DB-Kmeans:An Intrusion Detection Algorithm Based on DBSCAN and K-means

1st Gangsong Dong
Information & Telecommunication Co. of State Grid Henan Electric Power Company
Zhengzhou,China
489273711@qq.com

2nd Yi Jin
Information & Telecommunication Co. of State Grid Henan Electric Power Company
Zhengzhou,China
809259350@qq.com

3rd Shiwen Wang
Information & Telecommunication Co. of State Grid Henan Electric Power Company
Zhengzhou,China
wwangshiwen@163.com

4th Wencui Li
Information & Telecommunication Co. of State Grid Henan Electric Power Company
Zhengzhou,China
851746046@qq.com

5th Zhuo Tao
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications, Beijing, China
934501954@qq.com

6th Shaoyong Guo*
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications, Beijing, China
syguo@bupt.edu.cn

***Abstract*—Recently, with wide use of internet and rapid growth of computer networks, the problem of intrusion detection in network security has become an import issue of concern. In this paper, a new intrusion detection algorithm DB-Kmeans has been introduced which combines K-means with DBSCAN. DB-Kmeans uses a new selection method of initial cluster center in K-means and set the neighborhood radius in DBSCAN to dynamic. Compared to K-means algorithm, it overcomes the shortage of sensitivity to initial centers and reduces the impact of noise points. Compared to DBSCAN algorithm, it reduces the influence of fixed neighborhood radius. The experiments on the NSL-KDD data set indicate that the proposed method is more efficient than that based on MinMax K-means algorithm. Also, the method has higher detection accuracy and lower false alarm rate.**

Keywords—K-means, DBSCAN, intrusion detection, density function, dynamic neighborhood radius

I. INTRODUCTION

Computer and network security is gaining importance as increase in number of attacks targeting confidentiality, integrity, and availability of the data. Intrusions are targeting individual's or organization's network to steal their data. Many schemes and efforts have been done to detect the intrusions to the data [1]. As an important network supervision and control method, network intrusion detection and analysis is a key link in network monitoring and management. By monitoring and analyzing traffic and discovering abnormal phenomena in the network in time, it is of great significance for maintaining the normal operation of the network.

In order to further improve the efficiency of intrusion detection, this paper presents an improved algorithm based on the combination of K-means and DBSCAN. The algorithm DB-Kmeans first uses the improved DBSCAN algorithm to eliminate noise points; and then uses the improved K-means algorithm to divide the data to be detected into different clusters and mark it as normal or abnormal in order to achieve the purpose of improving the detection efficiency. Finally, the NSL-KDD data set is used to verify the superiority of the algorithm. Concretely, for DBSCAN, we set the neighborhood radius to dynamic. For K-means, we define a density function and use it to select initial cluster centers.

The remainder of this paper is organized as below. The related work will be described in Section II. The proposed algorithm DB-Kmeans is presented in Section III. In Section IV, we show the experiment and evaluation. Finally, we draw some conclusions and future works in Section V.

II. RELATED WORK

S.Varuna et al. [2] proposes a new hybrid learning method, that integrates k-means clustering and naïve bayes classification. A relation between the distances from each data sample to a number of centroids found by a clustering algorithm is introduced. This is used to form new features, based on the features of the original data set. These distance sum-based features are then used for classifier training and detection.

Mohsen Eslamnezhad et al. [3] proposes a network intrusion detection algorithm based on MinMax K-means clustering algorithm. This algorithm overcomes the shortcomings of k-means sensitive to the initial clustering center and improves the clustering performance.

Anand Sukumar J V et al. [4] uses the improved genetic K-means algorithm for intrusion detection, which improves the accuracy of detection.

Zhang Xiaofeng et al. [5] proposes an improved K-means and multi-level SVM fusion of semi-supervised learning network intrusion detection algorithm. The algorithm first uses the improved K-means to divide the data to be detected into different clusters and mark it as normal or abnormal; and then use the multi-level SVM to classify the clusters marked as abnormal in order to achieve the purpose of improving the detection efficiency.

Hatim Mohamad Tahir et al. [6] proposes an integrated machine learning algorithm using K-means clustering with discretization technique and Naïve Bayes Classifier. The algorithm improves the detection rate and accuracy.

In Vandana Shakya's et al. [7] propose work, classification of KDD intrusion dataset is proposed along with noise reduction, clustering and feature selection. DBSCAN algorithm has been applied to reduce noise present in KDD dataset. After noise removal genetic search approach is utilize to pick relevant feature. K-Means++ clustering method is utilized to cluster the dataset and resultant dataset is tested by SMO based classifier.

Yi Yi Aung et al. [8] proposes a hybrid method using K-means and Projective Adaptive Resonance Theory. Experimental results show that it can reduce model training time and maintains the accuracy of detections.

III. A NEW INTRUSION DETECTION ALGORITHM BASED ON DBSCAN AND K-MEANS

Since the K-means clustering center is greatly affected by noise points, and DBSCAN can identify discrete points of spatial data, this paper combines K-means and DBSCAN to make the clustering results more accurate. Because K-means is sensitive to the selection of the initial clustering center, this paper improves the selection method of the initial clustering center in K-means. The traditional DBSCAN algorithm is sensitive to two parameters: neighborhood density threshold MinPts and neighborhood radius Eps. When the data set has a category with uneven density distribution, the traditional DBSCAN algorithm is difficult to set two parameters. In response to this problem, this paper improves the DBSCAN algorithm, using fixed MinPts and dynamically adjusted Eps.

A. Improvements to the K-means Algorithm

K-means [9] clustering algorithm is a well-known technique of cluster data while it has a major drawback to select initial seed points. As initial centroids have a heavy impact in final cluster sets in depends exclusively on the selection of initial seed points. In this paper, we define the density function as the formula (1). We select the point with the highest density function value as the first initial cluster center, and then select the point farthest from the point as the second initial cluster center. For the selection of the sth center point is satisfied as

$$\max\left(\mathrm{d_{min}}(\mathrm{x_s}, \mathrm{C_1}), \mathrm{d_{min}}(\mathrm{x_s}, \mathrm{C_2}), \dots, \mathrm{d_{min}}(\mathrm{x_s}, \mathrm{C_{s-1}})\right) \quad (1)$$

Until k initial cluster center points are obtained.

In the data space R^d, there are data objects x and x', and the influence function of data point x' on data point x is defined as $\mathrm{Density}(\mathrm{x}, x')$. Gaussian functions are classical influence functions, which are defined as follows

$$\mathrm{Density}(\mathrm{x}, \mathrm{x}') = \mathrm{e}^{\frac{-\mathrm{d}(\mathrm{x},\mathrm{x}')^2}{2\delta^2}} \quad (2)$$

The density function of the data point x is the sum of the influence functions of all nearest neighbors in the range of the neighborhood parameter δ. That is, if n data objects $\mathrm{X} = (x_1, x_2, \dots, x_n)$ are given, the density function for the data point x can be defined as follows

$$\mathrm{Density}(\mathrm{x}, \mathrm{x}') = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\frac{-\mathrm{d}(\mathrm{x},\mathrm{x}')^2}{2\delta^2}} \quad (3)$$

In conclusion, we define a density function which can help us select K points as initial cluster centers.

B. Improvements to the DBSCAN Algorithm

Density Based Spatial Clustering of Applications with Noise(DBSCAN) [10] is the most used and a typical density-based clustering algorithm. It can discover clusters of arbitrary shape, can distinguish noise.

The traditional DBSCAN algorithm is sensitive to two parameters: neighborhood density threshold MinPts and neighborhood radius Eps. When the data set has a category with uneven density distribution, the traditional DBSCAN algorithm is difficult to set two parameters. In response to this problem, this paper improves the DBSCAN algorithm, using fixed MinPts and dynamically adjusted Eps.

The main idea of the algorithm is: when performing neighborhood search, the neighborhood radius is dynamically adjusted according to the density ratio of the neighbor object to the current core object to adapt to the local density change of the data set. It involves the following definitions:

Definition 1

Den(Eps): Eps-neighborhood density, number of data points in the Eps-neighborhood of data point p.

Definition 2

$\eta(\mathbf{q}, \mathbf{p})$: Neighborhood coefficient, given data points p and q, q belongs to the Eps-neighborhood of p, the neighborhood coefficient of data point q with respect to p is defined as:

$$\eta(\mathrm{q}, \mathrm{p}) = \frac{\mathrm{Den(q,Eps)}}{\mathrm{Den(p,Eps)}} \quad (4)$$

$$\mathrm{Eps_q} = \eta(\mathrm{q}, \mathrm{p}) * \mathrm{Eps_p} \quad (5)$$

where the same neighborhood radius is used to calculate the neighborhood density of p and q, which are the neighborhood radius Eps of the core point p. The neighborhood density measured by the same neighborhood radius is comparable.

The neighborhood coefficient reflects the density variation in the local range of the data set, and different density distributions can be clustered by different neighborhood radius. The algorithm uses the neighborhood coefficient to adjust the parameter Eps.

After the neighborhood radius is adjusted, when the density is expanded, from the high-density area to the low-density area, the neighborhood radius is gradually reduced, the density expansion speed is slowed, and when there is no object in the neighborhood, the expansion is stopped; from the low-density area In the high-density area, the neighborhood radius gradually increases (but cannot be greater than or equal to the set maximum value),and the density expansion speed increases. In this way, the algorithm can adapt well to its density change when performing local clustering on the density hierarchy representative set.

In conclusion, we set the neighborhood radius to dynamic so that the influence of the initial neighborhood radius is reduced.

C. DB-Kmeans

In this paper, we propose a new intrusion detection algorithm DB-Kmeans. we make some improvements on K-means and DBSCAN in order to overcome their respective shortcomings. Meanwhile, we combine K-means with DBSCAN so that can eliminate the effects of noise points. The DB-Kmeans algorithm is composed of the following steps:

Algorithm DB-Kmeans

1. input: data set D,
   initial neighborhood radius Eps_p,

neighborhood density threshold minPts, number of clusters k.
2. output: clustering result.
3. mark all point in D as unvisited.
4. for each p in D
5. if p.visited == unvisited
6. find the set M of all points whose distance from point p is not greater than Eps_p.
7. if M.size() < minPts
8. mark point p as a noise point
9. else
10. for each p_1 in M
11. if p_1.visit == unvisited
12. find the set M_1 of all points whose distance from point p is not greater than Eps_{p_i}.
13. calculate (5)
14. if M_1.size() >=minPts
15. add the point that the set M_1 doesn't belong to the set M to the set M.
16. end if
17. else
18. if p_1 is not clustered into a cluster
19. gather p_1 to the current cluster.
20. if p_1 is marked as a noise point
21. cancel the noise point mark of p_1.
22. end for
23. if point.getNoised()
24. put the point into D_{none}.
25. calculate the density function value for each point in D_{none}: (3).
26. select the point with the highest density function value as the first initial cluster center, select the point farthest from the point as the second initial cluster center, the sth initial cluster center is satisfied (1)
27. calculate the distance from each point in D_{none} to the center of the cluster.
28. add the data point to the closest cluster.
29. for each cluster
30. calculate the mean and update to the cluster center.
31. end for

IV. SIMULATION

A. Data preprocessing

This paper uses the NSL-KDD dataset, which contains three characteristics of each piece of data.1~9 is the basic information feature, 10~22 is the content feature, 23~31 is the time-based traffic feature and 32~41 is the host-based traffic feature. The training set contains 22 types of attacks. In order to simulate some new attacks in the actual situation, the test set contains 17 types of attacks that have not appeared in the training set. All attack data can be divided into the following four types of attacks: Denial of Service (DoS), Remote Network User Attack (R2L), Enhanced Permission (U2R), and Probe (Probe).

Since each piece of data in the NSL-KDD data set not only has digital attribute features, but also character-type attribute features, it is necessary to map the character-type features into digital-type attributes. In this paper, the character-type features are encoded by means of dummy variable coding. Since each feature dimension in the feature vector is different, the range of values is different, so standardization is required. The specific steps of data preprocessing are: first use dummy coding for character attribute mapping, then normalize the data attributes processed, finally Use PCA for dimensionality reduction.

B. Simulation results

The confusion matrix is used to depict the actual and predicted classes in cybersecurity attacks which is represented by the following terms: True positive, True Negative, False Positive and False Negative. Using the confusion matrix, the commonly used metrics for evaluation are Accuracy, Precision and FAR.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

$$\text{FalseAlarmrate} = \frac{\text{FP}}{\text{FP+}} \quad (8)$$

We choose 5000, 10000 and 20000 as different data sizes. We compare DB-Kmeans with MinMax K-means [3] . MinMax K-means is a new clustering algorithm which starts with randomly choosing of the initial centers of clusters and attempts to apply minimizing of the maximum internal variance of clusters instead of minimizing the sum of internal variance of clusters. For different data sizes and different intrusion detection algorithms, we calculate the values in the confusion matrix.

The comparison of various indicators is as follows:

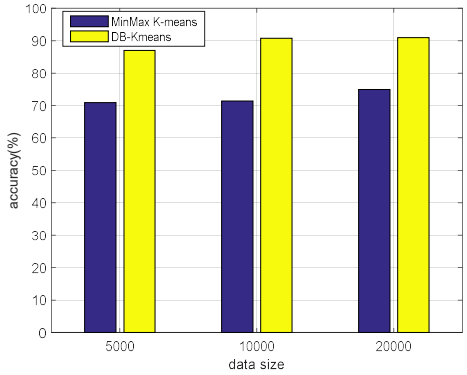

Fig. 1. Accuracy comparison between MinMax K-means and DB-Kmeans.

From Fig. 1, we can see that for different scales of NSL-KDD dataset, the detection accuracy of DB-Kmeans is higher than that of MinMax K-means. That indicates DB-Kmeans is more accurate than MinMax K-means in intrusion detection.

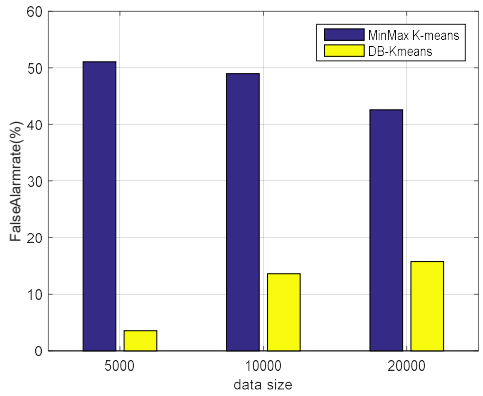

Fig. 2. FalseAlarmrate comparison between MinMax K-means and DB-Kmeans.

From Fig. 2, we can see that for different scales of NSL-KDD dataset, the false alarm rate of DB-Kmeans is lower than that of MinMax K-means. That indicates MinMax K-means is easier to make wrong judgments and cause unnecessary warnings. In comparison, DB-Kmeans' intrusion warning is more worthy of trust.

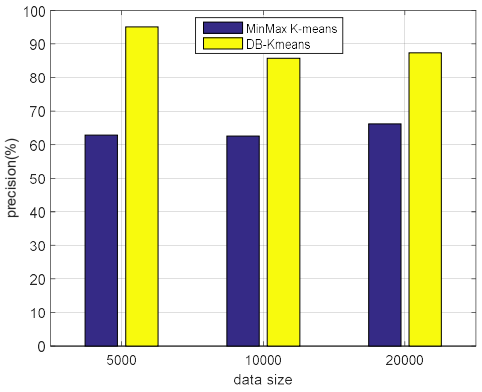

Fig. 3. Precision comparison between MinMax K-means and DB-Kmeans.

From Fig. 3, we can see that for different scales of NSL-KDD dataset, the precision rate of DB-Kmeans is higher than that of MinMax K-means. That indicates the discriminant result of DB-Kmeans is more credible than MinMax K-means.

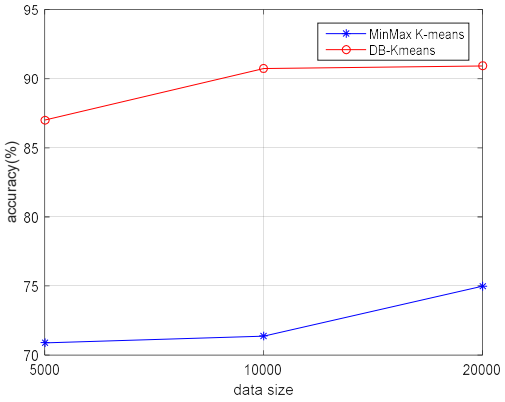

Fig. 4. Accuracy changes at different data sizes.

From Fig. 4, we can see that as the data size increases,the accuracy of various algorithms increases.

V. CONCLUSION

In this paper, we have proposed DB-Kmeans that is an improvement of DBSCAN density clustering and K-means distance clustering in terms of solving the disadvantages of the two algorithms. We set the neighborhood radius to dynamic in DBSCAN so that the problem of the global parameter can be relieved. For K-means, we define a density function and use it to select initial cluster centers. DB-Kmeans overcomes K-means' sensitivity to initial clustering centers and noise points, reduces the influence of fixed neighborhood radius in DBSCAN and improves detection efficiency. The experiments on the NSL-KDD data set indicate that the proposed method is more efficient than MinMax K-means.

ACKNOWLEDGMENT

This paper was supported by Information & Telecommunication Co. of State Grid Henan Electric Power Company "Data network security assessment strategy and status analysis research" project.

REFERENCES

[1] S. Ganapathy, K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh, A. Kannan, "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey", *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, pp. 1-16, 2013.

[2] S. Varuna and P. Natesan, "An integration of k-means clustering and naïve bayes classifier for Intrusion Detection," *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, Chennai, 2015, pp. 1-5.

[3] M. Eslamnezhad and A. Y. Varjani, "Intrusion detection based on MinMax K-means clustering," *7'th International Symposium on Telecommunications (IST'2014)*, Tehran, 2014, pp. 804-808.

[4] J. V. Anand Sukumar, I. Pranav, M. Neetish and J. Narayanan, "Network Intrusion Detection Using Improved Genetic k-means Algorithm," *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Bangalore, 2018, pp. 2441-2446.

[5] Z. Xiaofeng and H. Xiaohong, "Research on intrusion detection based on improved combination of K-means and multi-level SVM," *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, Chengdu, 2017, pp. 2042-2045.

[6] H. M. Tahir, A. M. Said, N. H. Osman, N. H. Zakaria, P. N. '. M. Sabri and N. Katuk, "Oving K-Means Clustering using discretization technique in Network Intrusion Detection System," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 248-252.

[7] V. Shakya and R. R. S. Makwana, "Feature selection based intrusion detection system using the combination of DBSCAN, K-Mean++ and SMO algorithms," *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, Tirunelveli, 2017, pp. 928-932.

[8] Y. Y. Aung and M. M. Min, "A collaborative intrusion detection based on K-means and projective adaptive resonance theory," *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Guilin, 2017, pp. 1575-1579.

[9] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceeding of the fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceeding of 2nd International Conference on Knowledge Discovery and, 1996, pp. 226–231.