Universe Scientific Publishing

Research and Application of Trajectory Stop Point Detection Algorithm for Time Series Clustering

Yuqi Wang

Hangzhou Normal University, School of Mathematics Hangzhou, Hangzhou 311121, China.

Abstract: In order to solve the problem of low accuracy of sampling irregular tracks, a time series clustering algorithm for detecting stops is proposed. Firstly, based on the data field theory, a hybrid feature density detection method considering temporal and spatial characteristics is designed. Secondly, according to the characteristic that the center density of the stop point is greater than the inlet density, the filtering and refining strategy is used to extract the stop point. In the filtration stage, the time duration and the minimum density threshold are selected as the candidate residence points. The maximum threshold is used to identify the actual residence point in the refining stage. The experimental results show that the proposed method can effectively detect the residence points on the irregular trajectories with higher accuracy and less time consumption than the existing methods.

Keywords: Time Series Clustering; Trajectory Stopping Point; Detection Algorithm

Introduction

With the wide application of positioning system and mobile terminal, it is more and more difficult to obtain trajectory data. The algorithm realizes the location and tracking of the stop point through spatial segmentation and analysis of the sampling curve. The user's behavior can be recorded in these trajectory data, which can reflect the user's behavior pattern and behavior law. However, when dealing with a large number of raw tracks, the efficiency of direct data mining is not satisfactory. Because the traditional trajectory clustering algorithm is very sensitive to noise, it is difficult to find hidden patterns behind the original data.

1. Concept and method

1.1 Basic Concepts

Definition 1: A trajectory is a collection of chronological location points numbered. Where,, and represent longitude, latitude, and time respectively.

Definition 2: is the locus point set table, which contains two stopping points, which represents the stopping state of the moving object at a certain position, and its value is

Definition 3: According to formula (1), we can calculate the spatial distance between trajectory points.

$$dist(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

1.2 Density function based on data field

Based on the idea of field theory, the interaction of matter particles and its field description method are introduced into the abstract number field space, and each object in the space is regarded as a particle with a certain mass, and described around a spherical virtual data field. In this field, each point has its corresponding energy value, and the coordinate corresponding to the point is the magnitude value of each physical quantity at the point, that is, the potential of its position represents the magnitude and direction of a variable at the point [1]. Similar to the vector intensity function and scalar potential number description of a physical field, the potential function is defined as a vector form describing a physical quantity. As shown in equation (2).

$$\varphi(x)=\sum_{i=n}^{n}\left(m_i\times e^{-\left(\|x-x_i\|/\sigma\right)^2}\right) \tag{2}$$

In formula (2) : the distance required for the target to reach the field point; The quality factors that control the interaction between objects are called impact factors.

In the trajectory, the trajectory points of the residence area show a highly dense distribution in time and space, and the interaction between these trajectory points is very obvious, resulting in a relatively high value of the potential function. Therefore, a density calculation method based on data field is proposed in this paper. Each locus point is regarded as a particle with a certain mass, and there is a virtual field around it, and the potential function is used to measure the density of the locus point. The greater the potential value, the denser the locus point and the higher the probability of stopping point. In order to ensure the accuracy of the results, appropriate methods must be adopted for processing. By directly using the potential function in equation (2) to calculate the density to distinguish between stops and moving points, we ignore the effect of the time interval on the detection of stops. In fact, the sampling interval of most trajectories is indeterminate. Because some objects move around the site in each measurement, it is impossible to accurately determine whether there is a retention point at the current position when the distance change is small. For the above situation. In this paper, we introduce a new distance measurement method, the time interval of an object's spot. The time interval of the object field point is mapped to 0~1 to represent the quality of the track point, which is equivalent to setting a time weight for each object to control the interaction between the track points, so as to realize the differential influence of different time intervals on the track density. When a user arrives at a specified location, it can accurately estimate which objects exist in the target area according to the distribution characteristics of all tracks near the location. Within a given period of time, the position coordinates of each distance measuring device are determined according to the set target time [2]. Locate the normal range of the device recording interval, that is, the interval with the highest frequency. Through this method, a set of better motion estimation results can be obtained, thus improving the tracking accuracy. The calculation method is shown in equation (3)

$$CT(j,x)=\begin{cases}\dfrac{t_{j-1}-t_j}{t_{re}^{f}} & \text{if} j<x\\[2ex] \dfrac{t_{j+1}-t_j}{t_{re}^{f}} & \text{if} j>x\end{cases} \tag{3}$$

In formula (3), it represents the position of the center point of the track segment; Represents the time required for track points; Record the time interval for the device.

Considering that the stopping points in the trajectory show a clustering distribution, and the short-range field is more conducive to revealing the characteristics of the data distribution, this paper adopts Gaussian function to define the potential function, that is, the density function. Using potential function and distance formula to discretize the data, a new parameter, the mean cluster radius, is obtained to describe the clustering property. By processing the index displacement of objects and field points at the same time, the clustering of stop point distribution can be better demonstrated. By analyzing and calculating the change of the data in the moving process, a stop point set generation algorithm based on the density formula is obtained. The calculation formula of density is shown in equation (4).

$$\varphi(p_i)=\sum_{j=low}^{up}\frac{1}{1+e^{-CT(j,i)}}\times e^{-\left(dist\left(p_i,p_j\right)/\sigma d\right)^2-\left(|i-j|/\sigma h\right)^2} \tag{4}$$

Where: is the ratio of standard deviation to distance; When the value of the objective function is given, its range can be determined as needed, that is, whether there is one or more optimal locations. Is the index standard deviation; In the search window, the positions represented by low and up are the lower and upper bounds respectively, which are calculated as shown in formula (5) and (6). The distance between the center and the center has a great influence on the density calculation, and the influence of the exceeded trace points on the density calculation is negligible.

$$low = \max\left(0, i - 3\sigma_h\right) \tag{5}$$

$$up = \min\left(n, i + 3\sigma_h\right) \tag{6}$$

When the distance standard deviation value is too large, it will cause the moving track point to be incorrectly identified as the stop point, and when the value is too small, the stop point position will become blurred and difficult to find. Therefore, the sensitivity of the index to changes at different scales in time and direction needs to be taken into account when determining the position. In order to find the best value, this paper uses the inflection point as the estimated value, which is the point with the greatest curvature and usually represents the inflection point between two states. After analyzing and comparing the existing algorithms, an improved method is proposed to determine the threshold value and estimate the optimal segmentation interval. Firstly, the spatial distance between adjacent trajectory points is derived and arranged in order from large to small. Secondly, according to the time interval of two different moments, two straight line fitting and cubic curve fitting are done for each key point respectively, and the curve intersection point is determined as the inflection point. Because of the high fluctuation frequency of the distance value in practical application, its smoothness is not ideal, which brings considerable challenges to the calculation of the inflection point. Then, the maximum-minimum normalization technique is used to normalize the sequence, and the point with the most significant curvature change is taken as the inflection point.

2. Stop point detection algorithm based on time series clustering

By calculating the density of trajectory points, it can be inferred that each trajectory point is likely to become a stopping point. When the distance between any two points in a given area is less than a certain threshold, there will be multiple stops in the area, and these stops may contain important information such as moving objects or pedestrians. However, considering that the density value at the entrance of the stopping point is similar to the density value at the moving point, a single breaking value is difficult to accurately distinguish the two. In order to solve this problem, it is necessary to record and analyze multiple trajectories in a certain period of time to obtain the number and distribution of stopping points in each area. In order to reduce unnecessary redundant operations and improve time efficiency, a new statistical calculation method can be used to determine the threshold and determine whether there are isolated aggregation points. A single trace ST is adopted, with the minimum threshold min_threshold and the maximum threshold max_threshold as the input. The same statistical rules apply to all trajectory points in more than two consecutive time intervals. candidateidx and candidate_den are used to represent the index and density of the current cluster respectively. The core process of candidateIDX and CANDIdate_DEN is to calculate the density of each track point on the timeline and set the minimum density value to include the track points that meet the conditions into the candidate stop points. If the conditions are not met, the maximum density max_threshold of the current cluster needs to be evaluated. If the value is greater than max_threshold, the current cluster needs to be marked as the actual stop point [3].

3. Experiment and analysis

In the process of experiment, a variety of real data sets with labels, which have different characteristics, are used to carry out simulation experiments. See Table 1 for details.

name	length	stops	moves	mode
Shopping trip	1 968	195	1 773	foot ,car
house visit	1 128	285	843	foot ,car
hike	9 524	932	8 592	foot , car

short walk	120	37	93	foot
ferry	7 612	2 651	4 961	foot , ferry
dog walk	833	98	735	foot

Conclusion

In this paper, a hybrid feature density measurement method based on spatiotemporal characteristics is proposed to deal with various noises in the real trajectory. Based on the analysis of space and time series, the mathematical model of the relationship between regions in different time periods is established, and it is used as a hybrid feature extraction method. Aiming at the characteristic that the density of the entrance of the stop point is lower than the density of the center in a trajectory, we propose a new stop point detection algorithm based on time series clustering. On this basis, the candidate residence points are preliminatively determined by combining regional growth method and spatial neighborhood information. A double clustering strategy is used to identify the candidate stops, and then select the actual stops from these stops. This method can eliminate the false stagnation points effectively and calculate the real stagnation points accurately.

References

[1] Yang QY, Weng XQ. Time series clustering based on LLE and Gaussian Mixture model [J]. Computer Technology and Development,202,32(8):33-41. (in Chinese)

[2] Li HL, Lin CP. Research results keyword analysis method based on time series clustering [J]. Science Research Management,202,43(1):176-183.

[3] Sun XL, Zhu CC, Li MN, Liang TT, Ma CQ. Research on passenger flow prediction of urban rail transit under time series clustering [J]. Railway Transportation and Economics,2023,45(3):149-157. (in Chinese)