Network Security Level Protection Evaluation Model Based on Large Language Model and Cluster Analysis

Bin Chen, Bin Li, Yuting Tang, and Xiaogang Xie
(Corresponding author: Bin Li)

Huaxin Consulting Co., Ltd.
Hangzhou 310052, China
Email: 18958016123@163.com

(Received Aug. 6, 2024; Revised and Accepted June 12, 2025; First Online Aug. 2, 2025)

Abstract

The progress of Internet technology makes the network security problem increasingly prominent. Traditional network security protection measures are no longer able to cope with complex and ever-changing network attacks, resulting in high security risks. Therefore, this study proposes a network security level protection evaluation model based on a large language model and cluster analysis. This model collects data from large language models and uses clustering analysis algorithms for analysis. In response to the problems with clustering analysis algorithms, relative entropy is introduced to improve them and enhance the performance of the model. The results indicated that when the dataset size reached 1000, the accuracy of the restricted Boltzmann machine was 0.98, and the contrast divergence was 0.21. The proposed model, based on improved clustering analysis, achieved recognition accuracies of 82.1%, 79.6%, 85.5%, and 88.8% for four different types of data. The experiment validates that the designed restricted Boltzmann machine has outstanding data processing capabilities, and the network security level protection model based on improved clustering analysis has excellent recognition performance. Research methods help promote the development of network security technology and enhance the overall protection level of information systems.

Keywords: Cluster Analysis; KL Distance; Network Security; Restricted Boltzmann Machine

1 Introduction

The network security level protection (NSLP) system is an important regulatory and standard system established to enhance the level of information system security protection. Its core goal is to ensure that the system can receive corresponding security protection at different security levels by classifying and evaluating the safety level of the information system [16]. However, the progress of information technique and the increasing complexity of network threats has made traditional security level protection assessment methods be inadequate when facing massive data and diverse attack methods [17]. However, the promotion of clustering analysis (CA) technology has solved this problem.

This technology performs well in processing unstructured data and discovering the intrinsic structure of data, effectively assisting in deep level data mining and pattern recognition. This study develops a network security level protection evaluation (NSLPE) model based on large language model (LLM) and CA. The model collects data through LLM, processes the collected data through restricted Boltzmann machine (RBM), and finally evaluates the data through CA. The research targets to lift the accuracy and effectiveness of security assessments, and provide new technical support for network security protection.

The research content mainly has four sections. Section 1 is a review of other scholars' research on network security topics. Section 2 is a brief description of the main methods used in this study. Section 3 is the model results obtained through the application of methods and the analysis of the results. Section 4 is a summary of all the above studies and prospects for future research.

2 Related Works

With the popularization and continuous expansion of the internet, network security has become increasingly important. Mohammed AB et al. proposed a counterattack strategy based on intelligent methods and diversified technologies to address security vulnerabilities and attacks in unmanned aerial vehicles (UAVs) and their networks. This strategy delved into the complexity of counterattack strategies, algorithm fusion, and integration with machine learning languages. These strategies could effectively enhance the security of UAV systems and provide important references for the development of UAV security technology in the future [13]. Computing C M proposed an image encryption algorithm based on optical information processing technology to meet the secure transmission needs of massive data and information in the information age. It achieved encryption by combining fractional Fourier transform and Arnold transform. This algorithm demonstrated good encryption and decryption performance in MATLAB simulation, providing new ideas for the field of image encryption [3]. Wang Y et al. designed a detection method based on outlier analysis to enhance network security protection, transforming intrusion detection into outlier recognition in network behavior datasets. This method efficiently identified outliers in simulation experiments, demonstrated excellent clustering ability, and had unique advantages compared to other schemes, as it did not require a training process [20]. Hang F et al. constructed a universal hybrid two-stage fusion model for network security situation. This model focused on studying diverse heterogeneous sensors to identify security risks that posed a threat to network systems, and had adaptability in corresponding network scenarios [7].

Das et al. developed the Lévy-Cauchy arithmetic optimization algorithm to handle the local optima in image segmentation caused by the rough K-means clustering algorithm (K-means). This method balanced exploration and development through Levy flight and Cauchy distribution, and introduced adversarial learning to maintain an effective population. This method outperformed other clustering algorithms in processing traditional color images, pathological images, and leaf images, achieving higher feature similarity index values [4]. Hengdong et al. proposed a new semi-supervised fuzzy K-means method to improve the performance and stability of unsupervised fuzzy K-means, which utilized annotated label information through dynamic adjustment and label discrimination. This method not only achieved the best performance, but also effectively reduced the impact of data noise, significantly better than traditional semi-supervised clustering methods [8]. Kodge et al. proposed a system based on K-means to deal with the problem of snow melting in the the Himalayas caused by global warming. The system used high-resolution snow satellite photos to extract snow cover area. Through the analysis of time series satellite images from 1984 to 2022, the change trend of snow cover in the the Himalayas was revealed, which provided an important reference for formulating relevant coping strategies [9]. Lim et al. proposed a method based on K-means to achieve instance segmentation under unsupervised conditions, which generates instance level pseudo object masks from unlabeled images through self supervised converters and cosine distance K-means. This model achieved excellent performance in unsupervised class agnostic instance segmentation tasks, significantly outperforming concurrent work on the COCO dataset, and successfully extended to tasks like unsupervised object detection [11].

In summary, the aforementioned methods still have the problem of insufficient consideration of the complexity of the actual network environment. Image encryption algorithms are challenging to adapt to the encryption requirements of dynamic network traffic. Furthermore, the adaptability of outlier analysis to novel attacks is limited, and its reliance on a substantial amount of labeled data is a significant drawback. The hybrid two-stage fusion model has real-time limitations. Levy-Cauchy algorithm is difficult to apply directly to network security. Semi-supervised fuzzy K-means is less efficient on large network data. Unsupervised object detection is excellent for instance segmentation tasks, but needs to be adapted to network security needs. To address these shortcomings, this study proposes a comprehensive method that combines a LLM and improved cluster analysis, and uses RBM to reduce and classify large data to improve the accuracy and efficiency of anomaly detection. The improved K-medoids-KL algorithm optimizes the cluster center (CC) selection process, solves the local optimal problem of the traditional method, and has better adaptability, especially in the context of diverse data types and dimensions.

3 Methods

The first section addresses the issue of difficulty in collecting network data, using LLM to collect it and processing the collected data through RBM. The second section addresses the issues with K-means and improves the selection of its clustering centers by proposing K-medoids, which are then improved using KL.

3.1 NSLP Data Processing Model Based on RBM Algorithm

NSLP aims to strengthen the security protection of information systems, standardize the standards and management methods for computer system security construction and use in accordance with laws, regulations, and technical standards. This system aims to safeguard national information security, protect the legitimate rights and interests of citizens, legal persons, and other organizations, and maintain national security and social public interests [10]. According to the security threats and importance faced by different information systems, the security assessment divides information systems into five levels. Each level has corresponding security protection requirements and technical standards. Evaluators need to conduct a comprehensive evaluation of the information system based on these standards to determine whether it meets the requirements of the corresponding level. Its structure is shown in Figure 1.

In Figure 1, NSLPE mainly includes security control evaluation and overall system evaluation. Among them, the former includes security technology evaluation and security management evaluation. The latter includes secu-

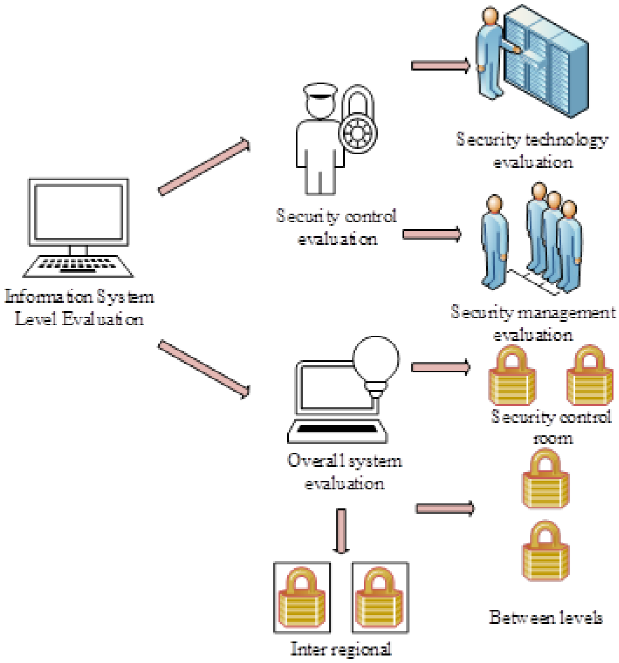
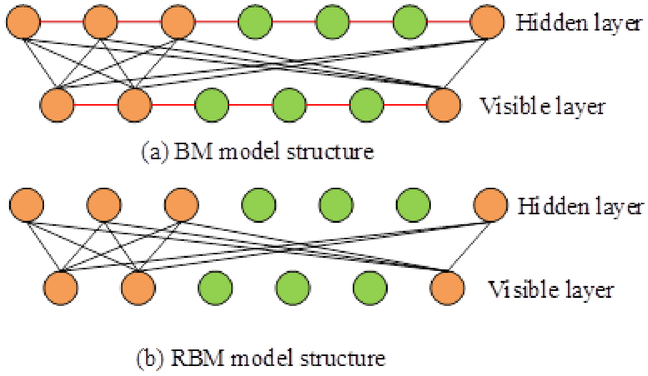

Security technology evaluation
Security control evaluation
Security management evaluation
Information System Level Evaluation
Overall system evaluation
Security control room
Between levels
Inter regional

Figure 1: Level evaluation of information system

rity control rooms, security level rooms, and security area rooms. Security technology assessment mainly evaluates the security control measures implemented by information systems at the technical level. Security management evaluation mainly evaluates the security control measures implemented by information systems at the management level. The security control room evaluates the coordination and coordination between various security control measures to ensure that each control item can effectively work together [6]. Security level rooms evaluate the protective measures between different security levels of information systems, ensuring that the system has appropriate security protection from the underlying hardware to the application layer. The security domain room evaluates the protective measures and isolation strategies between different security domains of the information system, ensuring that each domain of the system can be effectively isolated and protected from attack.

Through systematic grading and filing, security construction rectification, self-examination and self-evaluation, third-party evaluation, supervision and continuous improvement, enterprises can effectively ensure network security and ensure that the protection capability of information systems meets the corresponding level requirements. The evaluation process will generate plenty of data through LLM, so it is required to analyze a large amount of data. This study uses RBM for data learning, dimensionality reduction, and classification operations. The core concept of RBM is to improve and optimize the traditional Boltzmann machine (BM) to achieve more efficient training and prediction. Figure 2 shows the structure of BM model and RBM model.



(a) BM model structure

(b) RBM model structure

Figure 2: Network structure diagram of BM and RBM models

In Figure 2, BM is a neural network model containing visible layer (VL) and hidden layer (HL). Its primary role of the HL is to extract hidden information from the sample dataset, while for the VL is to reduce input costs. However, the existence of this relationship makes the training process of the BM model more difficult and the efficiency is not ideal [14]. Therefore, the BM model will be improved and the RBM model will be proposed. RBM is based on BM, except for the connections between neurons in the HL and VL. Through this modification, the training efficiency of the model can be greatly improved and the training time can be reduced. Assuming there are n and m are VL neurons and HL neurons in RBM. For known HL and VL states, the energy formula of RBM is Equation (1).

$$E(v,h\,|\theta) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i W_{ij} h_j \quad (1)$$

In Equation (1), v and a are the state vectors and bias vectors of the VL. h and b are the state vectors and bias vectors of the HL. i and j are the i-th and j-th neurons in the VL and HL. W_{ij} is the connection weight between the j-th neuron in the HL and the i-th neuron in the VL [2,15]. By using the energy formula of RBM, the state of the VL is v. The joint distribution probability of the HL state h is Equation (2).

$$P_\theta(v,h) = \frac{1}{z_\theta} e^{-E_\theta(v,h)} \quad (2)$$

In Equation (2), z_θ is the normalization factor. Due to the fact that neurons in the HL and VL are not interconnected in the RBM model. The activation probabilities of neurons in the VL and HL are shown in Equation (3).

$$\begin{cases} P_\theta(v_i = 1\,|h,\ \theta) = \sigma(a_i + \sum_{j=1}^{m} W_{ij} h_j) \\ P_\theta(h_j = 1\,|h,\ \theta) = \sigma(b_i + \sum_{i=1}^{n} W_{ij} v_j) \end{cases} \quad (3)$$

In Equation (3), h_j and v_i are the states of the j-th and i-th neurons in the HL and VL. σ is the Sigmoid function. To better apply the RBM model to data processing systems, the RBM model is improved as shown in Figure 3.

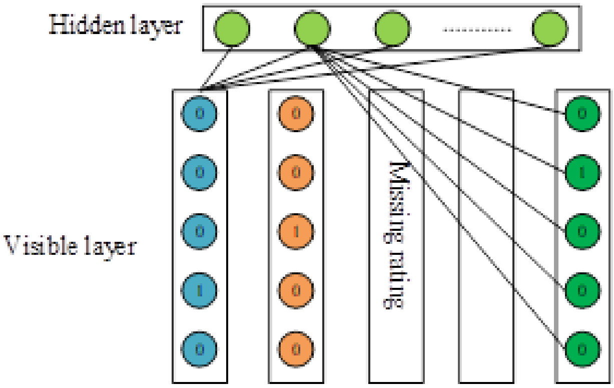


Figure 3: Network structure of improved RBM model

In Figure 3, softmax units are used to replace the VL units in RBM. When users rate the security level, the corresponding VL units are connected to the HL units. When the user does not rate the security level, missing units are used instead, and the VL units are not related to the HL units [18]. According to the energy formula of the RBM model, an improved formula can be obtained as shown in Equation (4).

$$E(v,h\,|\theta) = -\sum_{i=1}^{n}\sum_{k=1}^{K} a_i^k v_i^k - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K} v_i^k W_{ij}^k h_j \quad (4)$$

In Equation (4), a_i^k and b_j are biases between the VL and HL. v_i^k is the user's rating of security level i as k. The activation probabilities of neurons in the VL and HL of the improved model are shown in Equation (5).

$$\begin{cases} P(v_i^k = 1\,|h,\theta) = \frac{\exp(a_i^k + \sum_{j=1}^{m} W_{ij}^k h_j)}{\sum_{k=1}^{K}\exp(a_i^k + \sum_{j=1}^{m} h_j)} \\ P(h_j = 1\,|v,\theta) = \sigma(b_j + \sum_{i=1}^{n}\sum_{k=1}^{K} v_i^k W_{ij}^k) \end{cases} \quad (5)$$

In Equation (5), h_j and v_i^k are the states of the HL and VL. The performance of the model is evaluated by predicting and scoring it, and its prediction method is Equation (6).

$$R_l = \sum_{k=1}^{K} k P_\theta(v_l^k = 1\,|V,\theta) \quad (6)$$

In Equation (6), $\theta = \{W,a,b\}$ represents the connection weight, VL bias, and HL bias of the model, respectively. By improving the RBM model to analyze the data generated during the evaluation process, the subsequent training effect of the model can be better.

3.2 CA-based NSLPE Model

CA is an unsupervised learning method taken to divide a dataset into multiple groups, so that the similarity of data points within the same group is as high as possible, while that between different groups is as low as possible [1]. Among numerous clustering algorithms, K-means is the most classic, which has the advantages of easy implementation and good clustering performance. Figure 4 shows the main process of K-means.

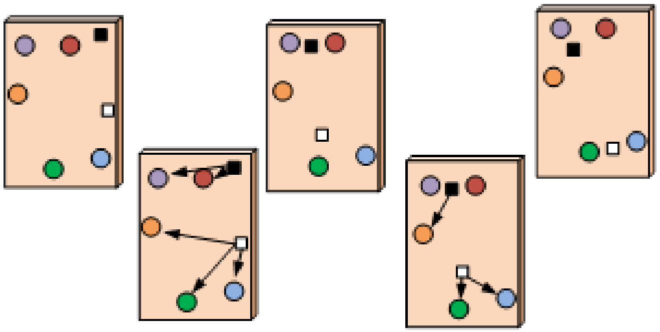


Figure 4: K-means algorithm diagram

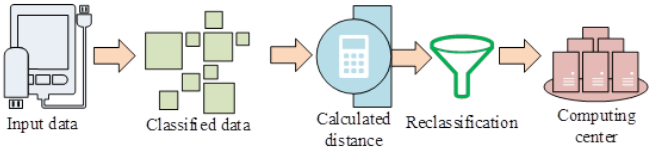


Figure 5: K-medoids clustering algorithm flow

In Figure 4, classify data objects into several various categories at first, and select an initial center from the category at random to calculate the distance from the data mentioned. Each classification object is assigned to the nearest center, and then the CC is calculated and re-assigned. Repeating these steps until the clustering results stops changing [19]. In K-means, Euclidean distance is commonly used as the distance measure between data points and CCs, as expressed in Equation (7).

$$d(x_i,\mu_j) = \sqrt{\sum_{k=1}^{n}(x_{ik}-\mu_{jk})^2} \quad (7)$$

In Equation (7), x_i is the data point, μ_j is the CC, and n is the dimension of the data point. For each cluster, the mean of all data points is calculated as the new CC, as shown in Equation (8).

$$\mu_j = \frac{1}{|C_j|}\sum_{x_i\in C_j} x_i \quad (8)$$

In Equation (8), C_j is the set of all data points in the j-th cluster, and $|C_j|$ is the number of data points in that cluster. However, the selection of initial CCs in K-means has a significant impact on the final clustering results, leading to local optima. Therefore, improvements are made to K-means and an NSLPE model based on K-medoids is proposed. Figure 5 shows the clustering process of the K-medoids algorithm.

In Figure 5, the data are classified into several distinctive types to choose original center from each category. Calculating them, and then each classified data is separated into the closest initial center [12, 21]. Then, given the classification, the CC of each cluster is re-calculated and redivided. These steps are repeated till stopping changing clustering results. To optimize the K-medoids performance, this study introduces KL distance to enhance it. The improved algorithm is divided into two stages, namely searching for CCs and generating clustering results. The search for CCs is Equation (9).

$$C_1 = \arg\min_{i\in N}\left[\sum_{j\in N\backslash\{i\}} D'(j\,|i\,)\right] \tag{9}$$

In Equation (9), i is the CC, j is a single data, and C_1 is the first CC. Firstly, randomly to select a center and calculate its KL distance from other data. Then, the sum of KL distances of all data and the data with the minimum value are established as the first CC, and the remaining CCs are found as shown in Equation (10).

$$C_k = \arg\max_{i\in N\backslash\{C_1,C_2,\cdots,C_{k-1}\}} [DEC(i)] \quad ,2\le n\le k \tag{10}$$

Equation (10) represents all CCs except for the first CC. Firstly, non-central data is specified to temporary clustering center (TCC), and the KL distance between each data and the TCC is calculated. If the calculated distance of KL< distance within data and CC, the data is assigned to a temporary CC [5]. The contribution of data to the TCC is Equation (11).

$$\max\left[0,\min_{t=1}^{n-1}(D'(j\,|C_t)) - (D'(j\,|i))\right] \tag{11}$$

For Equation (11), $n-1$ is the amount of CCs. The contribution degree between other non-central data and other data can be obtained through Equation (11) as shown in Equation (12).

$$DEC(i) = \sum_{j\in N\backslash\{i,C\}} \max\left[0,\min_{t=1}^{n-1}(D'(j\,|C_t)) - D'(j\,|i)\right] \tag{12}$$

In Equation (12), $DEC(i)$ is the gross of contributions, and multiple CCs are able to chose in the same way. Similar rating data can be assigned to the same cluster by the CC, and the target data can be directly searched for the same nearest neighbor data center within the cluster. Assuming the cluster containing the target data i is C_i, the distance between rating of i and other data in the cluster could be calculated using the KL distance. The data rating with the minimum KL distance is selected as the nearest neighbor combination of data i, as shown in Equation (13).

$$p_{ui} = \bar{r}_u + \frac{\sum_{j\in Cnei_i} sim_{KL}(j,i)\cdot(\bar{r}_j - \bar{r}_{Cnei_i})}{\sum_{j\in Cnei_i} sim_{KL}(j,i)} \tag{13}$$

In Equation (13), $Cnei_i$ is the nearest neighbor set. $\bar{r}_u$ is the mean rating of all levels that have been graded. $\bar{r}_{Cnei_i}$ is the average score obtained by rating all data in the $Cnei_i$. $\bar{r}_j$ is the average rating of the data. The calculation formula for $sim_{KL}(j,i)$ is Equation (14).

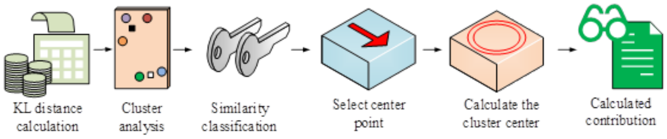


Figure 6: NSLPE process based on K-medoids

$$sim_{KL}(j,i) = \frac{1}{1+D'(j/i)} \tag{14}$$

From Equation (14), $sim_{KL}(j,i)$ is the similarity between data rating j and data rating i. According to the data rating obtained from analysis, the data without rating is rated to achieve the rating of network security level. The final model structure is Figure 6.

In Figure 6, the KL distance between each rating data is first calculated, and then CA is performed to determine the similarity for classification. Then, a single CC is selected, and other CCs are calculated. Finally, the contribution degree is calculated to obtain the network security rating of each model.

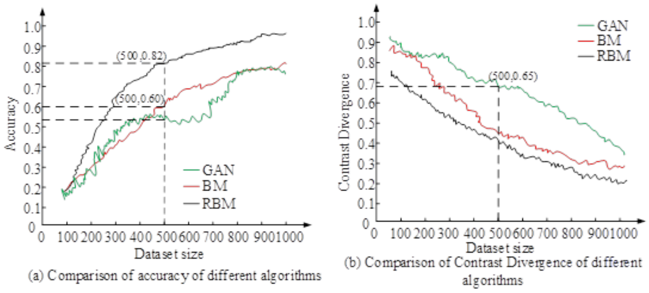
4 Results

In the first section, to verify the network data processing capability of the RBM algorithm, accuracy and contrastive divergence are used as reference indicators to analyze the comprehensive performance of the model. In the second section, to validate the performance of the K-medoids-KL model, accuracy, false alarm rate (FAR), and clustering performance are used to evaluate the model's performance.

4.1 Performance Analysis of Data Processing Model Based on RBM Algorithm

The hardware configuration includes an Intel Core i5-12600KF CPU, NVIDIA Geforce RTX4080 GPU, 16GB of VRAM, and 64GB of RAM. The dataset adopts the NSL-KDD public dataset, which contains network connection records from real network environments, covering normal network traffic and various types of network attacks. In this dataset, each network connection record has a clear label indicating whether the connection is normal or belongs to a specific type of attack. The generative adversarial network (GAN) is introduced and compared with the research model. Accuracy and contrast divergence are utilized as reference indicators and tested separately, as shown in Figure 7.

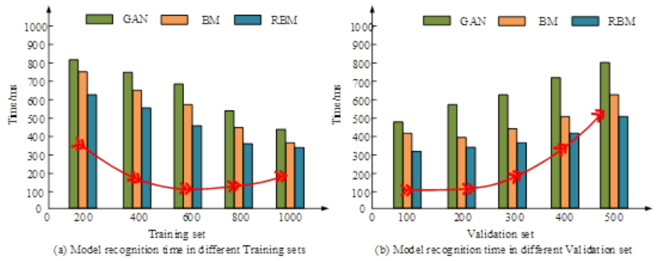
Figure 7(a) and (b) show the accuracy and contrast divergence of each model under two dataset sizes. In 7 (a), as the training set grows, the accuracy of each model continues to improve. When the dataset size reaches 1000,



(a) Comparison of accuracy of different algorithms

(b) Comparison of Contrast Divergence of different algorithms

Figure 7: Comparison of accuracy and contrast divergence of various models



(a) Model recognition time in different Training sets

(b) Model recognition time in different Validation set

Figure 8: Comparison of processing times for various models

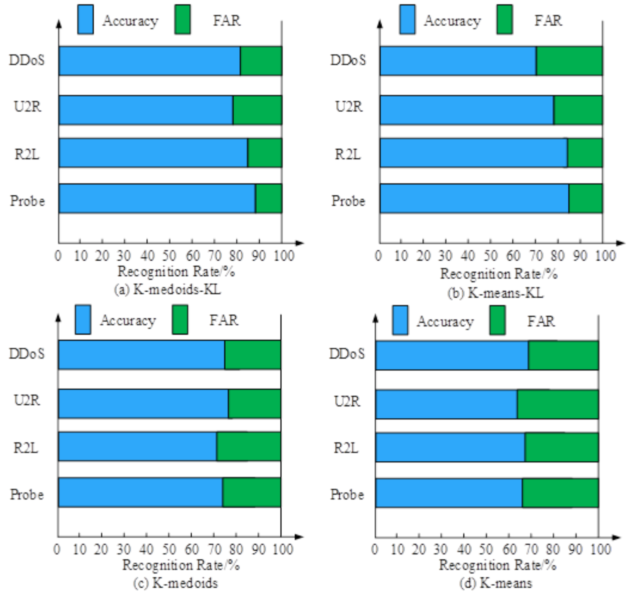
the accuracy of GAN, BM, and RBM are 0.78, 0.82, and 0.98, respectively. In Figure 7(b), as the training set increases, the contrast divergence of each model decreases. When the dataset size is 1000, the contrast divergence of GAN, BM, and RBM is 0.34, 0.29, and 0.21. The data shows that the RBM model not only exhibits excellent accuracy, but also performs well in contrast to divergence. The processing time of the model is compared using different types of network security threats, as shown in Figure 8.

Figure 8(a) compares the recognition time of different models after training on different training sets. Figure 8(b) shows the recognition time of different models in validation sets of different sizes. In Figure 8(a), when the training set is small, the recognition time of each model is longer, and when the training set is large, the recognition time of each model is significantly reduced. When the training set size is 1000, the recognition time of GAN, BM, and RBM is 442ms, 378ms, and 342ms. In Figure 8(b), when the validation set size is small, the recognition time of each model is shorter. However, as the validation set size increases, the recognition time of each model also increases. When the validation set size is 500, the recognition times of GAN, BM, and RBM are 804ms, 621ms, and 514ms. This indicates that the RBM model has higher efficiency compared to the other two models. Table 1 presents a comprehensive performance comparison of various models.

In Table 1, each model has the highest detection accuracy for DDoS and the lowest detection accuracy for Probe. Among the three models, the RBM model performed the best, with accuracies of 0.976, 0.981, 0.943, and 0.828 for DDoS, U2R, R2L, and Probe, and contrastive divergence of 0.155, 0.184, 0.168, and 0.178. Therefore, while the RBM model exhibits excellent performance, its contrastive divergence is also outstanding.

Table 1: Comprehensive performance of the model

Type	Accuracy			Contrastive divergence		
	RBM	BM	GAN	RBM	BM	GAN
DDoS	0.976	0.854	0.821	0.155	0.255	0.301
U2R	0.981	0.804	0.773	0.184	0.274	0.306
R2L	0.943	0.854	0.825	0.168	0.176	0.325
Probe	0.828	0.724	0.693	0.178	0.285	0.324



(a) K-medoids-KL

(b) K-means-KL

(c) K-medoids

(d) K-means

Figure 9: Accuracy and FAR analysis of various models on different types of data

4.2 CA-based NSLPE Model

To verify the performance of the proposed NSLP based on the K-medoids-KL, different types of intrusion attacks are selected from the NSL-KDD dataset, and the results are shown in Figure 9.

Figure 9(a) to (d) show the accuracy and FAR of K-medoids-KL, K-mean-KL, K-medoids, and K-mean models. In Figure 9, the K-medoids-KL model achieved recognition accuracy of over 79% for DDoS, U2R, R2L, and Probe. The recognition accuracy of this model for four types of attacks is 70.6%, 79.1%, 84.2%, and 85.3%. The recognition accuracy of K-medoids and K-means for four types of intrusions is much lower than that of K-medoids-KL. This indicates that K-medoids-KL has excellent recognition performance for various types of intrusions. Figure 10 shows the classification performance results of each model.

Figure 10(a), (b), (c), and (d) show the classification performance of K-medoids-KL, K-mean-KL, K-medoids, and K-mean models on different data. In Figure 10, K-

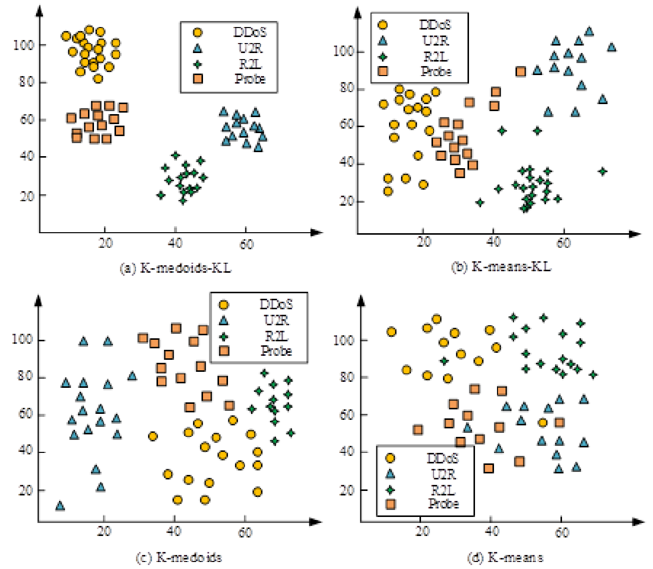


Figure 10: Comparison of classification performance of various models

Table 2: Comprehensive performance comparison of four models

Type	K-means		K-medoids		K-means-KL		K-medoids-KL	
	ACC/%	Time/s	ACC/%	Time/s	ACC/%	Time/s	ACC/%	Time/s
DDoS	81.1	5.3	86.8	3.2	93.7	2.2	95.8	0.9
U2R	83.8	5.2	88.7	3.1	95.3	1.9	97.9	0.8
R2L	79.3	5.9	83.2	3.8	90.2	2.6	92.3	1.5
Probe	71.5	6.1	76.6	4.1	83.6	2.8	85.7	1.7

medoids-KL exhibits the best classification performance, while K-mean has poorer classification performance. In K-medoids-KL, various types of data are basically aggregated together. The outcomes display that K-medoids-KL has excellent clustering performance for different types of data. Table 2 presents the results of comparing the comprehensive performance of various models.

In Table 2, the K-medoids-KL model has the highest accuracy, with recognition rates of 95.8%, 97.9%, 92.3%, and 85.7% for DDoS, U2R, R2L, and Probe, and recognition times of 0.9s, 0.8s, 1.5s, and 1.7s. The accuracy of the K-means model is the lowest, with recognition rates of 81.1%, 83.8%, 79.3%, and 71.5% for DDoS, U2R, R2L, and Probe, and recognition times of 5.3s, 5.2s, 5.9s, and 6.1s. The data proves that the K-medoids-KL model performs well among the four models.

5 Discussion

This study proposed a hierarchical network security protection evaluation model based on improved RBM and K-medoids-KL clustering algorithm, which had superior performance on multiple network security data sets. RBM significantly improved the feature extraction ability through efficient VL and HL connection structure. Compared with traditional BM and GAN models, RBM had better performance in training efficiency and processing time, especially in recognizing multi-category data. RBM improved the ability of data feature extraction by establishing an efficient connection structure between the VL and the HL. The introduction of KL distance enhanced the measurement of the difference between categories in the process of CC selection, which significantly optimized the clustering effect. Compared with K-means, K-medoids-KL reduced the dependence on the selection of the initial clustering center and had advantages in processing time. The comprehensive use of LLM and cluster analysis technology model not only solved the problem of low efficiency of traditional methods in the face of massive data, but also improved the accuracy of anomaly detection through RBM and improved clustering algorithm.

Although the experimental results showed that the model had excellent performance, it may face the following challenges in practical application. The current study was conducted only in the laboratory environment and could not verify the performance of the model in the actual complex network environment. The model placed high demands on the quality and labeling accuracy of the training data, which could be compromised when processing unlabeled or noisy data. Although the K-medoids-KL model was superior in terms of recognition time, it need further optimization in scenarios with higher real-time requirements.

6 Conclusion

The rapid development of the Internet makes it difficult for traditional network security measures to deal with complex and changeable network attacks. Therefore, this study proposed a NSLPE model based on LLM and CA. This model collected data through LLM and classified and recognizes the data through the K-medoids model. In response to the shortcomings of the K-medoids model, the KL distance algorithm was utilized to improve it. Experiments have shown that when the dataset size reached 1000, the accuracy of GAN, BM, and RBM was 0.78, 0.82, and 0.98, with a contrast divergence of 0.34, 0.29, and 0.21. When the training set size was 1000, the recognition times of GAN, BM, and RBM were 442ms, 378ms, and 342ms. When the validation set size was 500, the recognition times of these three models were 804ms, 621ms, and 514ms. The K-medoids-KL model achieved recognition accuracies of 82.1%, 79.6%, 85.5%, and 88.8% for DDoS, U2R, R2L, and Probe. Among the four models, K-medoids-KL showed the best classification performance, with recognition rates of 95.8%, 97.9%, 92.3%, and 85.7% for DDoS, U2R, R2L, and Probe, and recognition times of 0.9s, 0.8s, 1.5s, and 1.7s. Research has shown that the proposed K-medoids-KL model has superior performance. However, there are still uncertainties in this study. The experiment was conducted in a laboratory environment and was not tested in an actual network environment. If the experiment were conducted in an actual network environment, it could make the model of this study more practical.

References

[1] G. M. Borkar, L. H. Patil, D. Dalgade, and A. Hutke, "A novel clustering approach and adaptive svm classifier for intrusion detection in wsn: A data mining concept," *Sustainable Computing*, vol. 23, pp. 120–135, 2019.

[2] B. Chatterjee, I. Walulya, and P. Tsigas, "Concurrent linearizable nearest neighbour search in lockfree-kd-tree," *Theoretical Computer Science*, vol. 886, pp. 27–48, 2021.

[3] C. M. Computing, "Retracted: Computer network security management of data encryption technology," *Wireless Communications and Mobile Computing*, vol. 27, no. 6, pp. 214–223, 2023.

[4] A. Das, A. Namtirtha, and A. Dutta, "Lévy–cauchy arithmetic optimization algorithm combined with rough k-means for image segmentation," *Applied Soft Computing*, vol. 140, no. 12, pp. 110268–110272, 2023.

[5] M. Gheisari, H. Hamidpour, Y. Liu, P. Saedi, A. Raza, A. Jalili, H. Rokhsati, and R. Amin, "Data mining techniques for web mining: A survey," *Artificial Intelligence and Applications*, vol. 1, no. 1, pp. 3–10, 2023.

[6] Y. Gu, K. Li, Z. Guo, and Y. Wang, "Semi-supervised k-means ddos detection method using hybrid feature selection algorithm," *IEEE Access*, vol. 7, pp. 64351–64365, 2019.

[7] F. Hang, L. Xie, and W. Guo, "Pervasive hybrid two-stage fusion model of intelligent wireless network security threat perception," *International Journal of High Performance Systems Architecture*, vol. 10, no. 3/4, pp. 128–139, 2021.

[8] Z. Hengdong, X. Wenxiu, M. Yuanyuan, X. Juan, F. Wang, Y. Qu, and T. Hao, "A new semi-supervised fuzzy k-means clustering method with dynamic adjustment and label discrimination," *Neural Computing & Applications*, vol. 36, no. 9, pp. 4709–4725, 2024.

[9] B. G. Kodge, "Extraction and analysis of snow covered area from high resolution satellite imageries using k-means clustering," *Earth Science Informatics*, vol. 16, no. 4, pp. 4285–4291, 2023.

[10] W. Li, W. Meng, and M. H. Au, "Enhancing collaborative intrusion detection via disagreement-based semi-supervised learning in iot environments," *Journal of Network and Computer Applications*, vol. 161, p. 102631, 2020.

[11] S. T. Lim, J. E. Park, M. Lee, and H. Lee, "Unsupervised object discovery with pseudo label generated using k-means and self-supervised transformer," *Neurocomputing*, vol. 545, no. 8, pp. 1–12, 2023.

[12] M.R. Manesh and N. Kaabouch, "Cyber attacks on unmanned aerial system networks: Detection, countermeasure, and future research directions," *Computers & Security*, vol. 85, pp. 386–401, 2019.

[13] A.B. Mohammed, L.C. Fourati, and A.M. Fakhrudeen, "Comprehensive systematic review of intelligent approaches in uav-based intrusion detection, blockchain, and network security," *Computer Networks*, vol. 239, no. 2, pp. 110140.1–110140.28, 2024.

[14] L. Na, J. Zhou, X. Feng, and K. Chen, "A timeliness-enhanced traffic identification method in airborne network," *Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University*, vol. 38, no. 2, pp. 341–350, 2020.

[15] M. Roger, "Analytical modeling of modulated rotating-blade noise: the skipping rope and the darrieus wind turbine," *Journal of Physics Conference Series*, vol. 1909, no. 1, p. 12010, 2021.

[16] J. R. Saura, D. Ribeiro-Soriano, and D. Palacios-Marques, "Evaluating security and privacy issues of social networks based information systems in industry 4.0," *Enterprise Information Systems*, vol. 16, no. 10-11, pp. 1694–1710, 2022.

[17] F. Shaikh, M. Rahouti, N. Ghani, K. Xiong, E. Bou-Harb, and J. Haque, "A review of recent advances and security challenges in emerging e-enabled aircraft systems," *IEEE Access*, vol. 7, pp. 63164–63180, 2019.

[18] S. Wang, Y. Guo, W. Hua, X. Liu, and G. Song, "Semi-supervised polsar image classification based on improved tri-training with a minimum spanning tree," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, pp. 8583–8597, 2020.

[19] W. Wang, H. Zhou, K. Li, and F. Liu, "Cyber-attack behavior knowledge graph based on capec and cwe towards 6g," in *International Symposium on Mobile Internet Security*, vol. 1544, pp. 352–364. Springer, Singapore, 2022.

[20] Y. Wang, J. Ma, A. Sharma, P.K. Singh, G.S. Gaba, and M. Masud, "An exhaustive research on the application of intrusion detection technology in computer network security in sensor networks," *Journal of Sensors*, no. 3, pp. 1–11, 2021.

[21] H. Zeng and L. Qian, "Airborne gateway: key technology of earth observation system in the space-air-ground integration network," *Journal of Physics: Conference Series*, vol. 1345, no. 4, pp. 42001–42005, 2019.

Biography

Bin Chen received his Ph.D. in Electromagnetic Fields and Microwave Technology from Beijing University of Posts and Telecommunications in 2005. He is currently a Senior Engineer at Huaxin Consulting Co., Ltd., where he is engaged in cybersecurity consulting, solution design, and product development. He has published 15 academic papers and contributed to one national standard. Additionally, he co-authored an academic monograph. His research interests include Artificial Intelligence security, 5G network security , and cybersecurity evaluation.

Bin Li received his B.E. degree in 2007 from Hangzhou Dianzi University, majoring in Automation. He obtained his M.S. degree in 2008 from the University of Hertfordshire, UK, majoring in Network and Data Communications. He has been employed by Huaxin Consulting Co., Ltd. since 2009, working in the fields of data communication and network security.

Yuting Tang obtained her M.S. degree in Electrical and Computer Engineering from the University of California, Los Angeles (UCLA) in 2020. She received her B.E. degree in Information Engineering from Zhejiang University (ZJU) in 2017. She is currently a Cybersecurity Engineer at Huaxin Consulting Co., Ltd., where she was promoted to Senior Engineer in 2021. Her research interests include data security, cybersecurity, and IoT security.

Xiaogang Xie received his B.E. degree in Network Security from Hangzhou Dianzi University in 2021. He is currently working as a cybersecurity engineer at Huaxin Consulting Co., Ltd., where he is engaged in cybersecurity consulting, design, and product development. His research interests include Artificial Intelligence security and 5G network security.