Hindawi
Security and Communication Networks
Volume 2022, Article ID 9951609, 8 pages
https://doi.org/10.1155/2022/9951609

Research Article

DBSCAN Clustering Algorithm Based on Big Data Is Applied in Network Information Security Detection

Yan Zhang

Department of Information Engineering, ShiJiaZhuang University of Applied Technology, Shijiazhuang 050081, Hebei, China

Correspondence should be addressed to Yan Zhang; 320045119102@stu.suse.edu.cn

Received 30 April 2022; Revised 18 June 2022; Accepted 25 June 2022; Published 12 July 2022

Academic Editor: Mukesh Soni

In order to improve the certainty and clarity of information security detection, an application method of big data clustering algorithm in information security detection is proposed. The experimental results show that when the amount of data is close to 6000, the efficiency of the improved algorithm is nearly 70% higher than that of DBSCAN, and it is still very close to the efficiency of the BIRCH algorithm. The algorithm has a high processing speed for large-scale data sets without increasing the time complexity and can also accurately cluster clusters of any shape. When the data set increases from 9000 rows to 58000 rows, in turn, the time-consuming of the traditional DBSCAN algorithm increases sharply, while the time-consuming of the improved DBSCAN algorithm is still stable, and the time-consuming gap between the two is getting bigger and bigger. At the same time, the algorithm adopts a heuristic adaptive algorithm to estimate some threshold parameters of the clustering algorithm, which can avoid the direct setting of the threshold parameters by the user and can effectively estimate the relevant threshold parameters, extract clusters of any shape, and the clustering effect is obvious.

1. Introduction

The focus is on how to use data mining technology to develop and disseminate Internet technology and to improve information security detection and analysis during the development and implementation of the energy Internet. As a data mining method, a clustering algorithm can judge the similarity of samples and divide the samples with strong similarity into one class. In network information security, under normal network conditions, the user's operation is almost the same, and the network information security data are also similar [1–3]. In information security detection, the application of a clustering algorithm can classify similar network information into one class, and the information with large differences can be screened out. Once the information difference is large, which may be caused by the network attack, the system will automatically send an early warning to prompt the user. Therefore, the clustering algorithm is widely used in information security monitoring technology. Figure 1 shows the information security detection center. At present, the research on computer network information security detection and protection strategy is still in the initial stage, especially the correlation between different target attributes. The proportion of nonlinear relationship accounts for more than half. If conventional methods are adopted, it is difficult to fully reflect the actual relationship [4]. Internet technology is constantly updated and gradually applied in various fields, which has had a huge impact on people's life and work. A large amount of data information has exploded, and a large amount of private information is involved. As an extremely important information industry in the development of modern society, aerospace transportation is closely related to the security of classified information and aerospace and is also closely related to the stable development of social security.

DBSCAN (density-based spatial clustering of applications with noise) is a representative density-based clustering algorithm. Unlike partitioning and hierarchical clustering methods, it defines a cluster as the largest set of densely connected points, can divide regions with high enough density into clusters, and can find clusters of arbitrary shapes in noisy spatial databases. In the analysis

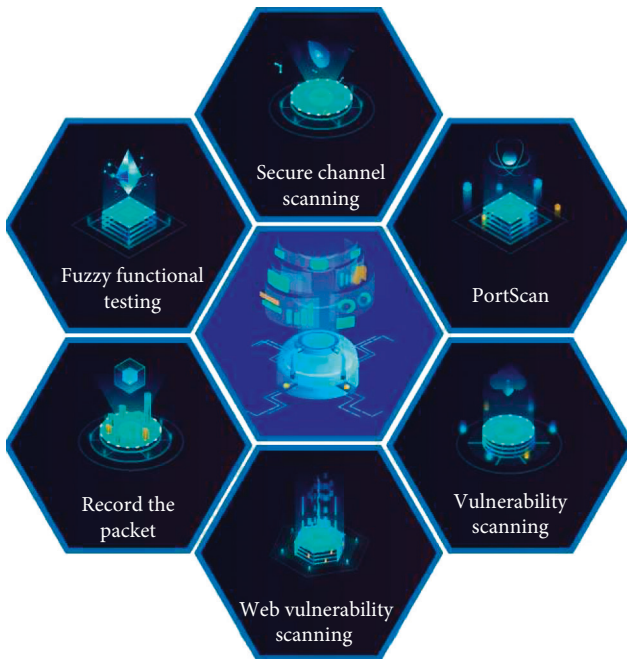

Figure 1: Information security detection center.

process, we found that the birch algorithm and DBSCAN algorithm can mine information and effectively realize information security detection. Birch algorithm has high efficiency. CF tree and CF vector can effectively describe the clustering-related information, but the clustering effect for nonspherical clusters is poor. Using the birch algorithm for clustering can effectively extract clusters of arbitrary shape and correctly identify noise points and outliers, but the space-time complexity is higher than using DBSCAN. Both methods require users to provide several threshold parameters, which increases the difficulty of clustering algorithm in practical application [5, 6]. Therefore, combined with the birch algorithm and DBSCAN algorithm, an improved clustering algorithm is proposed, which enables the algorithm to use a heuristic adaptive algorithm to estimate some threshold parameters of the clustering algorithm, avoids the setting of threshold parameters directly by users, and reduces the difficulty of clustering algorithm in practical application. The so-called cluster analysis is to divide things into clusters according to their own attributes, so that the similarity of things in different clusters is as small as possible, and the similarity of things in the same cluster is as large as possible. At present, the research on computer network information security protection strategy is still in the early stage, especially the correlation between different target attributes, and the proportion of nonlinear relationship accounts for more than half. If conventional methods are used, it is difficult to fully reflect the actual relationship. In the analysis process, there will also be contradictions, and it is likely that there will be unorganized situations. The article focuses on computer network security confidential information and conducts a comprehensive and accurate analysis of network information security.

2. Literature Review

The current improvement of computer computing power and price decline, as well as the continuous development of computer cluster technology, makes the cost of building, using, and maintaining computer clusters smaller and smaller. The computing power of a single computer is limited and cannot handle large-scale data sets, so we can use the high computing power of large-scale clusters for processing, so a large number of large-scale data clustering algorithms based on parallel ideas have been developed the resources to implement large-scale data clustering tasks. Guo and others suggest that cluster analysis is an important method of data mining technology and that the algorithm for clustering large data sets with rapidly growing data volumes is an important topic in today's data mining [7]. Bi and others proposed a birch algorithm, which is a clustering algorithm for large-scale data sets. It first stores the data set in a compact compression format, adopts the balanced tree structure, comprehensively considers the problems of system memory, time overhead, and clustering quality, has a high processing speed for large-scale data sets, and meets the scalability of data, so it is applied to many different fields [8]. Lieharyani and others believe that the BIRCH algorithm is thought to combine hierarchical and repetitive displacement methods, first using a bottom-up stepwise algorithm and then using repetitive displacement to improve results [9]. Jones and others believe that the birch algorithm integrates hierarchical aggregation and iterative relocation methods. First, the bottom-up hierarchical algorithm is used, and then, the iterative relocation is used to improve the results [10]. Its main idea is to scan the database. Birch algorithm has high efficiency. CF tree and CF vector can effectively describe the clustering-related information, but the clustering effect for nonspherical clusters is poor. The choice of parameters directly affects the effect of the cluster. Based on the advantages and disadvantages of the two cluster algorithms, an improved clustering algorithm is proposed that combines the BIRCH algorithm and the DBSCAN algorithm. First, the improved birch algorithm is implemented: the data set is sampled to obtain the estimated distance between clusters, which is used as the initial threshold T for establishing a CF tree. During tree replication, the new threshold is set by averaging the distance between adjacent entries of all leaf nodes. After generating the CF tree, the leaf node subclusters are studied to obtain the data set density parameters to calculate the radius parameter ε of the DBSCAN algorithm and the neighboring density threshold MinPts. According to the parameter estimation of the birch algorithm, DBSCAN clustering is carried out on the whole data set to obtain the final clustering result. The algorithm adopts a heuristic adaptive algorithm to estimate some threshold parameters of the clustering algorithm, which avoids the setting of threshold parameters directly by users, reduces the influence of parameters on the clustering effect, and reduces the application difficulty of the clustering algorithm in information detection [11, 12].

On the basis of this research, this paper proposes a method of applying a big data clustering algorithm in

information security detection. Experiments show that the algorithm can process large-scale data sets at a high speed without increasing the time complexity and can accurately cluster clusters of any shape, find noise points, and effectively estimate the correlation threshold. Parameters extract clusters of arbitrary shapes, and the clustering effect is obvious.

3. Research Methods

3.1. Birch Algorithm

3.1.1. Core Idea. Clustering is an important data mining task, and its purpose is to divide data into several data subsets according to certain criteria, in which the data within each subset are more similar and the data between the subsets are more different. The clustering task is an unsupervised learning method that is widely used in information retrieval, image segmentation, bioinformatics, and other fields. With the rapid development of storage technology, the cost of storing data is getting smaller and smaller, and the scale of available data accumulated in all walks of life is also increasing. Traditional clustering algorithms have achieved excellent results on small-scale data sets when faced with today's large-scale data, and these classical clustering algorithms are difficult to perform, or even unable to complete the task of clustering analysis. Birch algorithm is a kind of aggregation algorithm, and it is suitable for processing large data sets, whose time and spatial complexity are $O(n)$, where n is the number of clustered objects. Birch algorithm can establish a CF tree by scanning the database in a single pass, which can effectively identify noise points. However, it has a poor clustering effect for nonspherical clusters, which leads to low efficiency.

3.1.2. CF Vector. A cluster property (CF) is a three-dimensional vector that aggregates object cluster information. Defined as CF = , N is the number of points in the cluster, LS is the linear sum of N points, and SS is the square sum of N points. A cluster feature is basically a statistical summary of a cluster. Using the clustering feature, it is easy to obtain many useful statistics of the cluster.

$$\text{Centroid of cluster: } X = \frac{\text{LS}}{N}, \tag{1}$$

$$\text{Radius of cluster: } R = \sqrt{\frac{N \times \text{SS} - \text{LS} \times \text{LS} + N \times \text{LS}}{N \times N}}, \tag{2}$$

$$\text{Cluster diameter: } D = \sqrt{\frac{2 \times N \times \text{SS} - 2 \times \text{LS} \times \text{LS}}{N \times (N-1)}}, \tag{3}$$

$$\text{Distance between clusters: } D_2 = \sqrt{\frac{\text{SS}_1}{N_1} + \frac{\text{SS}_2}{N_2} - \frac{2\text{LS}_1 \times \text{LS}_2}{N_1 \times N_2}}. \tag{4}$$

Clustering features are additive. Assuming that $\text{CF}_1 = (N_1, \text{LS}_1, \text{SS}_1)$ and $\text{CF}_2 = (N_2, \text{LS}_2, \text{SS}_2)$ are characteristics of two class clusters, the new class cluster feature after merging will be $\text{CF}_1 + \text{CF}_2 = (N_1 + N_2)$, $\text{LS}_1 + \text{LS}_2$, $\text{SS}_1 + \text{SS}_2$). Using cluster functions to aggregate clusters avoids storing detailed information on a single object or point and requires only a certain amount of space to store cluster properties, which is the key to the efficiency of the BIRCH algorithm [13].

3.1.3. CF Tree. The CF tree is a highly balanced tree that retains the characteristics of a hierarchical cluster, as shown in Figure 2. Nonleaf nodes store the sum of their children's CFs and thus aggregate cluster information about their children. CF has two parameters: branching factor B (maximum number of children that can be a single node) and threshold T (maximum diameter of a leaf node subcluster), both of which affect the size of the clusters formed [14].

3.2. DBSCAN Algorithm

3.2.1. Core Idea. The DBSCAN algorithm is a density-based clustering method. The algorithm divides the region with certain density into clusters and it regards clusters as dense regions separated by sparse regions in the data space. This algorithm can effectively extract arbitrary shapes of clusters from noisy spatial data sets and correctly identify noise outliers. Its time complexity is $O(n_2)$, where n is the number of clustered objects.

3.2.2. Basic Concepts

Definition 1 (ε-neighbor). The ε-neighbor of an object is a space whose center is o and whose radius is a user-defined space ε.

Definition 2 (neighbor density). It is the number of nearby objects.

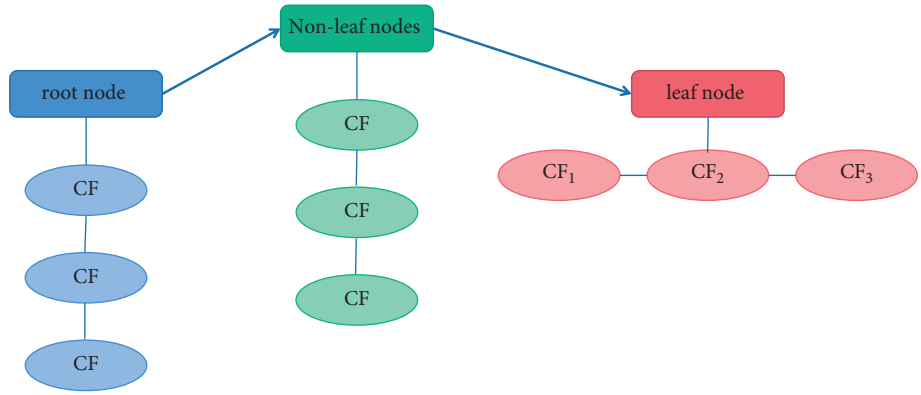

Figure 2: CF tree structure.

Definition 3 (basic object). If the ε-neighbor of object o contains at least MinPts (user-defined threshold parameters) objects, o is called the base object.

Definition 4 (possibility of direct density). p is a density that can only be achieved directly from q (for ε and MinPts) if q is the main object

Definition 5 (achievable density). It can be reached directly from q (for ε and MinPts) only if $p_1 = q$, $p_n = p$, and if there is a chain of objects p_1, p_2, p_3, ..., p_n, for any p_i, $p_i + 1$ is a density that can be reached directly from pi (for ε and MinPts).

Definition 6 (densely connected). Objects p_1, p_2 (for ε and MinPts) are densely connected (for ε and MinPts) only if there are q, p_1, p_2 objects that can reach the density [15–17].

3.3. Improved DBSCAN Algorithm

3.3.1. Algorithm Description. Birch algorithm has a poor clustering effect for nonspherical clusters, and because each node can only contain a certain number of subnodes, the final cluster may be very different from the natural cluster, which leads to low efficiency. The algorithm comprehensively considers various factors such as time/space efficiency, the sensitivity of data input, and the accuracy of final clustering results in the clustering process and is especially suitable for the processing of large data sets [18]. In addition, the birch algorithm does not give a specific setting method of the threshold T in the first stage but simply assigns $t = 0$. In the second stage, it does not give a specific method to raise the threshold T, so it can only rely on the user to specify the parameter value t. In this paper, an estimation method of threshold T is proposed. The threshold T is adaptively modified by the iterative method, so as to obtain the CF tree. DBSCAN algorithm can effectively extract clusters of arbitrary shape and correctly identify noise points and outliers, so the quality of extracted subclusters is relatively high. The selection of parameters directly affects the clustering effect. Considering the high efficiency of the birch algorithm, and through its data structure, clustering feature CF vector and CF tree, many key statistics of clusters can be easily deduced [19]. This paper proposes an estimation method of DBSCAN parameters. First, the CF tree is obtained by the birch algorithm. By analyzing the clustering feature CF vector on the tree, the density estimation value of the data set is obtained.

In conclusion, the improved BIRCH + DBSCAN method can be divided into two phases. Step 1: The CF vector and the CF tree are obtained using the enhanced BIRCH algorithm, so as to obtain the density information of the data set. The second stage used the density estimation value of the data set obtained in the first stage as the parameter of the DBSCAN algorithm clusters the density and obtains the clustering results.

The detailed algorithm of the first stage is as follows:

Step 1: input the expected number of clusters K and branching factor B to fix CF tree height.

Step 2: the data are sampled to obtain n samples. M takes the average value of the distance between N data pairs as the estimated value of the distance between clusters and initially takes $t = m/2$ as the diameter.

Step 3: run birch.

Step 4: if the tree is built successfully, the clustering ends. If the tree building fails, use the CF vector and CF tree to set the new threshold T to the average value of the distance between adjacent entries of the CF leaf node and turn to step 3 to rebuild the tree.

The detailed algorithm of the second stage is as follows:

Step 1: obtain the data set density information: use the CF vector and CF tree obtained in the first stage to obtain the subcluster characteristics of birch clustering.

Step 2: determine the analysis data set: for the obtained subcluster set, analyze its compactness, and select the subcluster with compactness above the average value as the subcluster set to be analyzed.

Step 3: parameter modeling: for the subcluster set to be analyzed, construct its minimum spanning tree, respectively.

Step 4: DBSCAn clustering: take MinPts = 1 and ε as the average value of the largest edge of all minimum

spanning trees, and then conduct DBSCAN clustering again. The flow chart of the clustering flow algorithm is shown in Figure 3.

3.3.2. Algorithm Analysis. The birch algorithm is a kind of aggregation algorithm, with low space-time complexity, but the clustering effect for nonspherical clusters is poor. The DBSCAN algorithm can efficiently extract clusters of arbitrary shape and correctly identify noise points and outliers, but with higher spatiotemporal complexity than BIRCH. In addition, both the birch algorithm and the DBSCAN algorithm require users to provide several threshold parameters, and the selection of parameters directly affects the clustering effect [20–22]. The improved birch + DBSCAN algorithm combines the advantages of the two and uses a heuristic adaptive algorithm to estimate some threshold parameters of the clustering algorithm. The CF vector and CF tree output by the birch algorithm are used as the basis of data set density estimation to obtain the threshold parameters of the DBSCAN algorithm, which avoids the direct setting of threshold parameters by users. The improved birch + DBSCAN algorithm is divided into two stages. Considering that the birch algorithm has two important data structures: these two structures can effectively represent the hierarchical structure of clustering and summarize the information of clusters, so as to correctly estimate the density information of data sets. Birch algorithm requires users to provide the threshold parameter t of subcluster diameter estimation, which has a great impact on the clustering effect [23]. This paper presents the initial setting method of T and the iterative lifting method. By sampling the data, several samples are obtained, and the average value of the distance between two samples is taken as the initial estimate of the distance between clusters. In the process of establishing a CF tree, the top-down search/addition method is adopted to add points to the corresponding subclusters one by one. A point is always inserted into the nearest cluster. If the diameter of the leaf node subcluster exceeds T, the tree is divided and balanced. If the CF tree exceeds the specified size, it is necessary to reestablish a tree, and the new threshold T is set to the average value of the distance between adjacent entries of the CF leaf node. Through such estimation, the threshold parameter t can be effectively improved, so as to continuously expand the size of the CF tree until the tree is successful. In the second stage, the CF vector and CF tree in the first stage are used to obtain the preliminary clustering characteristics and the basic information of the cluster, so as to obtain the density estimation value of the data set, which is used as the parameter of DBSCAN algorithm for density clustering and obtained the clustering results. For a small number of data sets, all subclusters can be regarded as the set to be analyzed; for a large number of data sets, the sampling method can be used to obtain subclusters and extract samples. The CF vector effectively summarizes the characteristic information of the cluster. According to equations (2) and (3), the average distance R to the centroid of the member objects and the average distance D of the pairs in the cluster can be obtained. Both R and D represent the density of the clusters around the center. Subclusters with compactness not less than the average of R and D can be selected as the set to be analyzed. The minimum spanning tree of the subcluster set to be analyzed is constructed, respectively. The maximum edge of the minimum spanning tree can effectively estimate the average value of the nearest neighbor distance of any two nodes in the subcluster. DBSCAN algorithm defines the concept of object neighborhood [24, 25]. If the number of objects in the ε-neighborhood is not less than the given threshold MinPts, that is, the density of its neighborhood is not less than the threshold MinPts, and then, this object is the core object. By aggregating small dense areas centered on the core object, we can get large dense areas, that is, clusters. Therefore, ε is selected as the average value of the maximum edge of all minimum spanning trees, and MinPts is set to 1 for DBSCAN clustering.

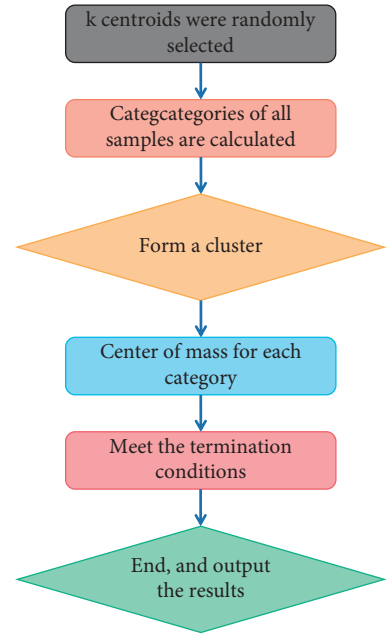

Figure 3: Flow chart of clustering flow algorithm.

4. Analysis of Experimental Results

We use the algorithm in this paper, the DBSCAN algorithm and birch to cluster the data set DB1 (200 data points), the generated data set test 1 (3000 two-dimensional data points, divided into 3 classes, with 1000 points in each class), pageblocks classification, and mushroom, respectively (5473 10-dimensional data sets), and analyze their running time (8124 22-dimensional nonnumerical attribute data sets. In the clustering process, nonnumerical values should be transformed into numerical values before clustering). The comparison results are shown in Figure 4.

As shown in Figure 4, the algorithm described in this article is slightly less efficient than the BIRCH algorithm but more efficient than the DBSCA algorithm. When the data size is close to 6000, the efficiency of the improved algorithm is almost 70% higher than that of DBSCAN and remains

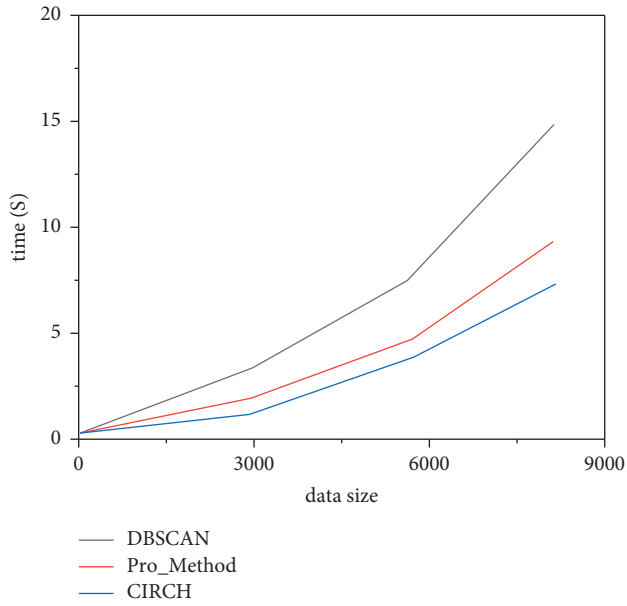

Figure 4: Comparison of clustering time.

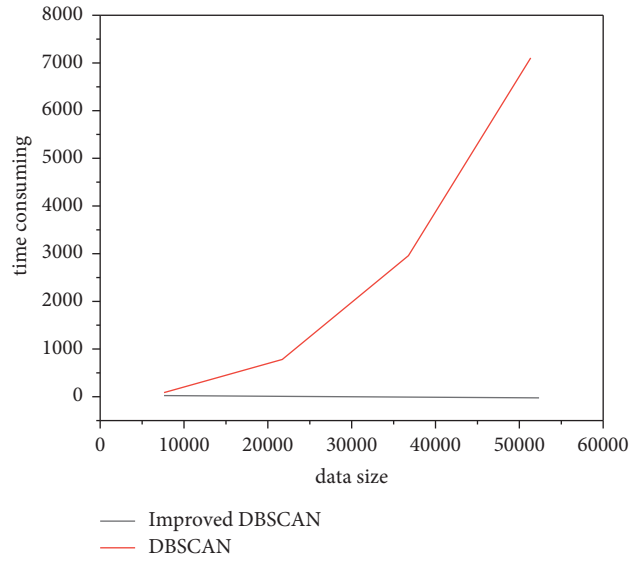

Figure 5: Comparison of algorithm time.

close to the efficiency of the BIRCH algorithm. The time complexity of the algorithm and the BIRCH algorithm in this article is the same, so the time trend is the same. Because the birch algorithm does not consider clustering analysis of subclusters with other clustering methods in the second step, the efficiency of the two algorithms is not much different. At the same time, the algorithm in this paper is compared with the DBSCAN algorithm. The tested data sets are from three data sets db1, DB2, and db3 of the DBSCAN algorithm. In the clustering results of this paper, some nonnoise points are classified as noise points. The main reason is that these points are far away from other data points and relatively isolated. From another point of view, it also shows the sensitivity of this algorithm to noise points; that is, it can accurately eliminate noise points. Then, set the parameters EPS = 100 m and MinPts = 60 and increase the data set from 9000 rows to 58000 rows in turn. The operation results are shown in Figures 5 and 6.

4.1. Clustering Time. As the amount of data increases, the time of the traditional DBSCAN algorithm increases dramatically, while the time of the improved DBSCAN algorithm remains constant, and as the amount of data increases, the time-consuming difference can be seen in Figure 5. These two are getting bigger.

4.2. Number of Clusters. When the two algorithms have the same amount of data, the number of clusters created by clustering is basically the same. When the amount of data is small and large, the number of clusters created by the improved DBSCAN algorithm is slightly less than the traditional DBSCAN algorithm, which may be due to the stricter division of data by the improved DBSCAN algorithm.

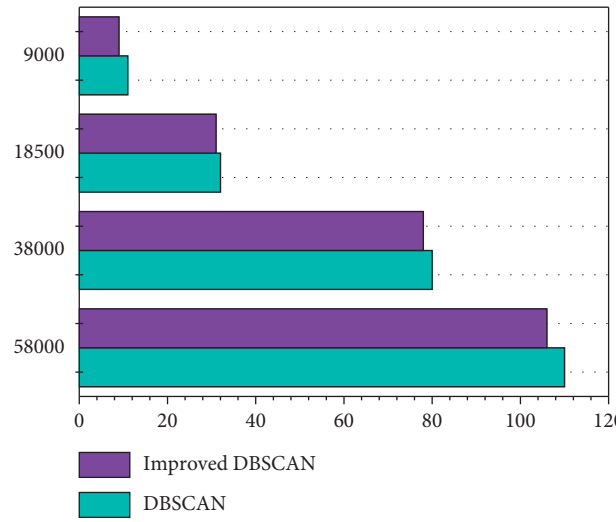

Figure 6: Number of clusters.

From Figures 5 and 6, we get the following conclusions:

(1) The improved DBSCAN algorithm significantly improves cluster efficiency on the basis that the quantity and quality of clusters are essentially consistent with the traditional DBSCAN algorithm. When the data size is close to 6000, the efficiency of the improved algorithm is almost 70% higher than that of DBSCAN and remains close to the efficiency of the BIRCH algorithm [26]. As the amount of data increases, the time of the traditional DBSCAN algorithm increases dramatically, while the time of the improved DBSCAN algorithm remains constant, and the time between the two increases.

(2) The improved DBSCAN algorithm is more sensitive to noise points and can accurately eliminate noise points. The two-dimensional data set after clustering

can effectively extract clusters of arbitrary shapes and correctly identify noise points and outliers.

5. Conclusion

Aiming at the problem of improving information security early warning analysis, this paper proposes an improved clustering algorithm by using data mining technology. The algorithm comprehensively considers the advantages and disadvantages of the BIRCH algorithm and the DBSCAN algorithm and proposes a "two-stage clustering model." The improved BIRCH + DBSCAN algorithm combines the advantages of both and uses a heuristic adaptive algorithm to estimate some threshold parameters of the clustering algorithm. The density estimation information of the data set is obtained through the first stage, and the CF vector and CF tree of the first stage are used in the second stage to obtain the preliminary clustering features and basic information of the cluster, so as to obtain the density estimation value of the data set, which is used as DBSCAN. The parameters of the algorithm perform density clustering and obtain the clustering results. It is verified that the improved DBSCAN algorithm can effectively estimate the threshold parameters in the two clustering algorithms. Experiments show that the algorithm can process large-scale data sets at a high speed without increasing the time complexity and can accurately cluster clusters of any shape, find noise points, and effectively estimate the correlation threshold. Parameters extract clusters of arbitrary shapes, and the clustering effect is obvious.

In addition to clustering algorithms, large-scale data processing methods based on bipartite graphs have also been widely used in fields such as hashing and manifold learning. Although this method has achieved very good results in practical applications, there is currently a lack of theoretical research on this method in the literature. Moreover, the selection of representative samples in this method is very important. When using this method in many kinds of literature, the selection of representative samples is obtained by random sampling or simply using the K-means algorithm to obtain some cluster centers as representative samples. The rationale for this choice of a representative sample is not given.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] I. O. Pappas, P. Mikalef, M. N. Giannakos, J. Krogstie, and G. Lekakos, "Correction to: big data and business analytics ecosystems: paving the way towards digital transformation and sustainable societies," *Information Systems and E-Business Management*, vol. 19, no. 4, p. 1355, 2021.

[2] G. P. Diller and H. Baumgartner, "Adäquate Versorgungsstrukturen und Bedeutung von Big-Data-Analysen bei EMAH-Patienten," *Aktuelle Kardiologie*, vol. 10, no. 05, pp. 403–407, 2021.

[3] S. V. Pronichkin and I. B. Mamai, "Research of the efficiency of scientific and technical results in the field of chemical safety based on big data analysis," *Journal of Physics: Conference Series*, vol. 1942, no. 1, pp. 012033–012038, 2021.

[4] F. Tang and G. Liu, "Application of folk art elements in packaging design based on big data analysis under the background of traditional culture," *Journal of Physics: Conference Series*, vol. 1744, no. 3, pp. 032065–032071, 2021.

[5] P. Li, Z. Guan, C. Han, S. Hu, and Y. Xiao, "Narrow density window and risk assessment based on big data," *Journal of Physics: Conference Series*, vol. 1757, no. 1, pp. 012105–012110, 2021.

[6] L. Chen, C. Lan, B. Xu, and K. Bi, "Progress on material characterization methods under big data environment," *Advanced Composites and Hybrid Materials*, vol. 4, no. 2, pp. 1–13, 2021.

[7] J. Guo, M. Zhang, Q. Shang, F. Liu, and X. Li, "River basin cyberinfrastructure in the big data era: an integrated observational data control system in the heihe river basin," *Sensors*, vol. 21, no. 16, pp. 5429–5434, 2021.

[8] K. Bi, D. Lin, Y. Liao, C. H. Wu, and P. Parandoush, "Additive manufacturing embraces big data," *Progress in Additive Manufacturing*, vol. 6, no. 6, pp. 1–17, 2021.

[9] D. Lieharyani, R. Ginardi, R. Ambarwati, and M. T. Multazam, "Assessment for good university governance in higher education focus on align strategy business with it at big data era," *Journal of Physics: Conference Series*, vol. 1175, pp. 7–12, 2020.

[10] C. Jones, DeR. Reinkensmeyer, and J. Morris, "Big data analytics and sensor-enhanced activity management to improve effectiveness and efficiency of outpatient medical rehabilitation," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, pp. 748–753, 2020.

[11] C. Zhou, A. Li, A. Hou, Z. Zhang, and F. Wang, "Modeling methodology for early warning of chronic heart failure based on real medical big data," *Expert Systems with Applications*, vol. 151, no. 5, pp. 113361–113365, 2020.

[12] J. Ruan, H. Jiang, J. Yuan et al., "Fuzzy correlation measurement algorithms for big data and application to exchange rates and stock prices," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1296–1309, 2020.

[13] M. R. Mathis, T. Z. Dubovoy, M. D. Caldwell, and M. C. Engoren, "Making sense of big data to improve perioperative care: learning health systems and the multicenter perioperative outcomes group," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 34, no. 3, pp. 582–585, 2020.

[14] J. Frcpc and F. M. Mcmullen, "Big data for a big problem: how can we enhance the implementation of perioperative cardiovascular guidelines sciencedirect," *Canadian Journal of Cardiology*, vol. 37, no. 1, pp. 11–13, 2020.

[15] M. Alipour and D. K. Harris, "A big data analytics strategy for scalable urban infrastructure condition assessment using semi-supervised multi-transform self-training," *Journal of Civil Structural Health Monitoring*, vol. 10, no. 2, pp. 313–332, 2020.

[16] B. Jka and B. Lga, "Big data and artificial intelligence: will they change our practice?" *Joint Bone Spine*, vol. 87, no. 2, pp. 107–109, 2020.

[17] R. C. Fitzgerald, "Big data is crucial to the early detection of cancer," *Nature Medicine*, vol. 26, no. 1, pp. 19-20, 2020.

[18] R. Ariyanto, R. H. Tjahjana, and T. Udjiani, "Forecasting retail sales based on cheng fuzzy time series and particle swarm optimization clustering algorithm," *Journal of Physics: Conference Series*, vol. 1918, no. 4, pp. 042032–042036, 2021.

[19] N. Xu, R. B. Finkelman, S. Dai, C. Xu, and M. Peng, "Average linkage hierarchical clustering algorithm for determining the relationships between elements in coal," *ACS Omega*, vol. 6, no. 9, pp. 6206–6217, 2021.

[20] R. Balamurugan, S. Ratheesh, and Y. M. Venila, "Classification of heart disease using adaptive Harris hawk optimization-based clustering algorithm and enhanced deep genetic algorithm," *Soft Computing*, vol. 26, no. 5, pp. 2357–2373, 2021.

[21] X. Xu, L. Li, and A. Sharma, "Controlling messy errors in virtual reconstruction of random sports image capture points for complex systems," *International Journal of Systems Assurance Engineering and Management*, vol. 2, no. 1 2021.

[22] I. E. Agbehadji, R. C. Millham, A. Abayomi, J. J. Jung, and S. O. Frimpong, "Clustering algorithm based on nature-inspired approach for energy optimization in heterogeneous wireless sensor network," *Applied Soft Computing*, vol. 104, pp. 107171–107175, 2021.

[23] M. Bradha, N. Balakrishnan, S. Suvi et al., "Experimental, Computational Analysis of Butein and Lanceoletin for Natural Dye-Sensitized Solar Cells and Stabilizing Efficiency by IoT," *Environment, Development and Sustainability*, vol. 24, no. 7, 2021.

[24] X. Liu, C. Ma, and C. Yang, "Power station flue gas desulfurization system based on automatic online monitoring platform," *Journal of Digital Information Management*, vol. 13, no. 06, pp. 480–488, 2015.

[25] R. Huang, S. Zhang, W. Zhang, and X. Yang, "Progress of zinc oxide-based nanocomposites in the textile industry," *IET Collaborative Intelligent Manufacturing*, vol. 3, no. 3, pp. 281–289, 2021.

[26] S. Bi, R. Xu, A. Liu, L. Wang, and L. Wan, "Mining taxi pick-up hotspots based on grid information entropy clustering algorithm," *Journal of Advanced Transportation*, vol. 2021, no. 1, pp. 1–25, 2021.