A Sensitive Trajectory Clustering Method Based on Spatiotemporal Density

Xinyi Wang, Yiwei Liao*, Lei Ji, Zhimin Zhang
School of Computer Science and Information Engineering, Harbin Normal University, Harbin, China
**Corresponding Author*

Abstract: Trajectory clustering method is a hot research topic in current trajectory data mining. Existing trajectory clustering methods have problems such as inflexible parameter selection, inaccurate clustering results, and inability to quickly obtain reliable clustering results. A sensitive trajectory clustering method based on spatiotemporal density is proposed to address this issue. Firstly, an ordering points to identify the clustering structure algorithm is used to identify the clustering structure to obtain a trajectory clustering decision map, and appropriate clustering parameters are selected. At the same time, considering the spatiotemporal properties of trajectories, a comprehensive analysis and clustering of the spatiotemporal trajectories of each participant were conducted using the Spatial Temporal-density-based Spatial Clustering of Applications with Noise method. The experimental results show that the proposed method can achieve good clustering results while maintaining good time overhead.

Keywords: Clustering Method; Spatiotemporal Trajectory; Trajectory Data Analysis

1. Introduction

According to the "2023 China Mobile Internet Annual Report" released by QuestMobile, the total number of mobile Internet users in China reached 1.227 billion by the end of 2023.The gradual increase of Internet users has led to a sharp increase in the number of Internet devices, and the extensive access of Internet devices has generated a large number of application data. How to find the general trajectory from these application data and mine information is a hot issue for current researchers [1].Cluster analysis is to divide data according to the degree of similarity to ensure the maximum distance between classes and the minimum distance within classes after clustering, so as to achieve the best clustering effect [2].Trajectory clustering is to comprehensively consider the similarity of a trajectory in time attributes and then space attributes to classify the trajectory, so as to realize the analysis of participants' behavior patterns and habits.Clustering methods can be roughly divided into four types, namely, division method, density method, hierarchy method and grid method [3].

In the partition-based clustering method, in order to achieve good clustering, the number of clusters should be considered first and the similarity function should be determined.For example, the common k-means clustering algorithm divides clusters by calculating the clustering of the initial center point and the non-initial center point, and calculates the average value of the clusters as the new center point, so that several different clusters are formed by iterating many times.To solve the problem of K-means clustering method's dependence on initial values, Liu Jing et al. [4] proposed an improved K-means algorithm based on Tukey rule and optimized initial center point selection, which divided the clustering process into two stages and effectively improved the clustering performance.Hierarchical clustering methods such as BIRCH clustering algorithm cluster by building a hierarchical structure between data points.Although this method does not need to determine the number of clusters in advance, it has high time complexity.Xia Xiaona [5] designed an improved BIRCH clustering algorithm, combining random walk model and BIRCH algorithm to design a clustering algorithm. The improved algorithm has obvious advantages in learning interactive activity clustering.Grid-based clustering methods, such as STING clustering algorithm, divide the space into data units, calculate the density and level of the units, identify the cluster center, and traverse the neighbor units to complete the clustering.Density-based clustering methods, such as DBSCAN algorithm, do not need to set the number of clusters in advance, it defines clusters as the maximum set of density-connected points, can divide regions with sufficient density into clusters, and can find clusters of arbitrary shape in noisy spatial data sets.Chen Xiaohui et al. [6] proposed an adaptive clustering algorithm based on DBSCAN, which used relevant statistical knowledge to obtain relevant knowledge among parameters, and proved the effectiveness of the method through experiments.

In order to solve the problems of presetting parameters and high time complexity existing in the above method.In order to solve the problem of inaccurate clustering results caused by inappropriate parameter selection, the proposed method adopts the sorting point algorithm used to identify the clustering structure and obtains the decision graph by taking into account the characteristics of the trajectory with time attributes.The optimal clustering parameter selection is obtained.In addition, the proposed method uses the spatiotemporal clustering algorithm based on spatiotemporal density noise to cluster tracks, which solves the problem that the clustering classification accuracy is reduced due to insufficient consideration of spatiotemporal attributes.

2. Related Concepts

This section mainly introduces the relevant models and algorithms that are needed: spatio-temporal trajectory model, sorting point algorithm for identifying cluster structure, and applied spatial clustering algorithm based on spatio-temporal density noise.

2.1 Space-Time Trajectory Model

Trajectory data set is a collection of trajectory data used to record the displacement of moving participants [7].Generally, trajectory data sets can be divided into spatio-temporal trajectory sets and spatial trajectory sets according to whether they contain time attributes [8].Spatio-temporal trajectory usually records information such as the location and direction of an object over a period of time, and is organized and recorded in a sequential manner.Space-time trajectory can be used to describe the movement behavior of various moving objects in geographical space, and can also be used to represent the evolution of natural phenomena and environmental changes in space-time dimension.The relevant definitions are as follows:

Define 1. Space-time trajectory points. Track points are indicated as in the following way $p_i = (time_i, lat_i, lng_i)$,among represents the timestamp of the sampling, lat_i, lng_i Represents the participants' latitude and longitude.

Definition 2. Space-temporal trajectory. The spatiotemporal trajectory is expressed as $T = \{p_1, p_2, \ldots, p_i, \ldots, p_n\}$ A series of sets consisting of time-stamp information trajectory points, where Represents the i-th point of this space-time trajectory.

Definition 3. Set of space-time tracks. The set of space-time trajectories is expressed as $TS = \{T_1, T_2, \ldots, T_i, \ldots, T_n\}$,Where Ti represents the i th trace and n represents the number of clusters.

2.2 Sorting Point Algorithm for Identifying Cluster Structure

the Ordering Points to Identify the Clustering Structure (OPTICS) algorithm is a density-based clustering algorithm [9].In order to solve the multi-dimensional problem, the researchers proposed Spatial Temporal Ordering Points to Identify the Clustering Structure, based on the OPTICS clustering algorithm.ST-OPTICS) This dimension can be analyzed differently depending on the actual situation, such as speed, time, etc.The clustering decision graph can be generated according to the reachable distance with a certain locus point as the core point, and the appropriate clustering parameters can be selected through the decision graph to ensure the better clustering effect.

2.3 Applied Spatial Clustering Algorithm Based on Spatiotemporal Density Noise

As one of the most representative algorithms of Density-Based Spatial Clustering of Applications with Noise, DBSCAN (Density-based Spatial Clustering of Applications with Noise, DBSCAN)Compared with Kmeans algorithm, DBSCSN does not need to pre-set the number of clusters in the data set and can find clusters of any shape.To solve the multidimensional clustering problem,Birant et al. [10] developed an applied Spatial Clustering algorithm ST-DBSCAN algorithm (Spatial Temporal-Density-Based spatial Clustering) based on the DBSCAN algorithmApplications with Noise, ST-DBSCAN).In this algorithm, the object is divided into core points, boundary points and noise points by considering both spatial and non-spatial values.

3. Design of Sensitive Trajectory Clustering Method Based on Space-Time Density

This section mainly introduces the specific steps of the system model and algorithm of the sensitive trajectory clustering method based on spatiotemporal density.

3.1 System Model

The trajectory data uploaded by the participants is a geographical coordinate system, which uses longitude and latitude to represent a unique geographical location on the Earth, ensuring the accuracy of geographic information.However, for the following privacy protection algorithm design and calculation requirements, it is necessary to combine the geographical information of each locus point of the participant to transform the geographical coordinate system.The proposed method first uses Mercator projection to project the geographic coordinates into a two-dimensional plane, and at the same time converts the time information of trajectory points into time stamps in order to more conveniently study the time attribute of trajectory.Secondly, in order to increase the accuracy of trajectory data, multiple trajectories of participants in a period of time are integrated to form a trajectory set composed of multiple trajectory points.In addition, in order to judge the characteristics of track points more clearly, this paper adopts ST-DBSCAN algorithm to cluster the track points. In order to make up for the poor clustering effect caused by the difficulty in determining the threshold value of ST-DBSCAN clustering, ST-OPTICS clustering method is adopted to cluster the track points to obtain the decision graph before ST-DBSCAN clustering.Select the appropriate clustering threshold and optimize the clustering result.As shown in Figure 1.

3.2 Sensitive Trajectory Clustering Algorithm Based on Space-Time Density

Inspired by the density-based clustering algorithm ST-DBSCAN clustering algorithm, this method follows the idea of the algorithm and judges the trajectory sensitive region combined with the time attribute of the trajectory. Meanwhile, considering that the spatial-temporal density-based sensitive trajectory clustering algorithm is greatly affected by the neighborhood distance threshold, that is, the trajectory clustering will be too sparse if a large neighborhood distance threshold is assigned.Assigning a small neighborhood distance threshold results in the problem of dividing a locus cluster into two locus clusters.A time-based ST-OPTICS decision graph is obtained by introducing the OPTICS clustering algorithm and considering the time attribute, and an appropriate distance threshold is selected.The sensitive trajectory clustering algorithm based on space-time density is as follows:

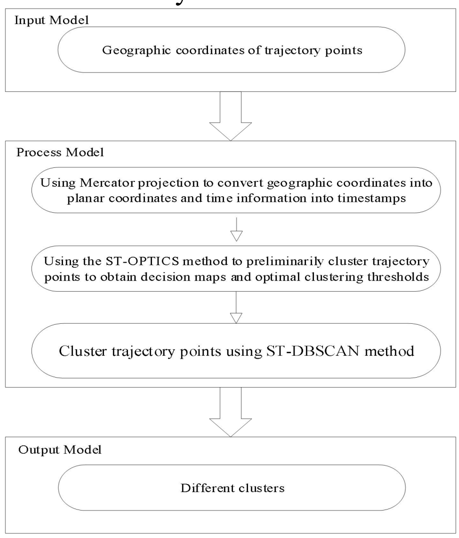

Figure 1. System Model of the Sensitive Trajectory Clustering Method Based on Spatiotemporal Density

Since the expansion process of density-based clustering method is arbitrary, it starts from any core point and does not expand from the densest place, the method first determines the density of the periphery of each trajectory point, and then performs clustering.

Firstly, taking into account the spatiotemporal characteristics of the trajectory based on the clustering algorithm, the decision graph of the cluster is obtained by randomly inputting the distance threshold eps and the given time threshold and density threshold. Then, the clustering situation when the distance threshold is specified can be observed and the appropriate value can be selected.As shown in Table 1.

Then, the distance threshold, time threshold and density threshold are initialized, an unprocessed point is selected, and all neighborhood points within the distance threshold eps and time threshold are found, and its track points are judged to be greater than the density threshold, and the core point set is added.From each locus point in the core point set, it is judged whether all neighborhood points that meet the distance threshold and time threshold are greater than the density threshold. If they meet the threshold, the points in their neighborhood are also added to the core point set.In this recursive execution method, a set of sensitive regions with similar density is developed.As shown in Table 2.

Table 1. Time-Based St-Optics Clustering Algorithm

Algorithm 1. Time-based ST-OPTICS clustering algorithm
Input: a collection of trace points TS,Time threshold minT, Density threshold minPts Output: Reachable clustering table reach_lists 1: Create a collection of core track points CPoint 2: Traverse the collection of tracks TS, Find all core points to the core trajectory point collection CPoint 3:if $\mathrm{CPoint} \neq \emptyset$ then 4:Walking through the CPoint set, randomly selecting an unprocessed core point P, calculate the accessible distance size of the core point P with all untreated trajectory points within the eps neighborhood and minT neighborhood, putting it in turn into the seed set Seet 5:Tragothrough the Seed Colleccollection 6: if $\mathrm{seed} \neq \emptyset$ then 7:Pick the nearest track point T 8:if T is the core point then 9:All unvisited trace points in T were added to the seed set Seed to recalate the reachable distance 10: end if 11: end if 12:end if 13:return reach_lists

Table 2. Sensitive Trajectory Clustering Algorithm Based on Spatiotemporal Density

Algorithm 2. Sensitive trajectory clustering algorithm based on spatiotemporal density
Input: trajectory point set TS, time threshold minT, density threshold minPts Output: the Clusters collection 1: Call the ST-OPTICS method to select the appropriate distance threshold eps 2: Mark all track points as UNprogress 3:do 4:The trace point T of an UNprogress trajectory is randomly selected in the trajectory set TS to mark the bit progress 5:if then 6: Add the trajectory point T to the Clustersi 7: $\mathrm{N} = \mathrm{number}(\mathrm{T_i.time} - \mathrm{T.time} < \mathrm{minminT}\ \&\&\ \mathrm{T_i.dis} - \mathrm{T.dis} < \mathrm{eps})$ 8: for N In each point 9: if then 10: $\mathrm{T}' = \mathrm{progress}$ 11: if then 12: $\mathrm{N} = \mathrm{N} + \mathrm{number}(\mathrm{T}'_\mathrm{i}.\mathrm{time} - \mathrm{T}'.\mathrm{t}\ \mathrm{T}'.\mathrm{time} < \mathrm{minT}\ \&\&\ \mathrm{T}'_\mathrm{i}.\mathrm{dis} - \mathrm{T}'.\mathrm{dis} < \mathrm{eps})$ 13: Add the trajectory point T'to $\mathrm{Clusters_i}$ 14: end if 15: end if 16: end for 17:else T is the outlier track point, looking for the next UNprogress track h track point 18:return Clusters gather

4. Experimental Analysis

This section focuses on the experimental analysis of the spatiotemporal density-based sensitive trajectory clustering algorithm, including the feasibility and time cost analysis of the algorithm.

4.1 Feasibility Analysis of the Algorithm

In order to avoid excessive cluster allocation, sparse regions may be mistakenly classified as hot spots, and too little cluster allocation may cause some hot spots to be misjudged as non-hot spots.ST-OPTICS decision plots are therefore used to select the appropriate search radius eps.Then the spatial and temporal density sensitive locus clustering algorithm is used for clustering.Through the comparison between FIG. 2 and FIG. 3, it can be clearly seen that the clustering method proposed in this paper is more accurate through comprehensive analysis of space-time attributes of trajectories.As shown in Figure 2, this method, by judging the trajectory density and comprehensively considering the temporal and spatial attributes of the trajectory, means that not only the spatial distance between two trajectory points should be close enough, but also the time between them should be close enough. This method will effectively avoid misjudging a large number of trajectory points as clusters and improve the clustering determination accuracy.

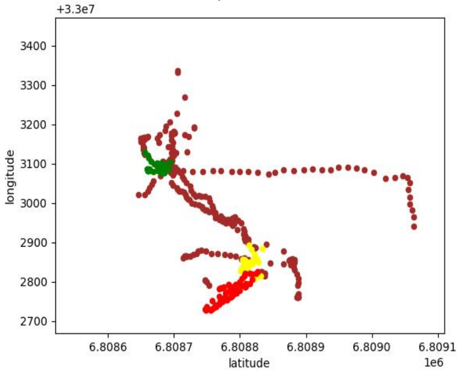

Figure 2. Cluster Plot of Sensitive Trajectories Based on Spatiotemporal Density

However, using the trajectory set of the same participants, Figure 3 of sensitive trajectory clustering based on density judged that the cluster is less accurate than Figure 2, which reflects the good effect of sensitive trajectory clustering based on spatiotemporal density.

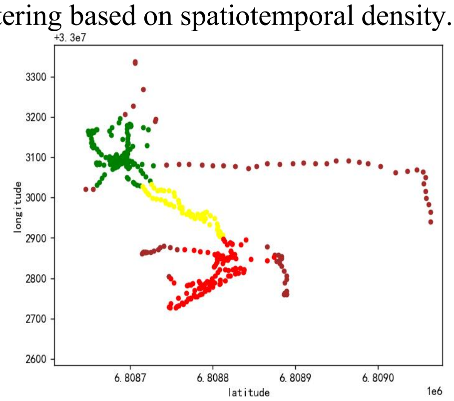

Figure 3. A Density-Based Clustering Plot of the Sensitive Trajectory

4.2 Time Cost Analysis of the Algorithm

Theoretically, the spatiotemporal density-based sensitive trajectory clustering method mainly combines ST-OPTICS algorithm and ST-DBSCN algorithm to determine the sensitive region of the participant's trajectory.Therefore, the time complexity of ST-OPTICS algorithm and ST-DBSCAN algorithm is analyzed to calculate the time cost of spatiotemporal density-based sensitive locus clustering method.

First, the time complexity of ST-OPTICS algorithm is analyzed. Assuming that a trajectory data set containing n data points is input, it can be seen by ST-OPTICS algorithm that the spatial distance and time distance between each pair of trajectory points need to be calculated. According to the time complexity of Euclidean distance, it is required at most $O(n^2)$ Time complexity.Then, the time complexity of ST-DBSCAN algorithm is analyzed. ST-DBSCAN algorithm mainly includes core point recognition, density reachable point calculation and noise point judgment. Therefore, the time complexity of the three methods needs to be analyzed respectively.

(1) Core point identification:

First, the algorithm needs to calculate the distance between trajectory points.At most, this step is needed $O(n^2)$ Time complexity.Then, check whether each track point has enough track points in its neighborhood to determine whether it is a core track point.This process has the most time complexity $O(n^2)$.

(2) Calculation of density reachable points:

For each core point, its clustering needs to be extended by density reachability.You need to go through all the points in its neighborhood to check if they are reachable in density.In the worst case, if all the data points are core points and are density-accessible to each other, then the time complexity may be reached $O(n^2)$,However, because the ST-OPTICS clustering algorithm optimizes the threshold selection of ST-DBSCAN, the time complexity is much less than that $O(n^2)$.

(3) Noise point judgment:

Finally, you need to process the noise points, assign them to the corresponding cluster or label them as noise.The time complexity of this step is usually low, independent of the size of the data set and the clustering structure.

Therefore, the time complexity of ST-DBSCAN algorithm can not exceed at most when identifying core points and calculating density reachable points $O(n^2)$,So the time complexity of the ST-DBSCAN algorithm is less than the$O(n^2)$.

In summary, the spatiotemporal density-based sensitive trajectory clustering method is performed sequentially by ST-OPTICS algorithm and ST-DBSCAN algorithm. In general, the time complexity of both algorithms is equal to the output of one algorithm serves as an input to another, then their joint time complexity remains$O(n^2)$.Since the adoption of the ST-OPTICS algorithm optimizes the choice of the ST-DBSCAN algorithm threshold, the sensitive trajectory clustering method based on spatio-temporal density has less time complexity than the $O(n^2)$,Have a good time overhead.

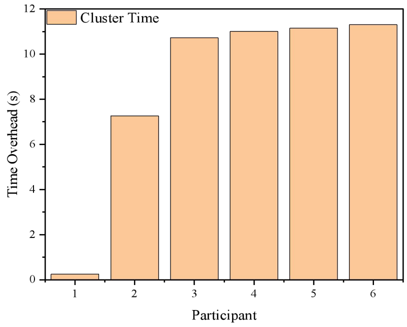

Figure 4. Time Overhead Diagram

Figure 4 shows the time cost diagram of sensitive trajectory clustering based on spatiotemporal density.As can be seen from the graph, as the number of participants increases, the overall time expenditure gradually increases, but finally levels off.This is because the proposed method adopts an improved spatiotemporal density-based noise applied spatial clustering method, which reduces the time complexity compared with k-means clustering method.

5. Conclusion

The proposed method takes into account the spatiotemporal properties of trajectories and the difficulty of parameter selection faced by traditional clustering methods.By combining the spatiotemporal sorting point algorithm used to identify the cluster structure with the spatiotemporal density noise based applied spatial clustering algorithm, the selection of clustering parameters is optimized, and good clustering effect is obtained.Because the proposed method can analyze the behavior characteristics of a certain participant, it can be used in the study of participant trajectory analysis and privacy protection.

References

[1] MOU Naixia, Xu Yujing, Zhang Hengcai et al.A review of moving trajectory clustering methods [J]. Bulletin of Surveying and Mapping,2018,(01):1-7.DOI:10.13474/j.cnki.11-2246.2018.0001.

[2] LIU Jingyi, Peng Ju, Tang Jianbo, et al.Adaptive clustering method of trajectory data with multi-feature fusion [J]. Journal of Geo-Information Science,2023,25(07):1363-1377.

[3] Ji Qiang, Sun Yanfeng, Hu Yongli et al.Review on deep clustering algorithms [J]. Journal of Beijing University of Technology,2021,47(08):912-924. (in Chinese)

[4] LIU Jing, Qiu Ziying, Guo Maozu et al.K-means clustering Improvement Algorithm based on Tukey rule and initial center point optimization [J].Data Acquisition and Processing, 2019,38(03):643-651.DOI:10.16337/j.1004-9037.2023.03.013.

[5] Xia X N. Random Walk and BIRCH fusion clustering statistical method [J].Statistics and decision, 2021, 5 (23) : 5-9. DOI: 10.13546 / j.carol carroll nki tjyjc. 2021.23.001.

[6] Chen Xiaohui, Xi Qinggang.Adaptive clustering algorithm based on DBSCAN is the research and implementation of [J]. Journal of huaiyin normal university (natural science edition), 2021, 20 (3) : 228-234. The DOI: 10.16119 / j.carol carroll nki issn1671-6876.2021.03.007.

[7] Huo Zheng, Meng Xiaofeng, Research on trajectory privacy protection technology, Journal of Computer Science, 2011, 34 (10) : 1820-1830.

[8] Qin Kun, WANG Yulong, ZHAO Pengxiang et al.Spatiotemporal clustering and analysis of behavior trajectory [J]. Nature Journal,2018,40(03):177-182. (in Chinese)

[9]Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure[J]. ACM Sigmod record, 1999, 28(2): 49-60.

[10]Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial – temporal data[J]. Data & knowledge engineering, 2007, 60(1): 208-221.

