processing results of network security defense dataset

Virus Dataset Name	Number of Attack Records/Piece	Virus Description
T-one	3000	BackOriffice data, NetSpy data, and glacier data
T-two	3000	Glacier data, ExeBind data, and KeyboardGhost data
T-three	3000	FluShot data, ransomware data, and WinNuke data
S-one	6000	BackOriffice data, NetSpy data, glacier data, Glacier data, ExeBind data, and KeyboardGhost data
S-two	6000	Glacier data, ExeBind data, KeyboardGhost data, FluShot data, ransomware data, and WinNuke data
S-three	6000	BackOriffice data, NetSpy data, glacier data, FluShot data, ransomware data, and WinNuke data
N	9000	BackOriffice data, NetSpy data, glacier data, Glacier data, ExeBind data, KeyboardGhost data, FluShot data, ransomware data, and WinNuke data

of the TDO-K-means algorithm in network security defense, 9000 network virus attack records were obtained by using the network packet capture tool, with nine types in total, from which data sources for subsequent experiments could be obtained, as shown in Table 1.

The study evaluates the application effectiveness of different algorithms using accuracy and false positive rate on the aforementioned virus dataset. Additionally, to mitigate experimental variability, 10 random samples from

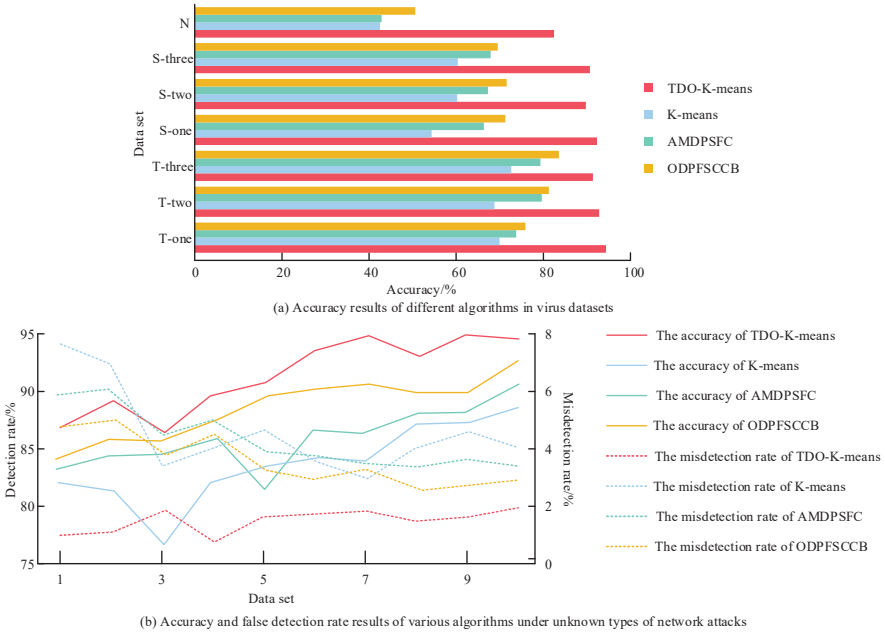


Figure 10 The application effects of different algorithms in network security defense.

the virus dataset were selected for comparative experiments, involving 10 participants.

Figure 10(a) compares the accuracy results of different algorithms in the virus dataset. It can be observed that the TDO-K-means algorithm achieves an accuracy rate of 94.38% in identifying attack threats in a simulated real network security environment, while the average accuracy rates for K-means algorithm, AMDPSFC algorithm, and ODPFSCCB algorithm are 61.07%, 67.55%, and 73.26% respectively. Figure 10(b) contrasts the accuracy and false positive rate of various algorithms under unknown network attack types. It is evident that the proposed algorithm demonstrates higher detection rates and lower false positive rates compared to other algorithms, with 94.63% and 1.32% respectively. In comparison, the average detection rates and false positive rates for K-means algorithm, AMDPSFC algorithm, and ODPFSCCB algorithm are 76.15% and 5.92%, 82.06% and 3.97%, and 85.34% and 4.27% respectively. The above results show that in the real network security attack environment, the accuracy rate of attack threat identification of the research method is as high as 94.38%, and it has the lowest false positive rate (1.32%), which indicates that the research method is conducive to INSM system to

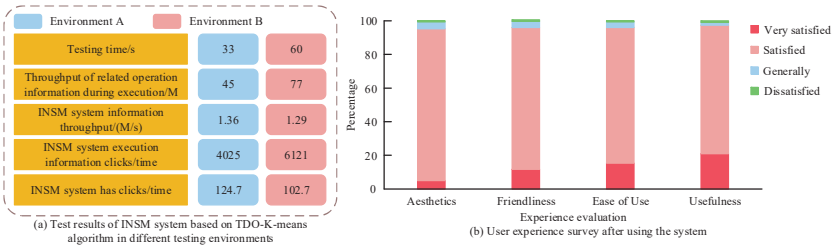


Figure 11 Test results and user experience of INSM supported by TDO-K-means.

generate early warning for unknown threats before network security events occur, and complete network security proactive defense. And then solve the potential threat problem of network security in reality. In traditional network security, there are often problems such as alarm overload and false positives, so that the security number can not be fully utilized, but the low false positives rate of the research method can help INSM system effectively avoid the above problems. In addition, in the face of the increasingly fierce network attacks and the trend of strong stealth, high professional degree and organization, the protection efficiency of traditional static signature database is faced with great adjustment. However, the research method can assist the subsequent practical application of INSM system to continuously observe the business network and analyze the abnormal changes in the accuracy recognition. By preventing attack threats before they invade, threat detection can effectively help INSM systems effectively deal with unknown threats, and show better results in the increasingly complex network environment. In summary, the proposed TDO-K-means algorithm exhibits superior performance, meeting the current demands for network security defense and proving to be more applicable in the context of the Internet.

4.2 Test Results of INSM System Based on TDO-K-means

The above outcomes validate the feasibility of the INSM system supported by TDO-K-means, setting the stage for further examination of the system's performance and application effects. Two testing environments are established for this purpose: Environment A involves randomly selecting 100 users to log in to the system and undergo testing, while Environment B entails selecting 200 users for login and testing, followed by a survey of the experience of 300 users post-testing.

Figure 11(a) presents the test results of the INSM supported by TDO-K-means under different testing environments. It can be observed that as

the number of users grows exponentially, the system's time consumption, execution time operation information throughput, INSM system information throughput, execution information click counts, and existing click counts are 60 s, 77 M, 1.29 M, 6121 times, and 102.7 times respectively, indicating that the system meets performance requirements such as interaction, high responsiveness, and excellent processing capabilities. Figure 11(b) depicts the user experience survey after using the system, revealing that over 90% of users express satisfaction with its aesthetics, user-friendliness, ease of use, and usefulness, indicating that the proposed INSM system has good application effects.

5 Conclusions

As people entered the high-speed information age, the Internet provided them with broader avenues for knowledge acquisition, making their lives more convenient. However, it also brought about significant threats to network data security. In addressing these challenges, research was conducted to enhance the K-means algorithm. The approach involved incorporating three decision-making components, feature weighting, and evaluating the effectiveness of the value of K. This led to the development of the TDO-K-means algorithm, which served as the foundation for constructing the INSM system. Experimental results demonstrated that, in more than half of the datasets, the accuracy of the TDO-K-means algorithm surpassed that of the ODPF-SCCB algorithm. Furthermore, the average accuracy of the TDO-K-means algorithm reached a maximum of 96.01%. The AMI and AMIN metrics for the TDO-K-means algorithm were the most outstanding, registering at 0.866 and 0.869, respectively. In a simulated real-world network security attack environment on the Internet, the TDO-K-means algorithm achieved a recognition accuracy of 94.38% for threat identification. This outperformed the K-means algorithm, AMDPSFC algorithm, and ODPFSCCB algorithm by 33.31%, 26.83%, and 21.12%, respectively. For unknown network attack types, the proposed algorithm exhibited higher detection rates and lower false positive rates at 94.63% and 1.32%, respectively. A survey conducted on the application effects of the INSM system after user usage revealed that over 90% of users were satisfied with its aesthetics, user-friendliness, ease of use, and usefulness. In summary, the research proposed a method capable of timely identifying and detecting network security threats, effectively enhancing the security performance of Internet usage. However, there are still some limitations in the research, and other optimization techniques or hybrid

methods can be explored in future studies to further improve the performance of K-means algorithm in network security applications.

References

[1] Guan S, Hu W, Zhou H, Lei Z and Liu G. Design and implementation of virtual experiment for complex control system: A case study of thermal control process. IET Generation Transmission & Distribution, 2021, 15(23):3270–3283.

[2] Alexakos C E, Votis K, Tzovaras D and Serpanos D. Reshaping the Intelligent Transportation Scene: Challenges of an Operational and Safe Internet of Vehicles. Computer, 2022, 55(1):104–107.

[3] Asvial M, Cracias A, Laagu M, Arifin A. Design and Analysis of Low Power and Lossy Network Routing System for Internet of Things Network. International Journal of Intelligent Engineering and Systems, 2021, 14(4):548–560.

[4] Choudhuri S, Adeniye S, Sen A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement for Partial Domain Adaptation. Artificial Intelligence and Applications. 2023, 1(1): 43–51.

[5] Hebbi C, Mamatha H. Comprehensive Dataset Building and Recognition of Isolated Handwritten Kannada Characters Using Machine Learning Models. Artificial Intelligence and Applications, 2023, 1(3): 179–190.

[6] Naveed Ahmed, N., & Nanath, K. (2021). Exploring Cybersecurity Ecosystem in the Middle East: Towards an SME Recommender System. Journal of Cyber Security and Mobility, 10(3), 511–536.

[7] Ozkok F O, Celik M. A Hybrid Validity Index to DetermineKParameter Value ofk-Means Algorithm for Time Series Clustering. International Journal of Information Technology & Decision Making, 2021, 20(06):1615–1636.

[8] Pullagura I. Towards Intelligent Machine Learning Models for Intrusion Detection System. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(5):643–655.

[9] Jian Y, Dong X and Jian L. Detection and Recognition of Abnormal Data Caused by Network Intrusion Using Deep Learning. Slovenian Association Informatika, 2021, 45(3):441–445.

[10] Gulganwa P and Jain S. EES-WCA: energy efficient and secure weighted clustering for WSN using machine learning approach. International Journal of Information Technology, 2022, 14(1):135–144.
[11] Ferrag M A, Shu L, Friha O and Yang X. Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-based Solutions, Datasets, and Future Directions. IEEE/CAA Journal of Automatica Sinica, 2021, 9(3):407–436.
[12] Shu L, Hancke G, Sheng V S and Wang L. Guest Editorial: Reliability and Security for Intelligent Wireless Sensing and Control Systems. IEEE Transactions on Industrial Informatics, 2022, 18(4):2651–2655.
[13] Elmorshidy A. M-Commerce Security: Assessing the Value of Mobile Applications Used in Controlling Internet Security Cameras at Home and Office – An Empirical Study. International journal of information security and privacy, 2021, 15(4):79–97.
[14] Wang P and Susilo W. Data Security Storage Model of the Internet of Things Based on Blockchain. Computer Systems Science and Engineering, 2021, 36(1):213–224.
[15] Sirisha A, Chaitanya K, Krishna K V S S R, Kanumalli S S. Intrusion Detection Models Using Supervised and Unsupervised Algorithms – A Comparative Estimation. International Journal of Safety and Security Engineering, 2021, 11(1):51–58.
[16] Vedavathi N and Kumar K M A. SentiWordNet Ontology and Deep Neural Network Based Collaborative Filtering Technique for Course Recommendation in an E-Learning Platform. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2022, 30(4):709–732.
[17] Wang H, Li X, Jhaveri R H, et al. Sparse Bayesian learning based channel estimation in FBMC/OQAM industrial IoT networks. Computer Communications, 2021, 176: 40–45.
[18] Abolfathi M, Inturi S, Banaei-Kashani F, et al. Toward enhancing web privacy on HTTPS traffic: A novel SuperLearner attack model and an efficient defense approach with adversarial examples. Computers & Security, 2024, 139: 103673.
[19] Goyal N, Joshi T and Ram M. Evaluating and Improving a Content Delivery Network (CDN) Workflow using Stochastic Modelling. Journal of Cyber Security and Mobility, 2021, 10(4), 679–698.

[20] Pasha A M, Moursi M S E, Zeineldin H H, Khadkikar V and Bendary A F. Enhanced Transient Response and Seamless Interconnection of Multi-Microgrids based on an Adaptive Control Scheme. IET Renewable Power Generation, 2021, 15(11):2452–2467.

[21] Jayshree, Seetharaman G and Pati D. Enhanced TACIT Encryption and Decryption Algorithm for Secured Data Routing in 3-D Network-on-Chip based Interconnection of SoC for IoT Application. NIScPR-CSIR, India, 2021, 80(6):520–527.

[22] Patil R A, Kavipriya P and Patil B P. Hybrid Meta-Heuristic Algorithms Based Optimal Antenna Selection for Large Scale MIMO in LTE Network. Journal of Interconnection Networks, 2022, 22(4):26–43.

[23] Yang, L. Internet of Things Security Design Based on Blockchain and Identity Re-encryption. Journal of Cyber Security and Mobility, 2024, 13(3), 369–392.

[24] Wang Y, Gao J, Xuan C, Guan T, Wang Y, Zhou G and Ding T. FSCAM: CAM-Based Feature Selection for Clustering scRNA-seq. Interdisciplinary Sciences: Computational Life Sciences, 2022, 14(2):394–408.

[25] Cai J, Han Y, Guo W and Fan J. Deep graph-level clustering using pseudo-label-guided mutual information maximization network. Neural Computing and Applications, 2024, 36(16):9551–9566.

Biographies

Cuijuan Liu obtained her Master's degree in Computer Application Technology from North China Electric Power University (Baoding) in 2006. Currently, she serves as the Deputy Director of the Computer Department at the School of Information Engineering, Hebei GEO University. Her areas of interest include data mining, machine learning, network security, innovative thinking, and more.

Liya Song, graduated from Hebei University of Technology in 2006, majoring in test measurement technology and instruments, with a postgraduate degree. She later worked in the Electronic Communication Experimental Center of Hebei GEO University. She has published articles based on Optical Opbidimetry, and her main research interests are photoelectric detection, image processing and information processing.