

以运动方向为主导的轨迹相似性度量方法

王海起, 翟文龙, 闫 滨, 费 涛, 李学伟, 陈海波, 李 建

(中国石油大学(华东)地球科学与技术学院, 山东 青岛 266580)

摘 要: 考虑到移动对象的行为趋势体现在行驶距离与前进方向上, 提出以运动方向为主导并兼顾形态距离特征的轨迹相似性度量方法。形态距离特征采用包围面积进行度量, 运动方向特征分别采用真实平均方向、线性平均方向、最长公共方向序列 3 种形式进行度量, 选取形态距离和运动方向的最优组合作为轨迹相似性度量的最终形式。北京市出租车 GPS 载客轨迹数据上的聚类应用结果表明, 该相似性度量方法能够有效区分移动对象的趋势方向, 且聚类准确率优于最长公共子序列方法。

关键词: GPS 轨迹; 方向角度; 包围面积; 最优组合; 相似性度量

开放科学(资源服务)标志码(OSID):



中文引用格式: 王海起, 翟文龙, 闫滨, 等. 以运动方向为主导的轨迹相似性度量方法[J]. 计算机工程, 2019, 45(11): 37-46.

英文引用格式: WANG Haiqi, ZHAI Wenlong, YAN Bin, et al. Motion direction dominated trajectory similarity measurement method[J]. Computer Engineering, 2019, 45(11): 37-46.

Motion Direction Dominated Trajectory Similarity Measurement Method

WANG Haiqi, ZHAI Wenlong, YAN Bin, FEI Tao, LI Xuewei, CHEN Haibo, LI Jian

(School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong 266580, China)

[Abstract] As the moving trend of an object is reflected by its driving distance and forward motion direction, this paper proposes a trajectory similarity measurement method based on motion direction. The features of motion direction are measured with actual average direction, linear average direction and longest common direction sequence. The method also considers the features of the shape and moving distance, which are measured with the area of regions bounded by trajectories. The optimal combination of the shape, moving distance and motion direction are taken as the final form of trajectory similarity measurement. The method is applied to GPS trajectory data clustering of occupied taxis in Beijing, and results show that the similarity measurement method can effectively distinguish moving trends and directions of objects with a higher clustering accuracy rate than the longest common subsequence method.

[Key words] GPS trajectory; direction angle; bounded area; optimal combination; similarity measurement

DOI: 10.19678/j.issn.1000-3428.0053022

0 概述

随着传感器、无线通信网络以及 GPS 定位等技术的快速发展, 各种基于位置的应用产生了海量轨迹数据^[1], 人们希望能够对这些海量数据进行分析, 发现潜在的分布特征和运动规律, 该需求促进了轨迹数据挖掘技术的产生与发展^[2]。

轨迹数据挖掘通常需要对相似性或距离进行定义, 即度量不同轨迹之间的相似或接近程度, 从而将

具有相似运动轨迹的个体划分成簇^[3]。例如, 轨迹分类通过计算当前轨迹与各类别的相似度来判定轨迹所属的类别; 在预测用户目的地时, 需要计算用户当前路径与历史路径的相似程度, 以此提供可能的目的地等。

国内外专家学者对轨迹相似性度量方法进行了大量的研究和探索, 对于两条轨迹之间的相似性度量, 按照两条轨迹的轨迹点匹配时的时间匹配条件从严到松的不同要求, 度量方法大体可分为 6 类^[4]:

基金项目: 国家自然科学基金(41471322); 山东省自然科学基金(ZR2012DM010)。

作者简介: 王海起(1972—), 男, 副教授, 主研方向为地理信息系统、轨迹数据挖掘; 翟文龙(通信作者)、闫 滨、费 涛、李学伟、陈海波、李 建, 硕士研究生。

收稿日期: 2018-10-30

修回日期: 2018-12-27

E-mail: zw1931025@qq.com

第 1 类是时间全区间相似性度量方法,前提是两条轨迹的轨迹点数必须相同,且对应轨迹点所处的时刻也相同,这类方法主要有欧氏距离^[5]和最小外包矩形距离^[6]等;第 2 类是全区间变换对应相似性度量方法,该类方法在第 1 类方法的基础上,放松了匹配的点时刻须完全相同的限制,这类方法的代表是动态时间规整(Dynamic Time Warping, DTW)距离^[7-8];第 3 类是多子区间对应相似方法,该类方法不要求对两条轨迹的所有轨迹点进行匹配,而是寻找不重叠的多个相似子区间,并将区间之间的相似性汇总成轨迹间的相似度,此类方法能发现局部相似的时空轨迹,其中最长公共序列^[9-11]和编辑距离^[12-13]是比较常用的方法;第 4 类方法仅寻找两条轨迹的最大相似子区间,用它来度量轨迹之间的相似性^[4],这类方法主要有子轨迹聚类^[14-15]、时间聚焦聚类^[16]、移动微聚类^[17]和移动聚类^[18]等;第 5 类方法是单点对应相似方法,该类方法是用某一匹配的点对的相似性代替轨迹之间的相似性,其中历史最近距离^[19]和 Frechet 距离^[2]是最主要的两种方法;第 6 类是无时间区间对应相似性度量方法,这类方法仅考虑空间位置的相似性,比如单向距离方法^[20]和特征提取方法^[21]等。

以上方法在相似性度量时大多是以时空距离为度量手段,很少从运动方向角度刻画轨迹的相似性。考虑到移动对象的行为趋势不仅体现在行驶距离上还体现在前进方向上,本文提出以运动方向为主导并兼顾形态距离特征的轨迹相似性度量方法,分别采用不同的指标对距离、方向特征进行度量,依据性能评价选取距离和方向的最优组合作为轨迹相似性度量的最终形式,并进行实际案例分析。

1 基于距离与方向特征的轨迹相似性度量

以运动方向为主导的度量方法采用方向特征与形态距离特征相结合的方式相似性度量。两条轨迹的方向特征相似性采用真实平均方向夹角、线性平均方向夹角、最长公共方向序列 3 种度量形式。轨迹的真实平均方向是该轨迹所有相邻时刻轨迹点之间真实方向的平均值,两条轨迹的真实平均方向夹角越小表示其方向相似性越高;轨迹的线性平均方向描述的是该轨迹的线性趋势方向,两条轨迹的线性平均方向夹角越小表示其方向相似性越高;轨迹的方向序列是基于轨迹相邻时刻轨迹点之间真实方向形成的方向序列,两条轨迹的最长公共方向序列是在两条轨迹中查找具有相同方向的最长且可不连续的方向子序列,最长公共方向序列越长表示两条轨迹的相似性越高。两条轨迹的形态距离特性相似性则采用两条轨迹的包围面积进行度量,包围面积越小表示其相似性越高。将形态距离特征与 3 种方向特征分别进行组合,通过评价多个测试数据的

聚类性能,从而选择最优的组合方式,并将该最优组合与其他相似性度量方法的聚类效果进行对比分析。下面分别阐述形态距离度量和运动方向度量方法。设轨迹 $Tr_i = \{(x_1^i, y_1^i, t_1^i), (x_2^i, y_2^i, t_2^i), \dots, (x_k^i, y_k^i, t_k^i), \dots, (x_n^i, y_n^i, t_n^i)\}$, 其中, (x_k^i, y_k^i, t_k^i) 表示第 i 条轨迹中的第 k 个轨迹点, (x_k^i, y_k^i) 与 t_k^i 分别表示该轨迹点的位置与所处时刻, n 为轨迹 Tr_i 的轨迹点总数。

1.1 形态距离度量方法

本文采用包围面积作为形态距离特征的相似性度量方法,包围面积直接使用两条轨迹 Tr_i 与 Tr_j 围成的面积 $S_{i,j}$ 来度量序列间的相似度, $S_{i,j}$ 越小两条轨迹越相似,接近程度越高,包围面积是一种比较直观的度量方法,如图 1 所示,灰色部分面积为轨迹 Tr_1 与 Tr_2 的包围面积 $S_{1,2}$ 。本文采用扫描线算法对两条轨迹 Tr_i 与 Tr_j 围成的多边形进行填充,从而得到包围面积。

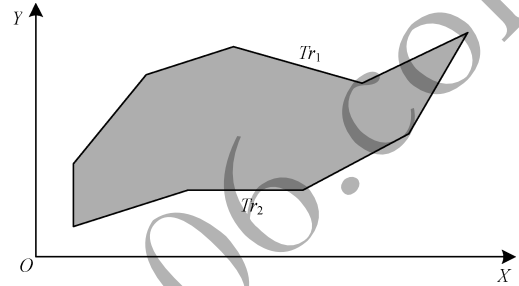


图 1 包围面积示意图

1.2 运动方向度量方法

1.2.1 真实平均方向夹角

单条轨迹的真实平均方向是该轨迹所有相邻时刻轨迹点形成的轨迹段真实方向的平均值。对于轨迹 Tr_i , 设 θ_k^i 表示轨迹段 $\{(x_k^i, y_k^i, t_k^i), (x_{k+1}^i, y_{k+1}^i, t_{k+1}^i)\}$ 以正东方向为基准,按逆时针旋转得到的方向角(如图 2 所示), $k=1, 2, \dots, n-1$ 且 $\theta_k^i \in [0^\circ, 360^\circ]$, 则 θ_k^i 计算公式如下:

$$\theta_k^i = \arctan\left(\left|\frac{y_{k+1}^i - y_k^i}{x_{k+1}^i - x_k^i}\right|\right) \quad (1)$$

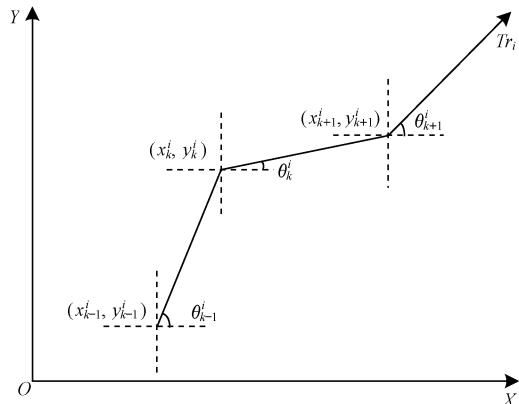


图 2 真实方向夹角示意图

θ_k^i 需根据以下不同情况进行调整:

$$\begin{cases} \theta_k^i \leftarrow \theta_k^i, x_{k+1}^i - x_k^i > 0 \text{ 且 } y_{k+1}^i - y_k^i \geq 0 \\ \theta_k^i \leftarrow 180^\circ - \theta_k^i, x_{k+1}^i - x_k^i < 0 \text{ 且 } y_{k+1}^i - y_k^i \geq 0 \\ \theta_k^i \leftarrow 180^\circ + \theta_k^i, x_{k+1}^i - x_k^i < 0 \text{ 且 } y_{k+1}^i - y_k^i < 0 \\ \theta_k^i \leftarrow 360^\circ - \theta_k^i, x_{k+1}^i - x_k^i > 0 \text{ 且 } y_{k+1}^i - y_k^i < 0 \end{cases} \quad (2)$$

轨迹 Tr_i 的真实平均方向角 θ_i 为各轨迹段方向角的平均值为:

$$\theta_i = (\sum_{k=1}^{n-1} \theta_k^i) / (n-1) \quad (3)$$

进一步地,两条轨迹 Tr_i 与 Tr_j 的真实平均方向夹角 $\theta_{i,j}$ 为:

$$\theta_{i,j} = |\theta_i - \theta_j| \quad (4)$$

1.2.2 线性平均方向夹角

线性平均方向用于描述一组线要素的趋势或平均方向。对于轨迹 Tr_i ,其线性平均方向 α_i 是所有轨迹段(每个轨迹段视为一个线要素)的趋势方向,计算公式如下:

$$\alpha_i = \arctan\left(\frac{\sum_{k=1}^{n-1} \sin \theta_k^i}{\sum_{k=1}^{n-1} \cos \theta_k^i}\right) \quad (5)$$

其中, θ_k^i 即轨迹段 $\{(x_k^i, y_k^i, t_k^i), (x_{k+1}^i, y_{k+1}^i, t_{k+1}^i)\}$ 的方向角,其含义与计算公式见 1.2.1 节, n 为轨迹点总数。 α_i 需根据以下不同情况进行调整:

$$\begin{cases} \alpha_i \leftarrow \alpha_i, \sum_{i=1}^{n-1} \sin \theta_k^i \geq 0 \text{ 且 } \sum_{i=1}^{n-1} \cos \theta_k^i > 0 \\ \alpha_i \leftarrow 180^\circ - \alpha_i, \sum_{i=1}^{n-1} \sin \theta_k^i \geq 0 \text{ 且 } \sum_{i=1}^{n-1} \cos \theta_k^i < 0 \\ \alpha_i \leftarrow 180^\circ + \alpha_i, \sum_{i=1}^{n-1} \sin \theta_k^i < 0 \text{ 且 } \sum_{i=1}^{n-1} \cos \theta_k^i < 0 \\ \alpha_i \leftarrow 360^\circ - \alpha_i, \sum_{i=1}^{n-1} \sin \theta_k^i < 0 \text{ 且 } \sum_{i=1}^{n-1} \cos \theta_k^i > 0 \end{cases} \quad (6)$$

两条轨迹 Tr_i 与 Tr_j 的线性平均方向夹角 $\alpha_{i,j}$ 为:

$$\alpha_{i,j} = |\alpha_i - \alpha_j| \quad (7)$$

1.2.3 最长公共方向序列

将 $[0^\circ, 360^\circ]$ 以 30° 间隔划分为不同的区间,并将各区间分别赋予不同的固定角度值(如表 1 所示),则轨迹 Tr_i 的各轨迹段方向角 θ_k^i 可根据其所属的区间转化为相应的固定角度,据此可获取轨迹 Tr_i 的方向序列 L_i 。

表 1 不同角度区间对应的固定角度值 ($^\circ$)

角度区间	固定角度值
$[0, 30)$	15
$[30, 60)$	45
$[60, 90)$	75
\vdots	\vdots
$[330, 360]$	345

对于轨迹 Tr_i 与 Tr_j 的方向序列 $L_i = \{L_1^i, L_2^i, \dots, L_n^i\}$, $L_j = \{L_1^j, L_2^j, \dots, L_m^j\}$, 使用最长公

共子序列(Longest Common Subsequence, LCSS)方法可获取最长公共方向序列个数 $l_{i,j}$, 其中 n, m 分别为方向序列 L_i, L_j 的方向值总数。

LCSS 用来查找两个字符串之间的最长公共子字符串,要求公共子字符串有序且连续,之后将字符串延伸为各种序列,查找两个序列之间的最长公共子序列,既可要求公共子序列有序且连续,也可要求公共子序列有序但不连续。本文使用 LCSS 方法在两条方向序列 L_i, L_j 中查找具有相同方向的最长且可不连续的方向子序列,进而得到 $l_{i,j}$, 其递归公式如下:

$$l(L_k^i, L_r^j) = \begin{cases} 0, k=0 \text{ or } r=0 \\ l(L_{k-1}^i, L_{r-1}^j) + 1, k, r > 0 \text{ and } L_k^i = L_r^j \\ \max(l(L_k^i, L_{r-1}^j), l(L_{k-1}^i, L_r^j)), \\ k, r > 0 \text{ and } L_k^i \neq L_r^j \end{cases} \quad (8)$$

其中, $k=0, 1, \dots, n, r=0, 1, \dots, m$ 。

例如, 设有 $L_k^i = \{15, 45, 105, 45, 135, 15, 475\}$, $L_r^j = \{45, 135, 105, 15, 45, 15\}$ 两个方向序列, 如图 3 所示, 使用 LCSS 递归方法得到最长公共方向序列个数 $l_{i,j}$ 为 4。

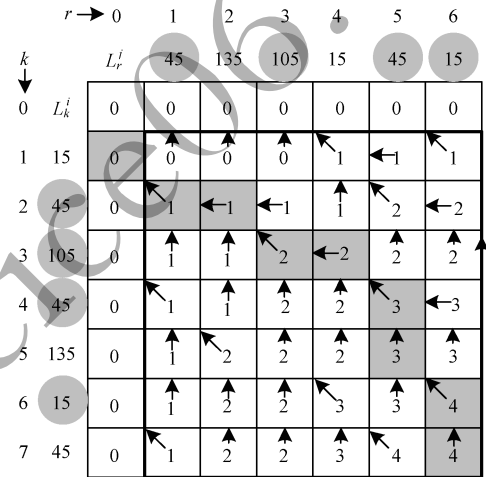


图 3 最长公共方向序列计算过程

1.3 相似性度量组合方法

设有 N 条轨迹, 对任两条轨迹 Tr_i, Tr_j ($i, j = 1, 2, \dots, N$) 分别计算包围面积 $S_{i,j}$ 、真实平均方向夹角 $\theta_{i,j}$ 、线性平均方向夹角 $\alpha_{i,j}$ 、最长公共方向序列 $l_{i,j}$, 其中, $S_{i,j}$ 值越小表示形态距离特征相似性越高, $\theta_{i,j}$ 和 $\alpha_{i,j}$ 值越小表示方向特征相似性越高, $l_{i,j}$ 值越大表示方向特征相似性越高。对于 N 条轨迹, 可分别形成包围面积矩阵 $S = [S_{i,j}]_{N \times N}$, 真实平均方向夹角矩阵 $\theta = [\theta_{i,j}]_{N \times N}$, 线性平均方向夹角矩阵 $\alpha = [\alpha_{i,j}]_{N \times N}$, 最长公共方向序列矩阵 $l = [l_{i,j}]_{N \times N}$, 且当 $i=j$ 时, $S_{i,i} = 0, \theta_{i,i} = 0, \alpha_{i,i} = 0, l_{i,i} = \text{len}(Tr_i)$, 其中, $\text{len}(Tr_i)$ 表示轨迹 Tr_i 的轨迹段个数。

对各矩阵分别进行规范化处理, 使 $S_{i,j}, \theta_{i,j}, \alpha_{i,j}, l_{i,j}$ 的判断准则保持一致, 即值越小, 两条轨迹的

相似性越高。

1) 包围面积的规范化处理, 规范化后的 S_{ij} 值在 $[0, 1]$ 范围内。

$$S'_{i,j} = \frac{S_{i,j} - \min_{i,j=1,2,\dots,N} (S_{ij})}{\max_{i,j=1,2,\dots,N} (S_{ij}) - \min_{i,j=1,2,\dots,N} (S_{ij})} \quad (9)$$

2) 真实平均方向夹角的规范化处理, 规范化后的 θ_{ij} 值在 $[0, 1]$ 范围内。

$$\theta'_{i,j} = \frac{\theta_{i,j} - \min_{i,j=1,2,\dots,N} (\theta_{ij})}{\max_{i,j=1,2,\dots,N} (\theta_{ij}) - \min_{i,j=1,2,\dots,N} (\theta_{ij})} \quad (10)$$

3) 线性平均方向夹角的规范化处理, 规范化后的 α_{ij} 值在 $[0, 1]$ 范围内。

$$\alpha'_{i,j} = \frac{\alpha_{i,j} - \min_{i,j=1,2,\dots,N} (\alpha_{ij})}{\max_{i,j=1,2,\dots,N} (\alpha_{ij}) - \min_{i,j=1,2,\dots,N} (\alpha_{ij})} \quad (11)$$

4) 最长公共方向序列的规范化处理:

$$l'_{i,j} = 1 - \frac{l_{i,j}}{(\text{len}(Tr_i) + \text{len}(Tr_j))/2} \quad (12)$$

其中, $\text{len}(Tr_i)$ 、 $\text{len}(Tr_j)$ 分别表示轨迹 Tr_i 、轨迹 Tr_j 的轨迹段个数。因为 $l_{ij} \leq \text{len}(Tr_i)$ 且 $l_{ij} \leq \text{len}(Tr_j)$, 所以规范化后的 l_{ij} 值在 $[0, 1]$ 范围内。

将度量形态距离特征的包围面积与度量方向特征的真实平均方向夹角、线性平均方向夹角、最长公共方向序列分别进行组合, 从而形成 3 种轨迹相似性度量的组合方式:

1) 包围面积与真实平均方向夹角的组合为:

$$\text{sim}_{ij}(\text{包围面积} + \text{真实方向}) = \frac{S_{ij} + \theta_{ij}}{2} \quad (13)$$

2) 包围面积与线性平均方向夹角的组合为:

$$\text{sim}_{ij}(\text{包围面积} + \text{线性方向}) = \frac{S_{ij} + \alpha_{ij}}{2} \quad (14)$$

3) 包围面积与最长公共方向序列的组合为:

$$\text{sim}_{ij}(\text{包围面积} + \text{方向序列}) = \frac{S_{ij} + l_{ij}}{2} \quad (15)$$

通过对多个测试数据集聚类结果的性能评价, 从而选择最优的相似性度量组合方式。

2 相似性度量方法性能评价

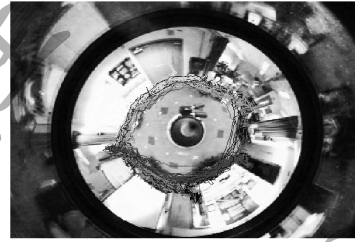
2.1 测试数据集

本文采用 CVRR 轨迹聚类数据集^[22]对基于不同轨迹相似性度量的聚类结果进行对比分析。使用 3 种类型的轨迹数据集: 1) I5 数据集, 该数据集是双向高速公路上的汽车行驶轨迹, 包含 806 条轨迹, 分为 8 类, 如图 4(a) 所示; 2) LABOMNI 数据集, 该数据集是人在室内的行走轨迹, 包含 209 条轨迹, 分为 15 类, 如图 4(b) 所示; 3) CROSS 数据集, 该数据集是模拟十字路口车辆直行与转弯的轨迹, 包含 1 900 条轨迹, 分为 19 类, 如图 4(c) 所示。所有数据集均标记了每个轨迹所属的类, 可以基于轨迹相

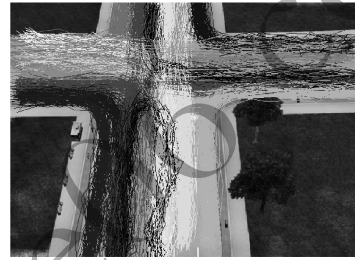
似性度量方法对这些数据集进行聚类, 进而将获得的聚类结果与数据集中标记的正确聚类进行对比, 从而评价不同轨迹相似性度量方法的聚类性能。



(a) I5 数据集



(b) LABOMNI 数据集



(c) CROSS 数据集

图 4 CVRR 轨迹聚类测试数据集

2.2 性能评价

2.2.1 3 种轨迹相似性度量组合方式的对比

本文分别采用 sim_{ij} (包围面积 + 真实方向)、 sim_{ij} (包围面积 + 线性方向)、 sim_{ij} (包围面积 + 方向序列) 3 种组合方式对测试数据集进行层次聚类分析。使用轮廓系数、类内距离平方和对聚类结果进行评价, 使用 F 值 (F-Measure) 对聚类类数识别的精度进行评价, 从而根据指标结果对 3 种组合方式的聚类效果进行对比。表 2 为轮廓系数、类内距离平方和指标公式及其含义。

表 2 轮廓系数、类内距离平方和指标公式及其含义

指标名称	指标公式	含义
轮廓系数	$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$ 其中, $a(i)$ 是类内的相似度, $b(i)$ 是类间的相似度	$s(i)$ 范围为 $[-1, 1]$, 值越大表示类内相似度越高, 组间距离越远, 即轮廓系数值越大, 聚类效果越好
类内距离平方和	$SS(i) = \sum_{k=1}^n (x - y)^2$ 其中, x, y 分别为类内的不同对象, n 为类的大小	$SS(i)$ 体现了同一类内对象之间的相似程度, 值越小, 类内对象的相似度越高

F 值计算公式如下:

$$F = \frac{2 \times P \times R}{P + R} \quad (16)$$

其中, P 为正确率, R 为召回率, 且 P 表示类标记识别正确的轨迹总数与识别出类标记的轨迹总数之比, R 表示类标记识别正确的轨迹总数与数据集包含的轨迹总数之比。

下文分别为 3 个测试集的层次聚类评价结果:

1) I5 数据集

图 5 统计了 3 种组合方式识别的各类中轨迹类标记与真实类标记不同的轨迹条数, 其中, 包围面积 + 真实方向、包围面积 + 线性方向两种组合类标记识别错误的轨迹条数较少。各组合方式评价指标结果见表 3。可以看出, 包围面积 + 真实方向、包围面积 + 线性方向两种组合方式的轮廓系数和 F 值较大, 类内距离平方和较小。

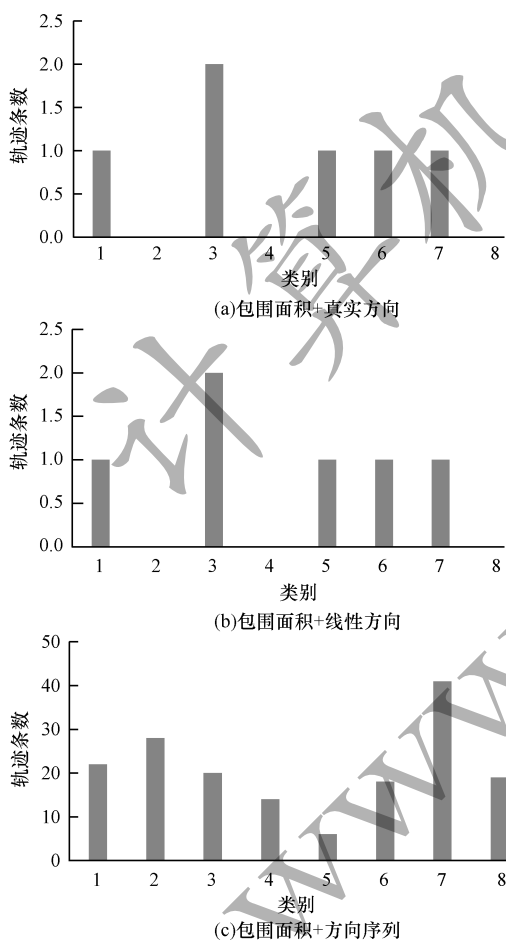


图 5 I5 数据集上 3 种组合方式类标记识别错误的轨迹条数

表 3 I5 数据集上各组合聚类结果评价指标

组合方式	轮廓系数	类内距离平方和	F 值
包围面积 + 真实方向	0.796	0.045	0.995
包围面积 + 线性方向	0.796	0.045	0.995
包围面积 + 方向序列	0.459	1.999	0.621

2) LABOMNI 数据集

从图 6 可以看出, 包围面积 + 线性方向组合方式类标记识别错误的轨迹条数最少。各组合方式评价指标结果见表 4。可以看出, 包围面积 + 线性方向的轮廓系数和 F 值最大, 类内聚类平方和最小。

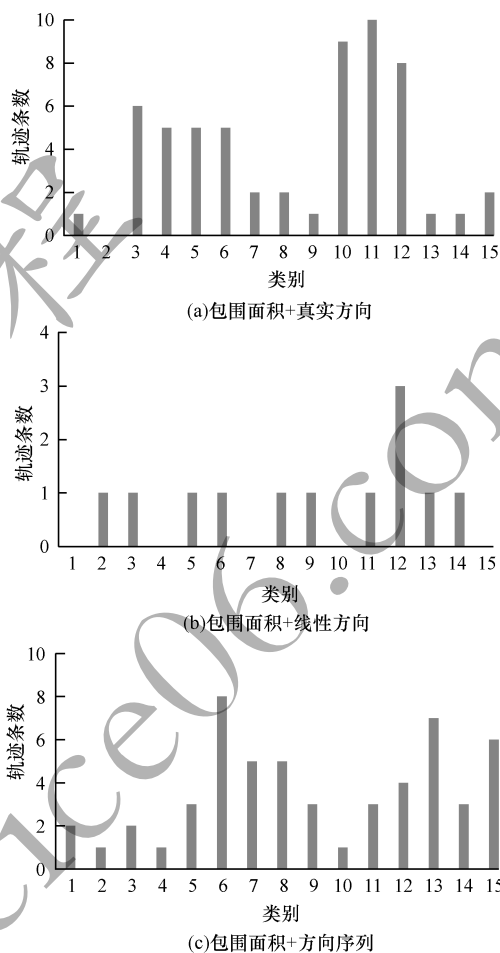


图 6 LABOMNI 数据集上 3 种组合方式类标记识别错误的轨迹条数

表 4 LABOMNI 数据集上各组合方式聚类结果评价指标

组合方式	轮廓系数	类内距离平方和	F 值
包围面积 + 真实方向	0.458	0.379	0.662
包围面积 + 线性方向	0.694	0.099	0.942
包围面积 + 方向序列	0.160	6.297	0.745

3) CROSS 数据集

图 7 表明包围面积 + 线性方向组合方式类标记识别错误的轨迹条数最少。各组合方式评价指标结果见表 5, 同样表明包围面积 + 线性方向的指标结果最优。综合 3 个数据集的最优组合, 本文选择包围面积 + 线性方向作为基于形态距离特征和方向特征的轨迹相似性度量的最优组合方式。

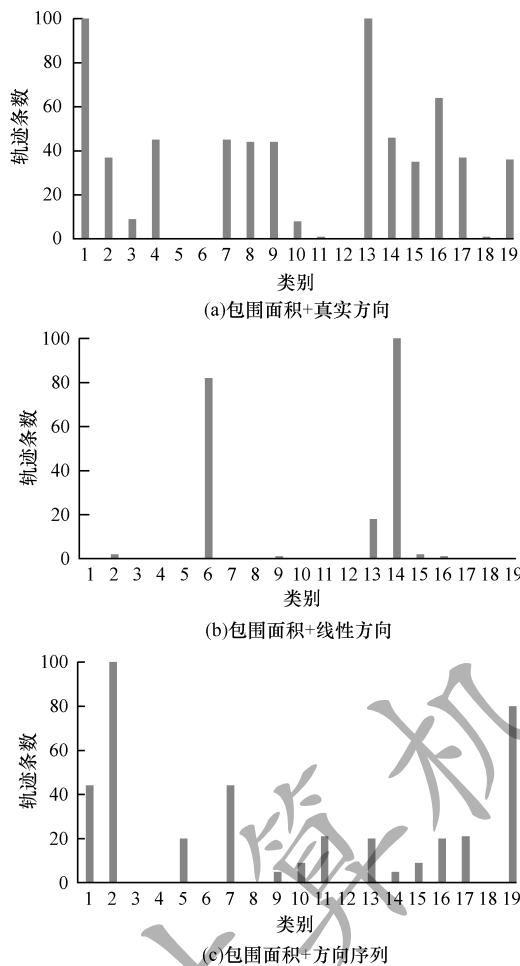


图 7 CROSS 数据集上 3 种组合方式类标记识别错误的轨迹条数

表 5 CROSS 数据集上各组合方式聚类结果评价指标

组合方式	轮廓系数	类内距离平方和	F 值
包围面积 + 真实方向	0.517	2.162	0.809
包围面积 + 线性方向	0.686	0.602	0.959
包围面积 + 方向序列	0.283	29.265	0.831

2.2.2 最优组合方式与 LCSS 方法的对比

使用 I5、LABOMNI、CROSS 这 3 个测试集,将最优组合方式(包围面积与线性平均方向夹角)和经典的轨迹相似度量 LCSS 方法进行对比。

1) I5 数据集

采用包围面积 + 线性方向、LCSS 方法分别对 I5 数据集进行层次聚类分析,图 8(a)为聚类结果的 F 值,图 8(b)为聚类得到的各类轨迹条数。可以看出,相比于 LCSS 方法,包围面积 + 线性方向的 F 值更优,聚类得到的各类轨迹条数更贴近真实的轨迹条数。

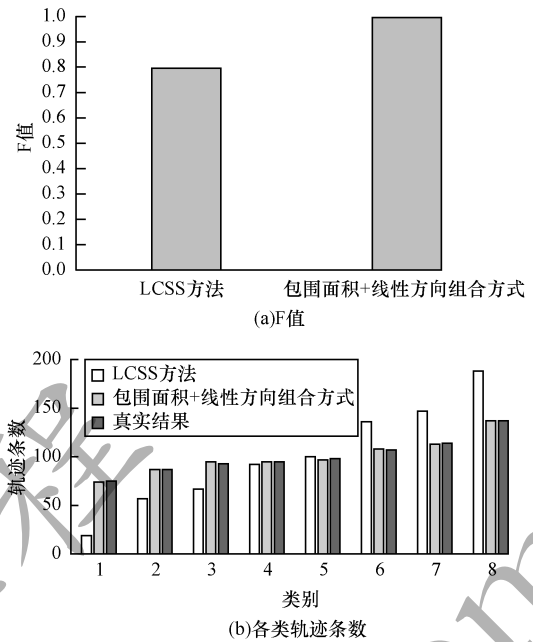


图 8 I5 数据集上包围面积 + 线性方向组合方式与 LCSS 方法的层次聚类结果对比

2) LABOMNI 数据集

图 9(a)、图 9(b)的结果表明,包围面积 + 线性方向的 F 值,聚类得到的各类轨迹条数均优于 LCSS 方法。

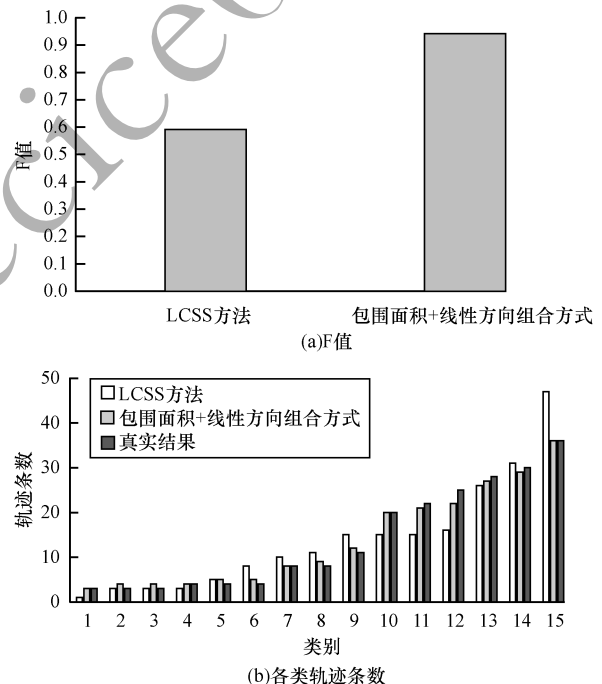


图 9 LABOMNI 数据集上包围面积 + 线性方向组合方式与 LCSS 方法的层次聚类结果对比

3) CROSS 数据集

图 10(a)表明包围面积 + 线性方向组合方式的 F 值优于 LCSS 方法,图 10(b)表明包围面积 + 线性方向组合方式得到的各类轨迹条数略优于 LCSS 方法。

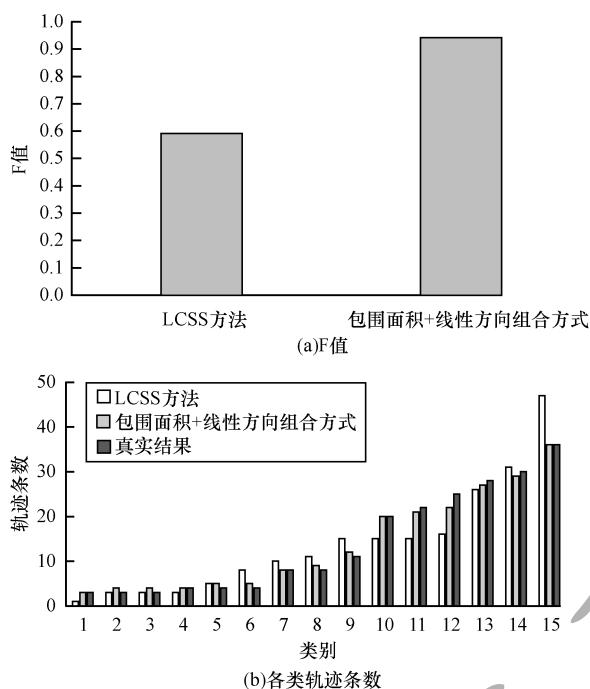


图 10 CROSS 数据集上包围面积 + 线性方向组合方式与 LCSS 方法的层次聚类结果对比

综合上述 3 个不同数据集的对比结果,包围面积 + 线性方向组合方式作为轨迹相似性度量方法的聚类结果相较于 LCSS 方法的聚类结果,准确率更优,并且每一簇内轨迹条数更接近于真实结果,因此认为以包围面积 + 线性方向组合方式作为轨迹相似性度量方法优于 LCSS 方法。

2.2.3 与最小外包矩形面积 + 线性方向的对比

本文使用最小外包矩形面积代替最优组合中的包围面积,对“包围面积 + 线性方向”和“最小外包矩形面积 + 线性方向”两种组合方式的层次聚类结果进行对比。

最小外包矩形面积指包含两条轨迹所有点集的最小外接矩形的面积(如图 11 所示),使用该矩形面积度量两条轨迹之间的距离,面积越小表示这两条轨迹的距离越小,接近程度越高,从而可通过最小外包矩形面积刻画两条轨迹整体的形态距离特征。对于轨迹 Tr_i 和 Tr_j 的最小外包矩形面积可记为 $S_{i,j}$ 。

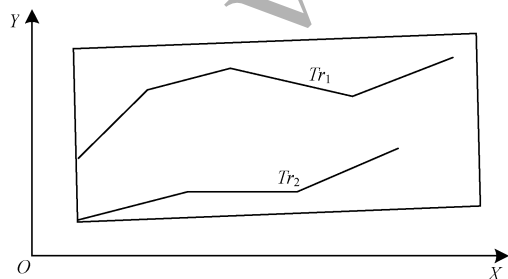


图 11 最小外包矩形面积示意图

两种组合方式的层次聚类结果的轮廓系数、类内聚类平方和、F 值见图 12,相比于最小外包矩形面积 + 线性方向组合方式,包围面积 + 线性方向组合方式具有更大的轮廓系数值、更小的类内距离平方和以及更接近于 1 的 F 值,即包围面积 + 线性方向组合方式的聚类效果更优。

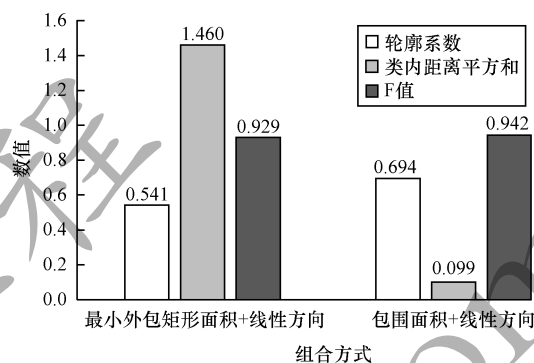


图 12 不同组合方式的层次聚类结果对比

3 北京市出租车 GPS 载客轨迹数据聚类应用

应用数据集为北京市 33 000 多辆出租车一周(2015-05-11—2015-05-17)产生的 GPS 轨迹数据,共计约 3.85 亿条点数据,采样间隔为 60 s,采样时间为 24 h。

轨迹数据主要字段名称和含义如表 6 所示,经过预处理(包括噪声剔除、非载客点剔除等),得到如图 13 所示的数据,该数据由若干载客轨迹构成,每条轨迹仅包含载客点数据,不同载客轨迹之间由[Flag]字段标志,同一辆车可含有多条载客轨迹。

表 6 部分重要字段名称及其说明

字段名称	字段说明
SUID	每辆车的唯一标志码
UTC	世界标准时间
LAT	纬度
LON	经度
SPEED	行驶速度
DISTANCE	行驶里程
TFLAG	定位描述,0 表示定位有效
VFLAG	空重车描述,0 表示没有载客
OSTDESC	车辆状态描述
Flag	载客轨迹标志

	SUID	UTC	LAT	LOn	SPEED	DISTANCE	TFLAG	VFLAG	OSTDESC	Flag
1	1663	2015-05-11 07:12:13.000	39.99328	116.48788	668	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
2	1663	2015-05-11 07:12:22.000	39.99315	116.48774	0	0	0	268435456	定位有效, ACC开, 重车	430
3	1663	2015-05-11 07:13:13.000	39.99215	116.4867	308	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
4	1663	2015-05-11 07:14:13.000	39.99149	116.48593	205	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
5	1663	2015-05-11 07:15:14.000	39.99146	116.48589	0	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
6	1663	2015-05-11 07:16:14.000	39.99146	116.48589	0	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
7	1663	2015-05-11 07:17:14.000	39.99047	116.48476	154	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
8	1663	2015-05-11 07:18:13.000	39.99009	116.48436	0	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
9	1663	2015-05-11 07:19:13.000	39.98989	116.48405	102	0	0	268435456	车门关, 定位有效, ACC开, 重车	430
10	1663	2015-05-11 07:20:14.000	39.9899	116.48395	257	0	0	268435456	车门关, 定位有效, ACC开, 重车	430

图 13 载客轨迹数据

从该数据集中提取行驶时间位于 2015-05-11 07:00—2015-05-11 08:00 的载客轨迹, 共计 11 385 条, 406 424 个载客数据点。对选取的载客轨迹 Tr_i ($i=1, 2, \dots, N, N=11\ 385$) 采用 sim_{ij} (包围面积 + 线性方向) 组合方式进行聚类。

使用 Calinski-Harabasz 分数值确定最佳聚类数量, 如图 14 所示。

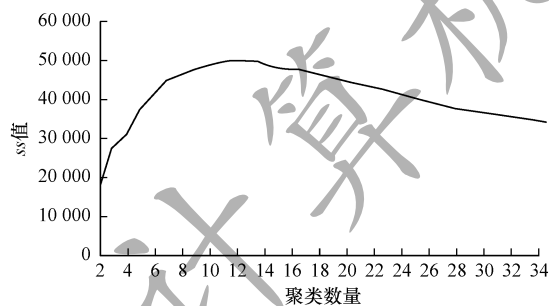


图 14 基于 Calinski-Harabasz 分数值的聚类数量

Calinski-Harabasz 分数值 ss 的数学计算公式如式(17)所示。

$$s(k) = \frac{\text{tr}(\mathbf{B}_k) m - k}{\text{tr}(\mathbf{W}_k) k - 1} \quad (17)$$

其中, m 为训练集样本数, k 为类别数, \mathbf{B}_k 为类别间的协方差矩阵, \mathbf{W}_k 为类内部的协方差矩阵, tr 为矩阵的迹。类内部的协方差越小越好, 类间协方差越大越好, ss 越大越好, 因此选择最佳聚类个数为 12。

以类为对象进行分析。绘制每一类的核密度图与线性方向平均值, 如图 15 所示, 黑色线表示线性方向平均值, 表明该类所有轨迹的平均方向, 深色区域表明该类主要分布区域 (行政区划)。由图 15 可见, 每一类轨迹的主要分布区域均以朝阳区、东城区、西城区、海淀区、丰台区为主要聚集区域, 且在东城区与西城区内, 各类的主要聚集区基本都集中在区域边界处。这些区域均在四环范围内, 北京市繁华地带, 属于二环到四环之间, 结合图 15 可以发现各类的分布基本符合实际情况。此外, 有很大一部分轨迹显示经由朝阳区去往顺义区, 这些轨迹基本以机场为目的, 由图 15(a)、图 15(c)、图 15(h)、图 15(i)、图 15(k)、图 15(l) 也可以验证。

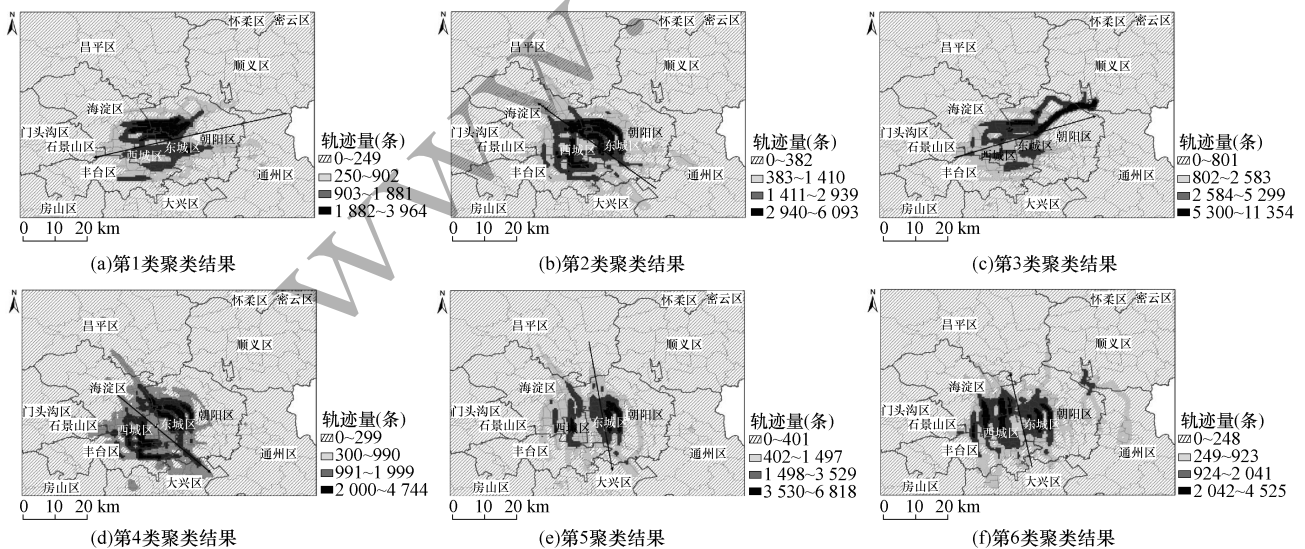


图 15 载客轨迹数据聚类结果的核密度图与线性方向平均值

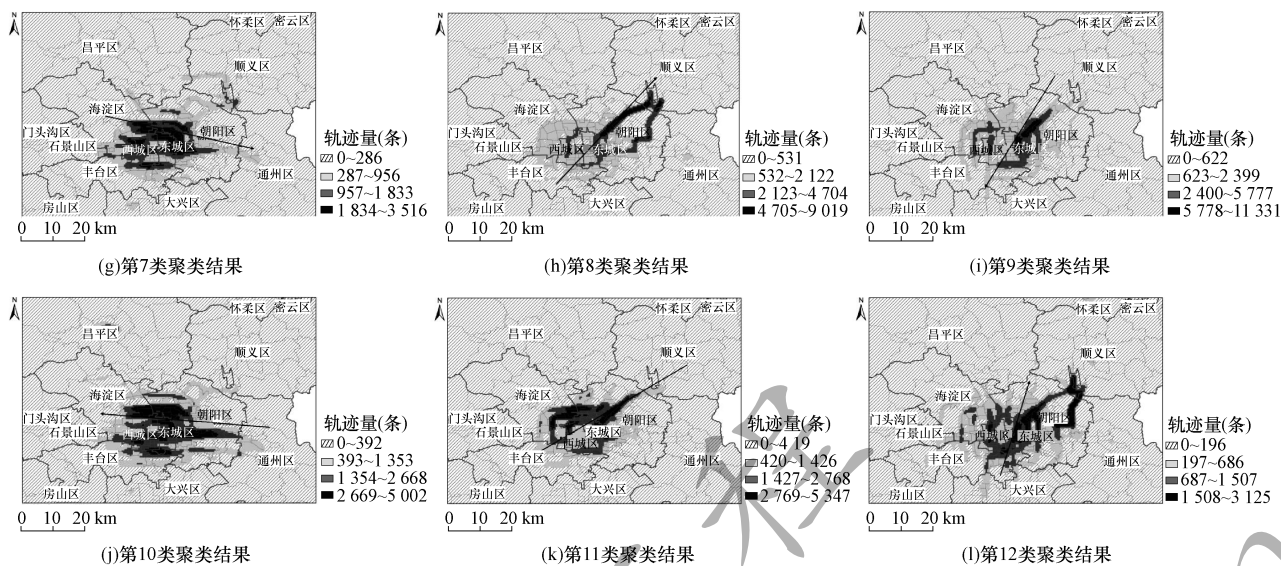


图 15 载客轨迹数据聚类结果的核密度图与线性方向平均值(续)

类间线性方向平均值差异较大,可以反映本文方法效果明显,能够对不同趋势方向的轨迹进行分类。将 12 类结果分为 6 组,具体组合如表 7 所示,分类依据为每一组内两类结果的线性方向相反。每一组内不同类别在乡镇区域分布情况不同,具体详见表 8。可以看出,组内两类结果主要分布区域在行政区划上有很相似性,在以乡、镇、街道为单位的区域分布上有较大差异,该差异体现在同一行政区划内的不同乡、镇、街道上,即同一组内不同类的主要分布区域集中在相同的行政区划上,但是并没有全部位于相同的乡、镇、街道中。

表 7 线性方向角度组合结果 (°)			
分组	类别	角度(正东方向 逆时针旋转)	角度差值
第 1 组	1	192.944	176.516
	3	16.428	
第 2 组	2	144.098	176.073
	4	320.171	
第 3 组	5	280.103	176.942
	6	103.161	
第 4 组	7	347.632	172.150
	10	175.482	
第 5 组	8	46.813	163.732
	11	210.545	
第 6 组	9	238.023	165.197
	12	72.826	

表 8 不同类别的主要区域分布情况

分组	类别	以区为单位的分布区域	以乡、镇、街道为单位的分布区域
第 1 组	1	朝阳区、海淀区、西城区、东城区	香河园街道、和平街街道、太阳宫乡、望京街道、和平里街道、花园路街道、德胜街道、中关村街道
	3	朝阳区、海淀区、西城区、顺义区、通州区	太阳宫乡、望京街道、花园路街道、德胜街道、天竺镇、空港街道、宋庄镇
第 2 组	2	朝阳区、海淀区、西城区、东城区、丰台区	麦子店街道、太阳宫乡、花园路街道、德胜街道、太平桥街道、和平里街道
	4	朝阳区、东城区、西城区、丰台区	建外街道、和平里街道、广安门街道、右安门街道、太平桥街道、卢沟桥乡
第 3 组	5	朝阳区、海淀区、西城区、东城区	三里屯街道、左家庄街道、太阳宫乡、望京街道、花园路街道、展览路街道、和平里街道
	6	朝阳区、海淀区、西城区、丰台区	麦子店街道、北太平庄街道、花园路街道、展览路街道、卢沟桥乡
第 4 组	7	朝阳区、海淀区、东城区、西城区、丰台区	左家庄街道、和平里街道、花园路街道、德胜街道、太平桥街道
	10	朝阳区、海淀区、西城区、东城区	太阳宫乡、高碑店乡、和平里街道、花园路街道、德胜街道
第 5 组	8	朝阳区、顺义区、通州区	麦子店街道、崔各庄乡、空港街道、天竺镇、宋庄镇
	11	朝阳区、海淀区、西城区、东城区	太阳宫乡、望京街道、北太平庄街道、花园路街道、中关村街道、德胜街道
第 6 组	9	朝阳区、东城区	左家庄街道、太阳宫乡、望京街道、东直门街道、北新桥街道
	12	朝阳区、东城区、顺义区、通州区	麦子店街道、太阳宫乡、望京街道、空港街道、天竺镇、宋庄镇

4 结束语

本文提出形态距离特征与 3 种方向特征组合的

轨迹相似性度量方法,选出最优组合方式为包围面积与线性平均方向。将该组合方式分别与 LCSS 方法、最小外包矩形面积及线性平均方向组合方式进

行对比验证,发现该组合方式在性能与准确率上均表现良好。使用该组合方式对北京市出租车 GPS 轨迹数据进行聚类应用,结果显示能够区分移动对象趋势方向,验证了以移动对象运动方向为主导的轨迹相似性度量方法的有效性与准确性。由于本文在生成方向序列过程中,采用固定阈值进行区间分割,灵活性较差,因此下一步可将方向序列分割阈值设置为可变参数,选择多个阈值影响下的最优组合,并且可结合轨迹数据与 POI 数据进行轨迹语义相似性度量,实现乘客流向、主要上下车点区域的分析与识别。

参考文献

- [1] 李正欣,张凤鸣,李克武,等. 一种支持 DTW 距离的多元时间序列索引结构[J]. 软件学报,2014,25(3): 560-575.
- [2] SKOUMAS G, SKOUTAS D, VLACHAKI A. Efficient identification and approximation of k-nearest moving neighbors [C]//Proceedings of ACM Sigspatial International Conference on Advances in Geographic Information Systems. New York, USA: ACM Press, 2013:264-273.
- [3] 袁正午,袁松彪. 流聚类模型及其统一表示[J]. 计算机工程,2009,35(16):76-77,80.
- [4] 龚玺,裴韬,孙嘉,等. 时空轨迹聚类方法研究进展[J]. 地理科学进展,2011,30(5):522-534.
- [5] KEOGH E, PALPANAS T, ZORDAN V, et al. Indexing large human-motion databases [C]//Proceedings of the 13th International Conference on Very Large Data Bases. New York, USA: ACM Press, 2004:780-791.
- [6] ZHENG Yu, ZHOU Xiaofang. Computing with spatial trajectories [M]. Berlin, Germany: Springer, 2011.
- [7] CHEN L, ÖZSU M, ORIA V. Robust and fast similarity search for moving object trajectories [C]//Proceedings of 2005 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2005:491-502.
- [8] VLACHOS M, GUNOPULOS D, DAS G. Rotation invariant distance measures for trajectories [C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2004:707-712.
- [9] HERMES C, WOHLER C, SCHENK K, et al. Long-term vehicle motion prediction [C]//Proceedings of 2009 IEEE Intelligent Vehicles Symposium. Washington D. C., USA: IEEE Press, 2009:652-657.
- [10] SAKURAI Y, YOSHIKAWA M, FALOUTSOS C. FTW: fast similarity search under the time warping distance [C]//Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York, USA: ACM Press, 2005:326-337.
- [11] KIM J, MAHMASSANI H S. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories [J]. Transportation Research Part C: Emerging Technologies, 2015, 9:375-390.
- [12] 曾万聃,周敏奇,刘云翔. 轨迹大数据的比较算法研究[J]. 吉林大学学报(信息科学版), 2016, 34(6): 792-799.
- [13] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases [C]//Proceedings of International Conference on Foundations of Data Organization and Algorithms. Berlin, Germany: Springer, 1993:69-84.
- [14] LEE J G, HAN J, WHANG K Y. Trajectory clustering: a partition-and-group framework [C]//Proceedings of 2007 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2007:593-604.
- [15] LIU Jinpeng, ZHANG Yanling, LIU Gang. Partition and density-based clustering for moving objects trajectories [C]//Proceedings of the 3rd International Conference on Computer Science and Education. Washington D. C., USA: IEEE Press, 2008:182-187.
- [16] NANNI M, PEDRESCHI D. Time-focused clustering of trajectories of moving objects [J]. Journal of Intelligent Information Systems, 2006, 27(3):267-289.
- [17] LI Yifan, HAN Jiawei, YANG Jiong. Clustering moving objects [C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2004: 617-622.
- [18] BEREZANSKY M, GREENSPAN H, COHEN-OR D, et al. Segmentation and tracking of human sperm cells using spatio-temporal representation and clustering [J]. Proceedings of SPIE, 2007, 6512:1-12.
- [19] GAO Yunjun, ZHENG Baihua, CHEN Gencai, et al. Algorithms for constrained K-nearest neighbor queries over moving object trajectories [J]. Geoinformatica, 2010, 14(2):241-276.
- [20] LIN Bin, SU Jianwen. One way distance: for shape based similarity search of moving object trajectories [J]. Geoinformatica, 2008, 12(2):117-142.
- [21] PERNG C S, WANG H, ZHANG S R, et al. Landmarks: a new model for similarity-based pattern querying in time series databases [C]//Proceedings of the 16th International Conference on Data Engineering. New York, USA: ACM Press, 2000:33-42.
- [22] Trajectory analysis datasets [EB/OL]. [2018-09-18]. http://cvrr.ucsd.edu/bmorris/datasets/dataset_trajectory_analysis.html.

编辑 陆燕菲