Research on Network Security Intrusion Detection System Based on Machine Learning

Yin Luo

(Corresponding author: Yin Luo)

Sichuan TOP IT Vocational Institute, China

No. 2000, Xiqu Avenue, High-tech District, Chengdu, Sichuan 611743, China

Email: yinjielan@21cn.com

(Received May 17, 2019; Revised and Accepted June 20, 2020; First Online Apr. 25, 2021)

Abstract

This paper mainly analyzed the application of the machine learning method in the intrusion detection system (IDS). The support vector machine (SVM) algorithm parameters were improved by the adaptive particle swarm optimization (APSO) algorithm and the APSO-SVM algorithm, which obtains for intrusion detection. In feature selection, we will compare the proposed method with Relief and InfoGain methods. Experiments were carried out on the KDD CUP 99. The results showed that the proposed method greatly reduced the running time of the algorithm and improved the performance to a certain extent after the dimensionality reduction of features selected by Relief and InfoGain. Comparatively speaking, the feature extracted by Relief performed better in the algorithm. The comparison between SVM, particle swarm optimization (PSO)-SVM, and APSO-SVM algorithms demonstrated that the APSO-SVM algorithm had higher accuracy and lowered false alarm rate and missing alarm rate, i.e.,,, it had better performance in intrusion detection. The results show that the machine learning method is effective on IDS, which contributes to the further realization of network security.

Keywords: Intrusion Detection System; Machine Learning; Network Security; Particle Swarm Optimization; Support Vector Machine

1 Introduction

With the popularity of the network [10], it not only facilitates people's study, work and life but also brings a lot of security problems. The emergence of various viruses, vulnerabilities, and attacks poses a great threat to the security of individuals, enterprises, and even the country. Network security generally needs to ensure the integrity, availability, confidentiality, and controllability of information and prevent information from being leaked, tampered, or destroyed [17].

The current technologies used include access control [5], firewall [20], identity authentication [7], data encryption [24], *etc.*, but they can only carry out passive defense, not real-time monitoring.; therefore, intrusion detection system (IDS) [23] appears. IDS can detect potential threats in time by analyzing network information [13], which has been widely concerned by researchers. Kang *et al.* [12] designed an IDS using a deep neural network (DNN) and used a deep belief network (DBN) to pre-train the initial parameters of DNN [2, 6]. Through experiments, they found that the method had a high detection rate and could make a real-time response to attacks.

Pham *et al.* [18] designed a lightweight IDS, which converted the original network traffic into image data and then used a convolutional neural network (CNN) for detection. The experiment showed that the improved method could achieve 95% accuracy. Muhammad *et al.* [16] designed an IDS based on DNN, reduced the feature width using the stacked automatic encoder (AE), carried out experiments on KDD CUP 99, NSL-KDD, and AWID datasets, and found that the accuracy of the improved method reached 94.2%, 99.7%, and 99.9%, respectively.

Lee *et al.* [14] designed a hybrid IDS combining the C4.5 decision tree with weighted K-means, verified the method, and found that it had a detection accuracy of 98.68%. In this study, the support vector machine (SVM) algorithm in machine learning was studied and improved by combining the particle swarm optimization (PSO) method. The method of feature dimension reduction was also analyzed, and the proposed method was tested on the KDD CUP 99. The present study is conducive to the further development of IDS and better realization of network security.

2 Intrusion Detection System

According to different data sources, IDS can be divided into (1) the host-based IDS, which finds out the intrusion behavior and respond through the analysis of the system and application log, but it can not detect other hosts, not suitable for the current complex network environment; (2) the application-based IDS, which is a refinement of the host-based IDS, mainly for an application; (3) the network-based IDS, which determines whether there are threats through the analysis of network packets, and it is most widely used because of its strong real-time performance and fast detection speed.

According to different detection principles, IDS can be divided into three categories. When the detection principle is anomaly detection [3], the IDS analyzes the characteristics of normal behavior, establishes a model, and determines an intrusion if there is a big difference between the behavior and normal behavior. The common methods include multivariate analysis and neural networks [1]. When the detection principle is misuse detection [8], the IDS analyzes the characteristics of abnormal behavior, establishes a feature library, and determines there is an intrusion if the feature that conforms to the feature library is detected. The common methods include pattern matching, expert system, *etc.*

The essence of IDS is a process of classification, i.e., distinguishing normal behaviors from intrusion behaviors. Therefore, the machine learning method has high availability in IDS [19]. This study mainly analyzes the application of the SVM method.

3 Feature Dimension Reduction Method

In intrusion detection, to reduce the dimension of data and improve the speed of detection, it is necessary to select features. A subset containing M features is selected from a set containing N features ($M < N$) to make the classification performance the best. Two common methods are introduced here.

Relief [22]: The method considers that good features can make the samples of the same class closer to each other and make the samples that do not belong to the same class farther away. The correlation between a feature and a class is represented by weight, and the weight lower than a threshold is removed. It is assumed that sample T is randomly selected from sample set S, and then samples X and Y are also selected and made closest to T; moreover, X and Y belong to the same class, and Y and T belong to different classes. For feature F, the distance between X and T and between Y and T on the feature is calculated. If the former is smaller than the latter, it indicates that the degree of distinction of the feature is good and the weight can be improved; otherwise, the weight is reduced. The updating formula of the weight is written as:

$$W(F) = W(F) + \frac{D(F,T,Y)}{n} - \frac{D(F,T,X)}{n}$$

where $W(F)$ refers to the weight of feature F. The distance between two samples and feature F can be written as:

$$D(F, I_1, I_2) = \frac{|value(F, I_1) - value(F, I_2)|}{\max(F) - \min(F)}$$

where $value(F, I_i)$ refers to the value of sample I_i on F. After n cycles, the feature with a larger weight has better classification performance, which can be used for intrusion detection.

InfoGain [4]: This method selects samples based on information entropy. It is assumed that there are s samples in sample set S, which can be divided into m classes, and class C_i contains S_i samples. The information entropy can be written as:

$$E(C) = -\sum_{i=1}^{m} \rho(C_i) \log_2 \rho(C_i)$$

where $\rho(C_i)$ refers to the probability that any sample belongs to class C_i, $\rho(C_i) = \frac{s_i}{S}$, and $E(C)$ represents the degree of uncertainty of classifying samples in C into m classes.

For feature F, when it is used for classifying S, the degree of uncertainty can be written as: $E(C|F)$. Suppose $F = \{F_1, F_2, \cdots, F_v\}$, then S is divided into: $S = (S_1, S_2, \cdots, S_v)$, and the conditional entropy can be obtained:

$$E(C|F) = \sum_{j=1}^{v} \rho(F_j) E(C|F = F_j),$$

where $\rho(F_j)$ refers to the occurrence probability of feature F_j. When the value of F is F_j, the conditional entropy can be written as:

$$E(C|F = F_j) = -\sum_{i=1}^{m} \rho_{i_j} \log_2 \rho_{i_j}$$

where $\rho_{i_j} = \frac{s_{i_j}}{s_j}$. After substitution, there is: .

$$E(C|F) = \sum_{j=1}^{v} \frac{S_{1_j} + S_{2_j} + \cdots + S_{m_j}}{S} \left(-\sum_{i=1}^{m} \frac{S_{i_j}}{S_j} \log_2 \frac{S_{i_j}}{S_j}\right).$$

The information gain of F is defined as $G(F)$, $G(F) = E(C) - E(C|F)$. The larger the $G(F)$ is, the larger the degree of distinction of F is, and the larger the contribution to the sample division is. In feature selection, the feature with larger $G(F)$ is selected for intrusion detection.

4 SVM Based Intrusion Detection Algorithm

4.1 Principle of SVM Algorithm

The SVM algorithm is a typical machine learning method, which can divide the data into two classes. In intrusion detection, the SVM algorithm can distinguish the normal behavior of the network and intrusion behavior, which has good usability. If there is a data set $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ and its classification hyperplane is $wx + b = 0$, the maximum class interval is calculated: $\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i$, such that $y_i[(wx_i + b)] - 1 + \xi_i \geq 0$, where C is the penalty factors and ξ_i is the slack variable ($\xi_i \geq 0$). The Lagrange factor is introduced to solve the above equation; then, $\min_a \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, y_j) - \sum_{i=1}^{N} a_i$, such that $\sum_{i=1}^{N} a_i y_i = 0$, where $K(x_i, x_j)$ is the kernel function. Finally, the classification function can be written as:

$$f(x) = sgn(\sum_{i=1}^{N} a_i y_i K(x_i \cdot x) + b).$$

In selecting kernel function, the radial basis function (RBF) with good nonlinear mapping ability is selected:

$$K(x_i, x_j) = \exp(-\frac{|x_i - x_j|^2}{\rho^2}$$

In the SVM algorithm, the performance of the algorithm is mainly related to two parameters: penalty factor C and kernel parameter ρ. It is an important problem for the SVM algorithm to find the best parameter value and make the performance of the algorithm the best.

The SVM algorithm is mainly used for binary classification. There are many kinds of intrusion behaviors. To solve the problem of multi-classification, it can be divided into multiple binary classification problems. In this study, the one vs. Rest (OvR) method is used. In each training, it is assumed that there are N classes, samples from one class were positive, and the other samples were negative. In the test, if only one classifier predicts positive, it can be used as the classification result; if multiple classifiers predict positive, the one with the highest confidence is selected. This method only needs to train N classifiers, which needs less time and space.

4.2 Parameter Optimization of the SVM Algorithm

For the parameter optimization of the SVM algorithm, an adaptive particle swarm optimization (APSO) algorithm was designed to select parameters. For the traditional PSO, the value of the inertia weight w has a great impact on the performance of the algorithm. In this study, the value of w is combined with the fitness value of particles. It is assumed that the relative variation rate of the fitness value of particles is: $k = \frac{f_i(t) - f_i(t-1)}{f_i(t-1)}$, where $f_i(t)$ refers to the fitness value of particle i at the t^{th} iteration. The adjustment formula of w is:

$$w_i(t) = (1 + e^{-k})^{-1}$$

The value of w is controlled in $(0, 1)$. When $k = 0$, $w_i(t) = 0.5$. With the increase of $f_i(t)$ value, the value of $w_i(t)$ also increases. Such a method can make the algorithm converge better.

5 Experimental Analysis

5.1 Experimental Data Set

Experiments were carried out on the KDD CUP 99, and 10% of data sets were selected, including the following four types of intrusion.

1) DOS, which makes the network unable to provide normal services, such as land, smurf, *etc.*

2) Probe, which monitors or scans ports to obtain open services, such as saint, ipweep, *etc.*

3) R2L, which is illegal access to remote machines, such as imap, multihop, *etc.*

4) U2R, which can make unauthorized users become privileged users, such as loadmodule, rootkit, *etc.*

In the selected data sets, the number of intrusion behaviors is shown in Table 1.

Table 1: Experimental data sets

Intrusion Behavior	Training Set	Testing set
Normal	97278	60593
Probe	4017	4166
DOS	391458	229853
U2R	52	228
R2L	1126	16189

In KDD CUP99, each record contained 41-dimensional features and intrusion categories, for example, 0, tcp, http, SF, 177, 1985, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 28, 119, 1.00, 0.00, 0.04, 0.04, 0.00, 0.00, 0.00, 0.00, normal. As the second, third, and fourth features were characters, they needed to be transformed into numbers. The second feature was represented by numbers 0-2. The third feature was represented by numbers 0-60. The fourth feature was represented by numbers 0-10. Then, all the values were normalized and transformed to numbers in the range of 0-2. The formula is:

$$x' = \frac{(y_{\max} - y_{\min})(x - x_{\min})}{x_{\max} - x_{\min}} + y_{\min}$$

where $y_{\max}$ and $y_{\min}$ are the maximum and minimum values of normalization and $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of feature attributes.

5.2 Evaluation Index

According to the confusion matrix, the performance of the APSO-SVM-based IDS was evaluated, as shown in Table 2.

The evaluation indexes include:

Table 2: Confusion matrix

		Classification Results	
		Normal Sample	Abnormal Sample
The Real Situation	Normal Sample	TP	FN
	Abnormal Sample	FP	TN

1) Accuracy (ACC):

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

2) False positive rate (FPR):

$$FPR = \frac{FP}{FP+TN} \times 100\%$$

3) False alarm rate (FAR):

$$FAR = \frac{FN}{TN+FP} \times 100\%$$

5.3 Experimental Results

Firstly, two feature selection methods, Relief and InfoGain, were compared. For KDD CUP 99, the top ten features were selected as the input of IDS, and the SVM algorithm was taken as an example to operate ten times. The operation time of the algorithm is shown in Table 3.

It was seen from Table 3 that the operation time of the algorithm became significantly shorter. When the 41-dimensional feature was used as input, the operation time of the algorithm was more than 30 s. After feature selection by Relief and InfoGain, the input was a ten-dimensional feature, and the operation time of the algorithm was less than 20 s. When Relief, InfoGain, and 41-dimensional feature were used as inputs, the average operation time of the algorithm was 14.39 s, 18.78 s, and 37.24 s, respectively. The ten-dimensional feature selected by Relief reduced the operation time of the algorithm by 61.36%, and the ten-dimensional feature selected by InfoGain reduced the operation time by 49.57%. In the aspect of the operation time, the feature selection result of Relief was better.

The ten-dimensional features selected by Relief and InfoGain were used as input, respectively. The operation repeated ten times, and the average value was taken. The performance of SVM, PSO-SVM, and APSO-SVM algorithms in detecting intrusions was compared, and the results are shown in Figures 1 and 2.

It was seen from Figure 1 that the ACC of the three algorithms was 87.42%, 92.34%, and 97.68%, respectively, i.e., the ACC of the APSO-SVM algorithm was the highest, which was 11.74% higher than the SVM algorithm and 5.78% higher than the PSO-SVM algorithm. The FRP of the three algorithms was 2.33%, 1.34%, and 0.17%, respectively, and the FAR was 21.22%, 12.36%, and 7.68%, respectively. It was found that the FPR and FAR of the SVM algorithm significantly decreased after optimization by the PSO algorithm and further decreased after further improvement by the PSO algorithm.

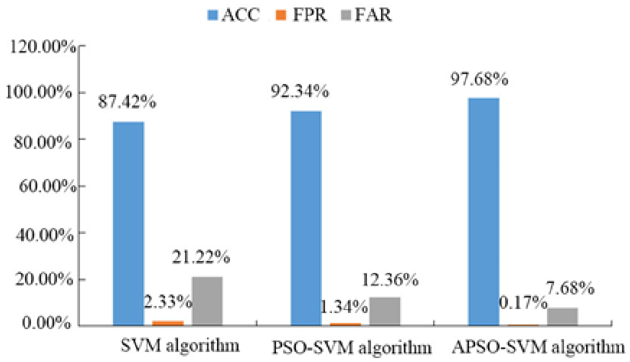

Figure 1: Comparison between algorithms when the feature selected by Relief is used

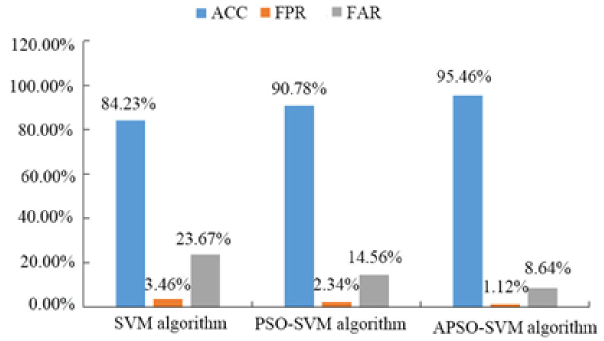

Figure 2: Comparison between algorithms when the feature selected by InfoGain is used

It was seen from Figure 2 that the ACC of the three algorithms was 84.23%, 90.78%, and 95.46%, respectively, the FPR was 3.46%, 2.34%, and 1.12%, respectively, and the FAR was 23.67%, 14.56%, and 8.64%, respectively. Compared with Figure 1, the ACC of the algorithm decreased, and the FPR and FAR increased, when the feature selected by InfoGain was used. It was concluded that the performance of the algorithm was better when the feature selected by Relief was used as the input.

Table 3: Comparison of the operation time of the algorithm

	1	2	3	4	5	6	7	8	9	10
Relief	13.24	14.68	13.68	15.12	14.11	15.18	14.26	15.41	13.96	14.21
InfoGain	19.33	19.12	18.78	18.59	19.03	18.72	19.33	17.64	19.21	18.07
41-dimensional Features	36.78	37.12	36.77	36.81	37.23	38.07	36.95	37.08	38.11	37.45

6 Discussion

The machine learning method is a simulation of human learning by computers [9], which is related to knowledge such as artificial intelligence, biology, and statistics. Its goal is to establish a learning machine from the existing data and classify or predict the unknown data. Up to now, it has been well applied in many fields, such as image processing [21], data classification [15], prediction [11], *etc.* This paper mainly analyzed the SVM algorithm in machine learning and its application in IDS.

Aiming at the problem of parameter optimization of the SVM algorithm, this paper selected the PSO algorithm and improved the SVM algorithm to obtain the APSO-SVM algorithm. Then, in feature selection, to reduce the feature dimension, the performance of Relief and InfoGain algorithms was compared, and the experiment was carried out on the KDD CUP 99 data set.

First of all, the features selected by Relief and InfoGain both significantly reduced the operation time of the algorithm, but the performance of Relief was better as it reduced the operation time of the SVM algorithm by 61.36%, greatly improving the efficiency of the algorithm. Then, in the aspect of the specific performance of the algorithm, when the feature selected by Relief was used, the accuracy of the algorithm became higher, and the false positive rate and false alarm rate became lower, which verified that Relief had a better performance in feature selection. Then, in the aspect of the optimization of the SVM algorithm, the performance of the algorithm significantly improved after optimization by the PSO algorithm and further improved after further optimization by the PSO algorithm. It was seen from Figure 1 that the ACC of the APSO-SVM algorithm was 5.78% higher, the FPR was 87.31% lower, and the FAR was 37.86% lower compared with the PSO-SVM algorithm. It was concluded that the APSO-SVM algorithm designed in this study presented an excellent performance in detecting intrusions.

Though this study has obtained some achievements from the research of the machine learning based-IDS, there are still some shortcomings. In future research, works, including studying more machine learning methods, verifying IDS in the real network environment, and further optimizing the performance of the SVM algorithm, need to be completed.

7 Conclusion

IDS was studied using the SVM algorithm in this paper, an APSO-SVM algorithm was designed for intrusion detection, and experiments were carried out on the KDD CUP 99. The results are as follows.

1) The features selected by Relief and InfoGain both reduced the operation time of the algorithm, and the performance of Relief was better.

2) In terms of accuracy, the performance of the feature selected by Relief was better than that by InfoGain, and the accuracy of the APSO-SVM algorithm was the highest, reaching 97.687%.

3) In terms of false positive rate and false alarm rate, the feature selected by Relief was better, and the false positive rate and false alarm rate of the APSO-SVM algorithm were lower.

It is concluded that the IDS that selects features with Relief and detects intrusions with the APSO-SVM algorithm has better performance, which can be further promoted and applied in practice.

References

[1] N. Abd, K. M. A. Alheeti, S. S. Al-Rawi, "Intelligent intrusion detection system in internal communication systems for driverless cars," *Webology*, vol. 17, no. 2, pp. 376, 2020.

[2] D. S. Abdul Minaam and E. Amer, "Survey on machine learning techniques: Concepts and algorithms," *International Journal of Electronics and Information Engineering*, vol. 10, no. 1, pp. 34–44, 2019.

[3] T. B. Adhi, R. Kyung-Hyune, "HFSTE: Hybrid feature selections and tree-based classifiers ensemble for intrusion detection system," *IEICE Transactions on Information & Systems*, vol. 100, no. 8, pp. 1729-1737, 2017.

[4] M. Alkasassbeh, "An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 22, pp. 5962-5976, 2017.

[5] M. Cheminod, L. Durante, L. Seno, F. Valenza, A. Valenzano, "A comprehensive approach to the automatic refinement and verification of access control

policies," *Computers & Security*, vol. 80, pp. 186-199, 2018.

[6] A. Dewanje and K. A. Kumar, "A new malware detection model using emerging machine learning algorithms," *International Journal of Electronics and Information Engineering*, vol. 13, no. 1, pp. 24–32, 2021.

[7] R. Divya, V. Vijayalakshmi, "Analysis of multimodal biometric fusion based authentication techniques for network security," *International Journal of Security & Its Applications*, vol. 9, no. 4, pp. 239-246, 2015.

[8] S. Elhag, A. Fernández, A. H. Altalhi, S. Alshomrani, "A multi-objective evolutionary fuzzy system to obtain a broad and accurate set of solutions in intrusion detection systems," *Soft Computing*, vol. 23, no. 4, pp. 1321-1336, 2019.

[9] M. I. Jordan, T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255-260, 2015.

[10] A. Kak, "Computer and network security," *Friend of Science Amateurs*, vol. 31, no. 9, pp. 785-786, 2017.

[11] U. Kanewala, J. M. Bieman, A. Ben-Hur, "Predicting metamorphic relations for testing scientific software: A machine learning approach using graph kernels," *Software Testing Verification & Reliability*, vol. 26, no. 3, pp. 245-269, 2016.

[12] M. J. Kang, J. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *Plos One*, vol. 11, no. 6, pp. e0155781, 2016.

[13] N. Khamphakdee, N. Benjamas, S. Saiyod, "Improving intrusion detection system based on snort rules for network probe attacks detection with association rules technique of data mining," *Journal of ICT Research & Applications*, vol. 8, no. 3, pp. 234-250, 2015.

[14] W. Lee, S. Oh, "Efficient feature selection based near real-time hybrid intrusion detection system," *KIPS Transactions on Computer and Communication Systems*, vol. 5, no. 12, pp. 471-480, 2016.

[15] N. Milosevic, A. Dehghantanha, K. K. R. Choo, "Machine learning aided Android malware classification," *Computers & Electrical Engineering*, vol. 61, pp. 266-274, 2017.

[16] G. Muhammad, M. S. Hossain, S. Garg, "Stacked autoencoder-based intrusion detection system to combat financial fraudulent," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1-1, 2020.

[17] E. U. Opara, O. A. Soluade, "Straddling the next cyber frontier: The empirical analysis on network security, exploits, and vulnerabilities," *International Journal of Electronics and Information Engineering*, vol. 3, no. 1, pp. 10–18, 2015.

[18] V. Pham, E. Seo, T. M. Chung, "Lightweight convolutional neural network based intrusion detection system," *Journal of Communications*, vol. 15, no. 11, pp. 808-817, 2020.

[19] N. A. H. Qaiwmchi, H. Amintoosi, A. Mohajerzadeh, "Intrusion detection system based on gradient corrected online sequential extreme learning machine," *IEEE Access*, vol. PP, no. 99, pp. 1-1, 2020.

[20] X. Song, "Firewall technology in computer network security in 5G environment," *Journal of Physics Conference Series*, vol. 1544, pp. 012090, 2020.

[21] S. A. Tsaftaris, M. Minervini, H. Scharr, "Machine learning for plant phenotyping needs image processing," *Trends in Plant Science*, vol. 21, no. 12, pp. 989-991, 2016.

[22] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, pp. S1532046418301400, 2017.

[23] G. B. White, E. A. Fisch, U. W. Pooch, "Cooperating security managers: a peer-based intrusion detection system," *IEEE Network*, vol. 10, no. 1, pp. 20-23, 2015.

[24] R. Yadav, V. Kapoor, "A hybrid cryptography technique for improving network security," *International Journal of Computer Applications*, vol. 141, no. 11, pp. 25-30, 2016.

Biography

Luo Yin, born on July 20, 1982, holds a master's degree and is an associate professor of Sichuan top information technology vocational college. He is interested in big data and artificial intelligence.