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

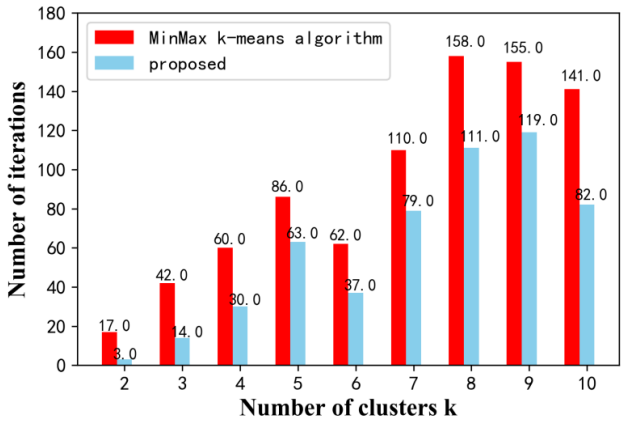

Fig. 2. Iteration numbers of MinMax K-mean algorithm and improved algorithm

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.

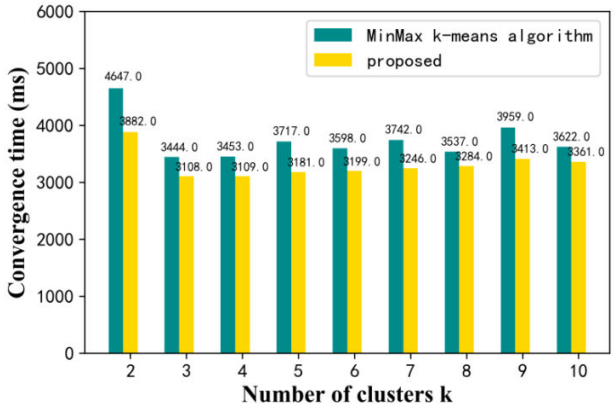

Fig. 3. Convergence time of MinMax K-means and improved algorithms

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.