Assessing the quality of trajectory data is an important preprocessing procedure in many data-driven research as it can significantly influence the knowledge derived such as travel time, traffic flow and route choice, etc. Designing an effective trajectory quality assessment method to handle different kinds of noise and outliers in trajectory data is a challenge. To alleviate this problem, this paper develops a trajectory scoring algorithm by firstly extracting different kinds of movement features then evaluating the quality of each point based on the local feature stability. Preliminary results on a real-world taxi GPS dataset are reported.

Keywords Quality assessment, Trajectory data cleaning, Feature evaluation

1 Introduction

The wide deployment of positioning sensors in vehicles makes it possible to derive various types of knowledge from GPS data such as travel time [1], traffic flow [2] and route choice [3], etc. The raw GPS observations are generally inaccurate due to sensor noise and other factors such as poor signals in urban canyon. Instead of developing a robust algorithm to handling noisy observations in knowledge discovery step, data cleaning in the preprocessing stage is more cost effective. Compared with conventional data cleaning, assessing the quality of trajectory data can not only be used for detecting outliers but also provide confidence for the knowledge derived. However, it is a challenge to design an effective and robust method to quantitatively assess the quality of trajectory in an unbiased way for both normal observations and various types of outliers.

To alleviate this problem, a trajectory scoring algorithm is developed in this paper, which is based on the observation that trajectories with high quality are likely to exhibit smooth movement features. Specifically, the algorithm firstly extracts four movement features including turning angle, length, duration and speed from a raw trajectory then assigning a score to each point based on the local stability of these features in a neighboring region. The algorithm is finally evaluated on real-world taxi GPS data.

2 Related work

GPS data quality assessment can be regarded as one type of data cleaning tasks. These relevant methods in the literature can be classified into three categories: filtering-based and rule-based and simplification-based [4].

The filtering-based approaches are designed primarily for correcting noisy observations, ranging from the simple mean filter to more complicated Kalman filter and particle filter [5]. The mean filter is ineffective in handling consecutive noisy observations whereas the latter two filters are computationally expensive in modeling complicated movement behavior in reality. Another limitation of filtering based approaches is that each observation is replaced with an estimation. In practice, it is more appropriate to directly keep those observations that are reliable.
Rule-based approaches firstly extract features such as speed and then empirically design rules to detect outliers, such as unrealistic speed or acceleration [4]. However, the potential of detecting more complicated outliers from these features and assessing the quality of all observations in a trajectory are not fully exploited.

Simplification-based approaches are largely devoted to compression of trajectory data by eliminating redundant or irrelevant movement information. A conventional line simplification algorithm is Douglas Peucker (DP) [7] which performs a top-down search to recursively divide a line and maintain the point furthest from a checked segment. A variant of DP is developed in [8] to compress trajectory data by integrating time into the distance measurement. In [9], minimum descriptive length principle (MDL) is adopted to find an optimal partitioning of a trajectory, which is solved approximately by considering characteristic points in a trajectory. In [10], a Paralleled Road-Network-Based Trajectory Compression (PRESS) framework is designed where the spatial path and temporal information of a trajectory are extracted and compressed separately. The spatial compression part needs to perform frequent subsequence mining, which could be computationally expensive. Three simplification approaches: incremental, sliding window and global are adopted for quality evaluation. An example with these features extracted from a single trajectory is displayed in Figure 1. It shows that the stabilities of these features provide useful insight into the quality of GPS observations.

The advantages of the proposed approach lie in several aspects. Instead of correcting the original observation, a score is assigned to each point in a trajectory, which maintains detailed movement information. The results can be potentially applied for quality evaluation of a single point, a sub-trajectory or the whole trajectory. Additionally, multiple features such as velocity and duration can be integrated in an extensible manner.

3 Methodology

3.1 Movement feature extraction

Let \( tr \) denote a trajectory stored as a sequence of points written as \( tr = \langle p_1, p_2, \cdots, p_N \rangle \) where \( p_i = (x_i, y_i, t_i) \) stores the coordinates and the timestamp of recording. The following point based local features can be extracted from \( tr \) including movement direction \( D \) (in radians), turning angle \( \alpha \) (in radians), distance \( L \), duration \( T \) and speed \( v \). Each of them is a feature array with the \( i \)-th element calculated as

\[
D[i] = \arctan2(y_{i+1} - y_i, x_{i+1} - x_i) \quad (1)
\]

\[
\alpha[i] = D[i+1] - D[i] \quad (2)
\]

\[
L[i] = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (3)
\]

\[
T[i] = t_{i+1} - t_i \quad (4)
\]

\[
v[i] = L[i]/T[i] \quad (5)
\]

The function \( \arctan2(y, x) \) in Equation (1) computes the angle in radians in the Euclidean plane, between the positive \( x \)-axis and the ray to \((x, y)\). Since movement direction is already recorded in turning angle, only the latter four features are adopted for quality evaluation. An example with these features extracted from a single trajectory is displayed in Figure 1. It shows that the stabilities of these features provide useful insight into the quality of GPS observations.

3.2 Movement feature stability evaluation

The feature stability evaluation step assigns a score in range of \([0, 1]\) to each point in a trajectory based on the stability of the extracted features. Let \( f[i : j] \) denote the sub-array of a feature array \( f \) from index \( i \) to \( j \). In the first step, the local standard deviation \( \sigma_f^l \) of feature \( f \) for point with index \( i \) with respect to a window size \( w \) can be calculated as

\[
\sigma_f^l[i] = \sigma(f[i - \lfloor 0.5w \rfloor : i + \lfloor 0.5w \rfloor + 1]) \quad (6)
\]

where \( \sigma( \cdot ) \) computes the standard deviation of the operand that is an array and \( \lfloor 0.5w \rfloor \) defines the index range from \( i \) to the border of the window. Given an expected standard deviation \( \sigma_f^e \) of feature \( f \), a preliminary score \( s_p \) is defined as

\[
s_p[i] = \min(\frac{\sigma_f^e}{\sigma_f^l[i]}, 1) \quad (7)
\]

where \( 0 \leq s_p[i] \leq 1 \). A larger standard deviation leads to a lower score. In reality, a large deviation may not represent a noisy observation. For instance, the moving direction generally exhibits a sharp change at road intersections as...
Figure 1: Illustration of features (right) extracted from a trajectory (left). The numbers labeled in the trajectory map represents the point index. The points highlighted in red color are outliers whose final quality score is smaller than 0.9.

Figure 2: Illustration of feature scoring based on stability displayed in Figure 1 (the segment around the point with index of 80). To handle this issue, a corrected score \( s_f \) is calculated to be the local maximum of \( s_p \) as

\[
s_f[i] = \max(s_p[i - \lfloor 0.5w \rfloor : i + \lceil 0.5w \rceil + 1])
\]

The principle is explained in Figure 2. A small score that is caused by a large \( \sigma_f \) at a sharp change point is likely to be overwritten by a higher score at its neighboring region. At the same time, outliers can be more accurately identified.

Finally, a score array \( s \) indicating the feature stability can be computed for \( tr \) as a weighted sum of the four feature score arrays:

\[
s = w_1s_\alpha + w_2s_L + w_3s_T + w_4s_v
\]

where \( w_i \) is the user-specified weight. For simplicity, the weights are denoted by \( w = (w_1, w_2, w_3, w_4) \).
Figure 3: Quality scores calculated for the trajectory in Figure 1 with configurations of \( w = 5 \), \( \sigma_{e^\alpha} = 3 \), \( \sigma_L = 600 \) meters, \( \sigma_T = 50 \) seconds, \( \sigma_v = 50 \) km/h and \( w = (0.6, 0.2, 0.1, 0.1) \).

Figure 4: More examples of trajectory quality scores

For demonstration purpose, the score for the same trajectory is presented in Figure 3. As can be observed, there are three segments that have a lower quality score than the others, which are around the point with index of 20, 60 and 90. The low quality mainly results from the unstable distribution of turning angles.

4 Preliminary results

The algorithm is evaluated using a public real-world taxi GPS dataset\(^1\) collected by DiDiChuXing, China. The configurations are empirically set as \( w = 5 \), \( \sigma_{e^\alpha} = 3 \), \( \sigma_L = 600 \) meters, \( \sigma_T = 50 \) seconds, \( \sigma_v = 50 \) km/h and \( w = (0.6, 0.2, 0.1, 0.1) \). Several trajectories with their quality scores are displayed in Figure 4, where the score provides a reasonable indicator of the trajectory quality.

---

\(^1\)The data can be downloaded from https://outreach.didichuxing.com/research/opendata/en/
From the scoring results, it is possible to detect outliers as points whose final score $s$ is smaller than a certain threshold $\theta$. Based on the score shown in Figure 3, outliers are detected with $\theta = 0.9$ and the result is highlighted in Figure 1.

5 Conclusion and future work

A trajectory scoring algorithm was developed in this paper to assess the quality of trajectory data. The algorithm firstly extracted four movement features from a trajectory then assigned a score to each point based on its local feature stability. Preliminary results on a real-world taxi GPS dataset were reported. Future work is planned in enhancing evaluation of the algorithm, investigating the spatial and temporal distribution of the quality scores and extending the algorithm to detect more complicated outliers.

Acknowledgements

The first author is funded by the Chinese Scholarship Council. The work is based on an internship taken by the first author in DiDiChuXing, which is also acknowledged.

References


