Research on Application of Improved K-means Algorithm in Network Intrusion Detection

Fengling Wang*

Hezhou University, School of Mathematics and Computer Hezhou Guangxi 542899, China
* Corresponding author. Tel.: 13945642902; email: wf1232983@163.com
Manuscript submitted February 9, 2018; accepted March 9, 2018.
doi: 10.17706/jsw.13.3.192-200

Abstract: In order to solve the problem of network intrusion detection, traditional k-means algorithm in the process of network intrusion detection application, there are some shortcomings, such as sensitivity to the initial value of clustering center, easy to fall into local optimal value, pre-set number of clusters k value, easy to be affected by noise and isolated points, not suitable for the discovery of non-spherical clusters or clusters of large size difference, etc. so that the network intrusion detection accuracy rate is low, high false detection rate. Aiming at the influence of isolated points on the clustering center of k-means algorithm, this paper firstly optimizes the data set itself, removes isolated points, and makes the data set as spherical as possible. For the selection of the initial clustering location, the maximum similarity distance within the class and the minimum similarity distance between classes are used to dynamically generate new classes, and then the data sets are merged into several classes according to the point density, and the unreasonable clusters are split by combining the minimum support tree clustering algorithm, so that the performance of the intrusion detection system is effectively improved. simulation results show that the improved k-means clustering algorithm is used in the network intrusion detection system to improve the detection rate of anomaly detection, reduce the false detection rate, and provide an effective reference for network detection optimization.

Key words: Intrusion detection; clustering analysis; k-means algorithm; minimum support tree.

1. Introduction

With the rapid development of network technology and network scale, the threat of network intrusion is more and more big, more and more network attacks and means, it is particularly necessary to effectively prevent all kinds of network intrusion. Intrusion detection system (ids), as a proactive security technology, provides real-time protection against network internal attacks, external attacks and misoperation, ensuring that intrusion is intercepted and responded to before the network system is compromised. Intrusion detection based on clustering algorithm is an unsupervised anomaly detection algorithm. by classifying the unlabeled data, new and unknown intrusion types can be found. The traditional k-means algorithm has many problems, such as: k points must be given in advance; The initial value has great influence on the algorithm. Sensitive to noise and isolated points, these problems greatly limit its application in network intrusion detection[1].

Many scholars have made a lot of analysis and improvement for the traditional k-means in intrusion detection. Leonid Portnoy et al. first proposed the application of clustering analysis in unsupervised anomaly detection. K-means clustering algorithm is the first dynamic clustering algorithm based on partition proposed by J.B.MACQueen in 1967. it is one of the most useful clustering algorithms in intrusion detection system. Sarle proposes a statistical criterion for evaluating the number of clusters called Cubic Clustering Criteria, which can

be used to estimate the number of clusters using k-means. An ISODATA algorithm is also proposed, which obtains a reasonable number of types k by automatic merging and splitting of classes. R.O.Duda et al. proposed that when searching for the optimal initial clustering center, the clustering process of randomly selecting the initial class center point is repeated many times, and under certain criteria, the optimal result is selected. J.T.Tou et al. first proposed a method to limit the distance between the initial class center point must exceed the specified threshold, but the disadvantage is that the method cannot avoid the class edge point or noise point as the initial class center point. Weiping proposed a new improved k-means clustering algorithm based on data sample density to select the initial clustering center. although the algorithm has made great progress in the detection effect, it still cannot avoid the randomness of the cluster result set. Fu Tao et al proposed a POS based k-means algorithm to solve the problem that the clustering results of traditional k - mean and algorithm are unstable due to different initial clustering centers, so that the clustering results will not fall into local optimal solution. Ma Xiaochun et al proposed the method of classification number, classification number is determined by the whole process of clustering, but also to make a limit on the size of the cluster, the main problem of this method is that the cluster size threshold selection is more difficult, and it is not easy to distinguish after clustering which class is abnormal data. Xie huihui et al. proposed an intrusion detection system based on ant colony for the deficiency of the existing intrusion detection on unknown attack detection rate and false detection rate. Cai weihong introduced the concept of cluster density and proposed his own density clustering algorithm based on the previous research. the main advantage of this method is to solve the unknown data set detection in the intrusion environment, to explore and analyze the abnormal data in a variety of shapes of data sets, there is a high detection rate, but the algorithm has high time complexity, to be applied to the detection and analysis of this mass data, further improvement is needed[2].

In view of these shortcomings, this paper first optimizes the data set itself, removes the isolated points, and makes the data set as spherical as possible; For the selection of the initial clustering location, the maximum similarity distance within the class and the minimum similarity distance between classes are used to dynamically generate new classes, and then the data sets are merged into several classes according to the point density. combining with the minimum support tree clustering algorithm to split the unreasonable cluster classes, arbitrary shape data classes can be found, which effectively improves the performance of the intrusion detection system. through simulation experiments, compared with the traditional algorithm, the improved algorithm can effectively improve the detection rate and reduce the false detection rate, and provide an effective reference for network detection optimization[3].

2. Intrusion Detection and Cluster Analysis

Intrusion detection system is a kind of network security equipment that monitors network transmission instantly, alerts when suspicious transmission is found or takes active response measures. It differs from other network security devices in that it is a proactive security protection technology. Intrusion detection system is mainly through the network data packets or host log information extraction, analysis, found that the invasion and attack behavior, and to respond to the invasion or attack. The core of intrusion detection system is to analyze whether there is any violation of security policy in the data, and it is particularly important to establish a standard to identify intrusion. Data mining can mine the correlation characteristics of intrusion through a large number of data, and establish accurate behavior standards. Therefore, the application of data mining in intrusion detection has great research value[4]. The data mining process of the intrusion detection system is shown in Fig. 1.

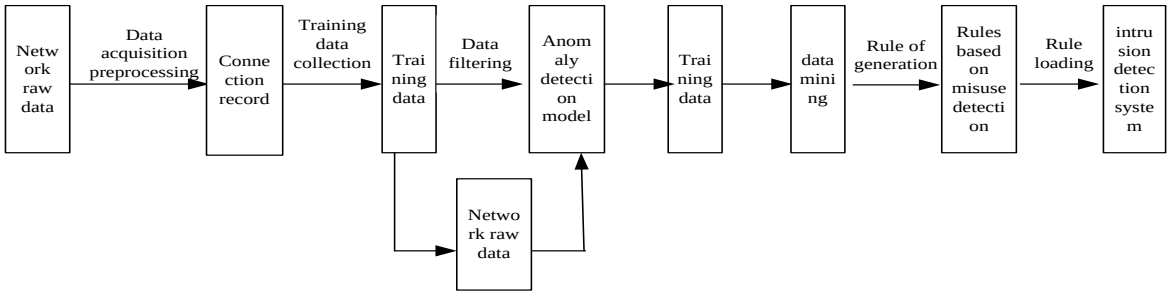

Fig. 1. Data mining of intrusion detection system.

3. Related Notion

Define 1 euclidean distance

$$d(i,j)=(\sum_{k=1}^{P}\left|x_{ik}-x_{jk}\right|^{2})^{1/2}$$

where d(i,j) represents the distance between object i and object j, and p represents the number of attribute values for each object.

The euclidean distance of multidimensional space is calculated as follows:

$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}$ Wherein, x and y respectively represent any two points in the multidimensional space, and any point in the space has k attributes. The coordinates of points x, y in multidimensional space are $(x_1,x_2,\cdots,x_k)$ and $(y_1,y_2,\cdots,y_k)$, respectively.

Define 2 C_i any C_j two clustering sum, the minimum similarity distance between SBC, m_i can m_j be used to cluster center and the minimum value of similarity distance.

$SBC=Min\ (\mathrm{Sim}\ (m_1,\ m_j))\quad (i\neq j)$

Define 3 any alarm record belongs to a class of the maximum similarity distance is:

$SWC=Max\ (\mathrm{SWC}_i)\quad (i=1,2,\cdots,\ K)$

Definition 4 an average SWC_i of the intra-class C_i data object similarity distances containing r data objects can be expressed as:

$$SWC_i=Avg\ (\sum_{h=1}^{r-1}\sum_{j=h+1}^{r}Sim\ (T_h,\ T_j))\quad (h\neq j)$$

Define 5 - point density: any point p in multidimensional space takes r as radius as hyperspace, and the number of data points in hyperspace is called the point density of point p within radius r, which is recorded as d(p,r).

Definition 6 the minimum support tree is a subgraph of the complete graph g(x)={v,e}, satisfies acyclic, has the smallest connection weight, and contains all nodes in g(x). In this paper, the euclidean distance of each edge is defined as the edge weight of the edge, and the minimal support tree is constructed by Prime algorithm.

4. Traditional K-means Algorithm

4.1. Introduction of k-means Algorithm

The k-means algorithm is first proposed by MacQueen. according to the final classification number k input by the user, k initial clustering centers are randomly selected. through iterative calculation, the final clustering results are obtained[5].

Mean square error is usually used as a standard measurement function, defined as formula (1):

$$J = \sum_{i=1}^{K} \sum_{x_j \in c_i} \left| x_j - m_i \right|^2 \tag{1}$$

where x_j is a data point in space and m is the mean of cluster C

The basic idea of that algorithm is: determine the clustering numb k, initializing the first k terms of the data set to a clustering center; The remaining data items are then compared with each clustering algorithm and aggregated into the cluster center with the smallest distance; A new clustering center is then obtained and the operation is repeated until the calculation error function is not significantly changed or the cluster members are no longer changed. Compared with other clustering algorithms, k-means algorithm is simple, scalable, fast, strong ability to deal with large-scale data, so it has advantages in intrusion detection technology[6].

The classical k-means algorithm processing flow is as follows:

(1) randomly selecting k objects as initial cluster centers;

(2) repeat;

(3) assigning each object (re) to the most similar cluster according to the average value of the objects in the cluster;

(4) updating the average value of the clusters, that is, calculating the average value of the objects in each cluster;

(5) until no longer changed.

4.2. The Shortcomings of K-means Algorithm

1. the k-means algorithm is easy to find spherical clusters, while the other shape clusters are difficult to find. And is very sensitive to noise points and isolated points, so that the clustering effect is poor.

2. k-means algorithm to determine the number of k in advance, continue to scan the entire data set, until the cluster center is no longer changed. The clustering results are directly related to the size of k, different numbers will produce different clustering results, in practice, it is often difficult to determine the optimal number of clusters[7].

3. k-means algorithm does not consider the different attribute characteristics in the sample may cause different effects on clustering algorithm, so the clustering effect is not ideal.

4. the efficiency of k-means algorithm depends heavily on the selection of initial k value. in intrusion detection, k value (normal or abnormal behavior pattern) is not known in advance, even uncertain, and whether k value is too small or too large is easy to fall into the trap of local optimum, rather than the overall optimum. therefore, the number of clusters in the algorithm and how to divide is a very big problem, which directly affects the efficiency of intrusion detection system.

5. the k-means algorithm is also affected by the selection of the initial cluster center, different data input order will produce different initial cluster center, random selection of the initial cluster center can not guarantee the selection of the cluster center is the overall optimal, so the efficiency of k-means algorithm depends on the selection of the initial cluster center. How to determine the initial cluster center is also key.

5. Improved K-means Algorithm

5.1. Improvement of Noise and Isolated Points

For each found point I and point I the sum of the appropriate distance, and calculate the distance and h, if the distance of the point and more than the sum of the normal distance h, the point is considered as an isolated point. Where n is the sample data and d is the dimension of the data[8]. Specific definitions are as follows:

$$s_j = \sum_{j=1}^{n} \sqrt{\sum_{h=1}^{d} (x_{ih} - x_{jh})^2}$$

$$H = \sum_{i=1}^{n} \frac{S_i}{n}$$

The algorithm is described as follows:

(1) scanning a data set a once, and calculating the distance sum S_i and the distance average sum h of each data point;

(2) for (each data point I) considers the point as an isolated point if S_i > h;

(3) removing isolated point data in A to obtain a new data set A.

5.2. Improved k-value Dependency

The selection of k value use that empirical algorithm of most scholars: use some reasonably estimated clustering values, compare the reasonable clustering values with the existing clustering values, and judge whether the clustering numbers should be merged or not by using the spatial distance, so as to judge the final k value.

Assume that that distance between the center point of cluster a and cluster b is recorded as d (a, b), the minimum distance between all the data object of cluster a and the center of cluster b is the distance between a and the center of cluster b, and is recorded as j (a, b); The class gap distance of classes a and b is the distance of the center line of classes a and b, expressed as z (a, b). D (a, b), j (a, b), z (a, b) satisfies the following relationship: d (a, b) = j (a, b) + j (b, a) - z (a, b). The average distance e (n) of a class is defined as the average adjacent distance of each of the various points in the class. The average neighbor distance can be obtained by the minimum spanning tree method. the weight of each edge is the distance between two samples, and the sum of the whole minimum spanning tree edges can be obtained. An algorithm description for finding a reasonable number of clusters k is given below[9].

(1) input the initial test value $K = \lfloor \sqrt{n} \rfloor$ of k, input sample set s;

(2) run k-means to obtain k clusters;

(3) calculate d (a, b), j (a, b), z (a, b);

(4) if e (n) > d (a, b) = j (a, b) + j (b, a) - z (a, b) then a, b into a class, k - 1;

(5) output the value of K.

5.3. Selection of Initial Clustering Center

(1) obtaining an initial clustering center

First, the distance between the two samples is calculated, and the two points closest to each other are found to form a sample set a1, which is deleted from the total sample set a'. The distance between each sample in a1 and each sample in a' is calculated. Find the closest point in a' to a1. Incorporate it into a1 and remove it from a'. Until the number of samples in a1 reaches a certain threshold. The two point of that sample closest to each other are found in a' to form a2, and the above proces is repeated until k point sets are formed. Finally, the arithmetic average of k point sets is carried out respectively. K initial cluster centers are for according to that density of points.

Euclidean distance is used for distance. The distance between the sample $X = (x_1, x_2, \ldots, x_n)$ and the sample $Y = (y_1, y_2, \ldots, y_n)$ is calculated according to formula (2):

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \quad (2)$$

Given radius r and r', x_i traverse d, the distance d between each point and any object $I_x \in D, \{D - \{I_x\}\}$ is calculated by formula 1, and the number of objects of d $\leqslant$r is counted. The data set d' is obtain by arranging that point density of all the object in descending order, the points with the maximum point density are adde into the

set c, the radius r of any point X_i in the c is not traversed by d, and the points with the maximum point density in the range are added into the c to obtain an initial $C\{O_1, O_2,\cdots O_m\}$ clustering center set, and c is really included in d, and the number of elements is far less than D.

The distance between a sample point and a sample set is defined as the closest distance between the sample point and all sample points in the sample set. The distance between one sample point x and one sample set v is defined as follows:

$D(X,Y)=\min(d(X,Y),Y\in V)$

Assuming that there are n samples in sample set w that are grouped into k classes and that the initial value of w is 1, the algorithm is described as follows:

1) calculating the distance $D(X,Y)$ between any two samples, finding the two closest points in the set a' to form a set $A_w(1\le W\le K)$, and deleting the two points from the set a';

2) finding the point closest to the set aw in a', adding it to the set aw and deleting it from the set a';

3) repeating step (2) until the number of sample points in the set is greater than or equal to $|\alpha N/K|(0\le\alpha\le 1)$;

4) if w < k, w + + from the set a' to find the nearest two points, form a new set $A_M(1\le M\le K)$ and remove the two points from the set a', return to step (2) execution;

5) respectively carrying out arithmetic averaging on the sample points in the finally formed k sets to form k initial clustering centers according to the point density;

(2) merging to generate an initial class

According to the center set $C\{O_1, O_2,\ldots O_x\}$, with points O_x as the center, will be within the radius of r' all points into a class, modify the class label of these points, and calculate the average value O_x, as the center of the class. Combine to obtain an initial class cent set $C'\{O_1', O_2',\ldots O_x'\}$.

(3) division to produce new classes

Traverse c' to obtain a subset of that data contained in each clas. $D_x'\{I_1', I_2',\ldots, I_K'\}$ represents a minimum support tree $T_x\{E_1, E_2,\ldots, E_h\}$ with O_x' as the center for all data records and a build point of D_x. Deleting an edge whose edge is longer than or equal to the average edge length L_{avg} of the tree to obtain a forest $F_x\{T_1', T_2',\ldots, T_x'\}$, calculating the center $O_y''=center(T_x')\ (T_x'\in F_x)$ of each sub - tree, and counting and changing the class mark of each data object to obtain a new cluster center set $C''\{O_1'', O_2'',\ldots, O_y''\}$.

$$L_{avg}=1/h\sum_{k=0}^{k=h}|E_k|$$

Where h represents the number of edges of the tree and $|E_k|$ represents the length of the edges E_k. The value of a depends on the actual situation. For a sample set a with more uniform data distribution, a larger value should be taken, whereas a smaller value should be taken. After a lot of experiments, when the isolated points are removed to obtain a more uniform distribution of the data set, a 0.8 can get a more accurate clustering center. In order to obtain better clustering results, the best choice is to make k initial agglomeration in the analysis of specific problems can be based on the actual needs, select the appropriate method.

5.4. Variable Correlation Analysis

K-means algorithm is sensitive to the correlation between variables, if the correlation of variables is not considered in the process of clustering, it will have a great impact on the clustering results. Since k-means algorithm uses euclidean distance to calculate the distance between samples, but euclidean distance has nothing to do with the correlation between variables, it is necessary to analyze the correlation between variables before clustering analysis. Principal component analysis method can be used to extract the principal components of the linear combination of the original variables related to each other, and these principal components and variables

not related to any variables are used as input variables of k-means cluster analysis[10].

5.5. Improvement of Clustering Algorithm

After studying the k-means algorithm and analyzing the problem, we get the idea of solving the problem. on the basis of the previous literature, we improve the existing algorithm by two ways. There are two main tasks completed: first, for the optimization of the data set itself, intrusion detection, the face of the data set is unknown, is not suitable for direct use of k-means algorithm to cluster, we must first optimize it, remove the isolated points, so that the data set as much as possible " spherical"; Secondly, the selection of the initial clustering location is no longer random, but by the method of maximum distance to ensure that the distance between class centers should be as large as possible.

Combined with the above analysis, the improved k-means algorithm flow is given as follows:

(1) scanning a data set a once, and calculating the distance sum S_i and the distance average sum h of each data point;

(2) for (each data point I) considers the point as an isolated point if S_i > h;

(3) removing isolated point data in a to obtain a new data set a', recording the number m of samples in a', and outputting isolated points;

(4) calculating the point density of each data record in the data set d

(5) obtaining a new data set containing n data objects after processing.

The set of each cluster D_x (including the set O_x of all elements represented by the center) is obtained, each D_x is split according to the idea of the third aspect of mstk-means algorithm, and the new class center update center set c is calculated. A new proces data set d (a set of centers of all clust classes, i. e., set c) is obtained and set c is emptied.

6. Experimental Results and Analysis

6.1. Test Dataset Selection

In order to verify that the improved algorithm is more efficient than the classical algorithm, KDDcup 99 data set is selected in this experiment. The data set is derived from LAN traffic data from a simulated U.S air force military network environment and is used in the third international knowledge discovery and data mining tool competition, held in conjunction with KDD99, and includes a variety of cyber attacks and intrusions. The whole data set consists of training data set and test data set.

6.2. The Establishment of Intrusion Detection System

The intrusion detection system model is divided into a training part and a detection part as shown in fig. 1. The train part inputs that train data into the system, and the system automatically generates a clustering result as a selected clustering center; The detection part is to standardize the new detection data object, and then compare with the selected clustering center to confirm whether the data object is normal data.

6.3. Experimental Results and Analysis

In the experiment, the detection rate and false detection rate are used as the criteria for judging the advantages and disadvantages of the detection system, and the calculation formula is as follows:

Firstly, cluster label class is carried out according to the training set, then four detection tables are sequentially detected by using the cluster result set, and the detection results of the traditional k-means algorithm and the improved clustering algorithm based on the point density are shown in table 1. The experimental results of mstk-means algorithm show that when radius = 0.01 and ROI = 2, the clustering effect is better, with higher detection rate and lower false detection rate. the detection results are shown in table 2.

Table 1. Test Data Sample Set

sample set	Number of normal records	Number of abnormal records	Number of attack types
Test1	3150	15	3
Test2	3189	12	4
Test3	3768	15	8
Test4	3864	16	9

In the experiment, 20 key attributes were selected for clustering, which were 10 discrete attributes and 15 continuous attributes. The selected data are tested by using the classical k-means algorithm and the improved k-means algorithm respectively. Both algorithm use that same training set to train the label class, and then input other test data sets again, and the result are averaged many times. The test results are shown in Table 2.

Table 2. Test Results

Number of clusters	Detection rate		False detection rate	
	Tradition k-means（%）	Improvement k-means（%）	Tradition k-means（%）	Improvement k-means（%）
10	18.3	22.1	0.42	0.31
20	42.5	52.3	0.53	0.51
30	55.8	76.4	0.81	0.75
40	60.2	87.5	1.23	0.92
50	68.4	90.2	1.45	0.94

From the above table, we can see that the improved k-means algorithm is higher in detection rate than the original algorithm as a whole, and the false detection rate is also significantly lower than the original k-means algorithm. Moreover, with the increase of data objects in the data set, the overall detection effect tends to a perfect value. Through the analysis of the experimental results, we know that the improvement of k-means algorithm in this paper is effective, and has achieved good results in the application of abnormal intrusion detection, the experiment also shows that this improved method is feasible. The improved algorithm has achieved certain results, which makes the new algorithm have a certain application value.

7. Summary

Compared with the traditional algorithm, the detection result obtained by applying the algorithm to the data mining module of the intrusion detection system is more stable. the improved algorithm improves the clustering effect of the data. the intrusion detection system based on the improved algorithm reduces the false detection rate and false alarm rate, and improves the quality of intrusion detection. Compared with the traditional algorithm, the improved algorithm has higher time complexity, to a certain extent, alleviate the initial cluster center selection sensitivity and avoid pre-set number of clusters and other issues, in the future work of learning, need further optimization, further detection and analysis of abnormal points and noise, improve the detection efficiency and reduce false detection rate, to ensure the detection effect while reducing the time complexity, and solve the algorithm itself sensitive to parameters, etc., clustering algorithm in real network intrusion detection system application adaptability and efficiency needs to be further studied.

Acknowledgment

The author thanks to the support from Hezhou University 2016 Professor of scientific research start fund project under Grant No.HZUJS201615.

References

[1] Wang, X., & Liu, S. H. (2014). Improved k-means algorithm in intrusion detection applications. *Computer Engineering and Applications*, (*2*),1- 5.

[2] Li, Y. H., & Zhang, J. (2013). Application of improved k-means algorithm in intrusion detection. *Computer Technology and Development*, *1*,165-168.
[3] Yi, Y. F., & Zhang, Z. P. (2013). K — Mean, algorithm in network intrusion detection application research. *Software Guide*, *12(2)*,124-126.
[4] Wang, F. L. (2010). Computer network security technology and prevention strategy. *Computer Security*. (*3*), 93-95.
[5] Wang, F. L. (2012). Research on the application of IPSec based VPN technology research on. *Computer Technology and Development*, 250-253.
[6] Du, Q., & Sun, M. (2011). Intrusion detection system based on improved clustering algorithm. *Computer Engineering and Applications*, *47(11)*,106-l08.
[7] Chen, Y. P. (2013). K — Mean algorithm in intrusion detection system application and optimization. *Network Security Technology and Application*, 7 - 9.
[8] Cai, W., & Hong, L. Z. (2005). Intrusion detection research based on density clustering algorithm. *Computer Engineering and Application*, *41(21)*, 149 - 151.
[9] Tan, J., & Yi, Y. F.(2008). Improved k-means clustering algorithm in network intrusion detection application research. *Zhongnan University of Nationalities Journal of Natural Science Edition*, 75-78.
[10] Ma, X. C., & Gao, X., & Gao, D. (2005). Cluster analysis in intrusion detection system application research. *Microelectronics and Computer*.

Wang Fengling was born in 1976. He is a professor of Jinxiang County of Shandong Province, graduated from Northeastern University in 2011. He is master of software engineering. He is a senior software designer, J2EE programmer and network engineer. He served as the director of computer center, the director of computer basic research and teaching, director of computer network professional teaching and research section, director of teaching and research section of information management and information system specialty. He is currently director of information management and basic teaching and research section of Heilongjiang Institute of Finance and economics. For many years, we have been engaged in the teaching and management of computer science and technology in the college. Research direction: data mining, educational technology.

In April July 2000 -2003, he served as a teacher in Harbin Information Engineering College. In August April 2003, Professor of Finance and economics in Heilongjiang was appointed professor in August 2016 -2016. More than 50 teaching research papers have been published in various journals such as computer application and software, computer engineering, computer technology and development. There are 22 teaching research papers, including the chief editor, the associate editor and the chief teacher.

Professor Wang Fengling, head of teaching steering group, director of Teaching Steering Committee of HeZhou University, Guangxi science and technology project evaluation expert, director of Guangxi society of artificial intelligence. Has repeatedly went to Shanghai briup software company full-time, digital network university, Shanghai Yangbang Information Technology Co. Ltd., CISCO network to participate in training and practice. We won the two prize and the three prize of teaching achievement in provinces and autonomous regions, the first prize of excellent academic papers, higher education scientific research achievement award and the excellent teaching and research achievement award of 30 provinces.