Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4699573, 10 pages
https://doi.org/10.1155/2022/4699573

Research Article

Partition KMNN-DBSCAN Algorithm and Its Application in Extraction of Rail Damage Data

Yujun Li[1,2] **Zhi Yang**[1] **Shangbin Jiao**[1,2] **and Yuxing Li**[1,2]

[1]*School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, Shaanxi, China*
[2]*Shaanxi Province Complex System Control and Intelligent Information Processing Key Laboratory, Xi'an 710048, Shaanxi, China*

Correspondence should be addressed to Zhi Yang; yangzhi1045549083@163.com

Received 24 May 2022; Revised 11 June 2022; Accepted 15 June 2022; Published 13 July 2022

Academic Editor: Saeid Jafarzadeh Ghoushchi

In order to realize intelligent identification of rail damage, this paper studies the extraction method of complete damage ultrasonic B-scan data based on the density-based spatial clustering of applications with noise algorithm (DBSCAN). Aiming at the problem that the traditional DBSCAN algorithm needs to manually set the Eps and Minpts parameters, a KMNN-DBSCAN (K-median nearest neighbor DBSCAN) algorithm is proposed. The algorithm first uses the dataset's own distribution characteristics to generate a list of Eps and Minpts parameters and then determines the optimal Eps and Minpts through an optimization strategy to achieve complete self-adaptation of the two parameters of Eps and Minpts. In order to further improve the clustering performance of the algorithm, the partition idea is introduced, and the partition KMNN-DBSCAN algorithm is proposed to solve the problem that the clustering results of the DBSCAN algorithm are inconsistent with the actual categories on datasets with uneven density. The experimental results show that the KMNN-DBSCAN algorithm has higher clustering accuracy and silhouette coefficient (SC) for the D_{037} dataset ultrasound information group (UIG) division; compared with the KMNN-DBSCAN algorithm, the proposed partition KMNN-DBSCAN algorithm has higher clustering accuracy, F-Measure, and SC values. The partition KMNN-DBSCAN algorithm achieves accurate division of all damage UIG on the damaged B-scan data with large density differences, and completes the effective extraction of complete damage data.

1. Introduction

Ultrasonic rail flaw detection vehicles are widely used in rail damage detection with the advantages of high detection sensitivity, good directionality, and accurate defect positioning [1]. The ultrasonic rail detection vehicle collects the damage B-scan data based on the multichannel ultrasonic probe, and the technicians classify the damage based on the damage B-scan data [2]. Rail damage identification is mainly divided into two parts: complete damage data extraction and damage identification. Effective extraction of damage data is an important prerequisite for accurate damage identification [3]. Complete damage data extraction refers to dividing the adjacent ultrasonic echo points in physical locations together to form an ultrasonic information group (UIG). UIG is the complete damage data and is the smallest unit of damage identification [4]. The original B-scan data are stored in the form of data stream, which is difficult to extract manually. In this paper, the clustering algorithm is used to complete the damage data extraction.

Clustering algorithms are divided into partition-based clustering, hierarchical-based clustering, model-based clustering, and density-based clustering according to different clustering criteria [5]. The partition-based clustering must determine the final classification of the dataset before clustering, and the number of damages in the B-scan data file is uncertain, so the partition-based clustering is not suitable for rail damage classification [6]. The hierarchical-based clustering is only suitable for finding spherical or spherical clusters, while UIG is irregular in shape, so this method is not suitable for UIG division [7]. The model-based clustering algorithm assumes that the input dataset has a potential probability distribution, and the clustering effect will be affected if the assumptions do not hold [8]. However, the distribution of B-scan data is random, so the model-based clustering algorithm is not suitable for UIG division of B-scan data. The density-based clustering algorithm uses density as the clustering criterion and can find clusters of any shape without presetting the number of clustering results. The representative algorithm is DBSCAN (density-based spatial clustering of applications with noise), which can filter out abnormal points as noise while clustering [9]. The DBSCAN algorithm satisfies the requirement that the number of damaged rails in ultrasonic B-scan data is unknown, the damage shape is irregular, and the noise needs to be filtered out. However, the traditional DBSCAN algorithm has two flaws. The first is that the neighborhood radius Eps and the minimum density threshold Minpts need to be manually set, which is prone to clustering failure. The second is that the traditional DBSCAN algorithm has errors when clustering datasets with large density differences. In view of the above two problems, some scholars have improved the DBSCAN algorithm. For example, the literature [10] proposed that the distance value corresponding to the region where the distance distribution curve of the input dataset "steepened" was taken as Eps. This method provides a criterion for determining Eps based on the characteristics of data distribution, which has certain guiding significance for the selection of Eps, but it needs to determine the "steepened" area through manual observation. The literature [11] proposes the AF-DBSCAN algorithm, which uses polynomial fitting to fit the distance curve corresponding to a certain K value in the input dataset, solves the inflection point of the fitted curve, and takes the maximum distance corresponding to the inflection point as the optimal value of Eps. However, the discussion on the selection of *K* value is lacking in the text. The literature [12] proposes the SA-DBSCAN algorithm, which uses inverse Gaussian fitting to fit the probability distribution curve of the dataset, and takes the distance value corresponding to the peak point of the distribution curve as the value of Eps. Under the same Eps, the noise points are the least when the corresponding Minpts are taken as the optimal value of Minpts. However, the SA-DBSCAN algorithm achieves complete adaptation of the two parameter values, but the algorithm makes assumptions about the distribution of the input dataset and is not applicable to all datasets. The literature [13–15] improves the DBSCAN algorithm based on the idea of grid division to achieve good clustering effect on datasets with uneven density distribution. However, the mesh size setting of the parameter adaptive algorithm based on mesh division lacks theoretical guidance and is difficult to set manually.

In view of this, this paper proposes a new parameter adaptive DBSCAN algorithm KMNN-DBSCAN, which automatically determines the optimal clustering parameters based on the distance distribution characteristics of the input dataset, and realizes the fully adaptive selection of Eps and Minpts parameters. After that, the partition idea was introduced to improve the KMNN-DBSCAN algorithm, and the partition KMNN-DBSCAN algorithm was proposed. The partition KMNN-DBSCAN algorithm is proposed to solve the problem of errors in the traditional DBSCAN algorithm when clustering datasets with large density differences, and realizes the effective extraction of complete damage data.

2. Algorithm Principle

2.1. Principle of DBSCAN Algorithm. The DBSCAN (density-based spatial clustering of applications with noise) algorithm is a clustering algorithm based on high-density connected regions, which can filter out noise points while discovering any cluster. The related concepts of DBSCAN algorithm are as follows:

(1) The Eps neighborhood of element points: for a certain element point p in a given dataset D, the Eps neighborhood of p refers to the set of all element points in the area with p as the center and Eps as the radius, denoted as Eps(p). Eps(p) expression is as follows:

$$Eps(p) = \{q \in D | distance(p, q) \leq Eps\}, \tag{1}$$

where distance (p, q) is the Euclidean distance between the element point p and the element point q in the dataset D.

(2) Density: DBSCAN defines the number of element points contained in the Eps neighborhood of an element point as the density of the point, and the expression is as follows:

$$Density = Num(Eps(p)). \tag{2}$$

(3) Core point: if the Eps neighborhood of an element point contains the number of element points greater than or equal to the given minimum density threshold Minpts, the point is called a core point, and the expression is as follows:

$$Num(Eps(p)) \geq Minpts. \tag{3}$$

(4) Boundary point: if the number of element points contained in the Eps neighborhood of an element point is less than the given minimum density threshold Minpts, the point is called a boundary point, and the expression is as follows:

$$Num(Eps(p)) < Minpts. \tag{4}$$

(5) Direct density reachability: for any two element points p and q in dataset D, if q is in the Eps neighborhood of p, and p is the core point, then point q is said to be directly density reachable from point p.

(6) Density reachable: for a set of element points p_1, p_2,. . ., p_i,. . .,p_n in dataset D, where $p = p_1$, and $q = p_n$. An element point q is said to be density-reachable from point p if it is directly density-reachable for any p_{i+1} to p_i.

(7) Density connection: for an element point r in the dataset D, if the element point p and the element

point q are both density-reachable from point r, then point q is said to be density-connected from point p.

(8) Cluster (C): for a nonempty subset C of the input dataset, if C meets the following conditions, C is called a cluster.

(1) For any element point q, there is an element point p belonging to C. If q is density-reachable from p, then q belongs to C.

(2) For any two element points in C, they are density connection.

Noise point: the element point that does not belong to any cluster is called noise point, denoted as noise, and the expression is as follows:

$$noise = \{o \in D | \forall i\text{: } o \notin Ci\}. \tag{5}$$

The clustering principle of DBSCAN algorithm is the maximum density connected sample set derived from the density reachability relationship [16].

2.2. Principle of KMNN-DBSCAN Algorithm. The core of KMNN-DBSCAN is to generate a list of Eps and Minpts parameter pairs based on the distance distribution characteristics of the dataset. The parameters are optimized based on the Eps and Minpts parameter lists, and the optimal parameter pair is obtained, so as to realize the complete adaptation of the Eps and Minpts parameters of DBSCAN. The algorithm flow of KMNN-DBSCAN is shown in Figure 1.

The steps of the KMNN-DBSCAN algorithm are as follows:

Step 1. Generate a list of Eps parameter values.

The Eps parameter list is generated by the K-median nearest neighbor (KMNN) algorithm. The main principle of this algorithm is to first calculate the K-nearest neighbor distance matrix of the input dataset and then find the median of the K-nearest neighbor distances of all element points, and constitute the K-median nearest neighbor set. Take the K-median nearest neighbor set as the Eps parameter list, and the specific steps are as follows:

(1) Calculate the distance distribution matrix $\text{Dist}_{n\times n}$

$$Dist_{n\times n} = \{dist\,(i, j) | 1 \leq i \leq n, 1 \leq j \leq n\}, \tag{6}$$

where n is the number of element points in the input dataset D, dist(i, j) is the Euclidean distance between element point i and element point j, and $\text{Dist}_{n\times n}$ is the distance matrix.

(2) Calculate the K-nearest neighbor distance matrix $\text{KNN}_{n\times n}$

$$KNN_{n\times n} = sort\,(Dist_{n\times n}). \tag{7}$$

Arrange each row of the distance distribution matrix in ascending order to obtain the K-nearest neighbor distance matrix $\text{KNN}_{n\times n}$. The ith row of $\text{KNN}_{n\times n}$ represents the distance between the element points closest to the ith element point 1~n, and the K-th column ($K = 1, 2, \ldots, n$) represents the K-nearest neighbor distance corresponding to all element points. Where the first column is the distance from the element point to itself, the first column of $\text{KNN}_{n\times n}$ is all zero.

(3) K-median nearest neighbor (KMNN) distance set $\text{KMNN}_{1\times n}$

Find the median of each column of $\text{KMNN}_{1\times n}$, and all the medians form the K-median nearest neighbor distance set $\text{KMNN}_{1\times n}$. Take the K-median nearest neighbor distance set $\text{KMNN}_{1\times n}$ as a list of Eps parameter values, where the Eps value corresponds to the K.

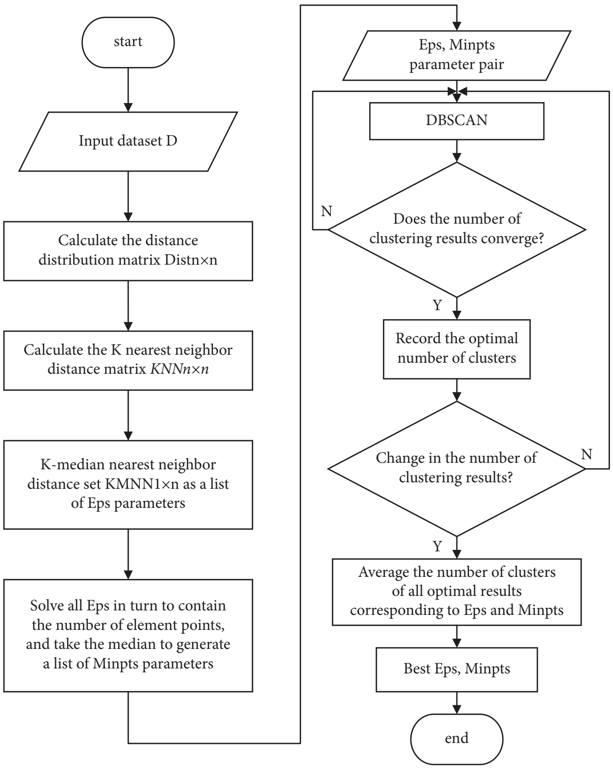

start
Input dataset D
Calculate the distance distribution matrix Distn×n
Calculate the K nearest neighbor distance matrix KNNn×n
K-median nearest neighbor distance set KMNN1×n as a list of Eps parameters
Solve all Eps in turn to contain the number of element points, and take the median to generate a list of Minpts parameters
Eps, Minpts parameter pair
DBSCAN
N
Does the number of clustering results converge?
Y
Record the optimal number of clusters
Change in the number of clustering results?
N
Y
Average the number of clusters of all optimal results corresponding to Eps and Minpts
Best Eps, Minpts
end

Figure 1: KMNN-DBSCAN algorithm flow.

Step 2. Generate a list of Minpts parameter values.

Generates a list of Minpts parameter values using the median method and a given list of Eps parameter values. For the current Eps parameter list, the number of element points contained in the Eps neighborhood of all element points under different Eps is obtained in turn. The median of the number of element points contained in the Eps neighborhood of all element points is taken as the Minpts value corresponding to the current Eps value. The Minpts values corresponding to all Eps values are obtained to form the Minpts parameter list, and the Minpts values correspond to the Eps values. Different K corresponds to different Eps, Minpts parameter value pairs.

Step 3. Eps and Minpts parameter pair optimization.

The Eps and Minpts parameter value pairs corresponding to different K values are selected in turn as the clustering parameters of the DBSCAN algorithm, and the number of clustering results under the Eps and Minpts parameter values corresponding to different K values is obtained. This paper considers that when the number of clustering results does not change for three consecutive times, the number of clustering results is stable, and the current number of clustering results is recorded as the optimal number of clustering results. After the number of clustering results is stable, we continue the above operations until the number of clustering results changes. The interval from the optimal number of clustering results to the change of the number of clustering results is called the stable area. For the stable area corresponding to all the Eps and Minpts parameter value pairs, the expected $\overline{\text{EPS}}$ and $\overline{\text{Minpts}}$ are obtained respectively, and $\overline{\text{EPS}}$ and $\overline{\text{Minpts}}$ are taken as the optimal Eps and Minpts value.

2.3. Partition KMNN-DBSCAN Algorithm Principle. In order to solve the problem that the traditional DBSCAN algorithm has errors in the clustering results on datasets with large density differences, the idea of partitioning is introduced to improve the KMNN-DBSCAN algorithm [17], and the partition KMNN-DBSCAN algorithm is proposed. The implementation steps of the partition KMNN-DBSCAN algorithm are as follows:

Step 1. Partition the input dataset.

Project the element points of the input dataset on the X-axis and the Y-axis, respectively, to obtain the distribution of the input dataset on the X-axis and the Y-axis. The partition points are selected at the places where the density is the most sparse among different density centers, and sub-datasets with different densities are obtained.

Step 2. Perform cluster analysis on sub-datasets separately.

KMNN-DBSCAN is used to adapt the parameters to the sub-datasets, respectively, and the local optimal parameters are used to cluster the sub-datasets to obtain the clustering results of the sub-datasets.

Step 3. Merge local clustering results.

Merge the clustering results of the sub-datasets and finally get the clustering results of the original datasets.

The flow chart of the partition KMNN-DBSCAN algorithm is shown in Figure 2.

3. Ultrasound Information Group (UIG) Division Experiment

As shown in Figure 3, the complete damage data extraction for the original B-scan data includes the division of UIC and UIG, and UIG is the complete damage data.

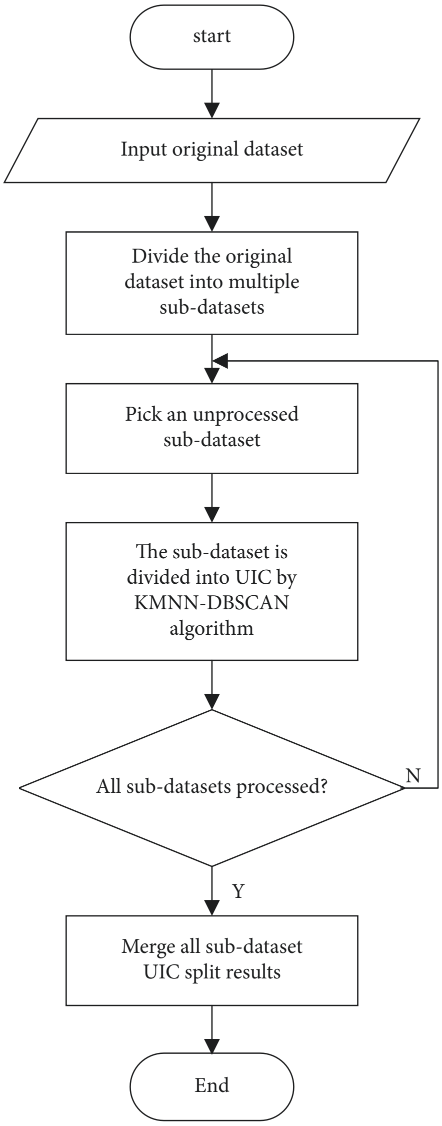

start
Input original dataset
Divide the original dataset into multiple sub-datasets
Pick an unprocessed sub-dataset
The sub-dataset is divided into UIC by KMNN-DBSCAN algorithm
All sub-datasets processed?
N
Y
Merge all sub-dataset UIC split results
End

Figure 2: Flowchart of the partitioned KMNN-DBSCAN algorithm.

3.1. Ultrasound Information Combination (UIC) Division. Ultrasonic steel rail flaw detection vehicle adopts ultrasonic probe wheel as sensor. The ultrasonic probe wheel has built-in ultrasonic sensors with four angles: 0°, 37°, straight 70°, and oblique 70°. Different angle sensors are used to detect different areas of the rail, and the combination of different ultrasonic sensor channels is used to detect different types of damage. If UIC is not divided, it will affect the complete data extraction of damage [18].

Figure 4 shows the detection range of ultrasonic sensors with different angles. The ultrasonic propagation range of the 0° probe is from the rail surface to the rail bottom, and the detection range is the projection area of the rail waist. The 0° probe is usually used to detect the level of crack from the rail surface to the rail bottom. The detection range of the 37° probe is from the rail surface of the rail waist projection area to the lower part of the rail waist. The 37° probe is usually used to detect the crack of the screw hole of the rail

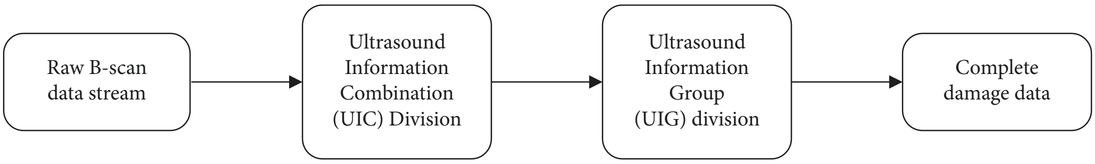

Raw B-scan data stream
Ultrasound Information Combination (UIC) Division
Ultrasound Information Group (UIG) division
Complete damage data

Figure 3: Complete damage data extraction process.

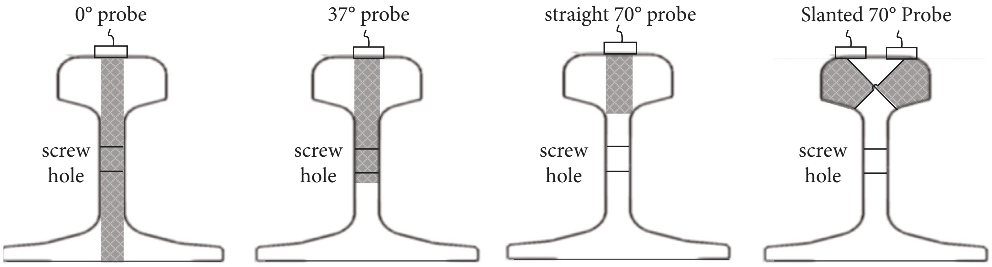


Figure 4: Detection range of ultrasonic sensors at different angles.

Table 1: Ultrasonic sensor combination corresponding to common rail damage.

Damage location	Damage type	Corresponding ultrasonic sensor combination
Rail surface	Oblique downward defect	37° probe
Rail head	Inner transverse hole	Slanted 70° probe
Rail head	Central nuclear injury	Straight 70° probe
Rail head	Outer lateral hole	Slanted 70° probe
Rail waist	Intact screw hole	0°, 37° probe
Rail waist	Horizontal cracks in screw holes	0°, 37° probe
Rail waist	Oblique crack in screw hole	0°, 37° probe
Rail bottom	Cross-hole	0° probe

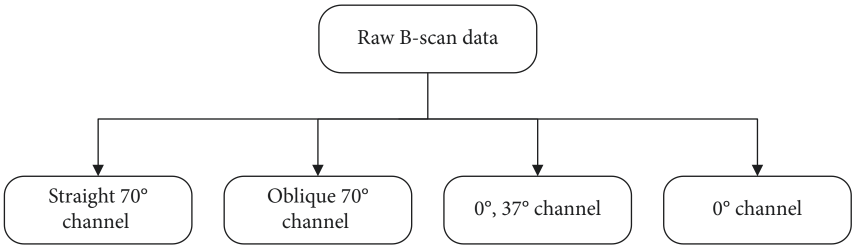


Figure 5: B-scan data UIC division.

waist. The detection range of the straight 70° probe is the central area of the rail head, which is generally used to detect nuclear damage in the middle of the rail head. The detection range of the oblique 70° probe is the area on both sides of the rail head and is generally used to detect the lateral holes on the inner and outer sides of the rail head.

As shown in Figure 4, there are related cases for sensors with different angles (such as 0° and 37° ultrasonic sensors), and there are also unrelated channels (such as 37° and oblique 70° ultrasonic sensors). If no distinction is made, the division of subsequent UIG will be affected, thereby affecting the effect of injury judgment. Table 1 shows the ultrasonic sensor combinations corresponding to seven common rail damages. From Table 1, it can be seen that different ultrasonic sensor combinations are used to detect different damages.

In this paper, the raw B-scan data are divided into UIC according to the ultrasonic sensor channel and position correlation, combined with the combination of ultrasonic sensor channels corresponding to each damage type. The UIC division results are shown in Figure 5.

3.2. KMNN-DBSCAN Algorithm Applied to Ultrasonic Information Group (UIG) Division. In order to verify the effectiveness of the KMNN-DBSCAN algorithm in this paper, the KMNN-DBSCAN algorithm is applied to the D_{037} UIC dataset for UIG division. The division results of the KMNN-DBSCAN algorithm are compared with those of the SA-DBSCAN and AF-DBSCAN algorithms. D_{037} is the 0° and 37° channel UIC data of the damage B-scan data collected in an experiment. The dataset contains 830 two-dimensional

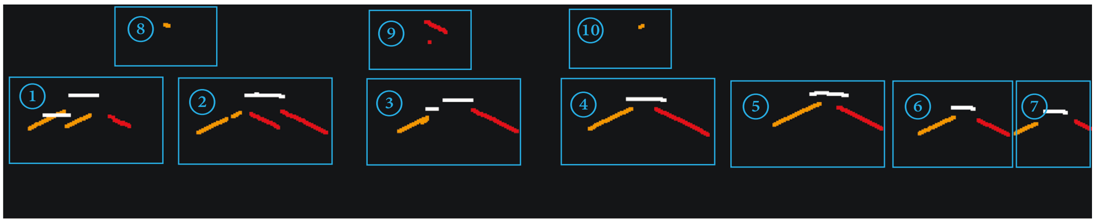
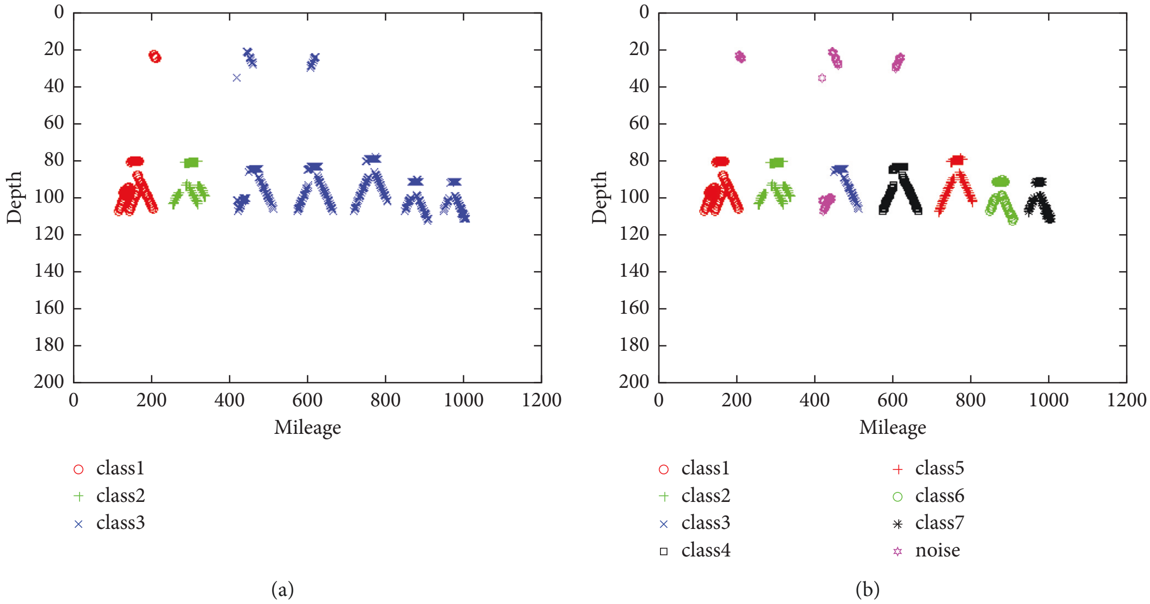


Figure 6: D_{037} dataset corresponds to B-scan.



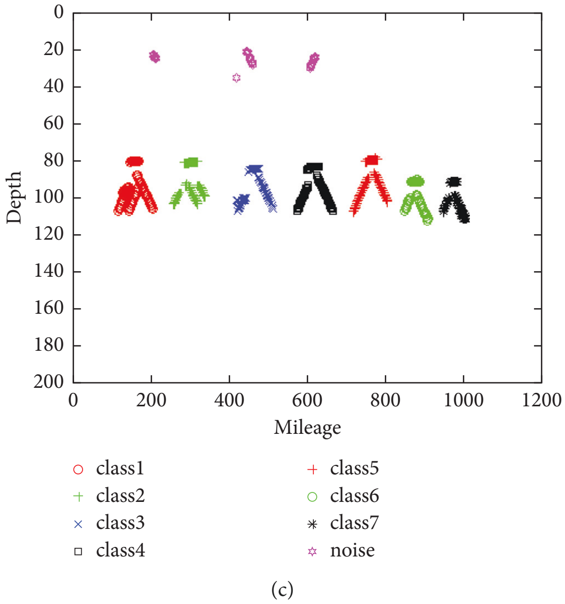


Figure 7: UIG division results of different parameter adaptive algorithms. (a) SA-DBSCAN. (b) AF-DBSCAN. (c) KMNN-DBSCAN.

ultrasonic echo data. Figure 6 shows the corresponding B-scan of the D_{037} dataset. The D_{037} dataset contains 0 damaged UIGs; that is, the D_{037} dataset contains 10 clusters.

Figure 7(a)–7(c) are the UIG division results of the D_{037} dataset by the SA-DBSCAN algorithm, the AF-DBSCAN algorithm, and the KMNN-DBSCAN algorithm. All elements with the same point type form a cluster class, elements with different point types represent different clusters, and noise points are points that do not belong to any cluster.

The UIG division results and evaluation indexes of different clustering algorithms are shown in Table 2.

Table 2: UIG division results and evaluation indexes of different clustering algorithms.

Algorithm name	Eps	Minpts	Actual number of classes	Number of clustering results	Number of correct clusters	F-Measure	SC
AF-DBSCAN	23.4	30.1	10	7	6	0.905	0.442
SA-DBSCAN	16.5	12.0	10	3	1	1	0.468
KMNN-DBSCAN	22.9	42.7	10	7	7	0.957	0.855

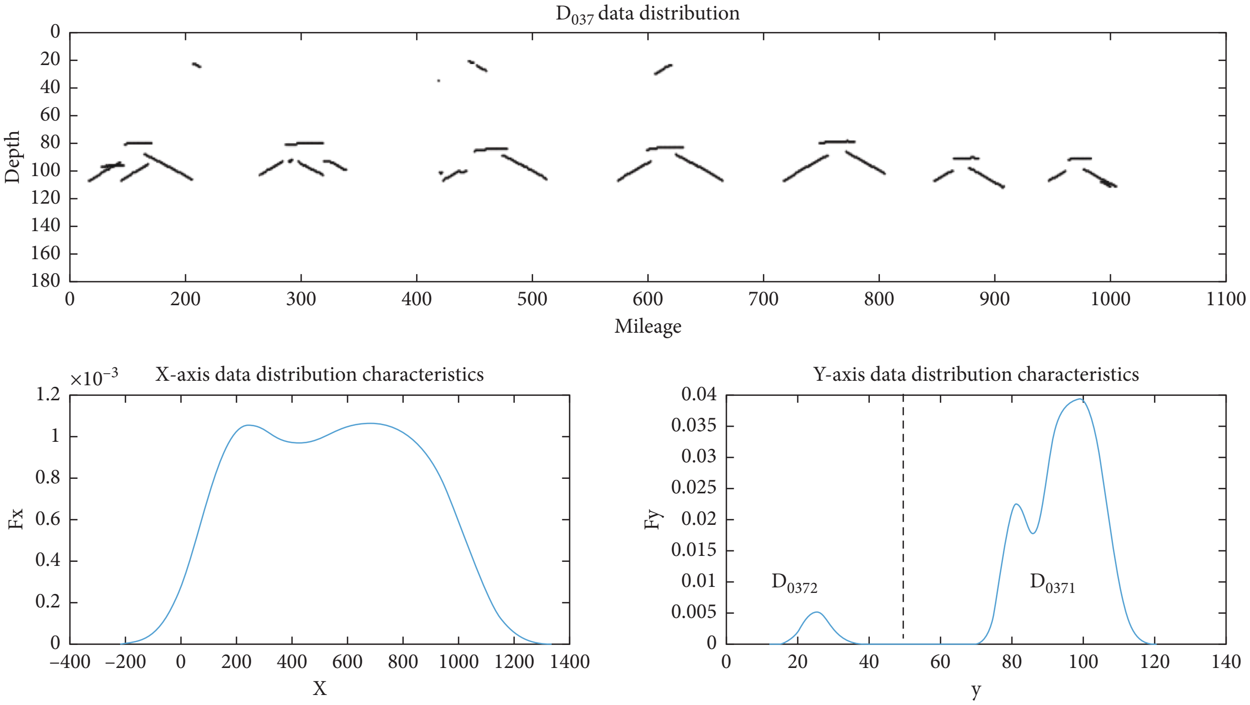


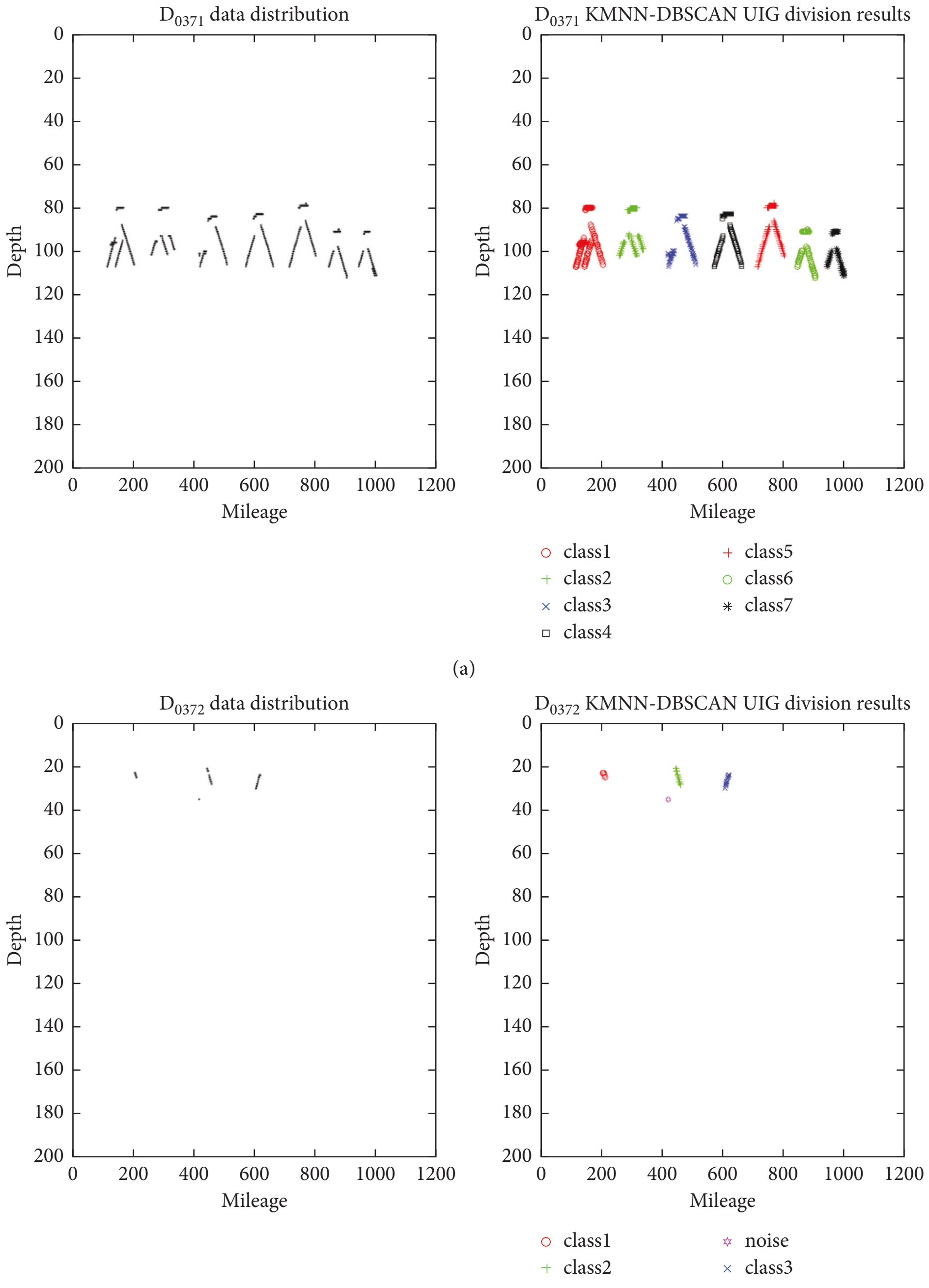
Figure 8: Data distribution characteristics of D_{037} dataset.

Among them, F-Measure reflects the ability of the algorithm to identify valid data points, and the larger the F value, the stronger the ability of the algorithm to identify valid data points [19]. The silhouette coefficient comprehensively reflects the compactness within the class and the separation between the classes. The value range of SC is [−1, 1]. Within the value range, the larger the value of SC, the better the clustering result [20].

Judging from the correct number of clusters in Table 2, the number of clustering results of the KMNN-DBSCAN algorithm is the closest to the real number of clusters in the D_{037} dataset, and the number of correct clusters is also the most, indicating that the KMNN-DBSCAN algorithm has the highest UIG division accuracy. From the F-Measure in Table 2, the F values of the AF-DBSCAN algorithm, the SA-DBSCAN algorithm, and the KMNN-DBSCAN algorithm are all above 0.9, indicating that the three algorithms have a strong ability to identify valid element points. According to the SC index in Table 2, the SC value of the UIG division result of the KMNN-DBSCAN algorithm is significantly higher than that of the comparison algorithm, indicating that the clusters divided by KMNN-DBSCAN have more compact elements within the cluster and more dispersed between different clusters, which shows that the UIG division of the KMNN-DBSCAN algorithm is more reasonable.

By comparing the UIG division results of the KMNN-DBSCAN algorithm and the comparison algorithm in terms of the number of clustering results, the number of correct clusters, F-Measure, and SC indicators, it can be seen that the KMNN-DBSCAN algorithm proposed in this paper not only has strong identification efficiency. The ability of element points, and the accuracy and rationality of ultrasound information group division are better.

3.3. Partition KMNN-DBSCAN Algorithm Applied to Ultrasound Information Group (UIG) Division. The D_{037} dataset contains a total of 10 rail damage UIG. The KMNN-DBSCAN algorithm divides the D_{037} dataset into 7 clusters and fails to correctly divide all the ultrasound information clusters. This is because the densities of the UIG ⑧~⑩ and the UIG ①~⑦ in the D_{037} dataset are quite different. The KMNN-DBSCAN algorithm uses the same clustering parameters to process the datasets with large density differences, resulting in the UIG ⑧~⑩ is divided into noise. To solve this problem, this paper proposes a partition KMNN-DBSCAN algorithm. In order to verify the effectiveness of



(a)

(b)

Figure 9: Sub-dataset KMNN-DBSCAN UIG division results. (a) D_{0371} dataset KMNN-DBSCAN UIG division results. (b) D_{0372} dataset KMNN-DBSCAN UIG division results.

the partition KMNN-DBSCAN algorithm, the algorithm was used in the D_{037} UIC dataset to divide ultrasound information groups, and the specific process is as follows:

(1) Partitioning the D_{037} dataset based on the density distribution characteristics of the dataset.

The D_{037} data points are projected on the X-axis and the Y-axis, respectively, and the spatial distribution characteristics of the data element points of the D_{037} dataset on the X-axis and the Y-axis are shown in Figure 8.

As shown in Figure 8, the density of the D_{037} dataset is continuous on the X-axis, and there are two dense regions on the Y-axis, and the two densities are quite different. Where the density is the most sparse between the two density centers, D_{037} is divided into two sub-datasets of D_{0371} and D_{0372}.

(2) The D_{0371} and D_{0372} datasets are divided into UIG by KMNN-DBSCAN.

Figure 9 shows the results of KMNN-DBSCAN dividing the D_{0371} and D_{0372} datasets by UIG.

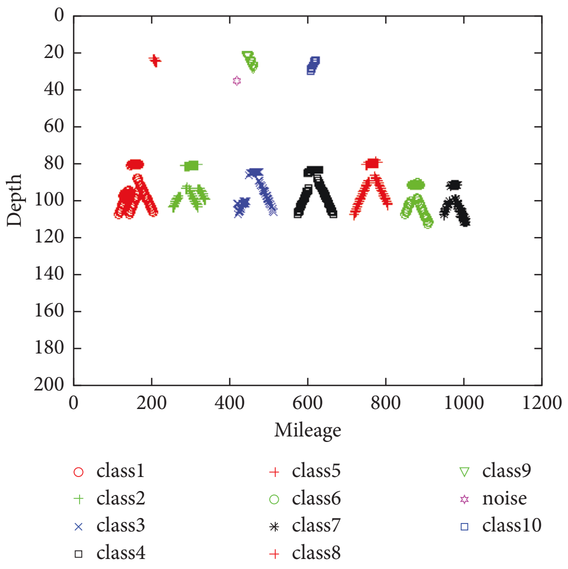


Figure 10: Partition KMNN-DBSCAN algorithm for UIG division of D_{037} dataset.

Table 3: Division results and evaluation indexes of UIG by partition KMNN-DBSCAN algorithm and KMNN-DBSCAN algorithm.

Algorithm name	Actual number of classes	Number of clustering results	Number of correct clusters	*F*-Measure	SC
KMNN-DBSCAN	10	7	7	0.957	0.855
Partition KMNN-DBSCAN	10	10	10	1	0.961

It can be seen from Figure 9 that the KMNN-DBSCAN algorithm is used to adapt the parameters of the two sub-datasets D_{0371} and D_{0372}, respectively, and the local optimal parameters are used for the division of the local UIG. Both sub-datasets realize the UIG accurate division.

(3) Merge the results of sub-dataset UIG division.

After completing the local clustering of each sub-dataset, all local clustering results should be merged to obtain the overall clustering result of the original dataset. By combining the division results of the UIG of the above two sub-datasets D_{0371} and D_{0372}, the result of dividing the UIG of the original input dataset D_{037} by partition KMNN-DBSCAN is shown in Figure 10.

Table 3 shows the comparison results of the UIG division of D_{037} by partition KMNN-DBSCAN and KMNN-DBSCAN. Judging from the number of clustering results and the number of correct clusters in Table 3, the partitioned KMNN-DBSCAN algorithm divides all UIGs of the D_{037} dataset correctly. From the F-Measure in Table 3, the F-Measure value of the partitioned KMNN-DBSCAN algorithm is 1, which indicates that the partitioned KMNN-DBSCAN algorithm has identified all valid data points. From the SC index in Table 3, the SC value of the UIG division result of the partitioned KMNN-DBSCAN algorithm is significantly higher than that of the KMNN-DBSCAN algorithm. This shows that the clusters divided by the partitioned KMNN-DBSCAN have more compact elements in the cluster and more dispersed between different clusters; that is, the partitioned KMNN-DBSCAN algorithm UIG division is more reasonable.

4. Conclusion

In this paper, according to the characteristics of damage B-scan data, the density-based DBSCAN clustering algorithm is selected to study the extraction method of complete damage B-scan data. The main conclusions are as follows:

(1) A parameter adaptive DBSCAN algorithm KMNN-DBSCAN is proposed, which solves the defect that the traditional DBSCAN algorithm needs to manually set the clustering parameters. KMNN-DBSCAN algorithm generates the list of Eps and Minpts parameter values according to the distance distribution characteristics of the input dataset, determines the optimal Eps and Minpts according to the relationship between the clustering results of the algorithm and the *K* under different parameter values, and realizes the complete adaptation of the two parameter values of Eps and Minpts.

(2) The KMNN-DBSCAN algorithm is applied to the D_{037} dataset for UIG partitioning. The experimental results show that, compared with SA-DBSCAN algorithm and AF-DBSCAN algorithm, KMNN-DBSCAN algorithm has stronger ability to identify valid element points and performs better in the accuracy and rationality of UIG division.

(3) The proposed partition KMNN-DBSCAN algorithm solves the problem that the traditional DBSCAN algorithm has errors in the clustering of datasets with uneven density distribution. The partition KMNN-DBSCAN algorithm is applied to the D_{037} dataset for UIG division. The experimental results show that the partition KMNN-DBSCAN algorithm can effectively realize the accurate UIG division, further improve the correctness and rationality of UIG division, and realize the effective extraction of complete damage data.

According to the principle of the partition KMNN-DBSCAN algorithm, the time complexity of the algorithm in this paper is relatively high. Therefore, how to effectively reduce the time complexity of the algorithm is the focus of our next work.

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 declared that they have no conflicts of interest.

Acknowledgments

This work was supported by the Key Scientific Research Program of Shaanxi Provincial Education Department (no. 20JY046), the Program of National Natural Science Foundation of China (no. 61871318), the Key Research Program of Shaanxi Province (no. 2021GY-259), and the Xi'an Science and Technology Project (21XJZZ0044).

References

[1] C. Sun, J. Liu, Y. Qin, and Y. Zhang, "Intelligent detection method for rail flaw based on deep learning," *Zhongguo Tiedao Kexue/China Railway Science*, vol. 39, no. 5, pp. 51–57, 2018.

[2] C. Lu, Y. Wei, and W. Xu, "Study on B-scan imaging detection for rail tread tilted cracks using ultrasonic surface wave," *Yi Qi Yi Biao Xue Bao/Chinese Journal of Scientific Instrument*.vol. 31, no. 10, pp. 2272–2278, 2010.

[3] X. Y. Huang, Y. S. Shi, and Y. H. Zhang, "BP neural network based on rail flaw classification of RFD car's B-scan data," *China Railway*, vol. 3, pp. 82–87, 2018.

[4] C. Tastimur, M. Karakose, E. Akin, and I. Aydin, "Rail defect detection with real time image processing technique," in *Proceedings of the 2016 IEEE 14TH INTERNATIONAL CONFERENCE ON INDUSTRIAL INFORMATICS (INDIN)*, pp. 411–415, IEEE, Poitiers, France, 19-21 July 2016.

[5] M. Namratha, "A comprehensive overview of clustering algorithms in pattern recognition," *IOSR Journal of Computer Engineering*, vol. 4, no. 6, pp. 23–30, 2012.

[6] Y. He, "Research and application of clustering algorithm based on partition," *Computer knowledge and technology*, vol. 13, no. 16, pp. 55-56, 2017.

[7] Z. G. Chen and D. M. Zhu, "Hierarchic clustering algorithm used for anomaly detecting," *Procedia Engineering*, vol. 15, no. C, pp. 3401–3405, 2011.

[8] B. Charles and B. S. Camille, "Model-based clustering of high-dimensional data: a review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, AAAI Press, Germany, 1996.

[10] J. Y. Song, Y. P. Guo, and B. Wang, "The parameter configuration method of DBSCAN clustering algorithm," in *Proceedings of the 2018 5th International Conference on Systems and Informatics (ICSAI)*, no. 5, pp. 1062–1070, IEEE, Nanjing, China, 10-12 November 2018.

[11] Z. P. Zhou, J. F. Wang, S. W. Zhu, and Z. Sun, "An improved adaptive and fast AF-DBSCAN clustering algorithm," *CAAI Transactions on Intelligent Systems*, vol. 11, no. 1, pp. 93–98, 2016.

[12] L. N. Xia and J. W. Sa-Dbscan, "A self-adaptive density-based clustering algorithm," *Journal of the Graduate School of the Chinese Academy of Sciences*, vol. 26, no. 4, pp. 530–538, 2009.

[13] B. Thapana, A. Xiang, L. Yang, Z. Weihong, Z. Fuzhen, and H. Qing, "Grid-based DBSCAN: indexing and inference," *Pattern Recognition*, vol. 90, pp. 271–284, 2019.

[14] S. Artur and C. Andrzej, "Grid-based approach to determining parameters of the DBSCAN algorithm," *Artificial Intelligence and Soft Computing*, vol. 24, pp. 555–565, 2020.

[15] L. M. Zhang, Z. G. Xu, and F. Q. Si, "IEEE.GCMDDBSCAN: multi-density DBSCAN based on grid and contribution," in *Proceedings of the 2013 IEEE 11TH International Conference on Dependable, Autonomic and Secure Computing (DASC)*, pp. 502–507, IEEE, Chengdu, China, 21-22 December 2013.

[16] J. B. Shen, X. P. Hao, Z. Y. Liang, Y. Liu, W. Wang, and L. Shao, "Real-time superpixel segmentation by DBSCAN clustering algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5933–5942, 2016.

[17] S. G. Zhou, A. Y. Zhou, and J. Cao, "A data-partitioning-based DBSCAN algorithm," *Computer Research and Development*, vol. 37, no. 10, pp. 1153–1159, 2000.

[18] W. Li and H. X. Zhang, "A FPGA based ultrasonic rail flaw detection system," in *Proceedings of the 2017 IEEE INTERNATIONAL SYMPOSIUM ON SIGNAL PROCESSING AND INFORMATION TECHNOLOGY (ISSPIT)*, pp. 150–155, IEEE, Bilbao, Spain, 18-20 December 2017.

[19] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proceedings of the KDD Workshop on Text Mining*, p. 223, Boston, MA, USA, August 20-23, 2000.

[20] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York, 2009.