ISSN: 2692-5079

IKM-NCS: A Novel Clustering Scheme Based on Improved K-Means Algorithm

Weipeng Wang, Shanshan TU and Xinyi Huang

***Abstract*—**Aiming at the problems of distorted center selection and slow iteration convergence in traditional clustering analysis algorithm, a novel clustering scheme based on improved k-means algorithm is proposed. In this paper, based on the analysis of all user behavior sets contained in the initial sample, a weight calculation method for abnormal behaviors and an eigenvalue extraction method for abnormal behavior set are proposed and a set of abnormal behaviors is constructed for each user according to the behavior data generated by abnormal users. Then, on the basis of the traditional k-means clustering algorithm, an improved algorithm is proposed. By calculating the compactness of all data points and selecting the initial cluster center among the data points with high and low compactness, the clustering performance is enhanced. Finally, the eigenvalues of the abnormal behavior set are used as the input of the algorithm to output the clustering results of the abnormal behavior. Experimental results show that the clustering performance of this algorithm is better than the traditional clustering algorithm, and can effectively improve the clustering performance of abnormal behavior.

***Keywords*—**clustering algorithms, k-means, feature extraction, behavior analysis.

I. INTRODUCTION

In recent years, with the explosive growth of data volume, the speed of data accumulation has reached an unprecedented level. In the field of information security, the behavioral data generated by abnormal users born out of these big data are complex and diverse. To build an abnormal behavior set for each user and distinguish the similarities between abnormal behavior sets is conducive to the analysis of similar behavior sets in the future. An excellent similarity distinguishing method can significantly improve the efficiency of analysis. Therefore, the study of abnormal behavior clustering algorithm has become a new method to solve the problem of similarity of abnormal behavior.

In cluster analysis, Fuzzy K-Means (FKM) algorithm is one of the most widely used methods. However, FKM algorithm is much more sensitive to the initialization, and easy to fall into local optimum [1]. To solve the problem of selecting initial cluster centers, Grigorios et al. [2] proposed MinMax K-means algorithm. First, this paper randomly selects a data point as the first initial center, and then selects the data point farthest from the first initial center as the second initial center. The selection of the remaining initial centers meets the following requirements: the k-th initial center is the largest one in the shortest distance from the k-1 initial center. Although this method can separate the initial cluster centers from each other, it cannot guarantee that the separated cluster centers are the optimal cluster centers. Zuo et al. [3] based on the properties of optimal cluster center, proposed the concept of compactness. By eliminating outlier regions, k initial cluster centers are evenly selected in the data compact region. Although this method makes the initial cluster center of the algorithm more reasonable, it adds additional time complexity to the algorithm. Song et al. [4] started with the sparsity of data points, put forward density parameters and distance theory, and determined the initial cluster center based on density theory and maximum distance. Although this method removes all outliers, it cannot guarantee the uniform distribution of the initial cluster centers. Liu et al. [5] proposed a clustering method of predetermined distance, which presets a clustering radius. The first data point is used as the first cluster center. The distance between the second data point and the first data point is calculated. If the distance is less than the preset cluster radius, the data point is divided into the first cluster and the cluster center is recalculated. If the distance is larger than the cluster radius, the data point will be taken as the new cluster center. The distance between the data points and the cluster centers will be calculated successively. If the distance is smaller than the cluster radius, the data points will be classified as the current class, and if the distance is larger, the new cluster center will be classified as the new cluster center. The method can be separated according to the clustering radius to achieve a uniform distribution to the greatest extent. However, the selection of clustering radius is uncertain, thus the clustering performance cannot be guaranteed. Olukanmi et al. [6] modified centroid update for outlier-robust k-means clustering. The classical k-means clustering algorithm is easily misled by outliers, so this method modifies its centroid update step so that outliers are avoided when new centroids are computed. It detects outliers automatically by means of a global threshold derived from the distribution of point-to-centroid distances but gives no considerations to real cluster centers. Wang et al. [7]

This work is supported in part by the National Natural Science Foundation of China (No. 61801008), National Key R&D Program of China (No. 2018YFB0803600), Beijing Natural Science Foundation National (No. L172049), Beijing Science and Technology Planning Project (No. Z171100004717001).

Weipeng Wang is with the Beijing Electro-Mechanical Engineering Institute, 100074, Beijing, China .

Shanshan Tu is with the Beijing Key Laboratory of Trusted Computing, Faculty of Information Technology, Beijing University of Technology, 100124, Beijing, China. (Corresponding author; e-mail: sstu@bjut.edu.cn).

Xinyi Huang is with the Faculty of Information Technology, Beijing University of Technology, 100124, Beijing, China.

analyzed the clustering method based on Kolmogrov-Smirnov test. With an unknown number of clusters, an algorithm capable of estimating the number of clusters and grouping the sequences was proposed and analyzed. But there are too many assumptions in this algorithm, which have a certain impact on the results.

The existing improved methods for K-means initial cluster centers optimize the selection process to a certain extent, but pay little attention to the location distribution of real cluster centers, so the improvement effect is not ideal. To solve this problem, this paper proposes an abnormal behavior clustering algorithm based on modified K-means. In this study, starting from the location distribution of the real cluster centers, the initial cluster centers are selected from the data points with high and low compactness by calculating the compactness of the data, which optimizes the selection process of the algorithm to obtain more reasonable initial cluster centers before the algorithm is executed. Based on this, a clustering algorithm for abnormal behaviors is proposed. In order to verify the advantages and disadvantages of the proposed improved method for selecting initial clustering centers, MinMax K-means algorithm is selected for comparative experiments. The experimental results show that the algorithm we proposed can effectively enhance the clustering performance for abnormal behaviors, and it performs better in terms of iteration times and convergence time.

The main contribution of this study is as follows:

1) To solve the problem of different risk degree of abnormal behaviors, a weight calculation method for abnormal behaviors is proposed.
2) To solve the problem of different similarities of abnormal behavior sets, a method of extracting eigenvalues from abnormal behavior sets is proposed.
3) To solve the problem of low clustering performance of abnormal behaviors, a clustering algorithm of abnormal behavior based on modified K-means is proposed.

II. Research Background

K-means algorithm divides a data set into several similar classes according to some measure (similarity or dissimilarity) to make the samples within the class as similar as possible. The core idea of this clustering algorithm is to divide N data objects into k clusters, in which each data object belongs to a cluster with the nearest cluster center.

K-means algorithm classifies N samples into k classes, $D=\{D_1, D_2, ..., D_k\}$, so that samples within a cluster have high similarity, while samples among different clusters have low similarity. Assume $c=\{c_1, c_2, ..., c_k\}$ are k clusters with corresponding cluster centers, where c_k is the average cluster center of the D_k-th cluster. Samples are divided by using the minimum square error (MSE) function in K-means algorithm, with the objective function defined as:

$$J=\sum_{k=1}^{K}\sum_{x_i\in D_k}\|x_i-c_k\|^2 \qquad (1)$$

If a sample has the minimum distance with center c_k of the k-th cluster D_k, then this sample belongs to cluster , which is described as:

$$D_k=\{x_i\in X\,|\,k=\arg\min_{j\in\{1,2,...,k\}}\|x_i-c_j\|^2\} \qquad (2)$$

$$c_k=\frac{\sum_{x_i\in D_k}x_i}{|D_k|} \qquad (3)$$

K-means algorithm is a kind of greedy algorithm. It obtains the optimal solution of the objective function by iteratively updating the cluster center and each cluster member.

Although K-means algorithm is simple and efficient, it still has some shortcomings: Since the initial cluster centers are chosen randomly, the final clustering results are uncertain. The initial k cluster centers are randomly selected when the algorithm performs iterative operations, which makes the search paths chosen by the algorithm different each time it executes. Therefore, the selection of the initial cluster center has a serious impact on the final clustering results, which makes the algorithm easy to fall into local optimal solution rather than global. If the initial cluster centers which are closer to the real cluster centers can be selected, the clustering performance will be significantly improved.

III. System Design

A. Model overview

The abnormal behavior clustering algorithm consists of three parts, namely weight calculation module, eigenvalue extraction module and clustering analysis module. The specific flow chart of the algorithm is shown in Fig. 1, and the detailed functions of each module are described below.

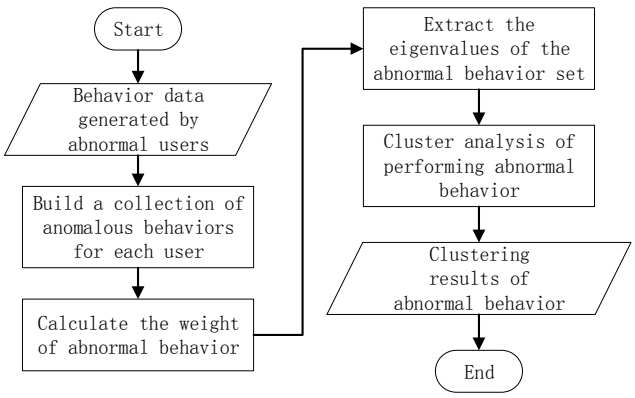

Fig. 1. Flowchart of abnormal behavior clustering algorithm

Weight calculation module: According to the abnormal behavior data, the numbers of occurrences of each abnormal behavior are recorded, and the corresponding weight are assigned according to the frequency.

Eigenvalue extraction module: Since the abnormal behavior often appears in the form of behavior set, according to the weight of each abnormal behavior, the weighted average value of the abnormal behavior set is extracted as the eigenvalue.

ISSN: 2692-5079

Clustering analysis module: the eigenvalues of the abnormal behavior set are read in, and the initial cluster center selection method is improved based on K-means clustering algorithm, to make the algorithm more suitable for the characteristic attributes of abnormal behaviors, improve the execution efficiency of the algorithm, and enhance the clustering performance, and finally the clustering results are output.

B. Module introduction

1) Weight calculation module

This module first uses abnormal behavior data to count the occurrence number N_{A_i} of each abnormal behavior A $(i=1,2,...,n)$, and assign corresponding W_{A_i} weight according to the frequency, with weight calculating equation defined as:

$$W_{A_i}=\frac{N_{A_i}}{\sum_{j=1}^{n}N_{A_j}} \quad (4)$$

2) Eigenvalue extraction module

This module considers that abnormal behaviors often appear in the form of behavior set $S_i=(A_1,A_2,...,A_n)\ i=1,2,...,n$. It is reasonable to extract the weighted average of abnormal behavior set S_i as the eigenvalue x_i, according to the weight W_{A_i} of each abnormal behavior, with eigenvalue extraction equation defined as (5)：

$$x_i=\frac{\sum_{i=1}^{n}W_{A_i}}{n} \quad (5)$$

3) Clustering analysis module

This module first reads in the eigenvalues of the abnormal behavior set, and finally realizes the clustering of abnormal behaviors through the improved K-means clustering algorithm.

1.Improving the selection rules of initial cluster centers

The first step of K-means is to randomly select k data points as the initial cluster center, and then obtain the final clustering results by constant iteration. However, due to the randomness of selecting initial cluster centers, the final clustering results are different. Using randomly selected initial cluster centers is not a good strategy for achieving better clustering results. How to select better initial cluster centers and reduce the algorithm complexity when enhancing clustering performance depends on the following rules.

(1) Avoid over-selection of data points in compact regions as initial cluster centers. Since the data point distribution in data compact area is very dense, there is no obvious regional division between data points, so it is more likely that the data points in such areas belong to the same cluster. If the data points in compact areas are over-selected as the initial cluster centers, the data points far from the compact areas will be neglected. Since the division of a sparse region and a compact region is clear, the data points in the sparse region are more likely to belong to a new cluster, and ignoring the sparse region will lead to lowered efficiency of the algorithm and clustering performance.

(2) Properly select sparse region data points as initial cluster centers. Sparse regions and compact regions have clear regional division and data points in sparse areas distribute sparsely, with obvious distances between sparsely distributed data points. Data points belong to different clustering to a certain extent. Selecting data points in sparse regions as the initial cluster center can better approximate the real cluster center, improve the efficiency of the algorithm, and enhance the clustering performance.

In order to ensure that the initial cluster center is closer to the real cluster center, the improved algorithm firstly divides the data points according to the compactness. The data points with the highest compactness are used as the first initial cluster center and those with the lowest compactness are used as the second cluster center. Then, data points with compactness higher than average compactness are removed and data points in sparse areas are retained. Next, the remaining initial cluster centers are randomly selected in the sparse region, which can enhance the clustering performance and improve the execution efficiency of the algorithm as much as possible.

The algorithm steps are detailedly introduced as follows.

Step1: For each data point x_i in data set $X=\{x_1,...,x_i,...,x_j,...,x_n\}$, calculate the compactness of x_i, $T(x_i)=\frac{n}{\sum_{j=1,x_j\in X}^{n}D(x_i,x_j)}$, where $D(x_i,x_j)$ is the Euclidean distance function of points x_i and x_j;

Step2: Select point x_i corresponding to the maximum $T(x_i)$ value as the first initial cluster center;

Step3: Select point x_j corresponding to the minimum $T(x_j)$ value as the second initial cluster center;

Step4: Remove compact data points in X with compactness $T>\frac{\sum_{x\in X}T(x)}{n}$, resulting in sparse data point set X';

Step5: If $K\geq 3$, then in sparse data point set X', the $k(3\leq k\leq K)$-th initial cluster center c_k satisfies the following condition $c_k=Random(X')$, where $c_k=Random(X')$ means randomly selecting a sparse data point in data set X' as initial cluster center c_k. Repeat Step5 until K initial cluster centers are selected.

Through the proposed improved method, the selection of initial cluster centers is more advantageous than random selection. Firstly, the points with the largest and the smallest degree of compactness are selected to ensure that the two points belong to different clusters, which meets the requirements of selecting cluster centers. Secondly, random selection of data points from sparse regions can ensure that there is a significant distance between selected data points, which conforms to the characteristics of cluster center. Moreover, random selection can reduce the complexity of the algorithm to the greatest extent when enhancing the clustering performance, improving

ISSN: 2692-5079

the algorithm efficiency.

2.Abnormal behavior clustering based on improved K-means

The eigenvalues extracted from the behavior set with similar abnormal behaviors have high similarity, which provides the characteristic attributes for the precise clustering of abnormal behaviors. In addition, K-means algorithm itself is sensitive to characteristic attributes. Therefore, based on improved K-means, an abnormal behavior clustering algorithm is proposed, which can effectively enhance the clustering performance for abnormal behaviors and better distinguish similar abnormal behaviors.

The implementation of the algorithm is introduced as follows.

Input: Eigenvalue set of abnormal behavior set $X = \{x_1,..., x_i,..., x_n\}$:

Output: Clustering results of K clusters after clustering $D = \{D_1, D_2,..., D_k\}$:

Step1: Read in eigenvalue set $X = \{x_1,..., x_i,..., x_n\}$ of abnormal behavior set;

Step2: For each data point x_i of data set $X = \{x_1,...,x_i,...,x_j,...,x_n\}$, calculate the compactness of x_i, $T(x_i) = \frac{\sum_{j=1, x_j \in X}^{n} D(x_i, x_j)}{n}$, where $D(x_i, x_j)$ is the Euclidean distance function of points x_i and x_j;

Step3: Select point x_i corresponding to the maximum $T(x_i)$ value as the first initial cluster center;

Step4: Select point x_j corresponding to the minimum $T(x_j)$ value as the second initial cluster center;

Step5: Remove compact data points in X with compactness $T > \frac{\sum_{x \in X} T(x)}{n}$, resulting in sparse data point set X';

Step6: If $K \geq 3$, then in sparse data point set X', the $k(3 \leq k \leq K)$-th initial cluster center c_k satisfies the following condition $c_k = Random(X')$, where $Random(X')$ means randomly selecting a sparse data point in data set X' as initial cluster center c_k. Repeat Step6 until K initial cluster centers are selected;

Step7: Calculate the Euclidean distance of each data point to each cluster center $D(x_i, c_j)$, where $i = 1,2,...,n$, $j = 1,2,...,K$. When the data point x_m satisfies $D(x_m, c_j) = \min(D(x_m, c_j))$, $j = 1,2,...,K$, is classified into cluster D_j represented by c_j;

Step8: When all data points are divided into corresponding clusters, the cluster center $c = \{c_1, c_2,..., c_k\}$ is updated and the clustering criterion function $J = \sum_{i=1}^{K} \sum_{d_j \in D_i} D(d_j, c_i)$ is calculated.

Step9: Repeat Step7 and Step8, until J is not changed, and then output the clustering result $D = \{D_1, D_2,..., D_k\}$ with K clusters.

The proposed abnormal behavior clustering algorithm is based on the improved K-means clustering algorithm. According to the fact that the behavior set with similar abnormal behaviors contains similar eigenvalues, the clustering is processed and finally the clustering for abnormal behaviors is realized. Since the used K-means algorithm is greatly influenced by the initial clustering center, optimizing the selection of the initial clustering center can not only improve the clustering quality, but also effectively enhance the clustering performance for abnormal behaviors.

IV. Performance Analysis

The experiment mainly focuses on the clustering performance of abnormal behaviors, and the main comparative evaluation indices include the number of iterations and the convergence time. In this experiment, the experimental configuration includes Ubuntu 16.04, IDEA, CPU 2.6GHz with 8.0GB memory. The experimental data is based on Yeast data set in UCI machine learning database [8].

The experiment analyzes some algorithms proposed recently. Huan et al. [9] proposed a clustering method based on KL divergence. Based on the maximum distance method, K data points with large distribution difference are selected as the initial cluster centers, and the similarity between the cluster centers and the sample data is obtained through KL divergence. Yu et al. [10] proposed a clustering method based on bootstrap sampling. Based on the bootstrap sampling, a new method is proposed to determine the best clustering number. Since there are many improved algorithms for K-means, in order to verify the advantages and disadvantages of the proposed improved method for selecting initial clustering centers, MinMax K-means algorithm is selected for comparative experiments.

A. Iteration number

In order to verify the clustering quality of the algorithm, the number of iterations is used to evaluate the experiment. If the number of iterations is less, it proves that the initial clustering center is closer to the real clustering center, and the selection result is more reasonable. In addition, as the iteration number decreases, the accuracy of clustering increases, and the algorithm is more efficient. The experimental results are shown in Fig. 2.

As shown in Fig. 2, with the increase of the cluster number K, the iteration numbers of MinMax K-means algorithm and the improved algorithm deviate gradually. Starting from $k = 8$, the difference of iteration times between the two algorithms increases significantly. This is because MinMax K-means algorithm does not start from the real distribution of clustering centers, but the improved algorithm takes it into account and pays more attention to data points with low compactness. When the cluster number K increases, the data points with low compactness are closer to the cluster center of the new cluster, and then the number of iterations of the improved algorithm is effectively reduced, resulting in an increase of iteration number difference between the two algorithms.

ISSN: 2692-5079

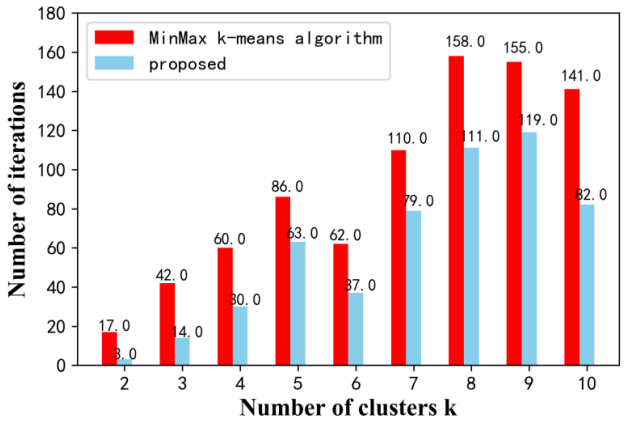

Fig. 2. Iteration numbers of MinMax K-mean algorithm and improved algorithm

Table1. Simulation parameters and Environment

Cluster number (k)	MinMax K-means (Times)	Proposed (Times)	Decrement (Times)
2	17	3	14
3	42	14	28
4	60	30	30
5	86	63	23
6	62	37	25
7	110	79	31
8	158	111	47
9	155	119	36
10	141	82	59

It is clear in Table 1 that the iteration number decreases by the smallest 14 times when $k=2$, while the iteration number decrement steadily increased with the increase of the number of clusters K. When $k=10$, the iteration number decrement reaches the largest 59 times, and the average number of iterations decreased by 43.2%. This is because the proposed improved algorithm first chooses the points with the greatest and smallest compactness, thus guaranteeing that the two initial clustering centers belong to different clusters. Secondly, random selection of initial cluster centers from data points with low compactness can ensure that there is a significant distance between the selected data points, and ensure to the maximum extent that they belong to different clusters. When the cluster number K increases, the initial cluster centers are closer to the real cluster centers, which makes the iteration number of the proposed algorithm decrease significantly and accelerates the convergence of the proposed algorithm. The iteration number of the improved algorithm is much lower than that of MinMax K-means algorithm.

B. Convergence time

In order to verify the efficiency of the algorithm, the convergence time is used for evaluation. The shorter the convergence time, the faster the algorithm runs and the higher the execution efficiency. In addition, due to the reduction of convergence time, the processing efficiency of the algorithm is increased, and the clustering performance is improved. The experimental results are shown in Fig. 3.

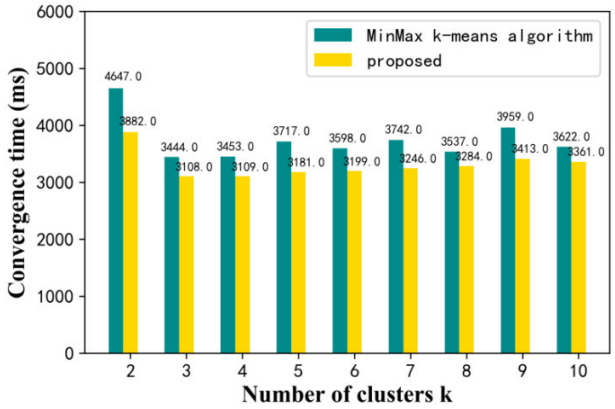

Fig. 3. Convergence time of MinMax K-means and improved algorithms

As shown in Fig. 3, the convergence time of MinMax K-means and improved algorithms approaches gradually with the increase of clustering number K. Starting from $k=3$, the convergence time of the two algorithms approaches gradually. This is because MinMax K-means algorithm uses uniform selection of initial clustering centers, resulting in a large number of distance operations in the selection process, which seriously affects the efficiency of the algorithm. The improved algorithm considers the distribution of real clustering centers, and adopts random selection of initial cluster centers in data points with low compactness, which greatly improves the efficiency of the algorithm while guaranteeing the clustering effect. Although the stochastic selection process becomes more complex when the cluster number K increases, resulting in a slight increase in convergence time, the convergence time of the improved algorithm is still lower than that of MinMax K-means algorithm. The convergence time is shown in Table 2.

It is clear in Table 2 that when $k=2$, the acceleration ratio of convergence time was 16.5% of the maximum at that time. With the increase of cluster number K, the acceleration ratio of convergence time decreased gradually, until it reached the minimum of 7.2% when $k=10$, and the convergence time decreased 11.5% on average. This is because the proposed improved algorithm randomly selects data points from sparse regions, avoiding the tedious operation of selecting initial clustering centers, and maximizing the execution speed of the algorithm. In addition, data points are selected from sparse areas to ensure that there is a significant distance between selected data points. Thus, under the condition of ensuring the algorithm efficiency, the selected data points can be divided into different clusters to the greatest extent, which not only improves the algorithm efficiency, but also has a better initial clustering center. This also reduces the convergence time of the algorithm, accelerates the convergence of the algorithm, and improves the algorithm efficiency. Although the acceleration ratio decreases, the convergence time of the improved

ISSN: 2692-5079

algorithm is still better than that of MinMax K-means algorithm.

Table 2. Comparison of simulation results with different number of SC base stations

Cluster number (k)	MinMax K-means (ms)	Proposed (ms)	Acceleration ratio (%)
2	4647	3882	16.5
3	3444	3108	9.8
4	3453	3109	10.0
5	3717	3181	14.4
6	3598	3199	11.1
7	3742	3246	13.3
8	3537	3284	7.2
9	3959	3413	13.8
10	3622	3361	7.2

V. CONCLUSION

By distinguishing the dangerous degree of abnormal behaviors from the similarity degree of abnormal behavior set, a weight calculation method for abnormal behaviors and an eigenvalue extraction method for abnormal behavior set are proposed. By distinguishing the compactness of data points, the selection process of initial cluster centers is optimized, so that K-means algorithm can obtain more reasonable initial cluster centers before execution. Based on this, a clustering algorithm for abnormal behaviors is proposed. The experimental results show that the algorithm can effectively enhance the clustering performance for abnormal behaviors, and it performs better in terms of iteration times and convergence time. With the development of abnormal behaviors towards diversification, machine learning algorithms adapted to higher dimensions and larger scale are the next direction of future study.

REFERENCES

[1] C. M. Emre, H. A. Kingravi, and P. A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm", *Expert Systems with Applications*, vol.40, no.1, pp.200-210, 2012.

[2] T. Grigorios, A. Likas, "The MinMax k-Means clustering algorithm", *Pattern Recognition*, vol.47, no.7, pp.2505-2516, 2014.

[3] J. Zhuo, Z. Chen, "Anomaly detection algorithm based on improved k-means clustering", *Computer science*, vol.43, no.8, pp.258-261, 2016.

[4] X. Song, Z. Gao, and L. Liu, "Research on network anomaly detection method based on data mining", *Electronic technology*, vol.45, no.11, pp.30-32, 2016.

[5] H. Liu, X. Hou, and Z Yang, "Research and design of intrusion detection system based on clustering and association", *Computer technology and development*, vol.23, no.7, pp.133-137, 2015.

[6] P. O. Olukanmi and B. Twala, "K-means-sharp: Modified centroid update for outlier-robust k-means clustering," *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics(PRASA-RobMech)*, pp. 14-19, 2017.

[7] T. Wang, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney, " Exponentially Consistent K-Means Clustering Algorithm Based on Kolmogrov-Smirnov Test," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2296-2300, 2018.

[8] A. Asuncion, D. Newman, "UCI Machine Learning Respository," *Electronic technology*, vol.40, 2015.

[9] Z. Huan, Z. Pengzhou, and G. Zeyang, "K-means Text Dynamic Clustering Algorithm Based on KL Divergence," *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp. 659-663, 2018.

[10] L. Yu, and C. Zhou, "Determining the Best Clustering Number of K-Means Based on Bootstrap Sampling," *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, pp. 78-83, 2018.

[11] S. Choi, Y. Choi, J. Lee, et al, "Network abnormal behaviour analysis system," *International Conference on Advanced Communication Technology*, 2017.

[12] L. Yang, F. Wang, T. Wang, "Analysis of dishonorable behavior on railway online ticketing system based on k-means and FP-growth," *Proceedings of the 2017 IEEE International Conference on Information and Automation*, pp.1173-1177.

[13] Y. Hu, L. Pang, Q. Pei, et al, "Instruction Clustering Analysis for Network Protocol's Abnormal Behavior," *2015 10th International Conference on P2P*, pp.791-793, 2015.

Weipeng Wang received his Master degree from School of Reliability and Systems Engineering at Beihang University in 2015. He is currently an engineer in Beijing Electro-Mechanical Engineering Institute, China. His research interests are in the areas of integrated avionics, cloud computing and information security techniques.

Shanshan Tu received the Ph.D. degree from the Computer Science Department, Beijing University of Posts and Telecommunications, in 2014. From 2013 to 2014, he visited the University of Essex for national joint doctoral training. He was with the Department of Electronic Engineering, Tsinghua University, as a Post-Doctoral Researcher, from 2014 to 2016. He is currently an Assistant Professor with the Faculty of Information Technology, Beijing University of Technology, China. His research interests are in the areas of cloud computing, MEC, and information security techniques.

Xinyi Huang received her B.Sc. from Beijing University of Technology, China in 2017, and is currently pursuing the M.Sc. in Beijing University of Technology. Her research interests are in the areas of pattern recognition and machine learning.