DOI:10.11918/j.issn.0367-6234.201811079

危险货物道路运输车辆出行链活动类型识别

赵慧英[1], 钱大琳[1,2], 张 博[1], 范爱华[1]

(1. 北京交通大学 交通运输学院, 北京 100044;2.综合交通大数据应用技术交通运输行业重点实验室(北京交通大学), 北京 100044)

摘 要: 针对危险货物运输(危货)车辆出行链上停留节点活动类型的识别问题,提出基于高斯混合模型-隐马尔科夫模型(GMM-HMM)的活动类型识别方法. 对车辆GPS数据构建基于决策树的车辆起停检测模型,提取出行链活动节点,并通过D-OPTICS算法对活动节点聚类得到活动热区;根据活动节点的个体特征、所在出行链的亲属特征和所处热区的群体特征进行多尺度特征体系构建,通过因子分析进行降维处理;利用GMM-HMM构建危货车辆活动类型识别模型,通过Baum-Welch算法进行参数估计,并使用Viterbi算法解码隐藏状态得到出行链各活动节点的类型识别结果. 在小规模实际活动数据集上直接验证所提方法的正确性,还结合活动节点的POI类别,间接评估所提方法对大规模GPS数据的车辆活动类型识别效果. 实验结果表明:在9种活动类型识别任务中,基于GMM-HMM的出行链活动类型识别方法的活动识别率超过80%. 识别结果可用于分析活动行为模式,及时发现异常活动,为危险品运输监管提供有效的决策支持.

关键词: 危险货物运输;GPS数据;出行链活动类型;多尺度特征体系;隐马尔科夫模型;高斯混合模型

中图分类号: U492.3+36.3 **文献标志码**: A **文章编号**: 0367-6234(2019)09-0193-08

Activity type recognition of trip chain for hazmat road transportation vehicle

ZHAO Huiying[1], QIAN Dalin[1,2], ZHANG Bo[1], FAN Aihua[1]

(1.School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China; 2.Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport(Beijing Jiaotong University), Beijing 100044, China)

Abstract: This paper proposes a method based on Gaussian-Mixture-Model Hidden Markov Model (GMM-HMM) to recognize the activity-node type of trip chain for hazardous-materials (hazmat) transportation vehicle. The GPS data of vehicle were pre-processed to identify the activity nodes of the trip chains using a Decision-Tree based move-stop detection algorithm. Then the activity nodes were grouped into the activity hotspots by a dropout-based OPTICS (D-OPTICS) algorithm. A multi-scale feature system was constructed according to the individual features of the activity nodes, the relative features of the corresponding trip chains, and the group features of the related hotspots. These feature vectors were further transformed into low-dimensional vectors using Factor Analysis method. Finally, a GMM-HMM based activity type recognition model for hazmat transportation vehicles was built where Baum-Welch algorithm was used for parameter estimation, and Viterbi algorithm for decoding the hidden state to obtain the recognition results of the activity-node type of trip chains. Not only the accuracy of the proposed method was directly verified based on the small-scale real-activity dataset, but also the effectiveness of the proposed method on the activity type identification of large-scale GPS data was evaluated indirectly using the Point-of-Interest (POI) information. The results demonstrate that the identification rate of the GMM-HMM based activity type recognition method was more than 80% in the task of nine-type activity recognition. The recognition results can help analyzing the activity behavioral patterns, discovering the abnormal activities, and providing effective decision-making support for hazmat transportation supervision.

Keywords: hazardous-materials transportation; GPS data; activity type of trip chain; multi-scale feature system; Hidden Markov Model; Gaussian Mixture Model

由危险货物运输所引起的泄漏事故频繁发生,对人身财产和环境安全造成了重大威胁,相关机构亟需改善危险货物运输安全现状[1]. 目前,运输企业普遍采用GPS卫星定位系统对危险货物车辆进行监控,一定程度上提高了运输的安全性. 但此类监管只停留在车辆定位追踪层面,并没有基于轨迹数据展开分析,难以清晰地掌握车辆在途活动状态,因而无法及时发现运输过程中存在的问题并采取有效的措施. 基于此不足,需要从GPS轨迹数据中识别车辆在途活动类型来提高危险货物运输行业监管水平.

收稿日期: 2018-11-11
基金项目: 北京市交通委科研项目(T18100100);
河南省交通运输厅科技项目(2019G-2-8)
作者简介: 赵慧英(1991—),女,博士研究生;
钱大琳(1963—),女,教授,博士生导师
通信作者: 钱大琳,dlqian@bjtu.edu.cn

近年来利用 GPS 数据挖掘交通信息的研究越来越普遍. 然而直接从 GPS 数据中推断活动类型仍面临技术瓶颈. 文献[2]直接利用时间间隔来区分活动类型,文献[3]利用 GPS 数据和辅助随车日志,基于线性 SVM 算法训练了活动分类器,可以将所有停车事件分为货运停车和非货运停车两类,然而现实中难以获得包含车辆真实活动的辅助信息,活动类型的识别还需依靠无监督学习方法和模型. 文献[4]对比了层次聚类和基于划分的聚类方法,选择 Ward's 方法来识别重复访问的目的地,以期从活动地点的空间分布特征来推断车辆出行行为. 文献[5]则使用基于密度的 DBSCAN 算法来识别货车锚点,DBSCAN 算法比 Ward's 方法更能切合 GPS 数据的空间特性,但是对参数十分敏感. 且文献[4-5]只是通过聚类显示了活动节点的空间特征,仅可看作是活动类型识别的前期工作. 而文献[6]则利用经纬度乘积生成的识别码对停车事件进行简单聚类,并引入熵的概念提出了基于货车 GPS 数据的主要停车活动和次要停车活动二分类识别方法. 既有研究仅限于解决简单的货运活动和非货运活动二分类问题,没有开展系统的特征体系构建,在缺乏辅助信息的条件下对活动类型的识别准确率低,无法满足精细化监管的需要.

本文针对 GPS 轨迹数据,除考虑活动的时空特点外,还引入出行链上前后活动类型的相互制约关系,提出一种基于 GMM-HMM 的多类危险货物运输车辆活动类型识别模型,以期有效提高活动类型识别的精度和实用性,为辅助运输监管决策提供科学依据.

1 数据预处理

在识别危货车辆出行活动类型之前,需预处理所收集的 GPS 数据,以提取车辆活动节点,并根据其地理空间分布聚类活动热区.

1.1 出行链活动节点提取

令 $St = \{Tr_l \mid l = 1,2,\cdots,n_T\}$ 为待处理的 GPS 轨迹数据集,其中 n_T 为待处理轨迹总数, Tr_l 为 St 中的第 l 条轨迹,由某辆车某日的 GPS 轨迹点组成,记作 $Tr_l = \{P_i^l \mid i = 1,2,\cdots,n_P^l\}$,其中 P_i^l 为第 l 条轨迹中按照时间顺序排列的第 i 个轨迹点, n_p^l 为该轨迹的轨迹点数量. 设轨迹点 $P_i^l = (x_i^l, y_i^l, d_i^l, t_i^l, s_i^l, h_i^l, e_i^l, v_i^l, f_i^l)$, 其中 $x_i^l, y_i^l, d_i^l, t_i^l, s_i^l, h_i^l, e_i^l, v_i^l, f_i^l$ 分别为该轨迹点的经度、纬度、日期、时间、速度、方位角、引擎状态、车辆编号和首点标记(当 $i = 1$ 时为 1,否则为 0). 由于 GPS 设备可能会发生定位错误和数据漂移等情况[7-10],本文事先对原始 GPS 数据进行了清洗,以保证数据属性完整、经纬度数据符合研究区域的实际情况,数据的时间属性唯一,数据误差小于规定的阈值.

轨迹划分是推断活动节点的前提,本文基于速度变化将轨迹划分为节点. 将轨迹 Tr_l 中速度持续为零的点或持续不为零的点列 $P_\alpha^l, P_{\alpha+1}^l, \cdots, P_\beta^l (1 \leqslant \alpha \leqslant \beta \leqslant n_P^l)$ 划分为节点 N_j^l, 其中 α, β 分别为构成节点的首末轨迹点的序号. 设节点 N_j^l 的平均速度和持续时间为 $\bar{s}_j^l, \Delta t_j^l$, 则 $\bar{s}_j^l = \sum_{i=\alpha}^{\beta} s_i^l/(\beta - \alpha + 1)$, $\Delta t_j^l = t_\beta^l - t_\alpha^l$. 至此轨迹 Tr_l 可转化为节点序列 $\{N_j^l \mid j = 1,2,\cdots,n_N^l\}$, 其中 n_N^l 为节点个数. 例如,图 1 中速度持续为零且坐标相近的点 P_1^l, P_2^l, P_3^l 被划分为节点 N_1^l, 同理 $P_{13}^l - P_{16}^l$ 划分为节点 N_3^l. 而速度持续不为零的 $P_4^l - P_{12}^l$ 则被抽象成节点 N_2^l, 节点序列表示为 $\{N_1^l, N_2^l, N_3^l\}$.

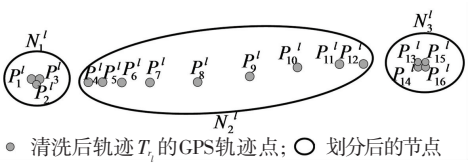

图 1 轨迹节点划分策略

Fig.1 Division strategy of trajectory nodes

节点序列理论上由行驶节点和活动(停留)节点组成,因此活动节点的识别问题实际上可转化为节点运动状态(停留/行驶)二元分类问题. 为了减少主观因素对识别的干预,本文基于决策树算法对节点的运动状态进行判断,步骤如下.

步骤 1 当 $j = 1$ 时,初始化节点 N_1^l 运动状态 ms_1^l 为"停留";

步骤 2 当 $j > 1$ 时,根据图 2 所示的分类规则依次判断节点的运动状态 ms_j^l: 对于当前节点 N_j^l, 设其前驱节点为 N_{j-1}^l, 令 $pre\_ms_j^l$ 为前驱节点运动状态,有 $pre\_ms_j^l = ms_{j-1}^l$; 将 N_j^l 的平均速度、持续时间和前驱节点的运动状态输入到图 2 的规则中,输出该节点运动状态的判断结果.

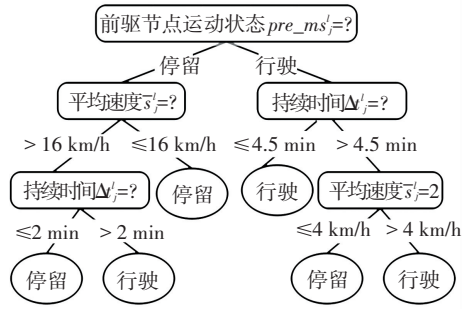

图 2 基于决策树算法的行驶/停留分类规则

Fig.2 Move-stop classification rules based on Decision Tree

步骤 3　将 ms_j^l 为“停留”的节点 N_j^l 提取为活动节点 A_k^l，此时 $A_k^l = N_j^l = \{P_\alpha^l, P_{\alpha+1}^l, \cdots, P_\beta^l\}$. 轨迹 Tr_l 上的所有活动节点组成第 l 条出行链活动序列 $Ch_l = \{A_k^l \mid k = 1,2,\cdots,n_A^l\}$，其中 n_A^l 为出行链 l 中活动节点个数.

1.2　活动热区识别

按照 1.1 的方法，处理所有的 GPS 轨迹，并将提取出的各条出行链的活动节点添加到同一个集合 Ac 中，即 $Ac = \{A_k^l \mid l = 1,2,\cdots,n_T; k = 1,2,\cdots,n_A^l\}$.

在提取群体特征之前，需将所有活动节点按照地理空间坐标聚类为活动热区. 文献[5]表明基于密度的聚类算法因其抗噪声能力强，能发现任意形状的簇等优点，在空间数据的聚类分析中比其他聚类方法更适用. OPTICS[11]是最具代表性的基于密度的聚类算法之一，它克服了文献[5]提到的 DBSCAN 算法使用全局密度易丢失簇的缺点，可以通过点排序来识别变密度聚类结构. 本文基于经典 OPTICS 算法，引入随机退出技术(Dropout)，设计了 D-OPTICS 算法对活动热区进行识别，该算法在继承 OPTICS 算法优点的同时，解决了 OPTICS 存在的两大问题.

1)最优参数确定.为确保大多数活动节点被有效聚类，在 D-OPTICS 中，根据活动节点数据集 Ac 的统计特性选择关键参数 ρ（形成团簇所需的最小点数）和 ε（邻域半径）的值. 依次令 $\rho = 2,3,\cdots,10$，对于任意活动节点 $A_k^l \in Ac$，计算出与 A_k^l 第 ρ 近的对象距离 $d_\rho(A_k^l)$，遍历 Ac 得到集合 $D_\rho = \{d_\rho(A_k^l)\}$，对 D_ρ 排序并绘制累积概率分布图，如图 3 所示，第 ρ 近邻距离累积概率分布曲线均是先急剧上升后趋于平缓的曲线. 将转折点（图 3 中空心点）的横坐标值记为 ε，计算对应于该参数组合 (ρ,ε) 的 Ac 数据集噪声比例. 取噪声比例最小时对应的参数组合作为最优参数取值[12].

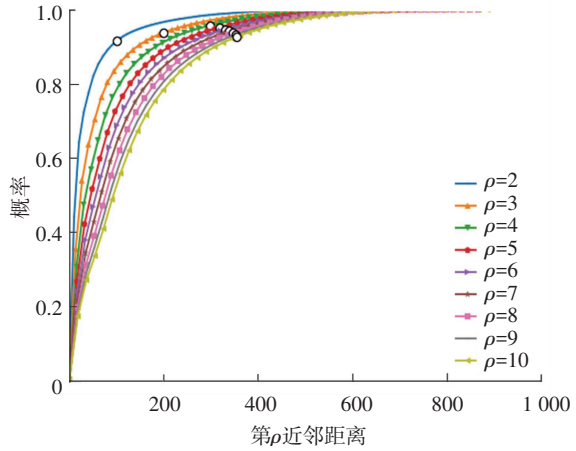

图 3　第 ρ 近邻距离累积概率分布曲线(CDF)

Fig.3　CDF of D_ρ corresponding to each $\rho = 2,3,\cdots,10$

2)避免过度聚类：D-OPTICS 算法中引入随机退出技术，使 OPTICS 有序队列中的扩展点具有随机不可连接性，可防止过度聚类，并利用高斯混合聚类对结果进行二次优化.

采用 D-OPTICS 对 Ac 中的活动节点进行聚类，得到的“类簇”即为活动热区，可表示为 H_u，设 $H_u = (\mathrm{avg}T_u, \mathrm{avg}F_u, En_u)$，其中 $\mathrm{avg}T_u, \mathrm{avg}F_u, En_u$ 分别为该热区的平均活动持续时间、平均首点率和熵指数. 若热区 H_u 包含活动节点 A_k^l，则 A_k^l 的所属热区编号 $ID_k^l = u$.

2　多尺度特征体系构建

2.1　多尺度特征

获取适合活动类型识别任务的有效特征，对正确揭示活动本质十分关键. 为了防止特征缺陷所可能导致的偏颇、歪曲、有缺陷的识别结论，需要尽可能完整且层次丰富地构建相关特征[13]. 遵循此原则，本文从活动节点自身、节点所处出行链以及节点所在热区 3 个尺度构建了个体特征、亲属特征和集群特征，图 4 展示了不同尺度下所提取的特征种类以及尺度间的关系.

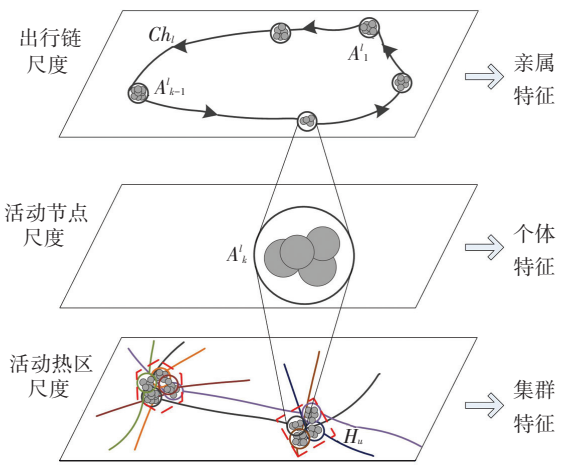

图 4　多尺度特征提取示意图

Fig.4　Illustration of multi-scale feature extraction

个体特征指的是当前所关注的活动节点 A_k^l 的个体属性，包括：所含轨迹点数 $F1_k^l$、活动开始时间 $F2_k^l$、持续时长 $F3_k^l$、结束时间 $F4_k^l$、轨迹点间平均时间间隔 $F5_k^l$、首末轨迹点距离 $F6_k^l$、轨迹点总移动里程 $F7_k^l$、轨迹点间平均距离 $F8_k^l$、迂回系数 $F9_k^l$、经纬度标准差 $F10_k^l$ 和 $F11_k^l$、各轨迹点与活动节点中心的平均距离 $F12_k^l$、最远距离 $F13_k^l$；活动节点的经纬距离 $F14_k^l$ 和 $F15_k^l$、长宽比 $F16_k^l$、面积 $F17_k^l$、点密度 $F18_k^l$；首末轨迹点方位角之差的绝对值 $F19_k^l$、方位角改变次数 $F20_k^l$、方位角改变度数均值 $F21_k^l$；平均速度 $F22_k^l$、速度增减趋势改变次数 $F23_k^l$；首末

轨迹点引擎状态 $F24_k^l$ 和 $F25_k^l$、引擎状态均值 $F26_k^l$、引擎状态改变次数 $F27_k^l$、引擎状态为熄火的轨迹点数 $F28_k^l$. 则有

$$F1_k^l = \beta - \alpha + 1, \tag{1}$$

$$F2_k^l = t_\alpha^l, \tag{2}$$

$$F3_k^l = t_\beta^l - t_\alpha^l, \tag{3}$$

$$F4_k^l = F2_k^l + F3_k^l, \tag{4}$$

$$F5_k^l = F3_k^l/(\beta - \alpha), \tag{5}$$

$$F6_k^l = \mathrm{dist}([x_\alpha^l, y_\alpha^l], [x_\beta^l, y_\beta^l]), \tag{6}$$

$$F7_k^l = \sum_{i=\alpha+1}^{\beta} \mathrm{dist}([x_{i-1}^l, y_{i-1}^l], [x_i^l, y_i^l]), \tag{7}$$

$$F8_k^l = F7_k^l/(\beta - \alpha), \tag{8}$$

$$F9_k^l = F7_k^l/F6_k^l, \tag{9}$$

$$F10_k^l = \sqrt{\sum_{i=\alpha}^{\beta} (x_i^l - \bar{x}_k^l)^2/(\beta - \alpha + 1)}, \tag{10}$$

$$F11_k^l = \sqrt{\sum_{i=\alpha}^{\beta} (y_i^l - \bar{y}_k^l)^2/(\beta - \alpha + 1)}, \tag{11}$$

$$F12_k^l = \sum_{i=\alpha}^{\beta} \mathrm{dist}([\bar{x}_k^l, \bar{y}_k^l], [x_i^l, y_i^l])/(\beta - \alpha + 1), \tag{12}$$

$$F13_k^l = \max_{\alpha \leqslant i \leqslant \beta} (\mathrm{dist}([\bar{x}_k^l, \bar{y}_k^l], [x_i^l, y_i^l])), \tag{13}$$

$$F14_k^l = \max_{\alpha \leqslant i \leqslant \beta} (x_i^l) - \min_{\alpha \leqslant i \leqslant \beta} (x_i^l), \tag{14}$$

$$F15_k^l = \max_{\alpha \leqslant i \leqslant \beta} (y_i^l) - \min_{\alpha \leqslant i \leqslant \beta} (y_i^l), \tag{15}$$

$$F16_k^l = \max(F14_k^l/F15_k^l, F15_k^l/F14_k^l), \tag{16}$$

$$F17_k^l = F14_k^l \times F15_k^l, \tag{17}$$

$$F18_k^l = F1_k^l/F17_k^l, \tag{18}$$

$$F19_k^l = \min(360 - |h_\alpha^l - h_\beta^l|, |h_\alpha^l - h_\beta^l|), \tag{19}$$

$$F20_k^l = \sum_{i=\alpha+1}^{\beta} (\max(|h_i^l - h_{i-1}^l|, 0)/|h_i^l - h_{i-1}^l|), \tag{20}$$

$$F21_k^l = F19_k^l/F20_k^l, \tag{21}$$

$$F22_k^l = \sum_{i=\alpha}^{\beta} s_i^l/(\beta - \alpha + 1), \tag{22}$$

$$F23_k^l = \sum_{i=\alpha+2}^{\beta} \frac{\min((s_i^l - s_{i-1}^l) \times (s_{i-1}^l - s_{i-2}^l), 0)}{(s_i^l - s_{i-1}^l) \times (s_{i-1}^l - s_{i-2}^l)}, \tag{23}$$

$$F24_k^l = e_\alpha^l, \tag{24}$$

$$F25_k^l = e_\beta^l, \tag{25}$$

$$F26_k^l = \sum_{i=\alpha}^{\beta} e_i^l/(\beta - \alpha + 1), \tag{26}$$

$$F27_k^l = \sum_{i=\alpha+1}^{\beta} (\max(|e_i^l - e_{i-1}^l|, 0)/|e_i^l - e_{i-1}^l|), \tag{27}$$

$$F28_k^l = \sum_{i=\alpha}^{\beta} \max(e_i^l - 4, 0). \tag{28}$$

式中 α、β 分别为构成活动节点 A_k^l 的首、末轨迹点在第 l 条轨迹上的序号；x_i^l、y_i^l、t_i^l、s_i^l、h_i^l、e_i^l 分别为第 l 条轨迹中第 i 个轨迹点的经度、纬度、时间、速度、方位角和引擎状态，而 x_α^l、y_α^l、t_α^l、s_α^l、h_α^l、e_α^l 和 x_β^l、y_β^l、t_β^l、s_β^l、h_β^l、e_β^l 分别为 $i = \alpha$ 和 $i = \beta$ 时对应的属性取值；$\bar{x}_k^l$、$\bar{y}_k^l$ 分别为 A_k^l 的中心经、纬度；dist、max、min、|·| 分别为求距离、最大值、最小值、绝对值的函数；

亲属特征围绕当前活动节点 A_k^l 在相应出行链活动序列 Ch_l 上所处的时空位置特点而展开，包括 A_k^l 所在物理位置的当日访问次数 $F29_k^l$，A_k^l 与前一活动节点 A_{k-1}^l 的时空距离 $F30_k^l$ 和 $F31_k^l$，A_k^l 与当日出行链起点 A_1^l 的时空距离 $F32_k^l$ 和 $F33_k^l$.

$$F29_k^l = \sum_{p=1}^{n_A^l} \frac{\min(\mathrm{dist}([\bar{x}_p^l, \bar{y}_p^l], [\bar{x}_k^l, \bar{y}_k^l]) - 50, 0)}{\mathrm{dist}([\bar{x}_p^l, \bar{y}_p^l], [\bar{x}_k^l, \bar{y}_k^l]) - 50}, \tag{29}$$

$$F30_k^l = F2_k^l - F2_{k-1}^l (k > 1); F30_k^l = 0 (k = 1). \tag{30}$$

$$F31_k^l = \begin{cases} \mathrm{dist}([\bar{x}_k^l, \bar{y}_k^l], [\bar{x}_{k-1}^l, \bar{y}_{k-1}^l]), & k > 1; \\ 0, & k = 1. \end{cases} \tag{31}$$

$$F32_k^l = F2_k^l - F2_1^l, \tag{32}$$

$$F33_k^l = \mathrm{dist}([\bar{x}_k^l, \bar{y}_k^l], [\bar{x}_1^l, \bar{y}_1^l]). \tag{33}$$

式中 $\bar{x}_p^l$、$\bar{y}_p^l$ 分别为 A_k^l 在同一出行链上的第 p 个活动节点的中心经、纬度.

集群特征刻画的是与 A_k^l 处于同一热区 H_u 内的所有活动节点所呈现的共性特征，提取为所在热区的平均活动持续时间 $F34_k^l$、平均起点包含率 $F35_k^l$ 和混合熵指数 $F36_k^l$.

$$F34_k^l = \mathrm{avg}T_u, \tag{34}$$

$$F35_k^l = \mathrm{avg}F_u, \tag{35}$$

$$F36_k^l = En_u = -\left(\sum_{c=1}^{C}\left(\left(\frac{n_c^u}{n_A^u}\right)\ln\left(\frac{n_c^u}{n_A^u}\right)\right) + \sum_{g=1}^{G}\left(\left(\frac{n_g^u}{n_A^u}\right)\ln\left(\frac{n_g^u}{n_A^u}\right)\right)\right). \tag{36}$$

式中 n_A^u 为热区 H_u 所含活动节点总数，n_c^u 为运营商 c 的车辆在该热区内产生的活动节点数，C 为在该热区停留车辆涉及的运营商总数，n_g^u 为在该热区运输 g 类货物的活动节点数，G 为该热区运输的所有货物种类数.

2.2 特征降维

为减少原始多维特征空间中冗余信息所造成的误差，需要对数据进行降维. 考虑到所用的数据缺乏标签，选用无监督降维方法. 主成分分析(PCA)和因子分析(FA)是目前最常用的无监督降维方法，

本文比较了分别使用 PCA 和 FA 方法将数据降到不同维度时的平均对数似然得分，对数似然值越大，表明效果越好. 由图 5 可知，随着维数的增加，两种方法的对数似然值均逐渐上升，但 FA 的降维效果始终优于 PCA；且降维到 20 维时，FA 的得分已接近收敛. 因此选用 FA 方法变换得到的 20 维特征向量作为后续模型的特征输入.

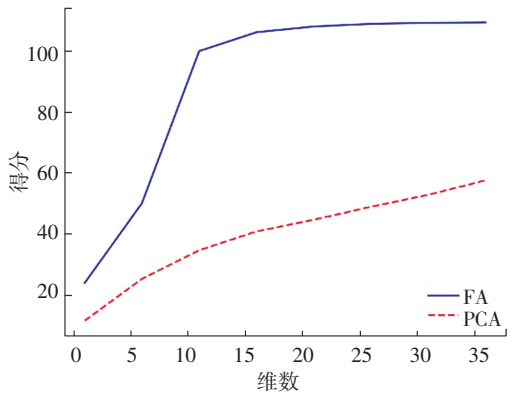

图 5　无监督降维方法比较

Fig.5　Comparison of unsupervised dimensionality reduction methods

3　活动类型识别

3.1　模型选择

隐马尔科夫模型(HMM)[14]作为序列数据处理和统计学习的一种重要概率模型，具有建模简单，物理意义明确等优点. 考虑到所构建的特征为多维连续特征，可假定每维特征均服从混合高斯分布. 因此本文选用基于混合高斯模型的隐马尔科夫模型(GMM-HMM)识别车辆出行链的活动类型.

识别模型定义如下：所有可能的活动类型集合为 $S=\{S_1,S_2,\cdots,S_N\}$，初始活动类型的概率向量为 $\boldsymbol{\pi}=(\pi_i),i=1,2,\cdots,N$，$N$ 为可能的活动类型数，$\pi_i=P(i_1=S_i)$；状态转移概率矩阵 $\boldsymbol{A}=[a_{ij}]_{N\times N}$，$a_{ij}=P(i_{k+1}=S_j\mid i_k=S_i),i=1,2,\cdots,N;j=1,2,\cdots,N$；$\boldsymbol{\pi}$ 和 $\boldsymbol{A}$ 决定出行链活动序列 Ch_l 内各活动的类型，记为活动类型序列 $Iq=(i_1,\cdots,i_k,\cdots,i_{n_A})$，$n_A$ 为待观测活动个数；各活动类型对应的可观测特征序列为 $Oq=(o_1,\cdots,o_k,\cdots,o_{n_A})$，由观测概率矩阵 $\boldsymbol{B}$ 来决定，且 $o_k=(F_1,F_2,\cdots,F_{n_F})^{\mathrm{T}}$，其中 F 为可观测特征变量，n_F 为特征变量个数，根据降维结果取 $n_F=20$；特别地，因本研究中提取的特征变量是连续值，$\boldsymbol{B}$ 无法用常规矩阵的形式表示，因此以混合高斯模型(GMM)来拟合 $\boldsymbol{B}$，即令 $\boldsymbol{B}=[b_j(o)]_{n\times 1}$，$b_j(o)=\sum_{m=1}^{M}\omega_{jm}b_{jm}(o)=\sum_{m=1}^{M}\omega_{jm}(2\pi)^{-n_F/2}\mid\Sigma_{jm}\mid^{-0.5}\times\exp\{-0.5(o-\mu_{jm})^{\mathrm{T}}\Sigma_{jm}^{-1}(o-\mu_{jm})\}$，其中 o 为特征向量可取的一组数值，ω_{jm} 为混合系数，M 为高斯个数，Σ_{jm} 为协方差矩阵，$|\Sigma_{jm}|$ 为协方差行列式，μ_{jm} 为特征均值向量，$(o-\mu_{jm})^{\mathrm{T}}$ 为 $(o-\mu_{jm})$ 的转置；将 $\boldsymbol{B}$ 输入到连续 HMM 模型中，最终建立的 GMM-HMM 模型为 $\lambda=(\pi,\boldsymbol{A},\boldsymbol{B})$.

3.2　参数估计

Baum-Welch 算法是一种广泛使用的方法来估计 HMM 的参数，算法步骤如下.

步骤 1　初始化 GMM-HMM 模型参数. 对于活动识别模型，参数的初值设定尤为重要，因为危险货物运输过程涉及的车辆车型复杂，承载的介质种类繁多，不同的运输出行链无法共享相同的初始参数矩阵. 本研究提出了如下 HMM 模型初始化思路：

1)采用岭回归方法初始化参数 $\boldsymbol{\pi}$ 和 $\boldsymbol{A}$：以调查收集的带活动类型标签的小规模出行链数据作为训练集，为每条出行链提取 V 维属性向量 $\boldsymbol{Z}=(z_1,z_2,\cdots,z_V)^{\mathrm{T}}$，如出行链长度、总耗时、活动总时长等；用岭回归方法训练一个带有系数矩阵 $\boldsymbol{w}$ 的线性模型来表征出行链属性和 HMM 参数 $(\pi,\boldsymbol{A})$ 之间的关系；对于活动类型未知且属性矩阵为 $\dot{\boldsymbol{Z}}$ 的任一出行链，则可按照系数矩阵 w 预测得到参数 $(\pi,\boldsymbol{A})$ 的初始值 $(\pi^{(0)},\boldsymbol{A}^{(0)})=\dot{\boldsymbol{Z}}\boldsymbol{w}$.

2)采用 K-Means 聚类[15]方法将训练集中的数据分为 N 类，计算均值及协方差矩阵以初始化参数 $\boldsymbol{B}$，记作 $\boldsymbol{B}^{(0)}$. 至此，模型参数可初始化为 $\lambda^{(0)}=(\pi^{(0)},\boldsymbol{A}^{(0)},\boldsymbol{B}^{(0)})$，令 $n=0$ 开始迭代.

步骤 2　利用 EM 算法迭代学习参数. 根据当前参数估计值 $\lambda^{(n)}$ 计算 Q 函数：

$$\begin{aligned}Q(\lambda,\lambda^{(n)})=&\sum_{Iq}\log\pi_{i_1}P(Oq,Iq\mid\lambda^{(n)})+\\&\sum_{Iq}\Big(\sum_{k=1}^{n_A}\log a_{i_k\,i_{k+1}}\Big)P(Oq,Iq\mid\lambda^{(n)})+\\&\sum_{Iq}\Big(\sum_{k=1}^{n_A}\log b_{i_k}(o_k)\Big)P(Oq,Iq\mid\lambda^{(n)}).\end{aligned}\tag{37}$$

然后分别极大化式(37)的各项，以确定模型参数的新估计值 $\lambda^{(n+1)}=\arg\max_{\lambda}Q(\lambda,\lambda^{(n)})$.

步骤 3　重复步骤 2，直到收敛，得到模型参数 $\lambda^{(n+1)}=(\pi^{(n+1)},\boldsymbol{A}^{(n+1)},\boldsymbol{B}^{(n+1)})$.

3.3　活动解码

在学习好的参数基础上，使用 Viterbi 算法求取最佳活动类型序列. 维特比算法实际是用动态规划求概率最大的出行链活动类型序列. 定义第 k 个活动的类型为 S_i 观测到特征 o_k 的概率为 $\delta_k(i)$，使 $\delta_k(i)$ 最大的第 $k-1$ 个活动的类型为 $\psi_k(i)$. 步骤如下：当 $k=1$ 时，初始化 $\delta_1(i)=\pi_i b_i(o_1),\psi_1(i)=0$，其中 $i=1,2,\cdots,N$；对 $k=2,3,\cdots,n_A$，依次令 $\delta_k(i)=\max_{1\leqslant j\leqslant N}[\delta_{k-1}(j)a_{ij}]b_i(o_k),\psi_k(i)=\arg\max_{1\leqslant j\leqslant N}[\delta_{k-1}(j)a_{ji}]$；

最终 $i_{n_A}^* = \arg\max_{1\leqslant j\leqslant N}\delta_{n_A}(i)$，回溯最优路径，可知对 $k = n_A - 1, n_A - 2, \cdots, 1$，有 $i_k^* = \psi_{k+1}(i_{k+1}^*)$，求得最优活动类型序列为：$Iq^* = (i_1^*, i_2^*, \cdots, i_{n_A}^*)$.

4 实验结果和分析

本文获取了 2015 年 11 月 1 日至 11 月 7 日期间，属于 108 家企业共计 1115 辆危险品车辆的大规模 GPS 数据，约有 400 万条记录. 另外还收集了相应车辆的属性数据，这些车辆的可搭载介质主要涉及第 1、2、3、4、6、8 类危险货物，代表爆炸品、压缩气体、易燃液体、易燃固体、有毒物质和腐蚀品[16]. 按照第 1 节的方法预处理 GPS 数据，最终识别出 26 176个活动节点，聚类得到 1 609 个活动热区，形成了 3 358 条出行链. 出行链活动数的概率分布如图 6，日均活动数为 8.6 个. 为了更好地初始化 GMM-HMM 模型参数，调查了其中的 182 条出行链数据的真实活动类型标签（称为小规模数据）. 调查可知出行链包含装载（L）、卸载（U）、归场（Y）、交通堵塞（J）、信号灯等待（T）、途中休息（R）、夜间住宿（H）、手续办理（O）和其他（E）九类活动.

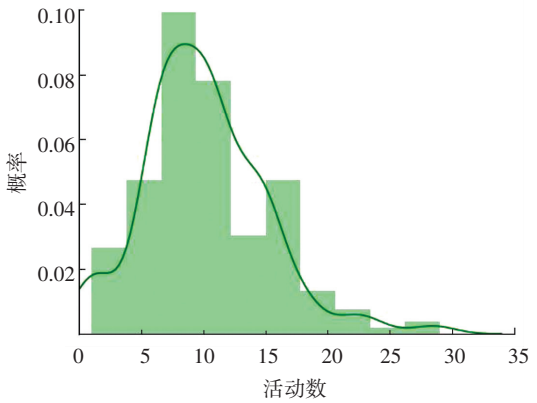

图 6 平均每日活动数的概率分布（PD）

Fig.6 PD of the average number of daily activities

利用式（1）~（36），对 26 176 个活动节点进行多尺度特征构建，并用 FA 方法降维得到 20 维特征向量. 设 HMM 模型的隐藏状态数 $N = 9$，在活动类型识别之前，先用第 3.1 节方法，以 60% 的小规模出行链数据初始化模型参数. 在此基础上，为大规模数据集中的每条出行链训练一个 GMM-HMM 模型推测活动序列，最终识别出所有出行链上活动节点的类型.

由于缺少大规模出行链数据的真实活动信息，因此本文一方面使用剩余 40% 的小规模出行链数据的真实活动标签来评估活动识别准确性；另一方面，利用 POI 信息间接评估大规模活动的识别效果.

4.1 直接评估

利用式（38）所定义的评估指标 $F1$ 分值来直接度量危险货物运输车辆的活动类型识别模型在小规模样本集上的性能，$F1$ 是对召回率 γ 和精确率 σ 的折中. 一般来说，$F1$ 分值越高，认为识别率越高.

$$F1 = 2(\sigma \times \gamma)/(\sigma + \gamma). \tag{38}$$

对于单类活动 $S_j \in \{L, U, \cdots, E\}$，$\gamma_j = TP_j/(TP_j + FN_j)$，$\sigma_j = TP_j/(TP_j + FP_j)$，其中 TP_j，FN_j 分别为活动类型 S_j 被正确识别和被错误识别的数量；FP_j 为其他活动类型被错误识别为 S_j 的数量；而对于整体活动类型（ALL），$\gamma_{\mathrm{ALL}} = \sum_{j=1}^{N} w_j \cdot \gamma_j$，$\sigma_{\mathrm{ALL}} = \sum_{j=1}^{N} w_j \cdot \sigma_j$，其中 w_j 为活动类型 S_j 在所有活动的样本占比.

对比调查所得的小规模真实活动数据与本文方法得到的活动类型识别结果，计算本文方法的性能对单一活动和整体活动的识别性能，结果见表 1. 从表 1 可以看出，本文方法对危险货物运输车辆不同类别活动的识别表现差异较大，$F1$ 分值最小低至 0.48，最大可达 0.88，识别错误偏向于样本占比较少的类别，存在类别不均衡现象：

表 1 本文识别方法的性能度量

Tab.1 Performance scores of the devised recognition method

活动类型	样本占比	召回率	精确率	$F1$
L	0.42	0.83	0.95	0.88
U	0.16	0.76	0.70	0.73
Y	0.10	0.79	0.70	0.74
J	0.05	0.62	0.50	0.56
T	0.13	0.80	0.67	0.73
R	0.04	0.59	0.53	0.56
H	0.03	0.46	0.50	0.48
O	0.05	0.74	0.65	0.70
E	0.03	0.38	0.99	0.55
ALL	1.00	0.76	0.79	0.77

一方面，装载（L）、卸载（U）、归场（Y）等货运相关活动类型的 $F1$ 分值普遍高于非货运活动类型（J、T、R、H、O、E）的 $F1$ 分值，这是因为货车出行主要是为了开展货运相关活动，此类活动发生频繁、特点突出，因此数据样本丰富且彼此的特征差异明显，识别效果较好；而另一方面，针对非货运相关活动，对信号灯等待（T）和手续办理（O）活动的识别效果良好；而对交通堵塞（J）和途中休息（R）、夜间住宿（H）和其他（E）活动识别效果不佳. 这是由于这些活动在运输途中不常发生，数据样本占比太小，导致所包含的特征过少，很难从中提取规律，活动类型无法被正确识别.

解决样本分布不均衡问题主要是通过某一途径来使不同类别的样本数据记录均衡. 如果条件支持，扩充小类数据集是最简单且高效的方法[17]. 进一步分析可知，识别效果较差的 4 类活动 J、R、H、E 更容易产生于长途运输过程中. 因此为了提高这四种活动类型的样本量，添加长途运输车辆的出行链数据至小规模活动数据集来优化样本分布. 优化前后的识别效果指标对比见表 2，调整样本分布后本文方法对各类活动的识别性能有所提升，识别效果稳定；且对所有活动类型的总体识别效果度量指标 $F1$ 分值由原来 0.77 提高到了 0.84，换言之，活动识别率超过 80%.

表 2　数据补充前后的模型效果对比

Tab.2　Comparison of model effects before and after data supplementation

活动类型	原样本占比	新样本占比	原 $F1$	新 $F1$
L	0.42	0.30	0.88	0.90
U	0.16	0.17	0.73	0.83
Y	0.10	0.10	0.74	0.84
J	0.05	0.07	0.56	0.72
T	0.13	0.13	0.73	0.86
R	0.04	0.06	0.56	0.76
H	0.03	0.05	0.48	0.74
O	0.05	0.07	0.70	0.77
E	0.03	0.05	0.55	0.73
ALL	1.00	1.00	0.77	0.84

4.2　间接评估

兴趣点(POI)数据库通常可以从开源地图或商业 API 中获得，由于 POI 数据描述的是地点周围的 POI 类别标签，可以反映地点和潜在的活动类型之间的相关性[18]，因此研究利用从百度地图 API 爬取的 POI 数据来间接验证本文方法对大规模出行链活动的识别效果，具体步骤如下：

首先设定 9 种活动类型对应的 POI 类别名称：活动类型是 L/U/Y 时，对应的 POI 类别可为公司企业、加油站(3 类适用)、加气站(2 类适用)、餐饮(2 类适用)、医疗和科研机构(2、8 类适用)；当活动类型为 J/T 时，POI 类别为交通设施；当活动类型为 R、H 时，POI 类别分别对应于餐饮、服务区；而活动类型为 O，POI 类别为公司企业或政府机构；活动类型为 E 时，POI 类别应为汽车服务或加油站(3 类除外). 使用本文方法对大规模出行链节点的活动类型进行识别，并为每一节点的活动类型识别结果匹配潜在 POI 类别名称. 然后根据经纬度信息为每一活动节点匹配其 50 米范围内从百度地图 API 爬取到的所有 POI 类别中标记出现次数最多 POI 类别名称. 若潜在 POI 类别名称与出现次数最多 POI 名称的交集不为空，则认为该节点的活动类型被成功识别. 对于大规模出行链数据，活动识别率定义为活动类型被成功识别的节点数占出行链总节点数的百分比，最终间接验证结果表明本方法可有效识别出大规模出行链数据中的活动类型，活动识别率(89.1%)在 80%以上.

对识别出的各类活动开始时间进行分析，如图 7 所示，“装载(L)”活动的典型开始时间呈现多峰值，分别在凌晨、上午和下午；“卸载(U)”的开始时间多数处于午后，峰值接近中午时分；“归场(Y)”活动有明显早出晚归的规律；“交通堵塞(J)”则主要发生于早高峰时段；白天全天均存在“信号灯等待(T)”；“途中休息(R)”集中于午间时段；“夜间住宿(H)”多数开始于午夜或是凌晨；“手续办理(O)”的开始时间与货运活动(L、U、Y)的开始时间关系密切；而“其他(E)”活动开始时间比较分散. 上述开始时间的分布符合实际情况，因此本文方法的识别结果可靠性高，可以进一步服务于运输安全监管工作.

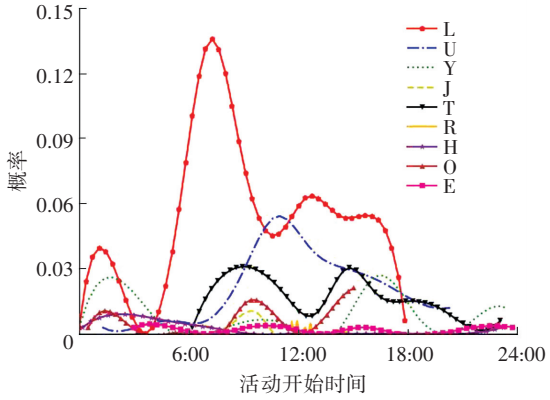

图 7　活动开始时间概率密度曲线(PDF)

Fig.7　PDF of activity start time

5　结　论

1)对危险货物运输车辆的 GPS 数据应用决策树起停分类模型提取出行链活动节点，设计 D-OPTICS 算法将活动节点快速聚类生成活动热区；根据个体、亲属和群体多尺度进行特征体系构建，充分考虑了车辆活动的时空特点. 在此基础上，应用 GMM-HMM 模型以超过 80%的识别率区分九类车辆活动.

2)不同于以前研究中货运/非货运活动的两类识别结果，多类识别结果可用于确定车辆在途装载状态，分析货运表现，识别异常活动等研究中，对于

促进危险货物运输安全有较强指导意义. 下一步研究可从以下两方面着手:基于 GMM-HMM 的活动类型识别方法的识别率与 HMM 模型的初始参数关系密切,需要进一步收集样本,优化参数矩阵,从而避免过拟合;还需结合轨迹模式挖掘技术对各类活动的规律进行深入剖析.

参考文献

[1] 王姝, 周侗, 孙健.基于时空数据分析的危化品车辆自动预警方案研究[J].物流工程与管理,2014,36(12):124
WANG Shu, ZHOU Tong, SUN Jian. Research on development of automatic warning system oriented vehicle conveyed dangerous goods based on spatio-temporal data analysis [J]. Logistics Engineering and Management, 2014,36(12): 124. DOI: 10.3969/j.issn.1674-4993.2014.12.051

[2] HESS S, QUDDUS M, RIESER-SCHUSSLER N, et al. Developing advanced route choice models for heavy goods vehicles using GPS data[J]. Transportation Research Part E: Logistics and Transportation Review,2015,77: 29.DOI:10.1016/j.tre.2015.01.010

[3] YANG Xia, SUN Zhanbo, BAN Xuegang, et al. Urban freight delivery stop identification with GPS data [J]. Transportation Research Record, 2014, 2411(2411):55.DOI:10.3141/2411-07

[4] SHARMAN B W, ROORDA M J. Analysis of freight global positioning system data: clustering approach for identifying trip destinations [J]. Transportation Research Record: Journal of the Transportation Research Board,2011,2246: 83.DOI: 10.3141/2246-11

[5] MA Xiaolei, WANG Yong, MCCORMACK E, et al. Understanding freight trip-chaining behavior using a spatial data-mining approach with GPS data[J]. Transportation Research Record: Journal of the Transportation Research Board,2016,2596: 44.DOI: 10.3141/2596-06

[6] GINGERICH K, MAOH H, ANDERSON W. Classifying the purpose of stopped truck events: an application of entropy to GPS data [J]. Transportation Research Part C Emerging Technologies,2016, 64: 17.DOI:10.1016/j.trc.2016.01.002

[7] TSUI S Y A, SHALABY A S. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems[J]. Transportation Research Record,2006,1972: 38.DOI: 10.3141/1972-07

[8] BOHTE W, MAAT K. Deriving and validating trip destinations and modes for multi-day GPS based travel surveys: an application in the Netherlands [J]. Transportation Research Board Meeting, 2008, 7(31): 813

[9] PLUVINETA P, GONZALEZ-FELIUA J, AMBROSINIA C. GPS data analysis for understanding urban goods movement[J]. Procedia-Social and Behavioral Sciences, 2012, 39: 450. DOI: 10.1016/j.sbspro.2012.03.121

[10] FU Yanhong, SHI Xiaofan. Research on freight truck operation characteristics based on GPS data[J].Procedia-Social and Behavioral Sciences,2013,96: 2320.DOI:10.1016/j.sbspro.2013.08.261

[11] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering points to identify the clustering structure[C]// Proceedings of the 1999 ACM SIGMOD international conference on Management of data.New York:ACM,1999:49.DOI:10.1145/304182.304187

[12] 夏鲁宁, 荆继武. SA-DBSCAN:一种自适应基于密度聚类算法[J]. 中国科学院大学学报, 2009, 26(4):530
XIA Luning, JING Jiwu.SA-DBSCAN: a self-adaptive density-based clustering algorithm[J].Journal of the Graduate School of the Chinese Academy of Sciences, 2009,26(4):530

[13] 杨东援.城市居民空间活动研究中大数据与复杂性理论的融合[J].城市规划学刊,2017(2):31
YANG Dongyuan. The integration of big data and complexity theory in the study of residents' activity space[J].Urban Planning Forum, 2017(2):31. DOI: 10.16361/j.upf.201702003

[14] YIN Mogeng,SHEEHAN M,FEYGIN S,et al.A generative model of urban activities from cellular data[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(6): 1682. DOI:10.1109/TITS.2017.2695438

[15] 王为凯.基于 GMM-HMM 的声学模型训练研究[D].广州:华南理工大学,2016
WANG Weikai. Research of the GMM-HMM based acoustic models [D].Guangzhou: South China University of Technology,2016

[16] 危险货物分类和品名编号:GB 6944—2012 [S]. 北京: 中国标准出版社, 2012
Classification and code of dangerous goods: GB 6944—2012 [S]. Beijing: China Standards Press, 2012

[17] 宋天龙. Python 数据分析与数据化运营[M]. 北京:机械工业出版社, 2017:100
SONG Tianlong. Data analysis and operations based on Python[M]. Beijing: China Machine Press,2017:100

[18] CHEN Chao, JIAO Shuhai, ZHANG Shu, et al. TripImputor: real-time imputing taxi trip purpose leveraging multi-sourced urban data [J]. IEEE Transactions on Intelligent Transportation Systems, 2018,19(10): 3292. DOI:10.1109/TITS.2017.2771231

(编辑 魏希柱)