Journal of Computers Vol. 30 No. 3, 2019, pp. 224-230
doi:10.3966/199115992019063003017

A DD-means Clustering Algorithm for Intrusion Behavior Analysis

Yongxin Feng, Yingyun Kang, Yuntao Zhao*, Wenbo Zhang

School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, Liaoning, China
fengyongxin@263.net, kangyingyun22@163.com, zhaoyuntao2014@163.com, zhangwenbo@sylu.edu.cn

Received 13 January 2019; Revised 8 March 2019; Accepted 12 March 2019

Abstract. With the rapid development of Internet technology, network security is becoming more and more severe. Data transferred on the Internet can be divided into two categories, normal data and intrusion data. Clustering is a popular data analysis technology to classify data according to the characteristics of the data and clustering can also be used for intrusion behavior analysis on the Internet. K-means is a simple and an efficient data clustering algorithm, but it has a tendency to converge to local optima and the optimization results depend on the initial values of cluster centers. Therefore, DD-means clustering algorithm, which selects the initial center points based on the density and the maximum distance of different data points, has been proposed to improve the performance of classical K-means. The performance of DD-means clustering algorithm is significantly better than that of classical K-means clustering algorithm in silhouette coefficient. The experimental results show that the proposed algorithm is effective and efficient for analyzing the intrusion behavior on the Internet.

Keywords: classification, DD-means, K-means clustering algorithm, network security

1 Introduction

With the development of the Internet, network security is becoming more and more severe and computer systems are prone to information theft [1]. Data transferred on the Internet can be divided into two categories, normal data and intrusion data. Clustering [2] is a popular data analysis technology which classifies the data according to their characteristics and can be used for intrusion behavior analysis. As a type of unsupervised classification technology, clustering, which partitions a set of objects in such a way that objects in the same clusters are more similar to one another than the objects in different clusters according to certain predefined criterion, is the most widely studied. The term unsupervised [3] means that grouping is established based on the intrinsic structure of the data without any need to supply the process with training items.

Clustering algorithm is usually used to divide the samples in a data set into several disjoint subsets, each of which is called "cluster". The algorithm automatically forms the structure of each cluster in the process of clustering, and each cluster may correspond to the same categories. K-means [4] is a classical and efficient method used in data clustering. Its main idea is to classify a given data set $D_n = \{x_1, x_2, ..., x_n, n \geq 1\}$ into k disjoint clusters. Firstly, the number of cluster centers k is selected randomly. The next step is to select k initial cluster centers randomly from the dataset. Then, each data point is assigned to the nearest center. Due to the possibility of local convergence of K-means clustering algorithm, the algorithm repeatedly updates the centers and assigns the data points until the centers do not change anymore.

Although, K-means is a classical clustering algorithm, and the performance of K-means clustering algorithm is very efficient. It also suffers from some major drawbacks [5]. First of all, in classical K-means clustering algorithm, the value of the initial clustering number k, which is the initial input

* Corresponding Author

parameter of the algorithm, needs to be given in advance. Different values of k lead to different clustering results. The most appropriate value of k cannot be given in advance. Meanwhile, there is no reasonable method to determine the most appropriate number of initial clustering. Secondly, classical K-means clustering algorithm may be easily affected by outliers and noisy data points. The outlier is a small number of remote data points. Due to the random selection of initial clustering centers in classical K-means clustering algorithm, selecting noisy data points and outliers as cluster centers may result in negative impact on the clustering results of classical K-means clustering algorithm. Furthermore, the clustering results of K-means clustering algorithm depend on the initial value of clustering centers.

Many heuristic clustering algorithms have been proposed to alleviate the drawbacks of classical K-means clustering algorithms. For instance, an improved algorithm called K-means++ [6] is proposed in order to tackle the initial centers sensitivity problem existed in K-means. DSKmeans [7] is another new kmeans-type algorithm proposed to discriminate subspace clustering. The Jaccard–Kmeans fast clustering method [8] is also an effective clustering algorithm proposed, which first computes the above multi-dimensional similarity, then generates multiple cluster centers with user behavior features and new content features, finally evaluates the clustering results according to cohesiveness. Besides, k-means clustering-based recommendation algorithm [9], which addresses the scalability issues associated with traditional recommender systems, is proposed.

Admittedly, K-means clustering algorithm has been developed in recent years, but there are still some limitations in the algorithm. Based on the analysis results, the DD-means clustering algorithm, is proposed to improve the performance of classical K-means, so as to provide a valuable method of the application for analyzing intrusion behavior. The rest of this paper is organized as follows. In Section 2, the definition of DD-means is given. In Section 3, DD-means clustering algorithm and the evaluation standard are proposed. A number of simulations are presented in Section 4. Finally, the summarization is given in Section 5.

2 Definition of DD-means

The DD-means clustering algorithm is proposed, in view of the improvement of classical K-means clustering algorithm. DD is short for density and maximum distance. The DD-means clustering algorithm means selecting the initial center based on the density and maximum distance of different data points.

Let $O_n = \{x_1, x_2, ..., x_n, n \geq 1\}$ be a data set with n data points, each data point has p dimensions, aiming to be split into k classes. The number of initial cluster centers is num and it is supposed that $num = 0$. Data set C is the container of initial cluster centers. x_s stands for X_{st}. X_{st} means the No. s attribution of the No. t dimension.

Definition I The distance among different data points is taken as Euclidean distance, and the Euclidean distance is defined as Eq. (1).

$$E(x_i, x_j) = (\sum_{a=1}^{p} (X_{ia} - X_{ja})^2)^{\frac{1}{2}} \quad i = 1,...,n; j = 1,...n \tag{1}$$

Definition II The total distance among all the data points is defined as Eq. (2).

$$D = \sum_{i=1}^{n} \sum_{j=1}^{n} E(x_i, x_j) \quad i = 1,...,n; j = 1,...n \tag{2}$$

Definition III The relative density for each data point is defined as Eq. (3).

$$V_s = \sum_{j=1}^{n} \frac{E(x_s, x_j)}{D} \quad s = 1,...,n \tag{3}$$

The relative density for data point x_s is distance between data point x_s and other data points divided by the total distances among all the data. The smaller V_s is, the nearer x_s is to other data points, and the more data points are around x_s. The smaller V_s is, the greater the density is. The relative density for each data point is corresponding to the density in DD-means.

Definition IV The farthest distance between x_s and other data points is defined as μ_s, which is defined as Eq.(4).

$$\mu_s = \max \sum_{j=1}^{n} (E(x_s, x_j)) \tag{4}$$

The farthest distance between x_s and other data points is corresponding to the maximum distance in DD-means.

Definition V The relative of density and maximum distance is defined as Eq.(5).

$$\theta_x = \frac{\mu_s}{V_s} \tag{5}$$

μ_s and V_s are the two factors of θ_x, which can determine whether the data point can be the center of a cluster. The larger θ_x is, the data point has more selection probability to be the center of a cluster.

3 DD-means Clustering Algorithm

3.1 The Descriptions of DD-means Clustering Algorithm

DD-means clustering algorithm is described as the following steps.

Step 1. Input a data set $O_n = \{x_1, x_2, ..., x_n, n \geq 1\}$ with n data points, aiming to split into k classes. Data set C is the initial cluster centers and $C = 0$. The number of data points in the initial cluster center is num and $num = 0$.

Step 2. Calculate the total distance among all the data points, which is defined as Eq. (2), and then calculate the relative density for each data point V_s.

Step 3. If $num \neq 0$, go to step 5.

Step 4. The data point with a larger V_s is selected as the first initial cluster center, and then go to step 7.

Step 5. Calculate the Euclidean distance among all data points and the data points selected as the initial cluster centers.

Step 6. The data point with the largest θ_x is selected as the initial cluster center.

Step 7. Remove the cluster center from the original data set and insert the cluster center into data set C, then let $num = num + 1$.

Step 8. If $num \neq k$, go to step 2, until $num = k$.

Step 9. The remaining data points are assigned to the nearest cluster, so as to obtain the final clustering results.

The flow chart of the DD-means clustering algorithm is illustrated in Fig. 1.

3.2 The Evaluation Standard of Clustering Algorithm

Clustering algorithm is one of the unsupervised machine learning algorithms. The data is not marked in unsupervised machine learning, so an evaluation standard is required to measure the clustering algorithm. Observing the degree of clustering dispersion is a good way to measure clustering algorithm.

The silhouette coefficient [10], which is defined to be used to measure the degree of the dispersion of clusters, combines two factors of cohesion and resolution. The clustering result is defined as Eq.(6). The silhouette coefficient is excellent with larger Sil average values. "$a(i)$" stands for the mean intra-cluster distance for each data point. "$b(i)$" stands for the distance between a data point and the nearest cluster.

$$Sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

The Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each data point. The best value of Silhouette Coefficient is 1 and the worst value is -1.

Negative values generally indicate that a data point has been assigned to the wrong cluster.

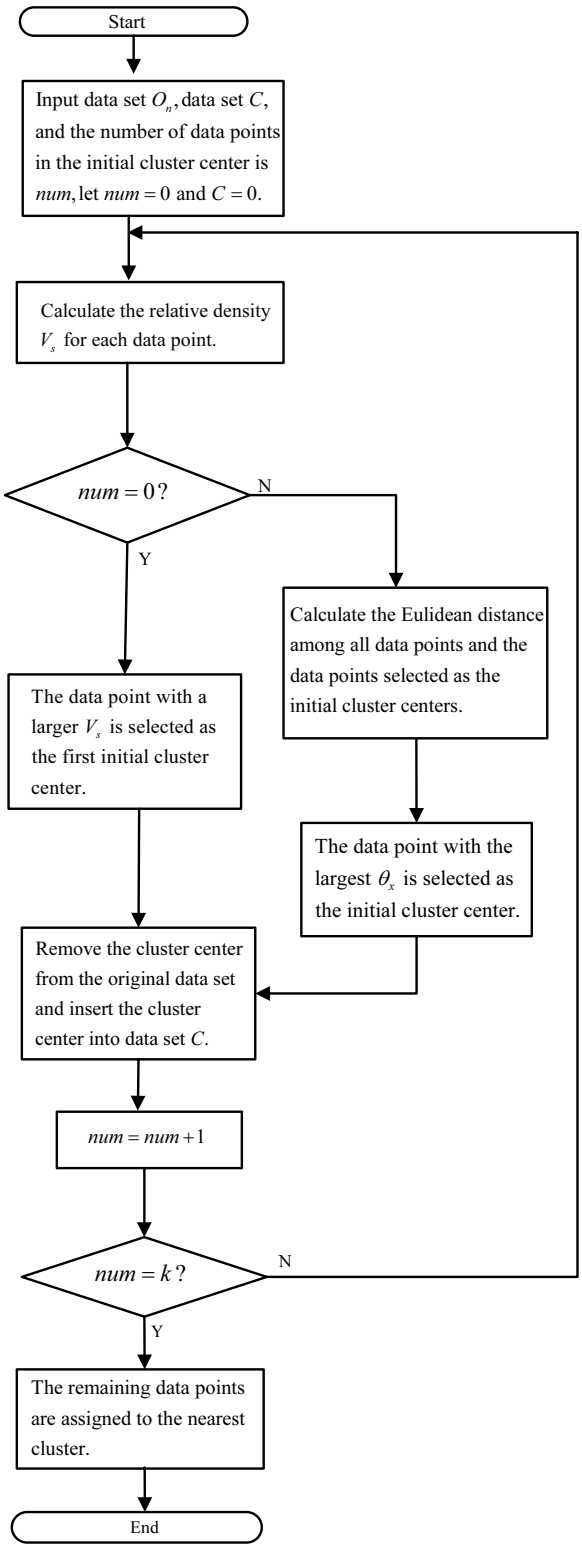

Fig. 1. The flow chart of DD-means clustering algorithm

4 Simulation and Analysis

Different experiment data are used to compare the results of traditional K-means clustering algorithm with those of DD-means clustering proposed in this paper. The value of initial cluster center in the

experiment is based on the classification number of the data set. Since the initial cluster center is selected randomly in K-means clustering algorithm, the average values of 100 times experiment results are used in this paper.

Iris and wine data set of the UCI database [11] are selected as the first part of test data set. UCI database, which is specially used to test machine learning and data mining algorithms, is a common database. The data in the UCI database are classified into a definite classification. Data in the iris data set are divided into three categories. Five groups of data, the number of which are 30, 50, 100,120, and 150, are selected randomly from the iris data set as test data. Three categories of data are contained in all the groups of test data. Data in the wine data set are divided into three categories. Five groups of data, the number of which are 40, 80, 120, 150, and 178, are selected randomly from the wine data set as test data. Three categories of data are contained in all the groups of test data.

KDDCup99 data set [12] is selected as the second part of test data set. The KDDCup99 data set, which contains intrusion and normal data, is used for the fifth event of the knowledge discovery and data mining. 41 attributes are contained in KDDCup99 data set. Continuous data type and discrete data type are two types of data attributes in KDDCup99. The three character types of data are also required to be numerically processed. Five groups of data, the number of which are 500, 1000, 1500, 2000, and 2800, are selected randomly from the KDDCup99 data set as test data. Two categories of data are contained in all the groups of test data.

Data from honeypot in an SOA system, which is established and simulated in the lab, are selected as the third part of test data set. As an active defense security model, honeypot can improve the security of an SOA system. A honeypot is established in SOA system and computers are used to attack the system in different ways. Data in the honeypot data set are divided into two categories. Five groups of data, the number of which are 1000, 2000, 3000, 4000, and 5000, are selected randomly from the honeypot in SOA system as test data. Two categories of data are contained in all the groups of test data.

4.1 Result Based on UCI Database

The Sil value of traditional K-means clustering algorithm and the result of DD-means clustering algorithm using different numbers of data from iris data set are shown in Fig. 2. The Sil value of traditional K-means clustering algorithm and the result of DD-means clustering algorithm using different numbers of data from wine data set is shown in Fig. 3.

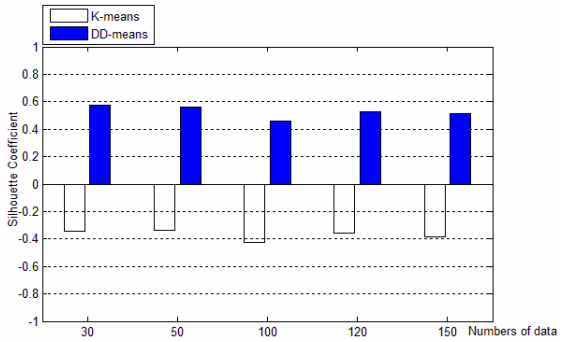

Fig. 2. Silhouette coefficient value in different K of Iris data set

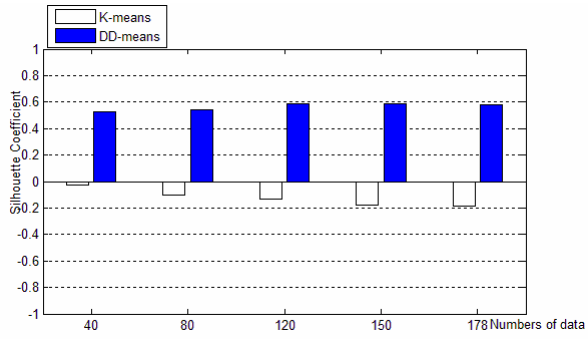

Fig. 3. Silhouette coefficient value in different K of wine data set

According to the results shown in Fig. 2 and Fig. 3, data classified can be implemented by DD-means clustering algorithm. The Sil value of DD-means clustering algorithm is slightly better than that of the K-means clustering algorithm regardless of the number of data set. DD-means clustering algorithm can implement the classification of data. The ability of clustering behaviors is improved by the DD-means clustering algorithm.

4.2 Result Based on KDDcup99 Data Set

The Sil value of traditional K-means clustering algorithm and the result of DD-means clustering algorithm using different numbers of data from KDDCup99 data set are shown in Fig. 4. Fig. 4 shows

that the ability of clustering behaviors is improved by the DD-means clustering algorithm and the DD-means clustering algorithm can be used to cluster data sets, which contain intrusion and normal data.

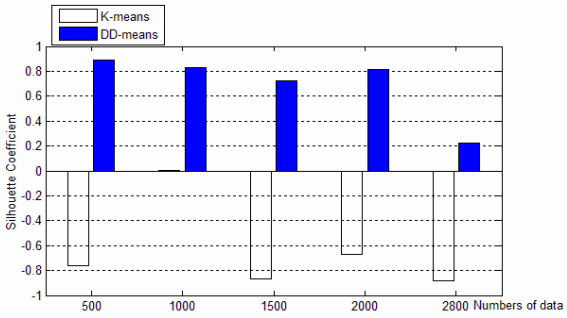

Fig. 4. Silhouette coefficient value in different K of KDDCup99 data set

4.3 Result Based on Data from Honeypot in SOA System

The Sil value of traditional K-means clustering algorithm and the result of DD-means clustering algorithm using different numbers of data from honeypot in SOA system are shown in Fig. 5. It can be seen from Fig. 5 that DD-means clustering algorithm can be applicated extensively. DD-means clustering algorithm can not only be used for normal dataset but also be used as a data analysis method for honeypot in SOA system, which indicates DD-means clustering algorithm can be used for intrusion behavior analysis in the Internet. Although the data collected from the real network situation is complex, the DD-means clustering algorithm has a better clustering result than classical K-means clustering algorithm in Silhouette Coefficient.

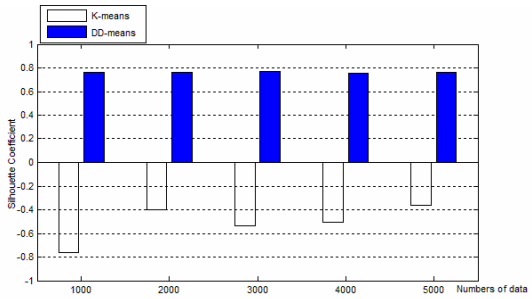

Fig. 5. Silhouette coefficient value in different K of data from honeypot in SOA system

4.4 Result Based on Different Full Test Data Set

The Sil value of traditional K-means clustering algorithm and the result of DD-means clustering algorithm using different full test data set are shown in Fig. 6. Since the initial cluster center is randomly selected in K-means clustering algorithm, the Sil value of K-means is of great randomness. Though the average value of 100 times experiment results is used in this paper, the clustering effect of DD-means clustering algorithm is much better than that of K-means as shown in Fig. 6.

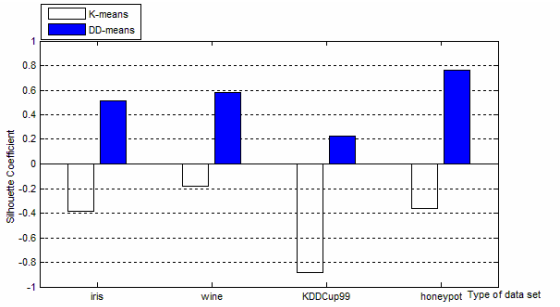

Fig. 6. Silhouette coefficient value in different K of different datasets

5 Conclusions

In this work, DD-means clustering algorithm is proposed to improve the performance of classical K-means clustering algorithm. Clustering is an important classification technology that gathers data into classes (or clusters) such that the data in each cluster shares a high degree of similarity while being very dissimilar from data of other clusters. The performance of DD-means clustering algorithm is compared with classical K-means clustering algorithm. The experimental results show that the DD-means clustering algorithm can be applicated to clustering different data set successfully, especially for data collected from honeypot in SOA system, which indicates that DD-means clustering algorithm is a valuable algorithm for analyzing intrusion behavior.

Acknowledgments

This work was supported by China Postdoctoral Science Foundation (2016M590234), General Project of Liaoning Provincial Department of Education (LG201611), Postdoctoral fund of Shenyang Ligong University, Project of Applied Basic Research of Shenyang (18-013-0-32), Natural Science Foundation of Liaoning Province (20180551066), Program for Liaoning Distinguished Professor, Program for Liaoning Innovative Research Team in University. The author declares that there is no conflict of interest regarding the publication of this article.

References

[1] M. Yu, L. Chao, Q Xinliang, Z. Shuang, Modeling and analysis of information theft trojan based on stochastic game net, in: Proc. 2015 International Conference on Information Science & Control Engineering, 2015.

[2] Q. Zhang, C. Zhu, L. Yang, Z. Chen, An incremental CFS algorithm for clustering large data in industrial Internet of Thing, IEEE Transactions on Industrial Informatics 13(3)(2017) 1193-1201

[3] Y. Li, M. Paluri, J. Rehg, P. Dollár, Unsupervised learning of edges, in: Proc. 2016 Computer Vision and Pattern Recognition, 2016.

[4] S. Ahmadian, A. Norouzi-Fard, O. Svensson, J. Ward, in: Proc. 2017 Foundations of Computer Science, 2017.

[5] O. Charles, D. Wang, A. Jain, Clustering millions of faces by identity, IEEE Transactions on Pattern Analysis & Machine Intelligence 40(2)(2016) 289-303.

[6] D. Arthur, S. Vassilvitskii, k-Means ++: the advantages of carefull seeding, Proceedings of the Eighteenth Annual ACM-SIAM Symposiumon Discrete algorithms, Society for Industrial and Applied Mathematics 11(6)(2007) 1027-1035.

[7] X. Huang, Y. Ye, H. Guo, Y. Cai, H. Zhang, Y. Li, A new kmeans-type approach to discriminative subspace clustering, Knowledge-Based Systems 70(2014) 293-300.

[8] M. Lu, Z. Qin, Y. Cao, Z. Liu, M. Wang, Scalable news recommendation using multi-dimensional similarity and Jaccard-Kmeans clustering, Journal of Systems & Software 95(9)(2014) 242-251.

[9] S. Zahra, M.A. Ghazanfar, A. Khalid, M.A. Azam, U. Naeem, A. Prugel-Bennett, Novel centroid selection approaches for KMeans-clustering based recommender systems, Information Sciences An International Journal 320(C)(2015) 156-189.

[10] M. Yesilbudak, Clustering analysis of multidimensional wind speed data using k-means approach, in: Proc. 2017 IEEE International Conference on Renewable Energy Research and Applications, 2017.

[11] V. Chouvatut, W. Jindaluang, E. Boonchieng, T. Rukkanchanunt, Efficiency comparisons between k-centers and k-means algorithms, in: Proc. 2016 Computer Science & Engineering Conference, 2016.

[12] O. Kaynar, A.G. Yüksek, Y. Görmez, Y.E. IŞIK, Intrusion detection with autoencoder based deep learning machine, in: Proc. 2017 Signal Processing & Communications Applications Conference, 2017.