2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD) | 979-8-3503-3168-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSCWD57460.2023.10152613

Traffic accident location study based on AD-DBSCAN Algorithm with Adaptive Parameters

Xijun Zhang
School of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
zhangxijun198079@sina.com

Jin Su
School of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
maydaysu1@163.com

Hong Zhang
School of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
493705332@qq.com

Xianli Zhang
School of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
46613307@qq.com

Xuan Chen
School of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
183620260@qq.com

Yong Cui
School of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
308322411@qq.com

***Abstract*—Aiming at the shortcomings of the traditional Density-Based Spatial Clustering of Applications with Noise - DBSCAN algorithm such as insignificant clustering effect and the choice of parameter combinations. This paper proposes an AD-DBSCAN algorithm with adaptive parameters, which makes the algorithm more difficult in the selection of the parameters. By establishing a DBSCAN algorithm model to adapt to finding the optimal distance threshold and the minimum number of neighbor points, the clustering is more accurate, and the noise point identified in the data is more accurate. Through the observation of the calculation model of the Calinski-Harabasz index, the evaluation index of the clustering algorithm, the selection of the optimal best distance threshold and the minimum number of neighborhood points, the accuracy of noise point recognition is improved by 5 times in the clustering algorithm, and the Calinski-Harabasz index improved by about 39.84%. The applicability of the algorithm in clustering the locations of urban road traffic accidents is verified.**

Keywords—Smart transportation, Complex network, DBSCAN algorithm, Calinski-Harabasz Index

I. INTRODUCTION

With the rapid growth of road mileage, motor vehicle ownership and the number of motorists, the total number of traffic accidents and the number of casualties remains high. According to the statistics of a large number of traffic accidents in my country, about 90% of traffic accidents are caused by human error, about 30% are caused by road traffic environmental conditions, and about 10% are caused by vehicle technical conditions. The characteristic distribution of regional traffic accidents is obviously related to the living area of residents. In the city area, there are frequent accidents with specific characteristics of some districts or counties or a village. The research on the geographical distribution of traffic accident characteristics has important theoretical significance and practical value, and can provide a reference for the traffic control department to formulate targeted accident prevention countermeasures.

In the analysis of the causes of past traffic accidents, the methods for identifying the accident-prone points (segments) mainly include the quality control method [1]. First, it is assumed that the number of accidents in each road segment obeys the Poisson distribution, and then the accident rate of the road segment is compared with that of similar road segments. Average accident rates for comparison. The matrix method [2] combines the number of accidents and the accident rate as a criterion for identifying accident-prone points. The accident frequency method [3], the identification standard of this method is a certain number of critical accidents. If the number of accidents on a road section is greater than this threshold, it can be judged as an accident-prone point.

At present, for the shortcomings of the DBSCAN algorithm, domestic and foreign scholars have made corresponding improvements and innovations to adapt to different types and sizes of data sets. sunita J et al [4] proposed a method of multi-dimensional parameter setting to derive the data distribution of each dimension according to the checked method, and the improved algorithm can deal with data sets with different density distributions more effectively. Wang et al [5] proposed an adaptive density clustering method to determine the appropriate neighborhood Eps parameter values based on the data distribution and improve the determination of the parameter threshold Minpts, and the improved algorithm can detect clusters of different density levels; Ahmad M. Bakr et al [6] proposed a kind of division-based incremental aggregation algorithm to restrict the search space to partitions rather than the whole data set, and this region incremental approach improves the efficiency of clustering; Kumar et al [7] proposed a fast DBSCAN clustering algorithm by establishing a graphical index

structure for fast neighbor search operation, and the improved algorithm can effectively deal with noisy points and improve the performance of the algorithm; Severino F et al [8] proposed a new region query strategy considering the most relevant region query strategy of DBSCAN, and the improved algorithm has a great improvement in efficiency;

In recent years, the clustering analysis in the cause analysis of traffic accidents and the identification of accident-prone sections has become a research hotspot for domestic and foreign scholars. A large number of K-means and DBSCAN improved algorithms have been proposed successively. In 2017, Pan Wei et al. [9] proposed a clustering algorithm based on semi-supervised PCA attribute weighting K-means, effectively removing the influence of irrelevant attributes and noise attributes. In 2019, Wei Kangyuan et al. [10] overcame the local optimum by introducing the adaptive concept and immune clonal selection mechanism, and proposed the A2-GSA Kmeans algorithm. In 2016, Cheng Jing et al. [11] proposed a clustering method combining distance measurement of time series and self-correlation of time series, which is of great significance for exploring urban functional areas and structural layout. In 2020, He Yue et al. [12] proposed a grid-based taxi-carrying hotspot clustering algorithm to provide information services for taxi operators and managers. In the same year, Liao Zhuhua et al. [13] proposed the taxi passenger carrying area recommendation algorithm based on sparse trajectory data. In 2017, Jiang Huijuan et al. [14] proposed a taxi-passenger data clustering algorithm based on RL-DBSCAN algorithm. In 2015, Liao Lvchao et al. [15] proposed a directed density fast clustering method (D-Optics) to extract structural information of complex road networks through clustering analysis of directed spatiotemporal data. Based on the above research, the improved K-means clustering algorithm proposed in this paper has high adaptability to the mining of urban traffic demand. In 2019, Kim et al [16] proposed the AA-DBSCAN algorithm, based on a new tree structure of quad trees to define the density layers of the data set to achieve clustering of uneven density data sets, but the algorithm still requires input of relevant parameters. In 2018, The AD-DBSCAN adaptive clustering algorithm proposed by Khan et al [17] requires the number of clusters to be specified in advance and cannot automatically identify the number of cluster classes. In 2016, Ajaykumar et al [18] used K-means clustering algorithm to cluster the probabilistic Hough transformed line segments and determine the optimal number of clusters using contour coefficients, due to the limitations of K-means in the algorithm, the clustering effect is easily affected. In 2019, Wang et al [19] proposed a hierarchical density-based binning algorithm OPTICS algorithm evolved from the idea of DBSCAN algorithm, which finally obtained the output data with optimal distance threshold , which is insensitive to parameters, but still need to manually input the clustering parameters , and the clustering results have large differences for different inputs.

The purpose of this paper is to find traffic accident hotspots through the accident location information recorded in the traffic accident information collection items, and the density clustering method is more applicable. The accident analysis method based on density clustering is generally used in the field of criminal investigation for the study of crime hotspot areas. In fact, the occurrence of traffic accidents is highly correlated with the residence of traffic participants, and regional traffic accidents tend to show a concentration of some areas, and traffic accidents with specific characteristics (e.g., following overtaking, presence of blind spots in vision, etc.) are especially obvious. The application of density clustering method to classify accident locations can facilitate the traffic management department to fine-tune the management of traffic accidents in the region and develop targeted accident prevention countermeasures. Therefore, this paper adopts density clustering algorithm to classify accident locations. The main innovation points of this paper are as follows.

(1) The Calinski-Harabasz index is proposed to verify the clustering effect and solve how to choose the optimal clustering result.
(2) The adaptive selection of two important parameters, Eps and MinPts, is proposed, which largely improves the clustering effect and enhances the identification of noise points.

II. Materials and Methods

A. Traditional DBSCAN algorithm

DBSCAN is a density-based clustering algorithm. The algorithm defines clusters as the largest collection of densely connected points, is able to divide regions with sufficient density into clusters, and can discover arbitrarily shaped clusters in noisy spatial datasets. The DBSCAN algorithm involves two important parameters, namely the minimum number of clustering points MinPts and the neighborhood radius Eps of the clustering. The DBSCAN clustering algorithm is shown in Fig.1.

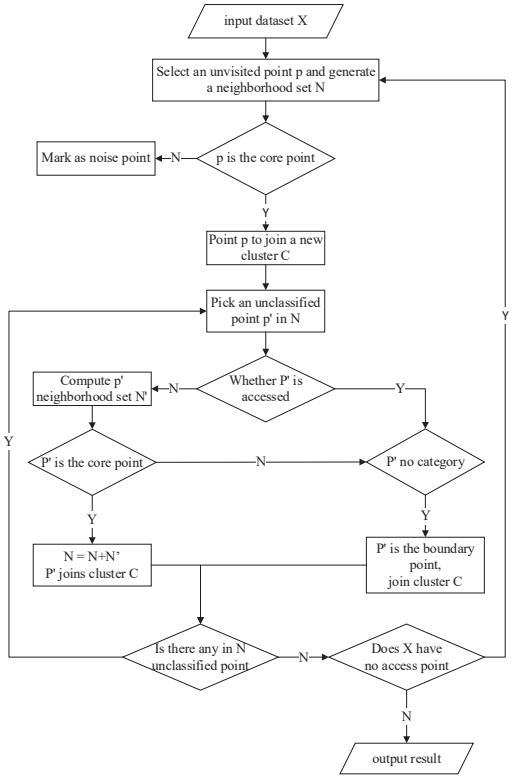

Fig. 1. DBSCAN clustering algorithm process

DBSCAN finds clusters by examining the ε-neighborhood of each object in the dataset, and if the ε-neighborhood of a point p contains m objects, a new cluster with p as the core object is created. Then, the DBSCAN algorithm goes to find the densities that these core objects can reach directly, and this process may involve the merging of density reachable clusters. The process ends when no new points can be added to any cluster. The intermediate values of ε and m for this algorithm are given according to prior knowledge.

B. Calinski-Harabasz Index

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

The Calinski-Harabasz index measures the intra-class tightness by calculating the sum of the squares of the distances between the points within the class and the center of the class, and the separation of a data set by calculating the sum of the squares of the distance between the center point of the class and the center point of the dataset, where tightness is used as the denominator and separation as the numerator, and the Calinski-Harabasz index is calculated by the ratio of separation to tightness, which is calculated at a much faster rate than the contour coefficient.

$$s(k) = \frac{tr(B_k)}{tr(W_k)} \frac{n-k}{k-1} \tag{1}$$

Where, n is the number of samples in the training set and k is the number of categories.

$$tr(B_k) = \sum_{i=1}^{k} n_i \, || \, m_i - m \, ||^2 \tag{2}$$

Where B_k is the covariance matrix between categories, m is the centroid of all points, and m_i is the center point of a class.

$$tr(W_k) = \sum_{i=1}^{k} \sum_{x \in c_i} || \, x - m_i \, ||^2 \tag{3}$$

Where W_k denotes the covariance matrix of the data within the category and tr is the trace of the matrix.

The analysis shows that the smaller the covariance of the data within a category the more accurate the classification result, and the larger the covariance between categories the more accurate the classification, which results in a higher Calinski-Harabasz score. A higher Calinski-Harabasz score means that the classes themselves are tighter and the classes are more dispersed from each other, then a better clustering result can be obtained.

C. AD-DBSCAN algorithm with adaptive parameters

Since the DBSCAN clustering algorithm is very sensitive to the uniformly set parameters clustering cluster minimum cluster points MinPts and neighborhood radius Eps, relevant scholars improve the DBSCAN algorithm for this problem. Jahirabadkar S et al [20] dynamically set the Eps parameter according to the density distribution of the data set in each dimension, but the MinPts parameter still needs to be input , and full automation of clustering is not achieved. In this paper, an AD-DBSCAN algorithm with MinPts and Eps adaptive selection is proposed for DBSCAN algorithm continues to be improved. The algorithm process is shown in Fig.2.

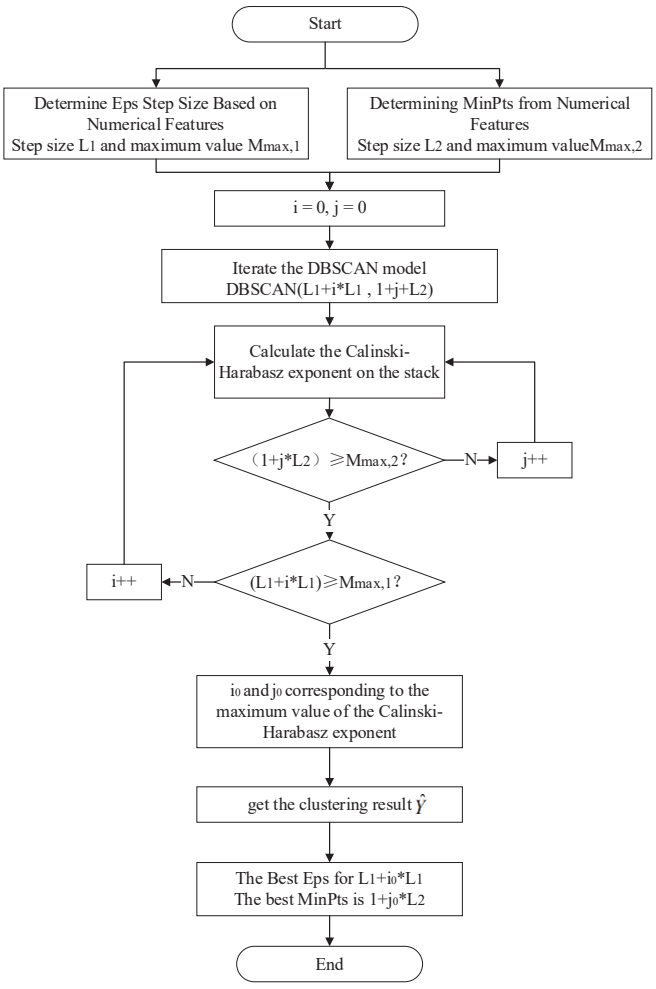

Fig. 2. AD-DBSCAN clustering algorithm process

The model of the AD-DBSCAN algorithm is represented as:

$$\theta = Eps, MinPts \tag{4}$$

Where, θ denotes the two parameters of neighborhood radius Eps and minimum neighborhood points MinPts.

$$\hat{Y} = DBSCAN_{\theta}(X) \tag{5}$$

where, $\hat{Y}$ denotes the clustering result derived from the DBSCAN algorithm.

$$\arg\max_{\theta} = CH(Y, \hat{Y}) \tag{6}$$

where, argmax denotes the traversal parameter to select the maximum value of the Calinski-Harabasz index. Y is the label.

It can be seen from 2.2 that the larger the Colinski-Harabasz index value is, the better the clustering effect of the DBSCAN clustering algorithm is. In this paper, for different input parameters, the larger the Calinski-Harabasz index value of the model, the better the clustering effect. Based on this principle, we propose the AD-DBSCAN clustering algorithm. First the data features determine the step size L_1 and the maximum value $M_{max,1}$ of Eps, $M_{max,1}$ is determined by the maximum value of the distance between any two nearest neighbors. Determine the step size L_2 and the maximum value $M_{max,2}$ of MinPts, $M_{max,2}$ is determined by the principle that all points are aggregated into one class when MinPts is greater than $M_{max,2}$. Construction of the initial DBSCAN model with an initial distance of the step L_1 and initialize the minimum number of points to 1; Iteration in distance steps and number of points steps until the Eps value and MinPts value parameters reach a predetermined maximum value. The optimal Eps and MinPts values are determined based on the maximum value of the calculated Calinski-Harabasz index, which is the AD-DBSCAN algorithm proposed in this paper, and it can be seen from Fig.2.

III. Results and Discussion

A. Experimental data

The experimental data in this paper uses the US National Traffic Accident Dataset. The dataset is a traffic accident dataset from February 2016, covering 49 states in the United States. To facilitate the experiment, Snapde is used in this paper to extract data from the original data set, and a data set of traffic accidents in Dallas, U.S.A., in February 2016 with a data sample size of 3,000 items was extracted, as shown in Fig.3. The first 2700 data are used as the training set and the last 300 data are used as the testing set.

Fig. 3. Traffic accident site in Dallas

B. Experimental results of the original DBSCAN algorithm

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

In order to more intuitively analyze the clustering results of various types of accident causes , we divide the data set into five categories of traffic accident causes, named "presence of blind spots", "ghost probes", "resting with foot off pedal", "following overtaking", and "overtaking by changing lanes left and right". The original DBSCAN algorithm is used to cluster the latitude and longitude points of the data set, and two parameters are set uniformly without any conditions: the minimum number of clustering points MinPts for clustering clusters is 5, and the neighborhood radius Eps is 0.03. The specific values and the results of accident division are shown in Table 1

Fig. 4. "Follow overtaking" traffic accident type clustering results

From the clustering results of the original DBSCAN algorithm in Table 1, we can see that the clustering results of the causes of traffic accidents in the categories of "following overtaking" and "left-right lane change overtaking" are more successful. The higher of the CH value in the proposed algorithm, the more accurate of the clustering algorithm results. Among them, there are 26 categories of clustering results for the causes of "following overtaking" accidents, and 12.6% of the data points are marked as noise points, and the corresponding CH index is 87.58. The clustering results are shown in Fig.4, from which we can see that the data points are not circled with circles indicate noise points, and the other points inside the circles are intra-cluster points.The accident causes of "left-right lane change overtaking" can be clustered into 22 categories, of which the percentage of noise points is 12.7% and the corresponding Calinski-Harabasz index is 97.56. The geographical map of clusters is shown in Fig. 5.

As can be seen from the figure, accidents are concentrated in downtown Dallas, Dallas to Farmers Branch, Dallas to Richardson, Dallas to Dallas Executive Airport, and the

highway from Farmers Branch to Richardson. The Calinski-Harabasz index is relatively low for all five categories, the clustering effect of "blind spots", "following overtaking" and "left-right lane change overtaking" is poor, and the clustering effect of "ghost probe" and "resting with foot off pedal" is relatively better., but the overall clustering result is rough. Due to few data points or lack of accurate input parameters, the clustering results of "ghost probe" and "foot off the pedal rest" are not obvious, which mark most points as noise points.

TABLE I. CLUSTERING RESULTS OF THE ORIGINAL DBSCAN ALGORITH

Causes of traffic accidents	Data points	Eps	MinPts	Distance solving model	Number of clusters	Noise Ratio	CH Index
presence of blind spots	630	0.03	5	Euclidean metric	32	26.8%	36.26
ghost probes	349				17	53.3%	8.43
resting with foot off pedal	338				18	60.9%	7.56
following overtaking	705				26	12.6%	87.58
overtaking by changing lanes left and right	552				22	12.7%	97.56

Fig. 5. "Change lanes left and right to overtake" traffic accident type clustering results

of clustering points and the neighborhood radius are also the optimal values, and it can be seen from the figure that the AD-DBSCAN clustering algorithm with this parameter value has the best clustering effect.

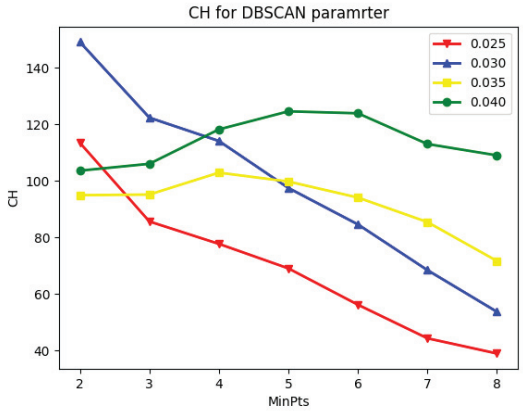

Fig. 6. Calinski-Harabasz Index

C. *Analysis of experimental results of adaptive parameter DBSCAN algorithm*

To verify the adaptability and accuracy of the proposed algorithm, the same data as the original DBSCAN algorithm is used in the experimental data of the proposed AD-DBSCAN clustering algorithm. By iterating the DBSCAN algorithm, the Calinski-Harabasz index is stacked, and then the maximum value is selected by observing the execution results of the algorithm, as shown in Fig. 6, from which it can be seen that the maximum value of the CH index is 137. The minimum number

D. *Analysis of experimental results of AD-DBSCAN clustering algorithm*

The AD-DBSCAN clustering algorithm was used to cluster the latitude and longitude of the accident location in Dallas, Texas, USA, in February 2016, and the Calinski-Harabasz index was used as the evaluation criterion of the clustering algorithm, and the clustering results are shown in Table 2.

TABLE II. AD-DBSCAN ALGORITHM CLUSTERING RESULTS

Causes of traffic accidents	Data points	Eps	MinPts	Number of clusters	Noise Ratio	CH Index
presence of blind spots	630	0.035	3	32	8.3%	82.82
ghost probes	349	0.04	2	46	8%	38.87
resting with foot off pedal	338	0.04	2	54	13.9%	24.89
following overtaking	705	0.03	2	55	1.8%	212.31
overtaking by changing lanes left and right	552	0.03	2	39	2.9%	149.16

As it can be seen from Table 2, the clustering effect using the AD-DBSCAN algorithm is greatly improved compared with the original clustering algorithm, and the best clusters are 55 ("following overtaking" traffic accident causes), and there is no

situation where a large number of data points are marked as noise points in the experimental results. Except from the two causes of traffic accidents, "foot off pedal rest" and "ghost probe", which have relatively few sample points among all features, the model scores above 80. Compared with the Calinski-Harabasz index of the original algorithm, the improvement in "blind spots" is about 43.78%, "ghost probes" is about 21.69%, "The improvement in "foot off pedal rest" was about 27.11%, 41.25% for "following overtaking", and 65.41% for "left-right lane change overtaking". The experimental results show when the model is applied to traffic accident geographic location clustering, the selection of parameters Eps and MinPts is more accurate, which makes both Calinski-Harabasz indices greater than the original algorithm results. The division of clusters is more accurate, and the identification of noise points in the data is more precise, which verifies the applicability of the algorithm in urban road traffic accident location clustering

IV. CONCLUSIONS

This paper mainly takes the spatial feature analysis of traffic accidents as the perspective, comprehensively analyzes the latitude and longitude of traffic accident locations, and solves the problems such as the difficulty of clustering accident spatial features. Finally, an AD-DBSCAN algorithm based on the Calinski-Harabaz index adaptively adjusting the input parameters of the original algorithm is proposed. In the experiment, clusters the accident data of different traffic accident causes in Dallas, Texas, USA, and compares it with the original DBSCAN clustering algorithm. The experimental results show that the mapping rate of accident locations reaches more than 98%, the improved algorithm performs more intelligently in parameter selection, the clustering effect is also more satisfactory, and the noise identification is more reasonable than the original algorithm. The scores of the Calinski-Harabaz index for the AD-DBSCAN algorithm when applying different amounts of data for clustering show that the effect of data size on the clustering effect is particularly significant. This paper analyzes a large amount of two-dimensional data, that is, only the clustering effect of accident locations (latitude and longitude) and accident causes are considered, and the geographic clustering analysis of multidimensional features of traffic accidents needs further study

ACKNOWLEDGMENT

This paper was supported by the Gansu Provincial Science and Technology Program (21ZD4GA028), Gansu Higher Education Innovation Fund Project (2021A-028), Gansu Provincial Science and Technology Program Funded Natural Science Fund Key Project (22JR5RA226), National Natural Science Foundation of China Project (62162040).

REFERENCES

[1] W.A. Shewhart, W.E. Deming. Statistical method from the viewpoint of quality control[M]. Courier Corporation, 1986.

[2] Y. Zhang,M. Roughan, C. Lund, et al. An information-theoretic approach to traffic matrix estimation[C]//Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications. 2003: 301-312.

[3] F. Yakar. Identification of accident-prone road sections by using relative frequency method[J]. Promet-Traffic&Transportation, 2015, 27(6): 539-547.

[4] Sunita Jahirabadkar，Parag Kulkarni.Algorithm to determine distance parameter in density-based clustering[J].Expert Systems with Applications，2014，41 : 2939-2946.

[5] WANG W T，WU YL，TANG C Y，et al.Adaptive density-based spatial clustering of applications with noise（DBSCAN）according to data[A].Proceedings of the 2015 International Conference on Machine Learning and Cybernetics[C].

[6] Ahmad M，Bakr，Nagia M，Ghanem，Mohamed A.Ismail.Efficient incremental density-based algorithm for clustering large datasets[J].Alexandria Engineering Journal，2015，54 : 1147-1154.

[7] K.Mahesh Kumar，A.Rama Mama Mohan Reddy.A fast DBSCAN clusteringalgorithm by accelerating neighbor searching using Groups method[J].Pattern Recognition，2016，1-10.

[8] Severino E.Galán.Comparative evaluation of region query strategies for DBSCAN clustering[J].Information Sciences，2019，502.

[9] Pan Wei,Zhou Xiaoying et al. Attribute Weighted Clustering Algorithm Based on Semisupervised K-means [J]. Computer Applications and Software, 2016, 34(3): 189-194.

[10] Wei Kangyuan, He Qing, Xu Qinshuai. K-means Clustering Based on Improved Gravity Search Algorithm [J]. Application Research of Computers, 2019, 36 (11):3240-3244.

[11] Cheng Jing, Liu Jiajun et al. Analysis of spatial-temporal characteristics of taxi travel volume in Beijing based on time series clustering method [J]. Journal of Geo-information Science, 2016, 18 (9):1227-1239.

[12] He Yue, Wang Chongchang. Research on Mining Taxi Passenger Hot Spot Based on Temporal and Spatial Clustering [J]. Geomatics and Spatial Information Technology, 2020, 43(1):99- 102.

[13] Liao Zhuhua, Zhang Jian et al. Taxi passenger carrying area recommendation based on coefficient trajectory data [J]. Acta Electronica Sinica, 2020, 11(11):2178-2185.

[14] Jiang Huijuan, Yu Yang. Improved DBSCAN Algorithm for Fine Extraction of Taxi Passenger Hot Spot [J]. Geospatial Information, 2017, 15(10):16-21.

[15] Liao Lvchao, Jiang Xinhua et al. Directed Density Method for Clustering Trajectional Data of Floating Vehicles [J]. Journal of Geo-information Science, 2015, 17(10):1152-1161.

[16] J. H. Kim, J. H. Choi, K. H. Yoo, et al. AA-DBSCAN: an approximate adaptive DBSCAN for finding clusters with varying densities[J]. The Journal of Supercomputing, 2019, 75(1): 142-169.

[17] M. R. Khan, M. B. Siddique, R. B. Arif, et al. ADBSCAN: adaptive density-based spatial clustering of applications with noise for identifying clusters with varying densities[C]//2018 4th international conference on electrical engineering and information & communication technology (iCEEiCT). IEEE, 2018: 107-111.

[18] A. Gupta, P. N. Merchant. Automated lane detection by k-means clustering: A machine learning approach[J]. Electronic Imaging, 2016, 2016(14): 1-6.

[19] J. Wang, D. K. Schreiber, N. Bailey, et al. The application of the OPTICS algorithm to cluster analysis in atom probe tomography data[J]. Microscopy and Microanalysis, 2019, 25(2): 338-348.

[20] S. Jahirabadkar, P. Kulkarni. Algorithm to determine ε-distance parameter in density based clustering[J]. Expert systems with applications, 2014, 41(6): 2939-2946