第 37 卷第 4 期 系统工程理论与实践 Vol.37, No.4
2017 年 4 月 Systems Engineering — Theory & Practice Apr., 2017

doi: 10.12011/1000-6788(2017)04-1046-10 中图分类号: TP181 文献标志码: A

基于灰色方法与结构距离的飓风轨迹聚类算法

江艺羡 [1,2], 张岐山 [1]

(1. 福州大学 经济与管理学院, 福州 350108; 2. 福建医科大学 公共卫生学院, 福州 350108)

摘 要 飓风轨迹的聚类分析以轨迹间的相似性为基础. 但是目前轨迹相似距离研究主要仅针对一个或两个轨迹属性, 使得得到的聚类结果无法充分体现轨迹的运动特征. 为充分利用轨迹的属性特征, 结合灰色相关理论, 提出将位置、方向、速度三个属性融合为结构相似距离, 并以此作为轨迹间相似距离度量函数进行聚类分析, 以达到更全面了解轨迹运动特点的目的. 实例分析表明该算法能充分体现轨迹运动特征, 聚类结果更具有实际意义.

关键词 聚类; 轨迹; 灰色 GM(1,1) 模型; 结构距离

The clustering algorithm of cyclone track based on grey theory and structure distance

JIANG Yixian[1,2], ZHANG Qishan[1]

(1. School of Economics and Management, Fuzhou University, Fuzhou 350108, China; 2. Public Health School, Fujian Medical University, Fuzhou 350108, China)

Abstract The clustering analysis of cyclone track is built on the similarity between them. Current studies on trajectory similarity distance were for one or two attributes only. So in the case, the results were unable to fully reflect the motion characteristics of trajectory. In order to make full use of the trajectory attributes, based on the grey theory, this paper proposed that the three attributes (position, direction and velocity) should be fused as structure distance for clustering analysis, helping to achieve a more comprehensive understanding of the purpose of trajectory motion. Example analysis shows that the algorithm can fully reflect the trajectory characteristics, and the results have more practical significance.

Keywords cluster; trajectory; grey GM(1,1) model; distance structure

1 引言

轨迹聚类的目的是根据轨迹的相似程度, 将相似轨迹进行归类. 因此轨迹间的相似度是聚类的依据, 相似度的定义影响聚类结果. 目前, 研究人员已经提出了很多轨迹相似性的度量方法. 比较典型的如欧几里德距离 [1−3], Hausdorff 距离 [4−5], 编辑距离 [6−7], 动态时间弯曲距离 (DTW)[8−9], 最长公共子序列 (LCSS)[10−11] 等. 其中, 欧几里德距离与 Hausdorff 距离计算方法简单直观, 但受噪声数据的影响较大; 编辑距离通过简单的插入、删除、替换等操作进行计算, 相对比较简单, 但其发现的相似时间区间是离散的, 不易被人们观察和理解; DTW 距离解决了序列间采样率不同和时间尺度不一的问题, 但对噪声较敏感, 计算复杂度较大; LCSS 不需要所有的记录点全部匹配, 可以有效地降低噪声影响, 但受参数影响且对于大样本数据, 时间复杂度较

收稿日期: 2015-10-12

作者简介: 江艺羡 (1983–), 女, 汉, 漳州龙海人, 讲师, 博士研究生, 研究方向: 灰色系统, 人工智能, 数据挖掘, E-mail: esian108@gmail.com; 张岐山 (1962–), 男, 汉, 黑龙江绥化人, 教授, 博士生导师, 研究方向: 数据挖掘、管理信息系统与决策支持系统、灰色系统、系统工程、物流管理与物流工程及人工智能, E-mail: zhang_qs@foxmail.com.

基金项目: 国家自然科学基金青年项目 (61300104); 福建省自然科学基金 (2013J01230)
Foundation item: National Natural Science Foundation of China (61300104); Natural Science Foundation of Fujian Province of China (2013J01230)
中文引用格式: 江艺羡, 张岐山. 基于灰色方法与结构距离的飓风轨迹聚类算法 [J]. 系统工程理论与实践, 2017, 37(4): 1046–1055.
英文引用格式: Jiang Y X, Zhang Q S. The clustering algorithm of cyclone track based on grey theory and structure distance[J]. Systems Engineering — Theory & Practice, 2017, 37(4): 1046–1055.

高. 此外, 动态轨迹的运动信息包含地理位置, 速度, 以及时间等, 目前较多文献均只考虑轨迹某一属性距离作为相似度进行聚类, 因此得到的聚类结果具有偏向性. 对于轨迹聚类应当综合考虑动态轨迹所包含的运动信息, 尽可能从多角度出发, 获得的聚类结果才能更符合实际情况, 对后续研究起到更好的指导作用.

灰色系统理论是我国著名学者邓聚龙教授于 1982 年创立. 经过 30 余年的发展, 该理论作为一门新兴学科已以其强大的生命力立于科学之林. 著名科学家钱学森教授、模糊数学创始人 Zadeh 教授等国内外著名学者以及众多实际工作者对灰色系统研究给予高度评价. 目前该理论已被广泛应用于工农业、经济生活以及科研等领域, 并取得显著应用成果 [12]. 灰色理论是解决不确定系统问题的有力工具之一. 利用灰色理论分析移动对象动态轨迹特征, 可以对动态轨迹当前或未来运行趋势进行更为科学准确地分析; 同时灰色理论在动态轨迹聚类中的应用研究可以促进灰色预测技术在移动对象轨迹挖掘中的发展.

综上所述, 借鉴文献 [13], 本文从三个方面考虑轨迹间的相似度: 地理距离, 运行方向, 速度变化等. 这三个属性的线性组合构成轨迹结构距离, 从而确定轨迹间的相似度. 本文算法在计算时, 首先通过文献 [14] 中的方法对轨迹进行划分处理, 之后利用灰色理论计算轨迹集间的结构距离, 最后采用 DBSCAN 聚类算法对轨迹进行聚类分析.

2 相关定义

2.1 结构距离定义

根据轨迹的结构特征, 定义轨迹在位置, 方向, 速度三个方面的相似距离, 分别为位置距离, 方向距离, 速度距离. 相关定义如下:

定义 1 (位置距离 L_{ij}) 设轨迹集为 $\boldsymbol{T}_{TR_i}=\{\boldsymbol{TR}_1,\boldsymbol{TR}_2,\cdots,\boldsymbol{TR}_N\}$, 轨迹集中的任意轨迹 $\boldsymbol{TR}_i=(tr_{i1},tr_{i2},\cdots,tr_{in})$ 与 $\boldsymbol{TR}_j=(tr_{j1},tr_{j2},\cdots,tr_{jm})$, 两轨迹间的 Hausdorff 距离为 $H_{ij}=\max\{h(\boldsymbol{TR}_i,\boldsymbol{TR}_j),h(\boldsymbol{TR}_j,\boldsymbol{TR}_i)\}$, 其中 $h(\boldsymbol{TR}_i,\boldsymbol{TR}_j)=\max_{p\in\boldsymbol{TR}_i}\left\{\min_{q\in\boldsymbol{TR}_j}\{dist(p,q)\}\right\}$, $dist(p,q)$ 为两点间的欧式距离函数, 则 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 的位置距离为 $L_{ij}=\frac{H_{ik}-\min(H_i^*)}{\max(H_i^*)-\min(H_i^*)}$, 其中 $H_i^*=\{H_{ik}|k=(1,2,\cdots,N)\}$, $H_{ii}=0$.

定义 2 (轨迹点方向角 $\theta_i(t)$) 轨迹 $\boldsymbol{TR}_i=(tr_{i1},tr_{i2},\cdots,tr_{in})$ 的曲线参数方程为 $f_i(x,y):\begin{cases}x_{it}=g_{i1}(t)\\ y_{it}=g_{i2}(t)\end{cases}$. 设轨迹 $\boldsymbol{TR}_i$ 在点 t 的运行方向沿着逆时针方向, 与 x 轴正方向的夹角为轨迹在该点的方向角 $\theta_i(t)$, 采用弧度制表示, 则

$$\theta_i(t)=\begin{cases}0, & y_{i,t+1}=y_{it}\ \text{ and }\ x_{i,t+1}\geq x_{it}\\ \pi, & y_{i,t+1}=y_{it}\ \text{ and }\ x_{i,t+1}<x_{it}\\ 0.5\pi, & x_{i,t+1}=x_{it}\ \text{ and }\ y_{i,t+1}>y_{it}\\ 1.5\pi, & x_{i,t+1}=x_{it}\ \text{ and }\ y_{i,t+1}<y_{it}\\ \arctan k_{it}, & y_{i,t+1}>y_{it}\ \text{ and }\ x_{i,t+1}>x_{it}\\ \pi+\arctan k_{it}, & y_{i,t+1}>y_{it}\ \text{ and }\ x_{i,t+1}<x_{it}\\ \pi+\arctan k_{it}, & y_{i,t+1}<y_{it}\ \text{ and }\ x_{i,t+1}<x_{it}\\ 2\pi+\arctan k_{it}, & y_{i,t+1}<y_{it}\ \text{ and }\ x_{i,t+1}>x_{it}\end{cases}\tag{1}$$

其中斜率 $k_{it}=\frac{dg_{i2}(t)}{dg_{i1}(t)}$, $t\in[1,n-1]$.

由于轨迹 $\boldsymbol{TR}_i$ 在点 n 后没有记录值, 因此轨迹在点 n 方向角不列入本文考虑范围. 长度为 n 的轨迹具有 $n-1$ 个点方向角.

定义 3 (轨迹点方向角序列 θ_i^*) 轨迹 $\boldsymbol{TR}_i=(tr_{i1},tr_{i2},\cdots,tr_{in})$ 的曲线参数方程为 $f_i(x,y):\begin{cases}x_{it}=g_{i1}(t)\\ y_{it}=g_{i2}(t)\end{cases}$, 轨迹在点 t 的方向角 $\theta_i(t)$, 称轨迹 $\boldsymbol{TR}_i$ 各点方向轨迹角按顺序所组成的序列为轨迹点方向角序列 θ_i^*, 则 $\theta_i^*=\{\theta_i(1),\theta_i(2),\cdots,\theta_i(n-1)\}$.

定义 4 (轨迹间方向角 $\widehat{\theta}_{ij}$) 轨迹 $\boldsymbol{TR}_i=(tr_{i1},tr_{i2},\cdots,tr_{in})$ 与 $\boldsymbol{TR}_j=(tr_{j1},tr_{j2},\cdots,tr_{jm})$, 两者长度

相等, 轨迹 $\boldsymbol{TR}_i$ 的点方向角为 θ_i^*, 轨迹 $\boldsymbol{TR}_j$ 的点方向角为 θ_j^*, 设两轨迹间的方向角为 $\widehat{\theta}_{ij}$, 则

$$\widehat{\theta}_{ij}=\frac{1}{n-1}\sum_{t=1}^{n-1}\theta_{ij}(t) \tag{2}$$

其中

$$\theta_{ij}(t)=\begin{cases}|\theta_i(t)-\theta_j(t)|, & 0\le|\theta_i(t)-\theta_j(t)|\le\pi\\ 2\pi-|\theta_i(t)-\theta_j(t)|, & \pi<|\theta_i(t)-\theta_j(t)|\le 2\pi\end{cases} \tag{3}$$

定义 5 (方向距离 D_{ij}) 设轨迹集 $\boldsymbol{T}_{\boldsymbol{TR}_i}=\{\boldsymbol{TR}_1,\boldsymbol{TR}_2,\cdots,\boldsymbol{TR}_N\}$, 对于轨迹集中的任一轨迹 $\boldsymbol{TR}_i=(tr_{i1},tr_{i2},\cdots,tr_{in})$ 与 $\boldsymbol{TR}_j=(tr_{j1},tr_{j2},\cdots,tr_{jm})$, 两轨迹间的方向角为 $\widehat{\theta}_{ij}$, 则 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 间的方向距离为

$$D_{ij}=\frac{\widehat{\theta}_{ij}-\min(\widehat{\theta}_i)}{\max(\widehat{\theta}_i)-\min(\widehat{\theta}_i)} \tag{4}$$

其中 $\widehat{\theta}_i=\{\widehat{\theta}_{ij}|k=(1,2,\cdots,N)\}$, $\widehat{\theta}_{ii}=0$.

以上关于轨迹 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 间的方向距离的定义, 是在两轨迹长度相等的前提下. 如果两轨迹长度不相等, 假设轨迹 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 长度分别为 n 和 m, 不失一般性, 令 $n>m$, 则从较长轨迹 $\boldsymbol{TR}_i$ 中按顺序截取长度为 m 的子轨迹集 $\boldsymbol{T}_{TR_i}=\{\boldsymbol{TR}_{i,1-i,m},\boldsymbol{TR}_{i,2-i,m+1},\cdots,\boldsymbol{TR}_{i,n-m+1-i,n}\}$, 计算子轨迹集 $\boldsymbol{T}_{TR_i}$ 中各子轨迹与轨迹 $\boldsymbol{TR}_j$ 的方向距离, 取所有距离中的最小值作为轨迹 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 间的方向距离. 下文对于长度不相等的两轨迹间的速度距离的求法类同.

轨迹速度是关于时间 t 的时间序列. 时间序列间的相似距离可根据序列曲线几何形状的相似程度进行分析. 因此本文借鉴灰关联分析中基于相似的角度 [2] 与基于接近的角度 [2] 两个方法衡量轨迹间的速度距离. 轨迹间的速度距离可分为基于相似的速度距离, 与基于接近的速度距离. 基于相似的速度距离是从几何形状上判断速度序列的相似程度, 即两速度变化趋势的相似程度. 基于接近的速度距离是从空间位置上判断速度序列的接近程度, 即两速度序列数值大小的相似程度.

定义 6 (基于相似的速度距离 sv_{ij}) 设轨迹 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 对应的速度序列 $\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{in})$ 与 $\boldsymbol{V}_j=(v_{j1},v_{j2},\cdots,v_{jn})$, 则两者间基于相似的速度距离 sv_{ij} 为

$$sv_{ij}=\frac{1}{n}\left|sv_i-sv_j\right| \tag{5}$$

其中, $|sv_i-sv_j|=|\sum_{k=2}^{n-1}(v_{ik}^0-v_{jk}^0)+\frac{1}{2}(v_{in}^0-v_{jn}^0)|$.

定义 7 (基于接近的速度距离 SV_{ij}) 设轨迹 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 对应的速度序列 $\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{in})$ 与 $\boldsymbol{V}_j=(v_{j1},v_{j2},\cdots,v_{jn})$, 则两者间基于接近的速度距离 SV_{ij} 为

$$SV_{ij}=\frac{1}{n}\left|SV_i-SV_j\right| \tag{6}$$

其中, $|SV_i-SV_j|=|\sum_{k=2}^{n-1}(v_{ik}-v_{jk})+\frac{1}{2}(v_{in}-v_{jn})|$.

定义 8 (速度距离 V_{ij}) 设轨迹 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 间基于相似的速度距离为 sv_{ij}, 基于接近的速度距离为 SV_{ij}, 轨迹 $\boldsymbol{TR}_i$ 与轨迹集 (轨迹总数为 N) 中各轨迹间基于相似的速度距离集合为 $sv_i^*=\{sv_{ik}|k=(1,2,\cdots,N)\}$, 其 $sv_{ii}=0$; 基于接近的速度距离集合为 $SV_i^*=\{SV_{ik}|k=(1,2,\cdots,N)\}$, 其中 $SV_{ii}=0$. 假设两轨迹间的速度距离为 V_{ij}, 则

$$V_{ij}=\frac{Vs_{ij}+VS_{ij}}{2} \tag{7}$$

其中, $Vs_{ij}=\frac{sv_{ij}-\min(sv_i^*)}{\max(sv_i^*)-\min(sv_i^*)}$, $VS_{ij}=\frac{SV_{ij}-\min(SV_i^*)}{\max(SV_i^*)-\min(SV_i^*)}$.

定义 9 (结构距离 SD_{ij}) 根据上述轨迹各属性间的距离比较, 得到两轨迹在三个属性上的相似距离比较值分别为 L_{ij}, V_{ij}, D_{ij}. 若三者的权重分别为 ω_L, ω_V, ω_D (其中 $\omega_L+\omega_V+\omega_D=1$). 则两轨迹加权距离为

$$SD_{ij}=\omega_L L_{ij}+\omega_V V_{ij}+\omega_D D_{ij} \tag{8}$$

若 $\omega_L=\omega_V=\omega_D=\frac{1}{3}$, 则目标轨迹 TR_i 与 TR_j 的结构距离为 $SD_{ij}=\frac{V_{ij}+D_{ij}+L_{ij}}{3}$.

定义 10 (结构相似度 $SSIM_{ij}$) 设轨迹集 $\boldsymbol{T}_{TR_i}=\{\boldsymbol{TR}_1,\boldsymbol{TR}_2,\cdots,\boldsymbol{TR}_N\}$, 对于轨迹集中的任一轨迹 $\boldsymbol{TR}_i=(tr_{i1},tr_{i2},\cdots,tr_{in})$ 与 $\boldsymbol{TR}_j=(tr_{j1},tr_{j2},\cdots,tr_{jm})$, $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 间的结构距离为 SD_{ij}, 则 $\boldsymbol{TR}_i$ 与 $\boldsymbol{TR}_j$ 间的结构相似度 $SSIM_{ij}$ 为

$$SSIM_{ij}=\frac{\max(SD_{ij})-SD_{ij}}{\max(SD_{ij})-\min(SD_{ij})} \tag{9}$$

2.2 结构距离分析

1) 方向距离对比分析

根据定义 4 可知, 方向距离 $\widehat{\theta}_{ij}$ 表示轨迹 $\boldsymbol{TR}_i$ 运行方向与轨迹 $\boldsymbol{TR}_j$ 运行方向的偏转角度. 起点与终点相同的轨迹, 采用公式 (2) 计算轨迹间的方向角, 并将结果与采用线段表示的轨迹方向角进行对比, 对比情况见图 1. 从图 1(a) 可看出虽然起点与终点相同, 不同的轨迹曲线间的方向角不同. 当起点与终点相同时, 采用线段表示的不同轨迹间的方向角均相同, 见图 1(b). 从两者的比较可看出, 采用公式 (2) 的方向角计算方法更符合实际.

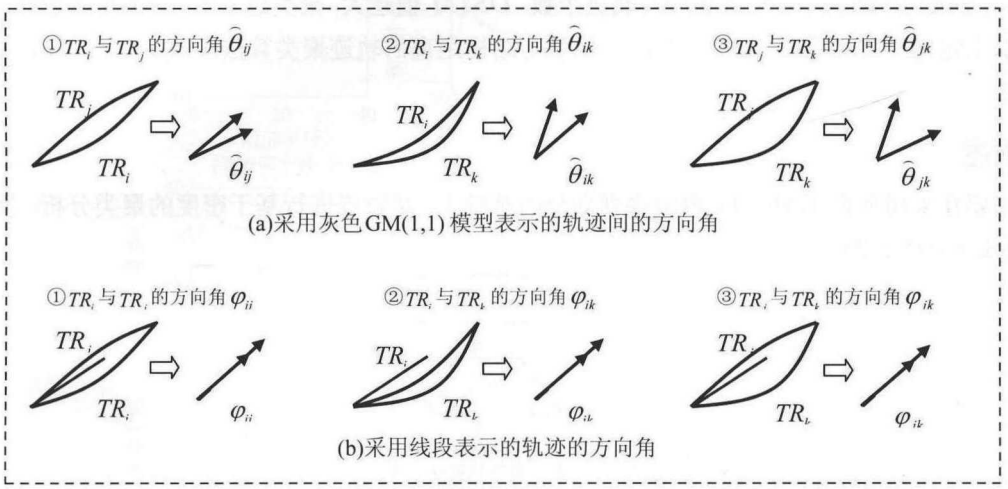

图 1　轨迹间方向角的对比分析

2) 速度距离对比分析

利用公式 (5) 与公式 (6) 计算得到轨迹间基于接近与基于相似的速度距离, 两者的示例见图 2. 图 2(a) 为基于接近的速度距离, 在区间 $[1,4]$ 中, $\boldsymbol{V}_i$ 与 $\boldsymbol{V}_j$ 之间所夹的阴影面积较大, 对应相同区间, 图 2(b) $\boldsymbol{v}_i^0$ 与 $\boldsymbol{v}_j^0$ 所夹的阴影面积则较小. 主要原因是, 在该区间两速度序列的数值差别较大, 因此基于接近的速度距离较大; 相反, 在该区间两速度序列数值变化的大小较相似, 因此基于相似的速度距离较小. 通过基于接近与相似视角的速度距离方法能更好体现不同轨迹间的速度差异.

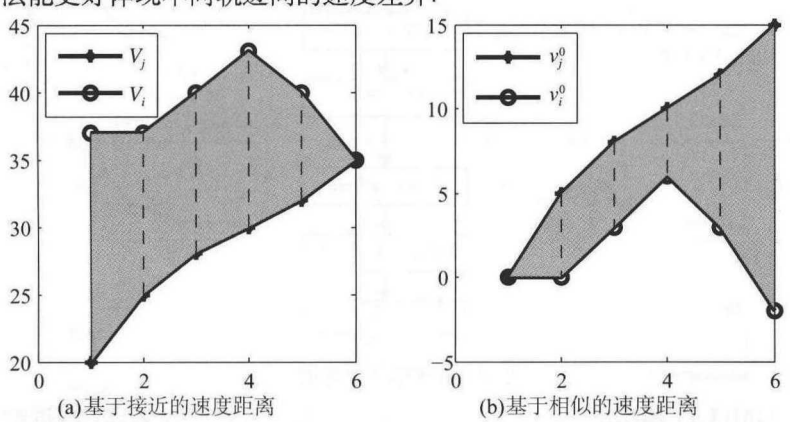

图 2　基于接近与基于相似的速度距离的对比分析

通过以上的对比可看出, 本文所定义的各属性距离函数是可行的, 利用该方法能更全面发现轨迹行为模式.

2.3 聚类效果精度分析

在聚类结果的效果评价指标中, 需要对聚类效果进行评价. 本文以精度分析的角度 [13] 对聚类结果进行评价. 采用聚类结果中轨迹间的聚集程度表示聚类的总体相似度, 相关定义如下:

定义 11 (聚类平均相似度 $Osim_i$) 对于任意第 i 个聚类 C_i, 该聚类的平均相似度表示为

$$Osim_i = \frac{\sum\limits_{x \in C_i} SSIM\,(x, c_i)}{m_i} \tag{10}$$

其中, m_i 是聚类 C_i 包含轨迹段的个数, x 是类中的任一轨迹段, c_i 是 C_i 的核心轨迹段.

定义 12 (轨迹集总体平均相似度 $OSIM$) 整个轨迹集聚类结果之间的总体平均相似度为各聚类平均相似度的加权和.

$$OSIM = \sum_{i=1}^{cnum} \frac{m_i}{m} \cdot Osim_i \tag{11}$$

其中, $cnum$ 是聚类个数, m 是轨迹集所有轨迹个数. $OSIM$ 值越大, 聚类越紧凑, 聚类结果越好.

根据以上定义, 给出基于灰色 GM(1,1) 模型与结构距离的轨迹聚类算法. 简记为 DBSCAN_GMPR_STRU 算法.

3 算法描述

本文主要在采用灰色 GM(1,1) 模型序列划分的基础上, 对轨迹进行基于密度的聚类分析. 算法描述如下, 对应描述示意图见图 3.

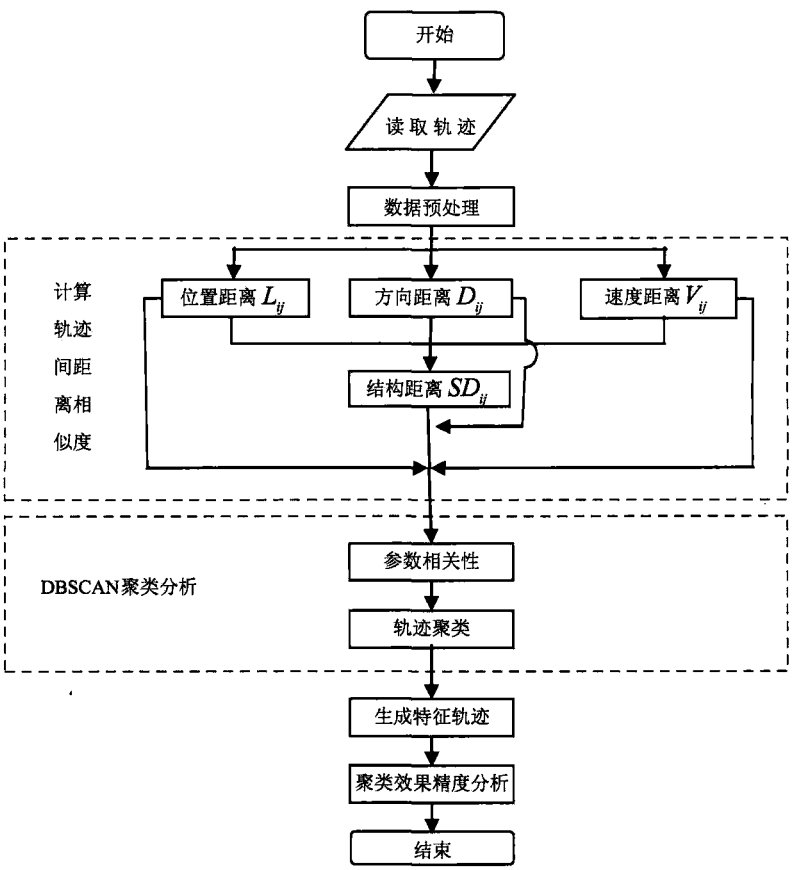

图 3 算法描述示意图

1) 采用基于灰色 GM(1,1) 模型的序列分段表示方法 [14] 对轨迹进行预处理. 根据给定的压缩率将轨迹集划分成具有单调性的子轨迹, 并利用模型获得各子轨迹的表达式.

2) 计算各轨迹间的相关距离, 确定轨迹相似度, 之后采用 DBSCAN 聚类算法对划分后的子轨迹集进行聚类. 其中, 可通过设置不同的 Eps 和 $MinPts$ 等算法参数 [13] 获得所需的聚类个数.

3) 确定各类的轨迹簇平均方位向量, 以轨迹簇平均方位向量为 x 轴进行坐标变换, 获取变换后轨迹与簇平均方位法向量的交点坐标. 若有多个轨迹对应同一 x 坐标, 则取这些轨迹与簇平均方位法向量交点的平均值, 所有的交点按顺序组成序列, 之后将序列变换回原始坐标系, 便获得各轨迹簇的特征轨迹.

4 实验及分析

为验证聚类算法, 实验数据采用 1950–2006 年的飓风轨迹 [13].

4.1 参数分析

数据进行聚类时, 需要根据不同的相似度量函数, 需要设置的相关参数不同. 以下分别针对位置距离、方向距离、速度距离以及结构距离的参数设置进行分析:

1) 以位置距离作为相似度量函数. 对于给定的近邻个数, 聚类个数与近邻半径的相关性见图 4. 当给定近邻个数 k 时, 飓风聚类个数 $class$ 随着近邻半径 Eps 的变大, 先变多后变少.

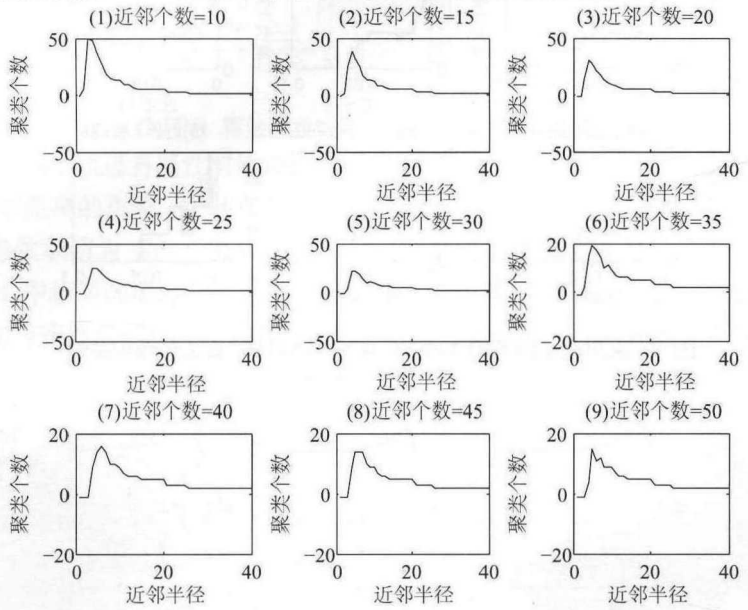

图 4　采用位置距离的 DBSCAN_GMPR 算法参数相关性

2) 以方向距离作为相似度量函数, 聚类分析所需参数的相关性见图 5. 对给定的近邻个数 k, 聚类个数 $class$ 随邻居领域 Eps 的变大先迅速变大后逐渐变小. 即 Eps 较小时, $class$ 对参数 Eps 值的设置较敏感.

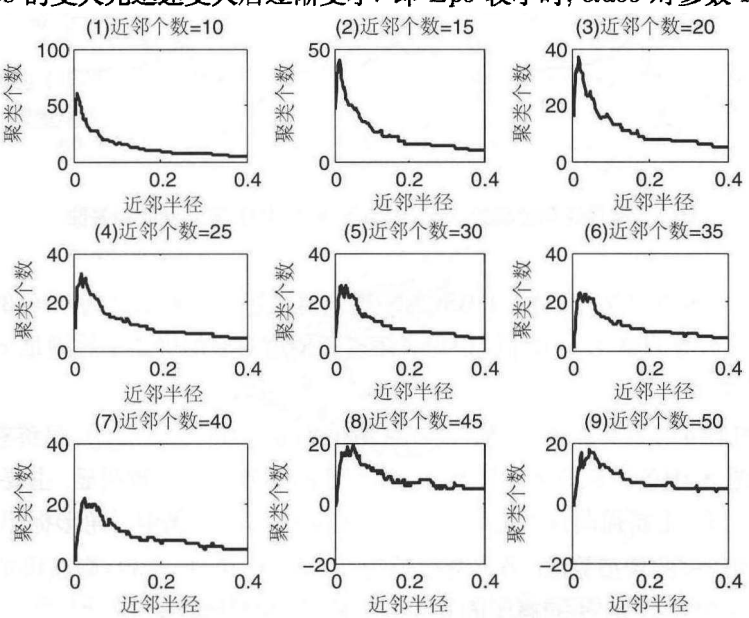

图 5　采用方向距离的 DBSCAN_GMPR 算法参数相关性

3) 以速度距离作为相似性度量函数, 则聚类分析所需参数的相关性见图 6. 给定 k, $class$ 逐渐变小. 给定 Eps, 对不同的 k 值, $class$ 值的变化较小.

4) 以三者的联合距离即结构距离作为相似性度量函数, 则聚类分析所需参数的相关性见图 7. 给定 k 值, 随着 Eps 逐渐变大, 聚类个数 $class$ 先变大后变小, 曲线形状类似对数正态分布图. 峰值左边的曲线较陡峭, 峰值后边的曲线较平缓.

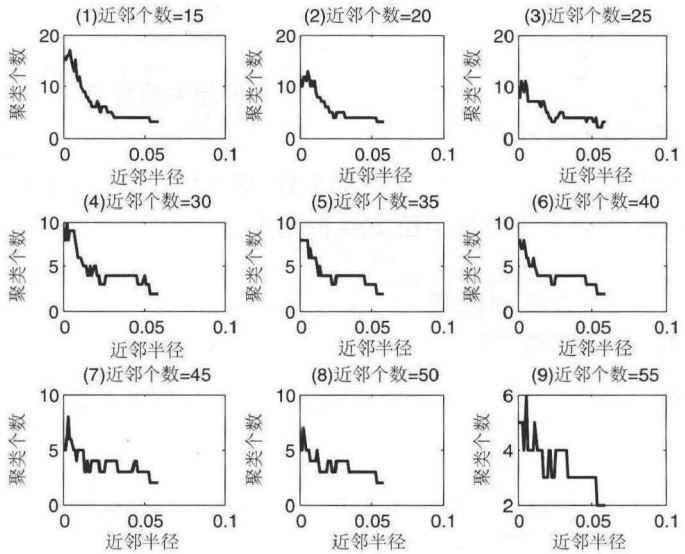

图 6 采用速度距离的 DBSCAN_GMPR 算法参数相关性

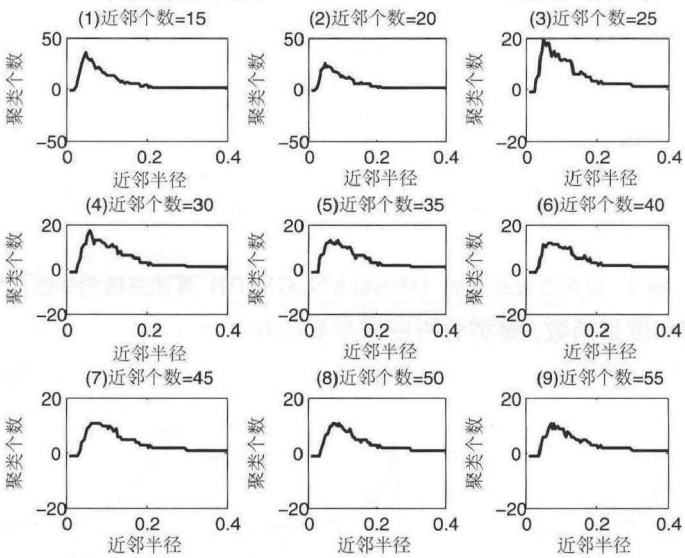

图 7 采用结构距离的 DBSCAN_GMPR 算法参数相关性

4.2 聚类结果分析

为方便与文献 [13] 中聚类结果相比较, DBSCAN 聚类算法近邻个数取值为 $k = 30$, 同时根据图 4~ 图 7 参数间的相关性, 通过适当调整 Eps 的值, 使得各距离相似度对应的聚类个数满足 $class \in [5, 8]$. 本文算法的轨迹聚类结果分析如下:

1) 以位置距离为相似度的聚类算法, 参数 k, Eps 取值为 $k = 30$, $Eps = 11$, 最终获得聚类个数 $class = 8$. 聚类结果见图 8. 图 8 中各类轨迹分布较集中, 类与类间的界限相对较明显, 主要原因是轨迹之间采用 Hausdorff 距离作为相似度, 主要强调各轨迹之间位置的相似性. 其中, 图中三角形标识的轨迹为噪声轨迹.

2) 以方向距离为相似度的聚类算法, 若参数取值为 $Eps = 0.35$, $k = 30$, 则获得聚类个数为 $class = 6$, 聚类结果见图 9. 从图 9 中可看出, 不同簇间的界限不明显, 但相对比较集中.

3) 以速度距离为相似度的聚类算法, 若参数值取为 $Eps = 0.01$, $k = 30$, 则获得聚类个数为 $class = 6$,

聚类结果见图 10. 速度相似的轨迹分布的位置非常分散. 对于同一区域的轨迹, 存在不同的聚类簇, 说明位置与方向相同的轨迹速度可能相差较大.

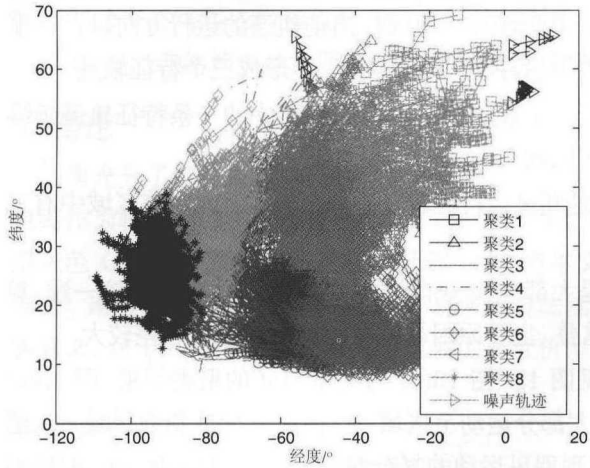

图 8　采用位置距离的 DBSCAN_GMPR 算法聚类结果

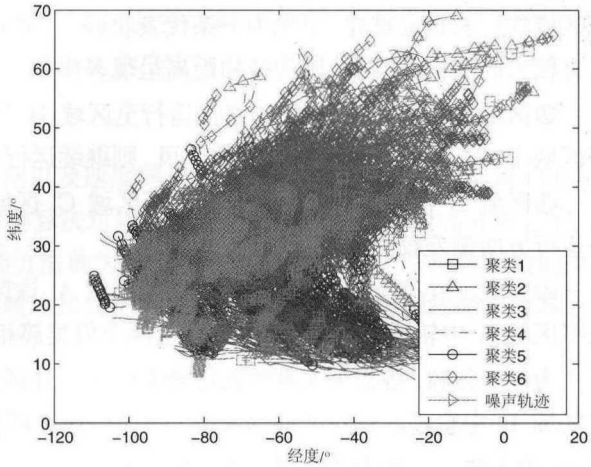

图 9　采用方向距离的 DBSCAN_GMPR 算法聚类结果

4) 通过以上分析可知, 轨迹各属性对轨迹聚类具有不同的影响. 因此若要挖掘轨迹的特征性质, 应充分考虑各属性对轨迹相似距离的影响. 所以有必要联合位置、方向与速度三个属性进行聚类分析. 以结构距离为相似性度量函数, 参数取值为 $Eps = 0.126$, $k = 30$ 时, 获得的聚类个数为 $class = 8$. 基于结构距离的聚类结果见图 11. 图 11 中相同区域可能包含两个甚至多个轨迹簇. 主要原因是同一区域的轨迹, 虽然位置距离很小, 但轨迹间方向与速度可能存在较大差别, 所以同一区域的轨迹可能分属于不同轨迹簇.

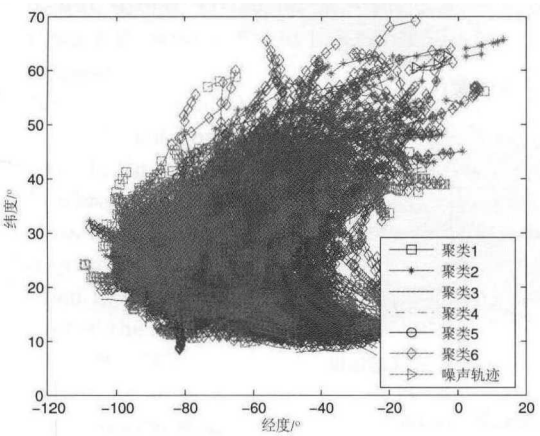

图 10　采用速度距离的 DBSCAN_GMPR 算法聚类结果

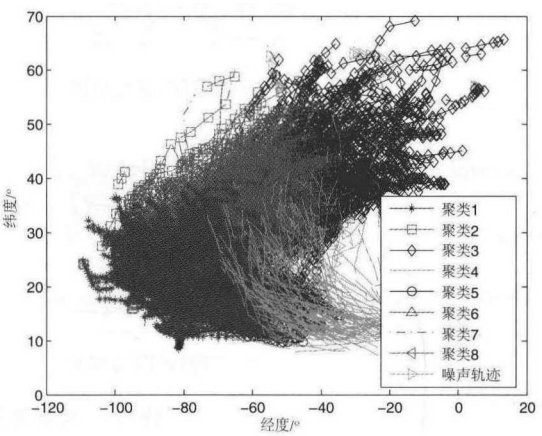

图 11　采用结构距离的 DBSCAN_GMPR 算法聚类结果

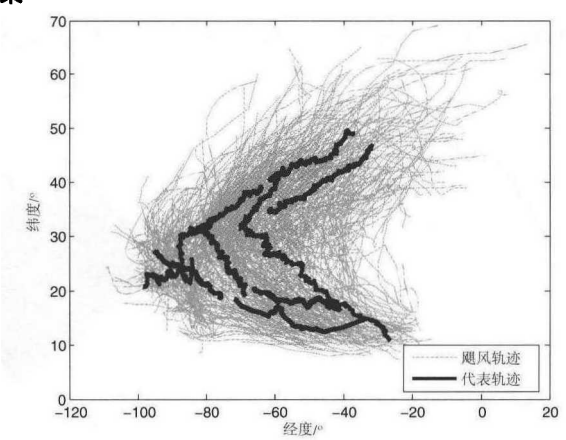

图 12　采用结构距离的 DBSCAN_GMPR 算法生成的代表轨迹

根据聚类结果生成数据集的代表轨迹见图 12. 若将轨迹图分成四个区域, 区域 D 为飓风起点区域, 区域

B、C 为飓风运动区域, 区域 A 为飓风运动末端区域.

①在区域 D 包含三个轨迹簇, 经度区间在 $[-30,-20]$ 的区域为飓风产生的主要区域, 因此不同轨迹簇在该区域的代表轨迹重叠, 呈现为一条代表轨迹. 当飓风前行靠近经度 -40 时, 当前轨迹的运行方向以及速度具有较大的变化, 不同轨迹的结构距离呈现多样化, 因此聚类后存在三个轨迹簇, 便形成三个特征轨迹.

②区域 D 中的一部分轨迹直接运行至区域 B, 同属于一个轨迹簇. 因此区域 D 中的一条特征轨迹延续至区域 B. 对于区域 B 没有消亡的飓风, 则继续运行, 转向区域 A.

③区域 D 中飓风一部分轨迹运动至区域 C. 区域 C 是飓风运行比较复杂的区域. 因此在该区域中有三条运行方向偏差较大的特征轨迹.

④区域 C 与区域 B 中部分轨迹进入区域 A, 该区域是大部分轨迹消亡的区域. 轨迹方向相对较一致. 算法将区域 A 中轨迹分为两个聚类簇, 这两个聚类簇相互重叠, 主要原因是该区域轨迹的速度相差较大.

为便于比较, 将相关文献的轨迹聚类结果一并画出, 见图 13. 图 13(a) 为文献 [15] 的聚类结果, 图 13(a) 在区域 B 中轨迹的运行方向与实际不相符, 该区域轨迹大部分运动至区域 A, 因此方向应指向区域 A. 图 13(b) 为文献 [13] 的聚类结果. 图 13(b) 的区域 C 不能体现飓风运动的复杂性. 图 13(c) 为文献 [16] 的聚类结果. 文献 [16] 在 C 区域能体现飓风运动的复杂性, 但特征轨迹主要集中在区域 B 与区域 C, 在区域 A 缺少特征轨迹的描述. 图 13(d) 为文献 [4] 的聚类结果. 文献 [4] 描述的特征轨迹起始点与实际相差甚远.

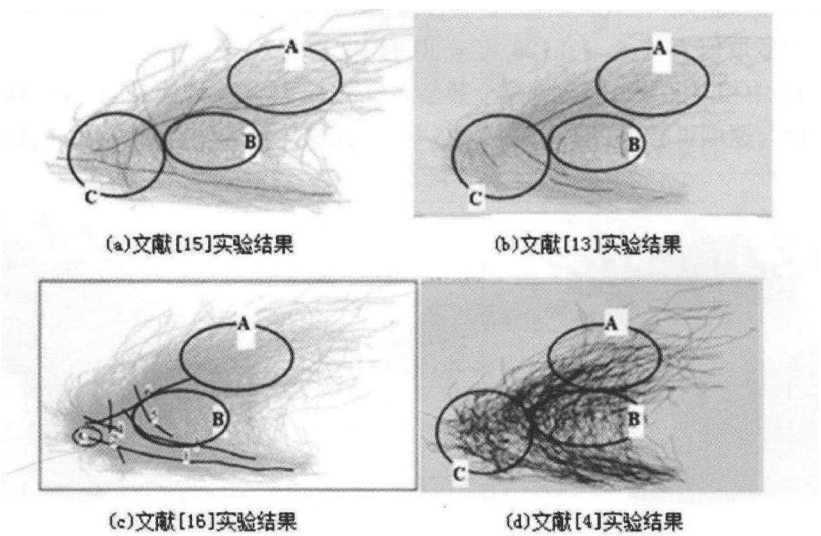

图 13 各参考文献飓风聚类结果对比图

4.3 算法精度分析

通过设置不同的参数对本文算法精度进行检验, 聚类结果平均相似度见图 14 与图 15. 在实验过程中, 当 $k=30$ 时, 聚类的总体平均相似度 $OSIM$ 随聚类个数 $class$ 的变化情况如图 14 所示. 总体而言, 随着聚

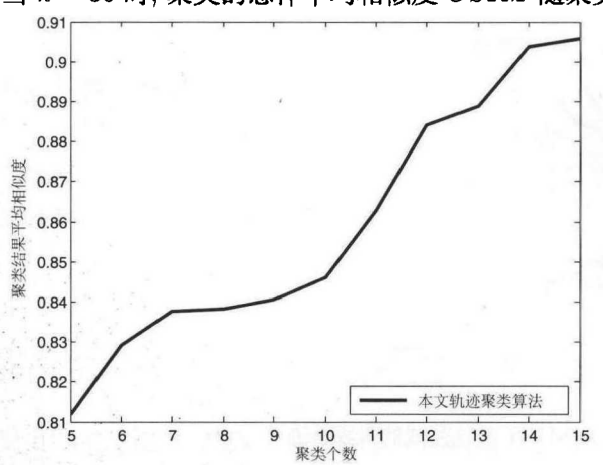

图 14 近邻个数 $k=30$ 时聚类效果随聚类个数的变化情况

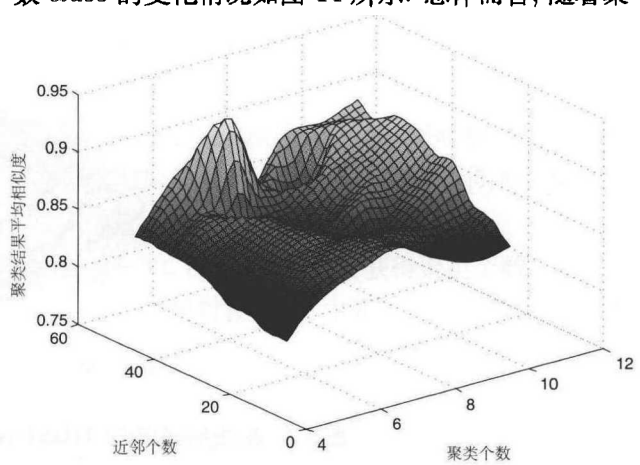

图 15 聚类效果随聚类个数与近邻个数的变化情况

类个数的增多, 各聚类间的紧凑性越大, 因此 $OSIM$ 值越大. 聚类算法采用 DBSCAN 聚类方法, 需要近邻个数以及近邻半径两个参数, 近邻半径的大小影响聚类个数, 同时影响边界轨迹的归属, 因此聚类后, 聚类个数相邻的聚类结果平均相似度会略有起伏的现象. 而近邻个数 k, 聚类个数 $class$ 以及聚类结果平均相似度 $OSIM$ 值三者的相关性如图 15 所示. 整体而言随着 k 与 $class$ 的变大, $OSIM$ 值变大.

5 结论

为充分了解轨迹运动特征, 本文将轨迹的位置、方向以及速度等三个属性距离联合成结构距离, 以结构距离作为轨迹间的相似性度量函数, 通过 DBSCAN 聚类算法对轨迹集进行分析. 轨迹在聚类分析之前, 采用灰色 GM(1,1) 模型进行分段预处理, 因此将本文聚类方法称为基于灰色 GM(1,1) 模型与结构距离的聚类算法. 通过算法有效性与精度分析, 表明该算法能够准确地描述飓风轨迹的运动特征, 得到的结果更具有现实意义, 对于天气预报等领域具有重要的研究价值.

参考文献

[1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases[C]// FODO1993: Proc of the 4th International Conference on Foundations of Data Organization and Algorithms, Chicago, Illinois, USA, 1993: 69–84.

[2] Keogh E, Palpanas T, Zordan V B, et al. Indexing large human-motion databases[C]// VLDB 2004: Proc of the 30th International Conference on Very Large Data Bases, Endowment, 2004: 780–791.

[3] Kong X Z, He W, Qin N, et al. Comparison of transport pathways and potential sources of PM_{10} in two cities around a large Chinese lake using the modified trajectory analysis[J]. Atmospheric Research, 2013, 122(3): 284–297.

[4] 陈锦阳, 宋加涛, 刘良旭, 等. 基于改进 Hausdorff 距离的轨迹聚类算法 [J]. 计算机工程, 2012, 38(17): 157–161.
Chen J Y, Song J T, Liu L X, et al. Trajectory clustering algorithm based on improved Hausdorff distance[J]. Computer Engineering, 2012, 38(17): 157–161.

[5] Junejo I N, Javed O, Shah M. Multi feature path modeling for video surveillance[C]// ICPR2004: Proc of 17th International Conference on Pattern Recognition, Cambridge, 2004: 716–719.

[6] Chen L, Ng R. On the marriage of lp-norms and edit distance[C]// VLDB2004: Proc of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, 2004: 792–803.

[7] Vries G K, Someren M V. An analysis of alignment and integral based kernels for machine learning from vessel trajectories[J]. Expert Systems with Applications, 2014, 41(16): 7596–7607.

[8] Chen L, Özsu M, Oria V. Robust and fast similarity search for moving object trajectories[C]// SIGMOD2005: Proc of the 2005 ACM SIGMOD International Conference on Management of Data, New York, USA: ACM Press, 2005: 491–502.

[9] Johnstone M, Le V T, Zhang J, et al. A dynamic time warped clustering technique for discrete event simulation-based system analysis[J]. Expert Systems with Applications, 2015, 42(21): 8078–8085.

[10] Agrawal R, Lin K I, Sawhney H S, et al. Fast similarity search in the presence of noise, scaling, and translation in time-series databases[C]// VLDB1995: Proc of the 21th International Conference on Very Large Data Bases, Zurich, 1995: 490–501.

[11] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories[C]// ICDE2002: Proc of the 18th International Conference on Data Engineering, Washington, DC, USA: IEEE Computer Society, 2002: 673–684.

[12] 刘思峰, 杨英杰. 灰色系统研究进展 (2004–2014)[J]. 南京航空航天大学学报, 2015, 47(1): 1–18.
Liu S F, Yang Y J. Advances in grey system research (2004–2014)[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2015, 47(1): 1–18.

[13] 袁冠, 夏士雄, 张磊, 等. 基于结构相似度的轨迹聚类算法 [J]. 通信学报, 2011, 32(9): 103–110.
Yuan G, Xia S X, Zhang L, et al. Trajectory clustering algorithm based on structural similarity[J]. Journal on Communications, 2011, 32(9): 103–110.

[14] 江艺羡. 基于 GM(1,1) 模型的时间序列分段表示方法 [J]. 系统工程, 2014, 32(7): 137–142.
Jiang Y X. Method of time series piecewise representation based on model GM(1,1)[J]. Systems Engineering, 2014, 32(7): 137–142.

[15] Lee J G, Han J W, Whang K Y. Trajectory clustering: A partition-and-group framework[C]// Proc of the 2007 ACM SIGMOD International Conference on Management of Data, New York, USA: ACM Press, 2007: 593–604.

[16] Zhang L, Yang G, Wang Z C, et al. Trajectory cluster based on spatial generalization[J]. Journal of Information and Computational Science, 2012, 9(2): 315–321.