**中图分类号：TP391**

**论文编号：10006ZY1706130**



硕士学位论文

**基于多源位置数据的人员特殊聚居区域识别模型研究与实现**

|  |  |
| --- | --- |
| 作者姓名 | 许伟 |
| 学科专业 | 计算机技术 |
| 指导教师 | 诸彤宇 副教授 |
| 培养院系 | 计算机学院 |

**The Research and Implementation of Personnel Special Area Recognition Model Based on Multi-source Location Data**

A Dissertation Submitted for the Degree of Master

**Candidate: Xu Wei**

**Supervisor: Associate Prof. Zhu Tongyu**

School of Computer Science & Engineering

Beihang University, Beijing, China

**中图分类号：TP391**

**论文编号：10006ZY1706130**

硕 士 学 位 论 文

基于多源位置数据的人员特殊聚居区域识别模型研究与实现

作者姓名 许伟 申请学位级别 工程硕士

指导教师姓名 诸彤宇 职 称 副教授

学科专业 计算机技术 研究方向 智能交通

学习时间自 2017年9月1日 起至 2019年11月27日止

论文提交日期 2019年11月27日 论文答辩日期 2019年12月10日

学位授予单位 北京航空航天大学 学位授予日期 年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：        日期： 年 月 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

# **摘 要**

随着城市化进程的不断推进，一些职住功能混合区域时有出现，这些区域内人群的社会属性较为特殊，且人员流动性低，人口密集，具有较为严重的安全隐患。交通信息化的飞速发展，产生了大量的人群位置数据，比如手机信令数据、共享单车数据、公交数据等。这些位置数据中蕴含了人群活动规律、生活习性等丰富的语义信息。对这些数据进行分析，挖掘人群的出行特征，利用出行特征辨别目标人群的社会属性，识别目标人群的聚居位置对相关部门进行预警防范工作具有重要的意义。

手机信令数据含有丰富的人群时空信息，具有覆盖用户面广，数据量大等优点。公共交通、共享单车作为大部分人群的出行选择，其生成的位置数据也蕴含了丰富的人群出行信息。本文设计并实现基于多源位置数据的人员特殊聚居区域识别模型，通过对手机信令数据进行预处理，提取个体出行链与职住地，实现个体出行特征的提取，并完成目标人群的社会属性辨别，最后完成目标人群的聚居区域提取，并结合公交数据与共享单车数据分析区域特征。本文取得的主要成果如下：

1. 通过对手机信令数据进行质量分析，证明了信令数据可以本文的后续分析提供有效的数据支撑。对信令数据进行预处理，剔除长距离基站抖动的数据，筛选出优质用户的数据。
2. 针对信令数据具有时空信息的特点，引入ST-DBSCAN算法对用户信令数据进行聚类分析，辨别用户的停留态/移动态，完成了用户出行链的提取，然后基于用户多日的停留点数据，完成了用户的职住地提取。
3. 基于用户出行链数据与职住地数据，采用LCSS算法计算用户出行轨迹相似性、出行距离等出行特征。根据人工标注的训练集，构建基于随机森林算法的预测模型，完成了个体社会属性的估计。
4. 根据个体社会属性的预测结果，将具有相同属性的人群的居住地聚类形成区域，并基于公交数据、共享单车数据计算区域的客流特征，采用K-means算法分析区域特征。

**关键字：**手机信令数据，出行特征，出行链，个人社会属性，区域特征

# **Abstract**

With the continuous advancement of urbanization, some mixed areas of occupational and residential functions appear from time to time. The social attributes of the people in these areas are special, the mobility of people is low, the population is dense, and there are serious security risks. The rapid development of traffic informatization has generated a large amount of crowd location data, such as mobile phone signaling data, shared bicycle data, and bus data. These location data contain rich semantic information such as crowd movement rules and living habits. Analyzing these data, digging out the travel characteristics of the crowd, using the travel characteristics to identify the social attributes of the target population, and identifying the location of the target population are of great significance to the relevant departments for early warning and prevention.

The mobile phone signaling data contains rich time and space information of the crowd, and has the advantages of covering a wide range of users and a large amount of data. Public transportation and shared bicycles are the travel choices of most people, and the generated location data also contains rich travel information. This paper designs and implements the special population area identification model based on multi-source location data, pre-processes the mobile phone signaling data, extracts individual travel chains and occupational residences, realizes the extraction of individual travel characteristics, and completes the social attribute identification of the target population. Finally, the population of the target population is extracted, and the regional characteristics are analyzed by combining the bus data with the shared bicycle data.

The main results achieved in this paper are as follows:

1. Through the quality analysis of the mobile phone signaling data, it is proved that the signaling data can provide effective data support for the subsequent analysis of this paper. Pre-processing the signaling data, eliminating the data of long-distance base station jitter, and filtering out the data of high-quality users.
2. According to the characteristics of spatio-temporal information of signaling

data, ST-DBSCAN algorithm is introduced to cluster analysis of user signaling data to distinguish the user's stay state/moving state, complete the user's travel chain extraction, and then the time period of the stay point. , completed the user's residence registration.

1. Based on the user travel chain data and occupational residence data, the LCSS algorithm is used to calculate the travel characteristics such as user travel trajectory similarity and travel distance. Based on the manually labeled training set, the random forest prediction model is constructed and the estimation of individual social attributes is completed.
2. According to the prediction results of individual social attributes, the residences of the people with the same attributes are clustered to form areas, and the K-means algorithm is used to analyze the regional features based on the bus data and shared bicycle data.

**Keywords:** cellular network data, travel characteristics, travel chain, personal social attributes, regional characteristics

**目 录**

[摘 要 I](#_Toc20736)

[Abstract II](#_Toc11430)

[第一章 绪论 1](#_Toc20702)

[1.1 研究背景及意义 1](#_Toc6682)

[1.2 课题来源 2](#_Toc29965)

[1.3 研究目标及研究内容 2](#_Toc16563)

[1.4 本文的组织结构 3](#_Toc1778)

[第二章 研究现状与相关技术 5](#_Toc15508)

[2.1 社会属性与出行特征相关研究 5](#_Toc14016)

[2.1.1 用户画像 5](#_Toc6773)

[2.1.2 特殊人群出行研究 6](#_Toc20056)

[2.2 手机信令数据的数据特征 7](#_Toc21656)

[2.2.1 手机信令数据的生成原理 7](#_Toc24403)

[2.2.2 手机信令数据的时间精度特征 8](#_Toc7458)

[2.2.3 手机信令数据的空间精度特征 8](#_Toc26214)

[2.3 基于位置数据的出行特征分析研究 9](#_Toc27764)

[2.3.1 出行链提取 9](#_Toc14057)

[2.3.2 出行目的研究 11](#_Toc31514)

[2.3.3 出行模式研究 12](#_Toc19520)

[2.4 相关聚类算法和预测模型分析研究 13](#_Toc333)

[2.4.1 聚类算法 13](#_Toc21717)

[2.4.2 预测模型 15](#_Toc15586)

[2.5 本章小结 17](#_Toc27841)

[第三章 基于多源位置数据的人员特殊聚居区域识别模型研究 18](#_Toc13631)

[3.1 特殊人群聚居区域研究相关问题的提出 18](#_Toc25311)

[3.2 手机信令数据预处理与质量分析 19](#_Toc22118)

[3.2.1 手机信令数据集介绍 19](#_Toc4494)

[3.2.2 数据质量分析 20](#_Toc15884)

[3.2.3 数据预处理 23](#_Toc288)

[3.3 用户出行链与职住地提取 24](#_Toc30441)

[3.3.1 用户出行链提取 25](#_Toc20081)

[3.3.2 用户职住地提取 29](#_Toc31586)

[3.4 用户个人属性预测 32](#_Toc32609)

[3.4.1 特征设计 33](#_Toc8866)

[3.4.2 训练集标注 37](#_Toc31760)

[3.4.3 用户个人属性预测模型 41](#_Toc29537)

[3.5 区域特征分析 43](#_Toc23308)

[3.5.1 公交数据与共享单车数据出行链 43](#_Toc29154)

[3.5.2 区域特征定义 43](#_Toc24525)

[3.5.3 区域特征分析 45](#_Toc27381)

[3.6 本章小结 45](#_Toc17971)

[第四章 人员特殊聚居区域分析系统的设计与实现 47](#_Toc3074)

[4.1 系统需求分析 47](#_Toc2794)

[4.2 系统架构设计 48](#_Toc22628)

[4.3 数据接口设计 49](#_Toc8148)

[4.3.1 手机信令数据接口 50](#_Toc20122)

[4.3.2 公交数据与共享单车数据接口 51](#_Toc13553)

[4.3.3 POI数据接口 52](#_Toc25635)

[4.3.4 用户出行链与职住地数据接口 52](#_Toc16732)

[4.3.5 用户出行特征数据接口 53](#_Toc29940)

[4.3.6 区域特征数据接口 54](#_Toc5459)

[4.4 关键模块设计与实现 55](#_Toc27630)

[4.4.1 数据预处理模块 55](#_Toc17769)

[4.4.2 出行链提取模块 57](#_Toc21441)

[4.4.3 个人属性预测模块 59](#_Toc79)

[4.4.4 区域分析模块 60](#_Toc15383)

[4.4.5 可视化模块 62](#_Toc11220)

[4.5 本章小结 62](#_Toc32010)

[第五章 系统测试与分析 63](#_Toc19221)

[5.1 实验环境 63](#_Toc29471)

[5.1.1 测试系统环境 63](#_Toc32495)

[5.1.2 测试数据环境 63](#_Toc24681)

[5.2 用户出行链与职住地提取结果测试与分析 64](#_Toc15148)

[5.2.1 停留点识别 64](#_Toc17283)

[5.2.2 职住地识别 65](#_Toc24075)

[5.3 个人属性预测结果测试与分析 66](#_Toc15224)

[5.4 区域特征分析 69](#_Toc1297)

[5.4.1 学校区域识别验证 69](#_Toc2008)

[5.4.2 工厂区域识别验证 71](#_Toc4414)

[5.5 本章小结 73](#_Toc28298)

[结论与展望 74](#_Toc7713)

[参考文献 75](#_Toc23834)

[攻读硕士学位期间取得的学术成果 79](#_Toc2659)

[致 谢 80](#_Toc2457)

**图 目**

[图 1 用户画像构建流程 5](#_Toc22513)

[图 2 低收入通勤者活动-出行模型 6](#_Toc1296)

[图 3 手机基站通信方式 7](#_Toc15509)

[图 4 北京市2018年人群出行目的分布 11](#_Toc20362)

[图 5 Kmeans算法流程图 14](#_Toc6705)

[图 6 DBSCAN聚类算法示意图 15](#_Toc5344)

[图 7 集成学习示意图 17](#_Toc9398)

[图 8 北京市基站分布图 20](#_Toc29469)

[图 9 信令数据更新周期分布图 21](#_Toc2613)

[图 10 长距离基站抖动样例 24](#_Toc9613)

[图 11 多个点簇分布在同一空间内的示例 26](#_Toc25064)

[图 12 相邻信令间的距离分布图 29](#_Toc3034)

[图 13 用户出行时间分布图 30](#_Toc3190)

[图 14 用户职住地识别算法流程图 32](#_Toc11381)

[图 15 用户出行时间算法流程图 33](#_Toc26995)

[图 16 用户出行特征热力图 37](#_Toc13879)

[图 17 北航区域划分图 38](#_Toc1574)

[图 18 Kmeans类簇数量选择 39](#_Toc27564)

[图 19 北航区域人群样本轨迹图 40](#_Toc18025)

[图 20 个人属性预测模型框架图 41](#_Toc11711)

[图 21 特征重要性 42](#_Toc26731)

[图 22 用户出行时段分布图 44](#_Toc26400)

[图 23 人员特殊聚居区域分析系统架构图 48](#_Toc21413)

[图 24 基础数据ER图 49](#_Toc5678)

[图 25 中间数据ER图 50](#_Toc27247)

[图 26 预处理模块结构图 56](#_Toc10794)

[图 27 预处理模块类图 56](#_Toc24909)

[图 28 出行链提取模块结构图 57](#_Toc8151)

[图 29 出行链提取模块类图 58](#_Toc24609)

[图 30 个人属性预测模块结构图 59](#_Toc19523)

[图 31 个人属性预测模块类图 60](#_Toc17311)

[图 32 区域分析模块结构图 61](#_Toc18916)

[图 33 区域分析模块类图 61](#_Toc4866)

[图 34 可视化页面展示 62](#_Toc15334)

[图 35 用户停留点提取结果实例图 65](#_Toc32253)

[图 36 用户停留点数量分布图 65](#_Toc23492)

[图 37 居住地热力图 66](#_Toc30453)

[图 38 工作地热力图 66](#_Toc1302)

[图 39 预测准确率与决策树数量关系图 67](#_Toc15725)

[图 40 预测准确率与不同参数组合关系图 68](#_Toc13226)

[图 41 学校区域识别样例 70](#_Toc27842)

[图 42 学生群体居住热力图 71](#_Toc10197)

[图 43 大兴区工人居住热力图 72](#_Toc25368)

[图 44 各类型区域位置分布 72](#_Toc9483)

[图 45 信令数据与企业台账区域人数对比 73](#_Toc5268)

**表 目**

[表 1 手机信令数据集概况 19](#_Toc14146)

[表 2 用户信令数量统计 21](#_Toc9600)

[表 3 春节期间本地用户与外地用户数量 22](#_Toc9238)

[表 4 工作期间本地用户与外地用户数量 22](#_Toc23914)

[表 5 活跃用户数量 23](#_Toc22298)

[表 6 用户出行特征说明表 36](#_Toc32347)

[表 7 标定人群属性定义 37](#_Toc30655)

[表 8 北航区域人群划分结果 39](#_Toc1980)

[表 9 安定化工厂区域人群划分结果 40](#_Toc17756)

[表 10 聚类特征说明表 45](#_Toc17911)

[表 11 手机信令数据接口 50](#_Toc3144)

[表 12 基站位置数据接口 51](#_Toc27290)

[表 13 公交数据接口 51](#_Toc13905)

[表 14 共享单车数据接口 52](#_Toc20951)

[表 15 POI数据接口 52](#_Toc28932)

[表 16 出行链数据接口 53](#_Toc13588)

[表 17 职住地数据接口 53](#_Toc11647)

[表 18 用户出行特征接口 54](#_Toc3179)

[表 19 区域特征数据接口 55](#_Toc5071)

[表 20 dataAnaly成员 57](#_Toc10363)

[表 21 dataProcess成员 57](#_Toc8872)

[表 22 stopPoint类成员 58](#_Toc28603)

[表 23 hwPoint类成员 59](#_Toc30553)

[表 24 hwPoint类成员 60](#_Toc1549)

[表 25 clusterUtil类成员 62](#_Toc20992)

[表 26 集群节点配置信息 63](#_Toc23982)

[表 27 节点组件信息 63](#_Toc21839)

[表 28 个人PC机配置 63](#_Toc25418)

[表 29 公交数据概况 64](#_Toc3897)

[表 30 共享单车数据概况 64](#_Toc6197)

[表 31 训练集概况 67](#_Toc9704)

[表 32 不同特征组合对预测结果的影响 68](#_Toc11050)

[表 33 个人属性预测结果对比分析 69](#_Toc20117)

[表 34 学校区域识别结果 71](#_Toc4276)

[表 35 大兴区工厂区域聚类结果 72](#_Toc25790)

# **绪论**

## 研究背景及意义

随着社会经济的持续发展和城市化进程的不断推进，各类商业、体育、娱乐、文化等大型活动越来越多，一些集购物、娱乐、餐饮、住宿为一体的综合性建筑不断出现，产生了大量的人员聚居区域。这些人员密集区域具有占地面积大、功能复杂、人员数量多、密集大等特征，一旦发生突发情况，有可能造成严重的人员伤亡和财产损失，甚至引发恶劣的社会影响。比如，2017年11月18日，北京市大兴区西红门镇新建村发生火灾事故共造成19人死亡，8人受伤。人员密集区域重大安全事故或事件的时有发生，给城市安全运行带来了巨大的挑战，引发了社会各界对人员密集区域安全风险问题的高度重视。

人员密集区域安全风险识别的关键是对人位置的采集与辨识，人员的行为轨迹数据能够清晰有效描述人的位置信息。在此基础上所提取的人群出行规律和画像可为人员聚居区域的识别提供数据支撑。近年来，随着以GPS导航技术和智能手机为代表的智能终端的普及与应用，加上公共交通出行方式便捷性以及共享出行方式的新起，人们已经能够以相对低廉的代价获得大量的用户时空数据。常用的时空数据集包括公交地铁刷卡数据、共享单车数据和手机信令数据等。

公交数据是指使用一卡通刷卡产生的数据，这些数据一般是人们乘坐城铁、公交时产生的。共享单车数据指人们骑乘单车时产生的数据。这两类数据所包含的信息类似，均记录了用户下车地点，下车时间，上车时间，上车地点。但公交数据与共享单车数据均只记录了用户交通出行时产生的数据，且部分用户会选择其他交通方式出行，这两类数据的时空连续性较差，因此在本文使用这两类数据进行辅助分析。

信令数据是指用户在平常使用手机的过程中，与手机所通信的基站记录下的数据，数据中包含所连基站的坐标，信令产生的时间，事件类型等信息。当用户携带手机移动，上网，收发短信，语音通话时，与手机所通信的基站均会记录相应的信令数据。手机用户的移动轨迹可以通过这些蕴含时空信息的信令数据来刻画。近几年来，随着智能手机与4G技术的普及，手机已成为人们必不可少的日常工具之一。根据工信部公布的调查结果显示，到2018年底，全国每100人拥有97.5部手机，手机用户的总数为13.78亿，其中北京市的用户拥有的手机数最多，每100人拥有189.3部手机。这些数据说明了在城市人群众，手机的普及率极高。大量人员具有长时间开机，随身携带手机的习惯，因此手机信令数据能够较为完整地反映用户一天内的出行轨迹。本文以手机信令数据为主进行后续分析。

手机信令数据、公交数据、共享单车这些位置数据中蕴含了人群活动规律、生活习性等丰富的语义信息。对这些数据进行分析，挖掘人群的出行特征，利用出行特征辨别目标人群的社会属性，识别目标人群的聚居位置对相关部门进行预警防范工作具有重要的意义。

## 课题来源

本课题来源于北京航空航天大学软件开发环境国家重点实验室承担的北京市科技计划项目：基于行为轨迹的人员密集场所安全风险动态辨识与预警关键技术研究与示范

## 研究目标及研究内容

本文旨在从城市安全生产部门的业务需求出发，结合机器学习技术，设计一个人员特殊聚居区域识别模型。本文的研究目标是基于手机信令数据，完成用户的出行链与职住地提取，以此作为本文的研究基础，识别人群的社会属性，提取特殊社会属性人群的聚居区域，结合公交刷卡数据与共享单车数据分析区域特征，完成人员特殊聚居区域的识别。对于社会安全生产部门，获取特殊聚居区域的位置及特征信息，有助于其进行预警防范工作，对社会安全维护具有重大意义。

本文的研究内容主要包括以下四个部分：

1. 手机信令数据质量分析与预处理

手机信令数据的产生方式与用户使用手机的习惯、用户的活动区域有较大关系，这就使手机信令数据存在一些先天的缺陷。部分用户生活中不经常使用手机，部分用户长期活动于信号不良的区域，这样会导致这些用户产生的信令量较少，不足以提取其一天的移动轨迹。本文制定了用户过滤规则，剔除产生信令数据较少的用户。由于手机信令的产生机制与基站定位问题，部分信令数据会出现数据格式错误、数据字段缺失、数据抖动等问题，本文制定了数据过滤规则，剔除错误数据，保证信令数据的可靠性。

1. 用户单日出行链、职住地提取

用户单日出行链、职住地数据是本文后续分析的基础。本文首先定义了停留点、移动点、出行链。结合手机信令数据包含了位置和时间两个维度信息的特点，引入ST-DBSCAN算法，识别用户的停留/移动态，完成了用户的停留点提取。根据用户多日的停留点数据，定义工作时间段与居家时间段，完成用户职住地的提取。

1. 用户社会属性估计

根据用户的出行链数据与职住地数据，结合POI兴趣点数据，提取出行特征：用户离家时间，用户归家时间，用户出行距离，用户出行次数，用户出行轨迹相似性，用户出行时间，POI特征等。最终依据用户出行特征数据，采用随机森林预测方法，完成个人社会属性的预测。

1. 人员聚居区域提取

将具有相同社会属性人群的居住地位置聚类形成区域。结合公交刷卡数据与共享单车数据分析区域特征，最终实现人员特殊聚居区域识别模型的构建。

## 本文的组织结构

本文的组织结构如下。

1. 介绍课题的研究背景及意义、课题来源、研究目标与内容以及本文的组织结构。
2. 介绍了目前关于个人社会属性与出行特征关系的研究。然后介绍了目前基于轨迹数据进行城市出行分析的国内外研究现状和热点研究方向，总结现阶段存在的问题和不足；最后介绍了本文涉及到的聚类、分类技术的实现原理和适用场景。
3. 提出基于多源位置数据的人员特殊聚居区域识别模型需要解决的问题，然后对这些问题依次展开讨论。首先对手机信令数据进行数据质量分析与预处理，然后基于经过预处理过后的信令数据，采用ST-DBSCAN聚类算法识别用户的停留点与移动点，提取用户的出行链。根据用户多天的停留点数据，计算其居住地与工作地。然后对已知聚居区域内的人群进行个人属性标注，并提取其出行特征，基于随机森林完成用户个人属性预测模型的构建。最后，将预测出的目标人群的居住地聚类形成区域，并结合公交数据，共享单车数据完成区域特征的分析。
4. 基于本文提出的人员特殊聚居识别模型，设计并实现了一个基于多源位置数据的人员特殊聚居区域分析展示系统。首先介绍了基于多源位置数据的人员特殊聚居区域分析与展示系统的需求及架构设计，然后依次从数据接口设计和关键模块设计等几个方面对系统的设计与实现进行了详细的介绍与说明。
5. 对本文提出的出行链与职住地提取、个体属性预测以及区域特征分析等方法的准确性进行验证。

最后对本文所述的内容进行总结，并对后续的研究方向与研究重点进行了展望。

# **研究现状与相关技术**

本章首先介绍了用户社会属性与出行特征相关的内容，包括低收入工作者，老年人，学生等特殊人群的活动规律的研究；然后介绍了手机信令数据的数据特征，基于位置数据的出行特征相关方面的研究，包括出行链，出行模式，出行目的等方面的内容，并总结了现阶段存在的问题与不足。最后，介绍本研究涉及到的聚类、分类技术的常见类型以及各自的适用场景。

## 社会属性与出行特征相关研究

### 用户画像

用户的社会属性包含个体的性别、年龄以及职业等信息，属于用户画像的子集。用户画像(User Profile)，一般也可称为用户建模(User Model)、用户标签(User Label)。一般可定义一些用于刻画用户行为的特征以对用户进行简单的描述。用户画像主要通过分析用户个体的轨迹数据、购物记录、出行等数据，进而抽象出用户的标签。图1展示了用户画像的构建流程。



图 1用户画像构建流程

G. Adomavicius等人[1]基于用户交易数据，使用分类方法，关联规则来发现用户的行为规则。Sugiyama等人[2]则从用户结果评分、浏览记录等信息，构建用户的偏好画像。Rui Li等人[3]基于社网络数据分析用户的家庭位置，捕捉用户跟踪用户的可能性或推送位置的可能性。

手机信令数据、公交数据中蕴含了丰富的人群出行信息，识别用户主要依赖于从这些轨迹数据中构建的用户特征属性。在文献[4]中，Zhu等人在用户手机中安装了记录轨迹数据的软件，并从这些轨迹数据中提取用户出行模式，用户长期停留点，用户频繁的移动路径等用户特征来描述用户的出行模式。Xihui Chen等人[5]提出了一种基于GPS数据来构建用户的移动行为画像并计算用户之间的移动相似度的方法。通过用户经常访问的地点计算停留点，通过层次聚类生成ROI（region of interest）轨迹，使用频繁序列模式挖掘技术来提取用户的移动特性。然后基于Ying等人[6]提出的相似性测量方法提出了最大语义模式（MSTP-相似度）和最长的公共序列（LCSS）计算用户之间的相似度。

黄文彬等人[7]使用手机话单数据，从用户移动速度，出行模式以活动规律这几方面构建用户画像，其结果可以为个性化服务提供更完整丰富的信息。其采用计算熵、频繁模式挖掘、构建概率矩阵等方法，基于手机话单数据中的基站位置信息完成了用户移动行为的分析。

### 特殊人群出行研究

特殊人群指社会属性，出行习惯有别于大部分通勤人员的用户，包括低收入者、学生、退休老人、工人等。出行作为一种派生性需求，个体基于活动，通过安排出行OD、出行时间和出行方式来组织日常生活，用户的出行习惯在很大程度上反映了该用户的属性[8-10]。

在文献[11]中，程龙基于南京市居民调查数据，使用结构方程建立低收入人群的活动-出行模型，结果表明相较于休闲型活动与维持型活动，低收人群会更多地进行生存型活动。其出行链以简单的居住-工作链为主，个体的社会经济属性对其出行方式的选择影响较大。图2展示了低收入通勤者的活动-出行的模型设定。

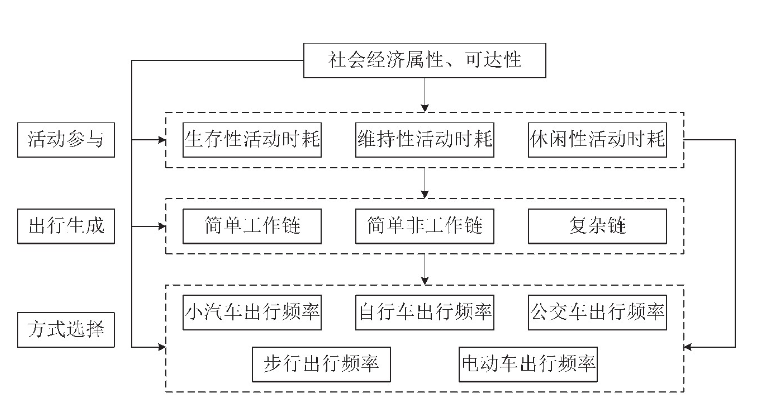


图 2 低收入通勤者活动-出行模型

廖琨在文献[12]中使用手机信令数据与区域住房价格数据分析住房价格与居民出行特性的相关性，结果表明相对于居住在低房价的居民，居住在高房价小区的居民具有更高的出行距离、出行频率、非基家出行比例和更低的高峰出行比例。杜影在文献[13]中，基于交通成本数据，构建了低收入人群出行方式选择MNL模型，确定了影响低收入人群出行方式选择的因素。

王竹影[14]利用调查数据分析了老年人的出行行为特征以及相关建成环境影响。结果表明在出行方式选择上，老年人一般偏向步行。在出行距离方面，老年人以娱乐为出行目的时，活动范围主要集中在1.5km内；以健身为目的的活动范围主要集中在3km内。

## 手机信令数据的数据特征

### 手机信令数据的生成原理

运营商为保证移动终端（例如手机，平板电脑，短信机）与基站能够顺利地完成通讯工作，会记录二者通讯期间产生的状态信息，手机信令数据即来源于这些状态信息。目前移动网络主要有2G，3G，4G这几种标准。这些通信标准主要是源于GSM(Global System of Mobile communication，全球移动通讯系统）这一系统发展而来。标准的GSM网络主要由移动终端，网络系统，基站系统这几部分组成。当用户使用手机等移动终端执行开关机、收发短息、上网等操作时，基站系统接收终端发送的信息，并将这些信息发送给网络系统，网络系统依据GSM制定的通信协议对终端与基站间的通信连接进行管理。其中手机信令主要产生于终端与基站间的通信过程期间，用户轨迹信息的获取主要源于信令数据中的时间信息与空间信息[15]，如图3所示。

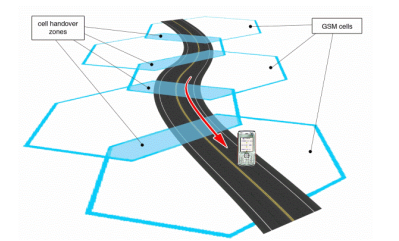


图 3手机基站通信方式

### 手机信令数据的时间精度特征

手机信令数据中所记录的时间信息为个体用户与基站之间产生通信时的时间，这一时间与一般与运营商通信网络的系统时钟为准。随着4G通信技术[16]的普及，用户产生的信令数据在时间精度上可以达到秒级别，这些原始信令数据由于数据量巨大，无法进行长期存储，因此运营商一般会将这些原始信令数据存储在流式计算平台上[17]，用于支撑需要实时数据的一些业务需求。一般而言，运营商会每隔一定的时间段(10分钟-1小时)对用户的移动终端进行一次采样，以确保用户的终端能够得到质量较高的通信服务，这些采样数据属于主动采样数据；开关机，基站位置切换以及通话数据属于用户终端状态产生变化时主动与基站产生的交互行为，这时产生的数据属于被动采样数据。被动采样数据与用户的行为特征有着密切的关系。由于用户使用手机习惯的差异，被动采样的数据具有在时间维度上分布不均匀的特点，比如用户一般在日间使用手机进行通话，收发短信，上网等活动，此时信令数据的数据量较大，时间精度较高；夜间部分用户有关机或者切断网络的习惯，此时信令数据的时间精度较低。

### 手机信令数据的空间精度特征

手机信令数据的空间精度主要由基站的定位技术与运营商设立的基站覆盖区域与这两个因素所决定。在定位技术方面，目前基站的定位技术主要有AOA(Angle of Arrival)、信号到达角法)、TDOA(Time Difference of Arrival，TDOA)、TOA(Time of Arrival，信号到达时间法)、COO(Cell of Origin，起源蜂窝区定位法)这五种。COO方法的定位原理较为简单，即根据与移动终端通信的基站位置确定终端的位置，其定位精度主要依赖于基站的覆盖面积，定位精度较低。AOA、TOA、TDOA这几种方法的思想比较相似，主要是结合分析信号衰减度、基站与终端间的角度或距离这几个特征来对移动终端进行定位，定位精度较高，但需要耗费较多的网络资源和计算资源。运营商记录信令数据的主要目的是保证移动终端与基站能够顺利地完成通讯工作，其业务场景对定位精度并没有太高的要求。尽管起源蜂窝区定位法的定位精度较低，但由于其具有资源耗费低，响应速度快等优势，运营商一般采用该方法对移动终端进行定位。

在基站覆盖区域方面，运营商为了增加一个基站的信号覆盖范围，一般会在一个基站上安装多个朝向不同的天线，这时基站的覆盖区域近似为圆形区域。也有部分基站上只安装了一个天线，这时基站的覆盖区域近似为扇形区域。一般来说基站的覆盖区域没有明确边界，在以往的研究中，学者为研究方便会规定基站区域的边界，比如泰森多边形或者类似蜂窝的六边形等[18]。本研究考虑计算方便，将基站的覆盖区域定义为矩形。

在基站覆盖半径方面，运营商往往结合区域内的业务需求量来设立基站。影响业务需求量主要有两个因素：一是该区域的人口密度，人口密度较高的区域，比如住宅区，商圈内一般汇聚了大量的用户，运营商一般会在这些区域设立大量的基站以保证用户使用的移动终端能够获得较好的通信质量，由于这些区域内的基站密度较高，相邻基站的间距也就较小；二是该区域的市政规划情况，一般待开发的经济区在未来可能会有大量的人员流入，对于这些计划内会产生大量业务需求的区域，运营商也会增加基站的布设。国内外研究方面，Ahas等人[19]所使用的信令数据在市区等人口密度较大的区域的空间精度为100米左右，在郊区为1000m左右，Xu等人[20]使用了上海与深圳的手机信令数据进行有关城市活动空间与个体的关联的研究，其中上海信令数据集的基站平均间距为200米，深圳信令数据的基站平均间距为190米。

## 基于位置数据的出行特征分析研究

### 出行链提取

用户的出行链一般由多个OD（Oriign-Destination，即用户一次出行的起止点）组成。对于公交数据与共享单车数据，由于其数据产生机制的原因，其原始数据本身即包含了OD信息，对于这两类数据无需再做处理。手机信令数据是由用户使用手机与基站通信时所产生的数据，其中并不包含停止、移动等有关用户移动状态的信息，因此需要对手机信令数据进行出行链信息的提取。

出行链提取主要涉及到停留、移动状态识别，职住地提取等问题。在文献[21]中，杨佩瑜基于3G手机信令数据，提出了基于斜率的用户停留/移动状态识别算法，通过设立合理的时间阈值与空间阈值来提取用户的停留点。在文献[22]中，林楠基于手机信令数据，采用设立一定宽度时间滑动窗口的思路，提出了一种基于滑动窗口的活动区域停留识别算法，综合考虑多个手机定位数据的时空临近程度来判断停留态与移动态。

Calabrese在一个项目中提到了一种更精细的移动状态分析方法，该项目使用手机信令数据实时监视城市道路网络[23] [24]。该方法首先设置速度阈值vt，该阈值区分慢速移动（步行或停止）和快速驾驶（驾驶或骑车）。然后，计算连续的两个信令数据间的瞬时速度vt，并结合预先设定的概率pt来区分缓慢移动和快速行驶两种状态。该方法的主要贡献在于通过预先设定的vt值与pt值提取驾驶车辆的快速移动用户，但车辆存在拥堵等情况，vt与pt的选取比较困难。

停留点在具体数据上体现为一系列时空相近的位置点的集合，因而也可以将提取停留点的过程看成是一种聚类的操作。目前适用于停留点提取的方法主要有两种：增长聚类算法与密度聚类算法。SMoT算法作为增长聚类算法的代表，在位置数据挖掘中获得了广泛的关注，在文献[25]中，Cik在SMoT算法的基础上，通过使用卡尔曼滤波的方式改善了从GPS位置数据中识别停留点的效果，并识别了GPS数据中的异常点；Ulm为了满足城市人群活动模式挖掘尺度研究和人群活动链构建，在SMoT算法的基础上采用低通滤波的方法完成了停留态/移动态的识别[26]。通过设立合理的时间阈值与空间阈值，SMoT算法可以判断用户的停留态/移动态。密度聚类算法的代表是DBSCAN算法[27],其根据密度相邻等空间上的特征将具有高密度的点簇聚集成一个点，并且能够较好地识别噪声数据。传统的DBSCAN算法要求数据需是单一维度的，因此其并不适用于蕴含时空信息的位置数据。针对上述缺点，Kut等在文献[28]中提出了ST-DBSCAN算法，该算法在时间维度上扩展了DBSCAN算法，解决了DBSCAN算法只适用于单一数据源的缺点。

出行链信息提取涉及到的另一个问题是人群职住地的提取。由于大部分通勤人员白天一般在工作地工作，晚上在居住地休息，人群的职住属性与时间关系较为紧密，因此目前关于提取职住地的研究多是基于人群停留区域时间段的研究。

在文献[29]中，赵阳定义了工作时间段与居家时间段，并通过计算个体停留点的停留时长与工作时间段和居住时间段的交集时长来判断个体的职住地属性。Yang在文献[30]中通过计算个体移动的速度得出了早高峰与晚高峰的时间段，并根据个体多天的停留点数据完成了个体的职住地提取。在基于公共交通卡数据的分析中，Long和Zhang[31]等人使用公交卡一天的出行数据，通过当天的出发地点和停留超过6小时的位置来识别单日居住地，然后整合一周的个人旅行数据并通过了决策树标识了持有该卡用户的最终居住地。 在获得持有该一卡通用户的住所后，论文根据规则确定该一卡通用户是否属于同一天的通勤行程，并通过与问卷数据的比较证明了该方法的可靠性和准确性。

### 出行目的研究

国内外一般将出行目的分为上班、公务、上学、文化娱乐、购物、回家(含回程)、探亲访友、和其他[32]。根据2018年北京交通发展年报[33]中展示的调查结果，如图4所示，2018年全年的出行中，上下学与上下班活动占比为51%，其他活动包括购物，休闲娱乐与接送人活动等占比48.6%

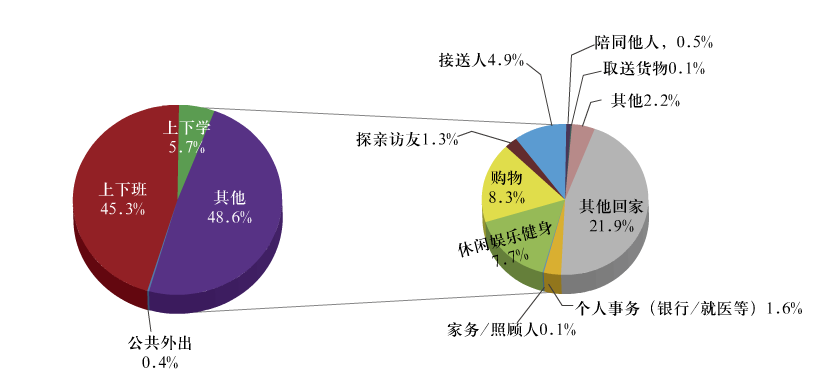


图 4 北京市2018年人群出行目的分布

目前国内外关于出行目的的识别主要有两种方法，一是基于经验规则方式的识别，在文献[34]中，Bohtd将个体出行地点周边100m范围内的POI（兴趣点）与出行地点位置进行关联并进行判断，该方法分析出行目的的准确度主要依赖于POI能否与出行地点建立正确的关系；曹劲舟在文献[35]中通过定义居家时间段与工作时间段并计算时间段内占比较高的位置作为用户的居住地或工作地。许宁在文献[36]中提出利用土地特征并结合手机话单数据分析用户的职住信息，但由于城市的部分用地特征比较复杂，该方法的准确性有待商榷。从目前基于经验规则识别出行目的的研究来看，利用时空特征可以较为准确地推断居家、工作这两种活动目的，但对于多种活动目的推断存在局限性。

关于出行目的识别的第二种方法是基于概率模型推断活动目的。在以往的研究中，大部分学者均是基于GPS数据采用有监督的方式进行活动目的的识别。Moiseeva使用贝叶斯网络[37]，Mc Gowen使用决策树[38]完成了基于GPS轨迹数据的出行目的识别。不同于手机信令数据，GPS轨迹数据具有定位精度高，时间采样丰富等优点，因此基于概率的出行目的识别模型在应用于类似GPS轨迹数据的这种具有高时空分辨率的数据集时具有较好的效果。在基于手机信令数据的出行目的识别研究中，Liu等人通过定义了一些时空维度上的特征（上一次活动的开始/结束时间，本次活动的持续时间）采用支持向量机对活动目的进行判别，结果显示预测精度为69.7%，效果一般。手机信令数据中一般只记录了用户终端与基站通信的时空信息，并无关于目的标签的属性。因此部分学者考虑引入辅助数据结合手机信令数据进行分析用户的出行目的，在文献[39]中，Tu使用微博签到数据训练HMM模型，并结合手机定位数据推断个体用户的休闲，购物等除居家工作以外的活动目的。Gao在文献[40]中使用FaceBook中的签到数据与POI兴趣点数据，对用户的出行目的进行了预测分析。

Lin等人[41]采用加权概率的方法，提出了一种基于手机信令数据的活动目的推断算法，结合活动转移概率等时间特征与POI信息等空间特征，采用概率加权的方式计算活动概率矩阵，并基于计算出的概率矩阵使用蒙特卡洛模拟推断个体活动目的。该研究基于活动目的的推断结果，结合个体停留区域信息构造个体活动链，在此基础上支持居民的活动模式分析。

### 出行模式研究

用户的出行模式一般指用户出行所使用的交通工具或采用的方法。出行模式的基本识别规则是各种交通工具具有不同的运动规律，比如出租车与地铁的速度远远大于个体步行的速度，地铁和公交一般具有固定的停靠点与路线。出行模式的基本研究对象是个体的出行链，通过确立精确的表示轨迹时空信息的统计量来建立判别规则，不同的规则设定方法和不同的统计量可以延伸出不同的出行模式判别方法。

Zheng Y和Li Q在文献[42]中使用原始的GPS数据通过监督学习的方法推测数据产生者的运动模式,包括行走、驾驶、乘坐公交车或者骑自行车,文中提出了一系列对交通环境鲁棒的特征,并基于地图将轨迹点通过DBSCAN的算法聚成节点,通过节点表示移动轨迹,这样的方式既保留了真实世界的限制条件,又容许用户个人行为的特性,使得预测结果准确性比以前的方法高8%。

Liu L等人[43]在2009年同样对出行模式的时空特性进行了统计分析,他们使用了深圳市一个月的一卡通数据,对比了周六、周日以及工作日情况下公交车和地铁两种交通方式一天客流的变化趋势,发现工作日情况下的客流曲线的早晚高峰现象比较显著。通过对比一天中乘客出行在不同站点中的分布情况,表明乘客出行主要集中于几个站点。这说明了乘客出行在时空上分布的不平衡特性。

Kim K等人[44]在2014年提出了区域出行模式MZP(movement patterns between zones)的概念,不同于传统站点到站点之间的出行移动模式,他们关注于区域与区域之间大规模范围内的出行移动模式,简化了出行模式的表达。主要的思路是通过将相邻的区域及区域间的出行模式进行循环合并,在发现区域的同时挖掘区域之间的出行移动模式。两个出行模式合并的依据主要基于它们之间的相似程度,通过考虑两个模式涵盖的站点接近程度,同时保留移动方向,循环地进行合并操作。

## 相关聚类算法和预测模型分析研究

聚类算法属于无监督学习领域，其主要的目标是将特征相似的对象划分到一个簇集中，特征不相似的对象划分到不同的聚集中。本文主要在用户出行链提取，区域特征分析部分使用聚类算法。预测模型属于有监督学习领域，其主要的目标是通过事先标注好的训练集构建预测模型，完成对无标注数据集的标签预测。本文主要在个人属性预测方面使用预测模型。本节将对文中涉及的预测模型和聚类方法做相应的介绍。

### 聚类算法

聚类属于非监督学习算法，其目的是在数据集中寻找相似的数据点，把各自相似的数据点划分到同一簇集中。其中数据点间的相似程度可以采用距离等概念来度量。

Kmeans算法[45]由于具有运行速度快，划分结果较为准确等优点，在聚类算法中应用较为广泛。其目的是在数据集中找出k个簇群，这里的k值需要事先指定。Kmeans算法的流程图如图5所示：



图 5 Kmeans算法流程图

DBSCAN算法是Hans-Peter Kriegel，Martin Ester等在1996年由提出的一个具有代表性的空间密度聚类算法[46]。这类密度算法通过判断样本分布的密集程度决定样本属于哪一个簇集，算法可以把标记出低密度区域的离散点，并把彼此紧邻的点划分为一个簇集，算法需要输入两个参数：Eps和minPts。

其具体定义与概念如下：

1. Eps邻域：给定对象半径为Eps内的区域为该对象的Eps领域；
2. 核心点：对象的Eps领域包含至少minPts个对象，则该对象为核心对象；
3. 边界点：不是核心点，在核心点的Eps领域内；
4. 噪声点：既不是核心点也不是边界点；
5. 直接密度可达：如果点p位于点q的Eps领域中，且q为核心对象，则称p由q直接密度可达；
6. 密度可达：当存在一个路径p1,p2,…,pn，其中p1=p，pn=q，且任意一个pi+1由pi直接密度可达；
7. 密度相连：对于样本集合中的一个点o，若点o到点p和点q都是密度可达的，则p和q密度相连。

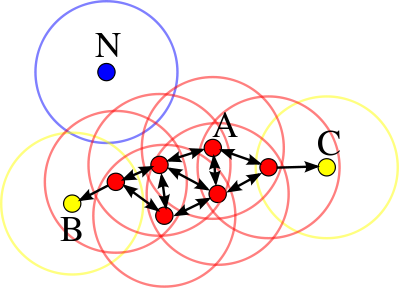
簇集是由一个核心点和与它密度可达的点构成的。噪声点由于与其他的非噪声点均密度不可达，其并不会被划分到簇集中。DBSCAN算法的示意图如图6所示，其中minPts=4。红色圆圈内的点属于一个簇集，其中红色的点为核心点，点B和点C虽然在Eps范围内没有包含4个点，但由于它们由A点（核心点）密度可达，因此也被划分到了簇集中。N是一个噪声点，因为它既不能由其他点密度可达也不是核心点。

图 6 DBSCAN聚类算法示意图

DBSCAN通过检查数据集中每个点的Eps领域来搜索簇，如果该点的Eps领域包含的点大于MinPts个，则创建一个以P为核心对象的簇。然后迭代地聚集核心对象直接密度可达的点，以及合并一些密度可达的簇。当没有新的点添加到簇中时结束。

### 预测模型

决策树（decision tree）是预测算法中的一种。其理解起来和推理相对直观。我们在日常生活中使用的推理方法一般都体现了决策树的思想。比如对于分类问题，我们想判断样本是否属于A类别，根据A类别样本的所体现特征，我们会进行一系列的“是/否”判断，这些判断过程就是决策树算法所体现的预测过程

顾名思义，决策树是一种树形结构，其包含了一个根节点，若干个叶节点与内部节点。其中根节点对应样本全集；内部节点对应一个特征判断，节点包含的样本数据根据特征判断的结果被划分到各个子节点中；叶节点对应预测结果。决策树的思想为分治策略，其基本的学习算法如下：

| **算法：决策树学习的基本算法** |
| --- |
| **输入：训练集属性集**  **输出：以node为根节点的一棵决策树** |
| **流程：函数TreeProduce(D,A)**  1: 生成结点node；  2: **if** D中样本全属于同一类别C **then**  3: 将node标记为C类叶结点; **return**  4: **end if**  5: **if** A= **OR** D中样本在A上取值相同**then**  6: 将node标记为叶结点，其类别标记为D中样本数最多的类; **return**  7:  **end** **if**；  8: 从A中选择最优划分属性;  9: **for** 中的每一个 **do**  10: 为node生成一个分支;令表示中在上取值为的样本子集;  11: **if**  为空 **then**  12: 将分支节点标记为叶结点，其类别标记为D中样本最多的类; **return**  13: **else**  14: 以为分支结点  15: **end if**  16: **end for** |

集成学习(ensemble learning)有时也被称为基于委员会的学习(committee-based learning)，多分类器系统(multi-classifer system)等,其主要思想为通过构建并结合多个学习器来完成学习任务。图7展示了集成学习的两个主要构成部分：结合模块与若干个个体学习器，结合模块通过一些组合策略将个体学习器组合起来，一般的结合策略有Boosting，Bagging等[47]。个体学习器一般由BP神经网络算法、C4.5决策树算法这些现有的学习算法组成。集成学习可分为同质学习与异质学习两种类别。同质学习指所有个体学习器的类型相同，比如决策树构成决策树集成算法，神经网络构成神经网络集成算法等，其中对应的算法称为基学习算法，个体学习器称为基学习器。异质学习则指个体学习器的类型不同，其不再有基学习器的概念，相应的个体学习器被称为组件学习器。

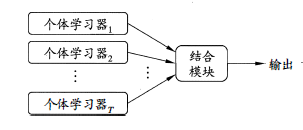


图 7 集成学习示意图

随机森林[48](Random Forest)是基于Bagging算法来扩展实现的。随机森林在决策树的训练过程中引入了随机属性选择，其主要以决策树为基学习器，采用Bagging算法来集成。一般决策树是基于在当前结点的属性集合（假设有d个属性）中选择一个最优属性来进行划分属性选择的；在随机森林中，对于基决策树的每个结点，先从该结点的属性集合中随机选择一个包含k个属性的子集，然后再从这个子集中选择一个最优属性用于划分。若令k=1，则是随机选择一个属性用于划分；若令k=d，则基决策树的构建与传统决策树相同；k值在一般情况下被设为。参数k控制了随机性的引入程度。在大数据的背景下，随机森林相对于一般机器学习算法具有较为明显的优势。Spark和MapReduce这样的大数据技术在设计之初了就考虑了处理数据的并行问题，由于决策树往往是单独构造的，在Spark或Hadoop中的数据块上可以独立计算总体答案的每个部分。这使得决策树构造的并行化较为容易，因为随机森林中决策树可以只在特征子集或输入数据子集上进行训练。

## 本章小结

本章首先介绍了个人社会属性与出行特征相关的研究，然后介绍了手机信令数据的数据特征以及基于位置数据的出行特征相关的研究。在目前国内外针对特定人群的出行规律研究中，大部分学者采用调查问卷或辅助数据的方式研究影响特殊人群出行方式选择的因素，较少有通过海量手机信令判别特殊人群社会属性的研究。最后介绍了常用的聚类方法与预测模型。

# **基于多源位置数据的人员特殊聚居区域识别模型研究**

特殊人群受其职业、经济能力的影响，其出行规律与一般通勤人员存在较大的差异。本章提出基于多源位置数据的人员特殊聚居区域识别模型需要解决的问题，然后对这些问题依次展开讨论。首先对手机信令数据进行数据质量分析与预处理，然后基于经过预处理过后的信令数据，采用ST-DBSCAN聚类算法识别用户的停留点与移动点，提取用户的出行链。根据用户多天的停留点数据，计算其居住地与工作地。然后对学生、工人这两类特殊人群进行个人属性标注，并提取其出行特征，基于随机森林完成用户个人属性预测模型的构建。最后，将预测出的目标人群的居住地聚类形成区域，并结合公交数据，共享单车数据完成区域特征的分析。

## 特殊人群聚居区域研究相关问题的提出

通过第二章的分析研究可知，目前关于特殊人群社会属性的研究主要是采用调查问卷的方式获取目标人群的出行信息与社会属性信息，并采用相关的概率模型研究特殊人群的出行行为与其个人属性的关系，较少有通过海量手机信令判别特殊人群社会属性的研究。出行作为一种派生性需求，个体基于活动，通过安排出行OD、出行时间和出行方式来组织日常生活，用户的出行习惯在很大程度上反映了该用户的属性。

物理类聚，人以群分。相对于社会属性差别较大的用户，具有相同社会属性的用户聚居在一起的可能性较大。因此特殊人群聚居区域识别问题的关键在于用户个体社会属性的识别。手机作为目前大部分人群随身携带的工具，其产生的手机信令数据中蕴含了丰富的人群轨迹信息。从海量的手机信令数据中挖据出特殊人群的出行特征，建立人群属性的预测模型是本文研究的第一个关键点。公交、共享单车作为大部分人群选择的交通工具，其生成的出行链数据可以更加准确地反映区域内人群的出行规律。基于手机信令分析得出的特殊人群聚居区域位置，结合公交刷卡数据，共享单车订单数据分析区域特征是本文研究的第二个关键点。

基于多源位置数据的特点，总结提炼特殊人群的出行规律以及其聚居区域的特征，本研究主要针对以下四个方面展开：

1. 手机信令数据的产生方式与用户使用手机的习惯、用户的活动区域有较大关系，这就使手机信令数据存在一些先天的缺陷。部分用户生活中不经常使用手机，部分用户长期活动于信号不良的区域，这样会导致这些用户产生的信令量较少，不足以提取其一天的移动轨迹。由于手机信令的产生机制与基站定位问题，部分信令数据会出现数据格式错误、数据字段缺失、数据抖动等问题。分析数据质量，制定信令数据预处理规则是本文首先要解决的问题。
2. 用户单日出行链、职住地数据是本研究后续分析的基础。手机信令数据包含了位置和时间两个维度信息，选取合适的聚类算法，识别用户的停留/移动态，完成用户的停留点信息与用户职住地信息的提取是构建特殊人群识别模型的基础。
3. 根据用户的出行链数据与职住地数据，结合POI兴趣点数据，分析并提取能够反映特殊人群出行规律的特征，结已知聚居区域内人群的轨迹数据，完成个人社会属性预测模型的构建，是特殊人群聚居区域识别问题的关键。
4. 将具有相同社会属性人群的居住地位置聚类形成区域。结合公交刷卡数据与共享单车数据分析区域特征，完成人员特殊聚居区域的提取是本文的最终目标。

## 手机信令数据预处理与质量分析

### 手机信令数据集介绍

本文所使用的手机信令数据来源于中国联通北京地区用户产生的4G信令数据，其中每条数据包含IMSI码（用户手机号加密），IMEI（用户设备号加密），与基站通信发生的时间，通信基站的坐标，用户手机号前7为等信息。手机信令数据的总体情况如表1所示：

表 1 手机信令数据集概况

| 属性 | 数值 |
| --- | --- |
| 数据时间间隔 | 2019年1月1日-2019年7月31日 |
| 数据记录条数 | 1216亿条记录（6.1亿条记录/天） |
| 数据量 | 28TB（约110GB/天） |
| 数据包含的用户数 | 1023万（北京2018年总人口约2154万） |
| 数据包含的基站个数 | 13.1万 |
| 数据涉及区域面积 | 16411平方公里 |

从数据集概况来看，联通4G信令数据集具有用户量多、时间跨度长、覆盖范围广等特点。在覆盖区域方面，数据集基本完成了北京市全区域的覆盖，包括市中心的城区以及远郊区县；在时间跨度方面，数据集包含了1月1日-7月31日完整7个月的信令数据，时间跨度较长，可以认为该数据集能够反映北京市用户日常的作息习惯。在用户覆盖方面，数据集覆盖了约1000万用户，约占北京市总人口的50%，考虑到部分用户具有多个手机号的情况，该数据集的用户占比基本与联通公司的市场占有率相当。从上概况可以看出，该数据集具有良好的样本代表性与广泛性。

### 数据质量分析

数据的质量在很大程度上影响了后续分析的结论。手机信令数据的质量分析主要包括两部分，即空间精度与时间精度。在空间精度方面，联通4G信令数据采用GSM定位方法，信令数据中的位置信息即为基站的坐标，因此基站的分布情况对空间精度有较大的影响。在时间精度方面，不同的用户具有不同的使用手机的习惯，因此时间精度具有较高的不确定性，本节从基站空间分布层面与用户层面对手机信令数据进行质量分析。

1. 基站空间分布

北京市内联通公司共计布设约有26万个基站，其中4G基站约有13万个。所有基站的位置分布如图8所示，可以看到西南角与东北角基站分布较为稀疏，市区较为密集，市区部分的基站分布满足后续计算要求。



图 8 北京市基站分布图

1. 更新周期

信令数据的更新周期指同一用户产生相邻2条信令的时间间隔，以10分钟为单位统计信令数据的更新周期，具体数据如图9所示。可以看到大部分信令数据的更新周期都在20分钟以内，占比达到了80%以上，有约8%的信令数据更新周期在60分钟以上，占比较小。4G信令数据的更新周期满足后续分析的要求。

图 9 信令数据更新周期分布图

1. 信令数量

信令数量指同一用户一天内使用手机产生的信令数量，可以反映用户的活跃度与使用手机的频繁程度。表2展示了2019年3月1日至2019年3月5日信令数据的数量，可以看到约有85%左右的用户一天使用手机产生的信令数据量大于100条，平均每名用户约15分钟产生一条信令数据，与信令更新周期的统计结果相符。4G信令数据的信令数量满足后续分析要求。

表 2 用户信令数量统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表目  时间 | >=120条 | [100条,120条) | [80条,100条) | <80条 | 大于100条用户占比 |
| 3月1日 | 6059064 | 3354782 | 909847 | 796980 | 84.65% |
| 3月2日 | 8191697 | 1369755 | 761320 | 744577 | 86.06% |
| 3月3日 | 7964527 | 1467329 | 756132 | 735754 | 86.34% |
| 3月4日 | 6789012 | 2764521 | 803726 | 751504 | 86.00% |
| 3月5日 | 7788073 | 2645116 | 764356 | 742412 | 85.7% |

1. 本地用户数与外地用户数

从4G信令数据中的手机号前7位字段中我们可以获取用户的号码归属地，通过判断用户的号码归属地我们可以将用户分为本地用户与外地用户。表3展示了2019年2月1日至2019年2月6日春节期间本地用户数量与外地用户数量，表4展示了2019年3月1日至2019年3月6日工作日期间本地用户数量与外地用户数量。由表3和表4可以发现，相较于正常工作期间，春节期间的总用户数量，北京市用户数量，外省市用户数量相对减少，北京用户占总用户比重相对增多。春节期间，大量人口从北京流出，其中包含大部分回家过年的外地用户。对信令数据的用户数量统计符合这一社会现象。

表 3 春节期间本地用户与外地用户数量

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 时间 | 总用户数量 | 北京市用户数量 | 外省市用户数量 | 北京用户占比 |
| 2019年2月1日 | 8301483 | 6094526 | 2206957 | 73.41% |
| 2019年2月2日 | 7685929 | 5677979 | 2007950 | 73.87% |
| 2019年2月3日 | 7014873 | 5263701 | 1751172 | 75.04% |
| 2019年2月4日 | 6704276 | 4829071 | 1875205 | 72.03% |
| 2019年2月5日 | 6202471 | 4628024 | 1574447 | 74.62% |
| 2019年2月6日 | 6238036 | 4539983 | 1698053 | 72.78% |

表 4 工作期间本地用户与外地用户数量

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 时间 | 总用户数量 | 北京市用户数量 | 外省市用户数量 | 北京用户占比 |
| 2019年3月1日 | 11123259 | 7770352 | 3352907 | 69.86% |
| 2019年3月2日 | 11120673 | 7769271 | 3351402 | 69.86% |
| 2019年3月3日 | 10937641 | 7689015 | 3248626 | 70.30% |
| 2019年3月4日 | 10719763 | 7290878 | 3428885 | 68.01% |
| 2019年3月5日 | 10826754 | 7345669 | 3481085 | 67.85% |
| 2019年3月6日 | 11065743 | 7709367 | 3356376 | 69.67% |

通过上述分析可知，约有80%以上的用户平均每天产生100条以上的信令数据，且数据的更新周期在20分钟以内。对比分析春节期间与工作日期间北京市用户数量与外地用户数量可知，信令数据准确地反映了春节期间人群迁徙这一社会现象。4G信令数据质量较好，可以支撑本文后续的分析。

### 数据预处理

1. 手机用户筛选

不同的用户可能具有不同的使用手机的习惯，为保证后续分析过程的顺利进行，需过滤出使用手机频度较高的用户。针对这些活跃度较高的用户才能精确地分析其出行特征，为此制定活跃用户判别规则如下：

* 用户在0点-7点间每两个小时产生一条信令记录
* 用户在8点-19点间每个小时产生一条信令记录
* 用户在20点-24点每两个小时产生一条信令记录

表5展示了2019年3月1日-2019年3月5日信令数据中的活跃用户数，可

看到约有75%的用户满足活跃用户的判别规则，活跃用户量足以支撑本文后续的分析。

表 5 活跃用户数量

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 总用户数量 | 活跃用户数量 | 活跃户占比 |
| 2019年3月1日 | 11123259 | 8342444 | 75.16% |
| 2019年3月2日 | 11120673 | 8341504 | 76.36% |
| 2019年3月3日 | 10937641 | 8203030 | 74.85% |
| 2019年3月4日 | 10719763 | 8039522 | 75.34% |
| 2019年3月5日 | 10826754 | 7345669 | 73.95% |

1. 基站抖动处理

基站抖动主要指由数据传输、存储错误、信号不稳定等因素造成的手机在不同基站间来回切换的现象。具体可分为两种情况：短距离基站抖动与长距离基站抖动。

短距离基站抖动一般发生在基站覆盖较为密集的区域，如居民区、商场等。当用户处于基站覆盖区域的重叠部分时，由于信号强度不稳定等原因，手机会在基站间来回切换，在信令数据上表示为用户在短时间内在多个临近的位置点间来回移动。短距离基站抖动一般会造成300m-1000m左右的定位误差，可以在后续的分析过程中采用聚类算法处理这一现象。

长距离基站抖动一般发生在郊区等基站覆盖较为稀疏的区域。其具体表示为用户在短时间内进行长距离的往返移动，图10展了长距离基站抖动的一个样例，其中某用户在7:31:10出现在基站1附近，在7:31:49出现在基站2附近，在7:32:17又回到了基站1附近。两次行程的速度均超过了2000km/h，此速度值远超过正常交通载具的运行速度，属于异常数据。

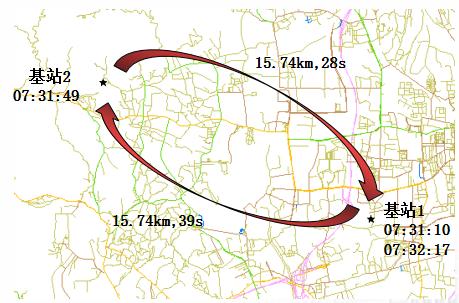


图 10 长距离基站抖动样例

长距离基站抖动会造成较大的定位误差，具体误差值在几千米至几十千米不等。对发生此类现象的数据需予以处理，长距离基站抖动数据的特征如下：

|  |  |
| --- | --- |
|  | （3.1） |

其中表示距离阈值，表示速度阈值，表示信令数据i-1与i之间的距离，表示信令数据i-1与i之间的速度。考虑一般交通载具的运行速度与手机信令数据的定位精度，速度阈值取180km/h，距离阈值取2km。对于满足上述特征的信令数据i予以删除处理。

## 用户出行链与职住地提取

手机信令数据预处理过程主要完成了活跃用户筛选与长距离基站抖动数据剔除。经过预处理过程后，我们可以得到质量较高的手机信令数据。本章的目标是在预处理过后数据的基础上，完成用户单日出行链与职住地的提取。

### 用户出行链提取

1.相关定义

* + 1. **原始信令数据**

原始信令数据是经过预处理过程后的信令记录，主要包含了id（用户手机号加密），通信事件发生的时间，位置点信息等：

|  |  |
| --- | --- |
|  | （3.2） |

其中，id表示经过imsi码（手机号码加密）唯一标识的用户，time表示通信事件发生时间，lng表示位置点的经度，lat表示位置点的纬度。

* + 1. **停留点**

用户在某个区域停留会产生一系列时空相近的信令数据。一个停留点sPoint由一簇信令数据cD生成：

|  |  |
| --- | --- |
|  | （3.3） |

其中，cD(lng,lat)表示sPoint的中心位置，cD(st,et)表示sPoint的起止时间：

|  |  |
| --- | --- |
|  | （3.4） |
|  | （3.5） |

* + 1. **移动点**

用户处于移动状态会产生一系列时空不临近的信令数据，其由聚类算法中无法被归类到簇集中的离群点组成：

|  |  |
| --- | --- |
|  | （3.6） |

其中lng表示移动点的经度，lat表示移动点的纬度,time表示移动事件发生的时间。

* + 1. **OD（Origin-Destination）出行**

OD出行表示用户在某一时间从一个停留点移动到另一个停留点，其包含了一个起始停留点(Origin)，一个终止停留点(Destination)以及若干的移动点：

|  |  |
| --- | --- |
|  | （3.7） |

* + 1. **OD出行链**

OD出行链由一组首尾相连的OD构成，表示用户一天内多次出行的集合。其中一个OD的终止停留点是下一个OD的起始停留点，起始停留点是上一个OD的终止停留点：

|  |  |
| --- | --- |
|  | （3.8） |

2.停留点提取

停留点、移动点的提取是识别用户出行链的关键。用户在某一个区域停留了一段时间在信令数据上表示为一系列时空维度上相近的点的集合。通过对聚类算法的研究，密度聚类算法DBSCAN比较适用于解决此类问题。传统的DBSCAN算法只考虑了空间这一单一维度，其不适用于处理手机信令这种具有多个维度的数据，如图11所示，用户在T1时刻位于某区域，产生了簇集A，T2时刻停留在另一个区域，产生了簇集B，T3时刻又返回原区域，产生了簇集C。DBSCAN算法由于没有考虑时间维度，在此场景下会将簇集A与簇集C识别成一个停留点，但由于A与C在时间上不连续，因此应将A与C识别成两个停留点。为此引入ST-DBSCAN算解决上述问题。

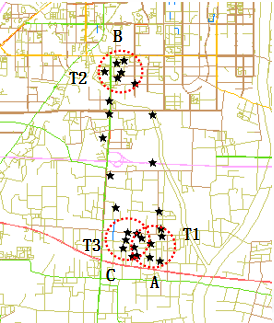


图 11 多个点簇分布在同一空间内的示例

ST-DBSCAN算法需要输入3个参数：D={}，即数据集；Eps，即邻域的半径；MinTimeRange，即最短时间跨度阈值。算法的流程如下所示：

| **算法：**面向时间序列数据的空间密度聚类算法（ST-DBSCAN） |
| --- |
| **输入：**D={x1,x2,.....,xn}；Eps：ε-邻域值；MinTimeSpan：时间阈值  **输出：**Trj：停留/移动点序列  **流程：函数ST-DBSCAN(D,Eps,MinTimeSpan)** |
| 1：初始化核心对象集合：，将D中样本按时间排序  2：**for** j=1,2,...,n **do**  3：令xj的领域为N(xj);  4：**if** Interval(N(xj))≥MinTimeSpan **then**  5： 将样本xj入核心对象集合：  6： **end if**  7：**end for**  8：初始化聚类簇数：k=0  9：初始化未访问样本集合：  10：**while**  **do**  11： 记录当前未访问样本集合：;  12： 随机选取一个核心对象，初始化队列;  13： ;  14： **while**  **do**  15： 取出队列中的首个样本q;  16： if Interval(N(xj))≥MinTimeSpan then  17： 令;  18： 将中的样本加入到队列;  19： ;  20： **end if**  21： **end while**  22： k=k+1，生成聚类簇  23：  24：**end while**  25： 将聚类簇中的点生成停留点，没有被加入到簇集中的点生成移动点，输出Trj |

其中，获取xj 的ε-邻域N(xj)的算法流程如下：

| **算法：获取样本x的ε邻域算法** |
| --- |
| **输入：**D={x1,x2,.....,xn}；Eps：ε-邻域值；样本xk  **输出：**样本xk的ε-邻域N(xk) |
| **流程：**  1: **for** j=k+1,k+2,...,n **do**  2: 计算xj 与xk 间的距离dj ;  3: **if** dj < Eps **then**  4: 将xj 加入到N(xk)中;  5: **else**  6: **break**;  7: **end if**  8: **end for**  9: **for** i=1,2,...,k-1 **do**  10: 计算xi 与xk 间的距离di ;  11: **if** di< Eps **then**  12: 将xi 加入到N(xk)中;  13: **else**  14: **break;**  15: **end if**  18:**end for**  19:输出N(xk); |

由上述算法描述可以看出，相较于传统的DBSCAN算法，面向信令数据的ST-DBSCAN算法主要有两点改进：即时间阈值的加入与选择样本邻域方法的改变。时间阈值限制了簇集中样本点的最短时间跨度值，选择样本邻域的方法由随机扩展方式改为按时序扩展方式，这样保证了一个簇集中的样本点在时间上是连续的。

ST-DBSCAN算法需要输入两个参数：Eps值与MinTimeSpan值。为了选择合适的Eps值，本文以100米为采样点，绘制了相邻信令记录间的距离分布图，如图12所示，相邻信令的间距的峰值在300m-400m之间，考虑到一般人群的停留时长远大于移动时长，因此信令记录间的峰值说明了用户处于停留态时基站的抖动误差在300m-400m之间。为了消除短距离基站抖动对识别停留点的影响，Eps取值应大于400m。本文将Eps值设定为500m，MinTimeSpan值设定为30分钟。

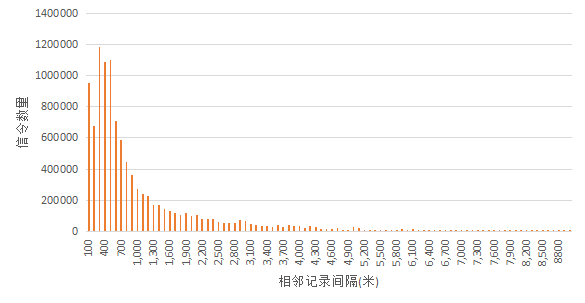


图 12 相邻信令间的距离分布图

### 用户职住地提取

完成了用户的停留点提取后，可以识别用户的停留态与移动态，进而完成用户单日出行链的提取。为了给用户的出行链赋予更丰富的语义，需识别用户的居住地、工作地与其他种类的停留点。用户的职住地定义如下：

|  |  |
| --- | --- |
|  | （3.9） |

其中lng与lat表示职住点的经纬度，kind表示职住点的属性，分为居住地，工作地与未知三种。

根据第二章有关用户停留点属性的调研发现，职住地的提取思路是根据用户停留点的停留时间长度与停留时间段来分析用户的停留点属性。根据人群日常生活规律可知，用户一天的作息习惯为早上从居住地出发去往工作地，晚上从工作地出发前往居住地。本文将用户单日出行链中第一个停留点的结束时间作为用户的离家时间，最后一个停留点的开始时间作为用户的回家时间。图13展示了用户的出行时间分布图，可以看到大部分用户离家时间在7点到8点间，回家时间在19点到20点间。因此本文定义0点-7点，20点-23点为居家时间段；7点-20点为工作时间段。

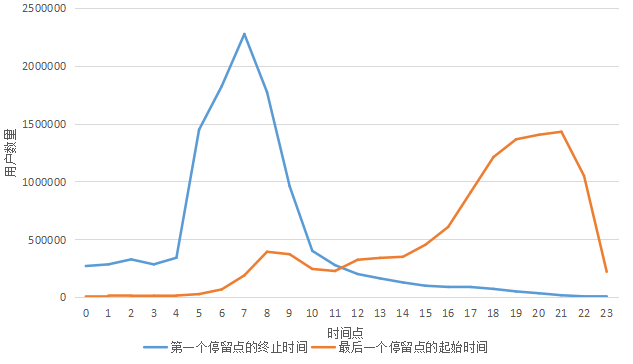


图 13 用户出行时间分布图

用户单一停留点的职住属性规则如下：对于用户出行链中的某一个停留点sPoint，如果其停留时间段与工作时间段的交集时长超过该停留点停留时长的50%，且交集时长大于3小时，则把该停留点标记为工作地，即：

|  |  |
| --- | --- |
|  | （3.9） |

且

|  |  |
| --- | --- |
|  | （3.10） |

其中表示该停留点与工作时间段交集的起止时间，表示该停留点的起止时间。

对于用户出行链中的某一个停留点sPoint，如果其停留时间段与工作时间段的交集时长大于2小时，且交集时长超过该停留点停留时长的50%，则把该停留点标记为居住地，即：

|  |  |
| --- | --- |
|  | （3.11） |

且

|  |  |
| --- | --- |
|  | （3.12） |

其中表示该停留点与工作时间段交集的起止时间，表示该停留点的起止时间。

人群的生活作息习惯一般都具有周期性，对于一般通勤人群，其在工作日期间晚间在居住地休息，日间在工作地工作，为更加准确地提取用户的职住地，本文采用用户多日的停留点数据计算其职住地信息，具体思路如下：

* 采用DBSCAN算法将用户一周内工作日期间的停留点聚类形成簇集。
* 对于一个簇集中每个停留点，根据上述单一停留点职住属性判别规则判断其职住属性。
* 若簇集中属性为工作地的停留点数量大于属性为居住地的停留点数量，且具有工作属性的停留点集合的时间跨度大于3天，则计算该簇集的中心点作为该用户的工作地位置；若簇集中属性为居住地的停留点数量大于属性为工作地的停留点数量，且具有工作属性的停留点集合的时间跨度大于3天，则计算该簇集的中心点作为该用户的居住地位置；若上述条件均不满足，则判断该簇集的属性为未知，计算其中心点的位置作为该用户的未知属性点位置。

算法的流程图如图14所示：

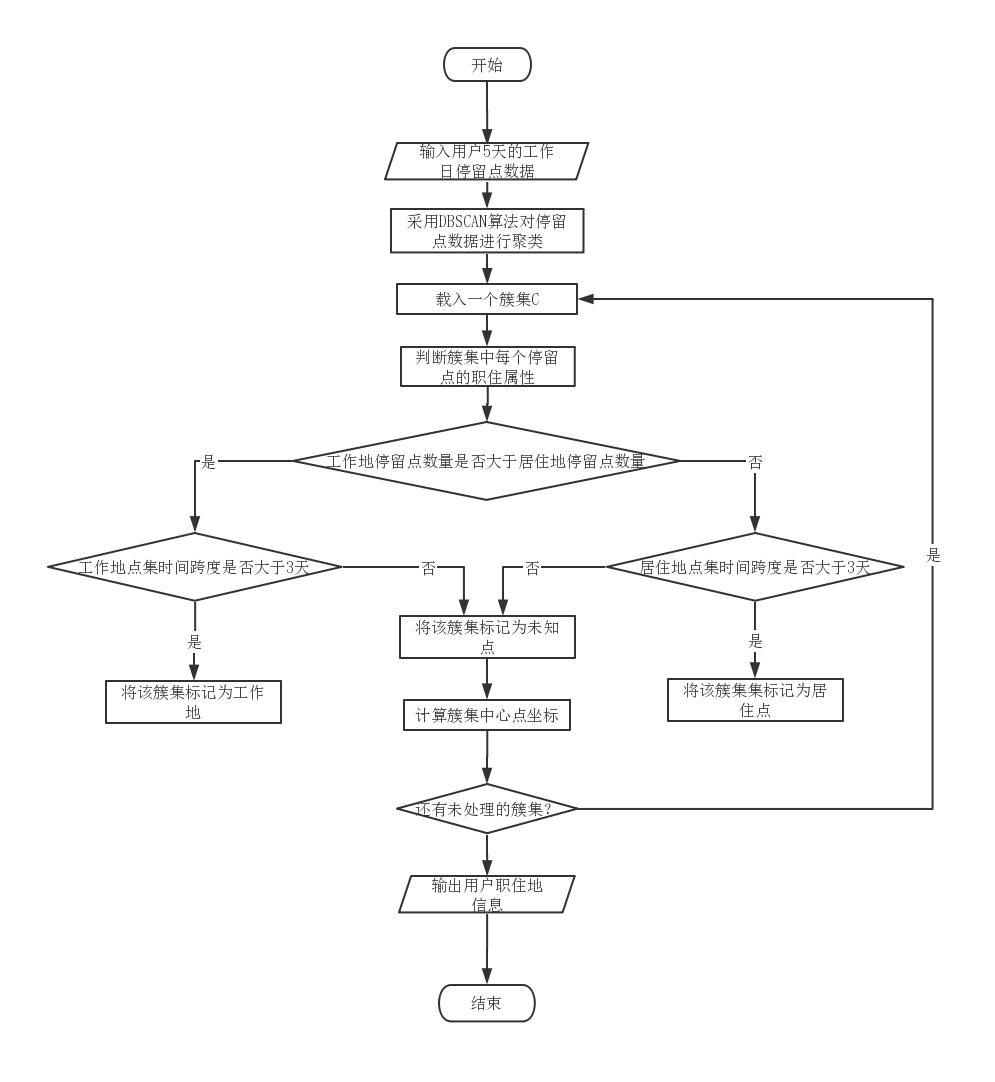


图 14 用户职住地识别算法流程图

经过上述处理，可以得到用户的职住地信息。

## 用户个人属性预测

完成了基于手机信令数据的出行链与职住地信息提取后，我们可以获得语义相对丰富的出行数据。经过第二章关于特殊人群出行规律的研究可知，受职业，经济条件等诸多因素的影响，其出行规律与一般通勤人员差异较大。本节的目标是根据人工标注的目标人群的训练集，构建预测模型，完成学生、工厂工人这两类特殊人群的属性预测。

### 特征设计

1. 单日出行特征
2. 离家时间，回家时间

离家时间指用户一天中第一次离开居住地的时间，回家时间指用户一天中最后一次回到居住地的时间，本文以天为单位计算上述两个特征。经过对数据研究发现，部分用户一天当中没有出行行为，此时将离家时间与回家时间均设为0；部分用户一天当中没有回家行为，此时将回家时间设为该用户当天最后一个停留点的结束时间。计算离家时间与回家时间的算法流程图如图15所示：

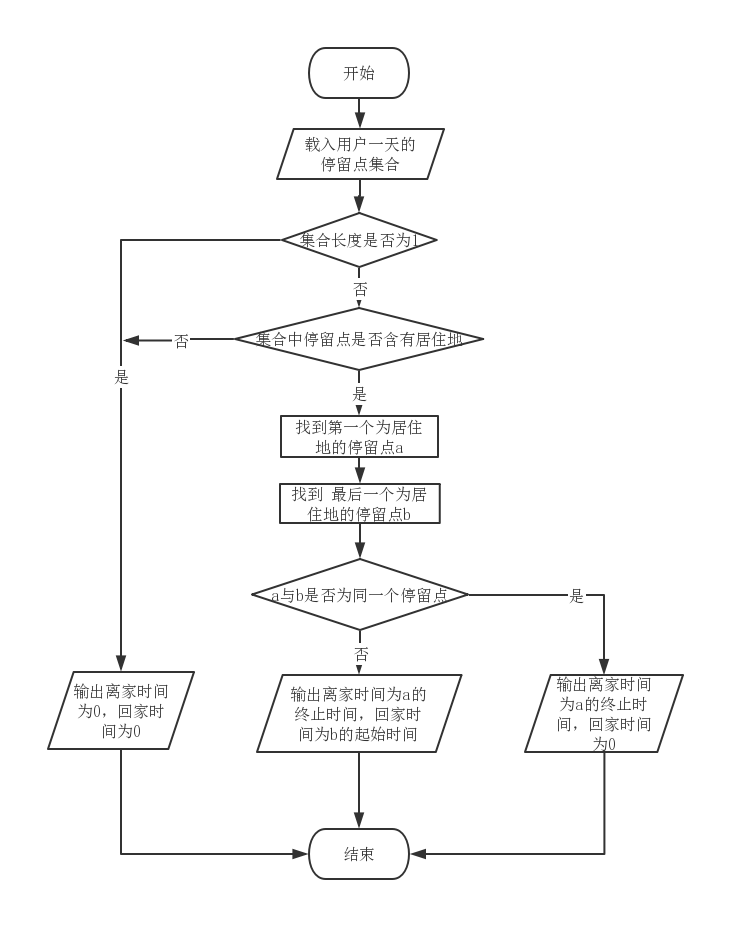


图 15 用户出行时间算法流程图

1. 移动距离

移动距离指用户单日出行链中各个OD间的球面距离之和，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.13） |

1. 出行时间

出行时间指用户一天内处于移动状态的时间之和，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.14） |

1. 移动范围

移动范围指用户一天内停留位置的中心点到各个停留点距离的均值，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.15） |

1. 移动频率

移动频率指用户一天内的出行次数，具体为用户单日出行链中的停留点个数减1。

1. 多日出行特征
2. 轨迹相似性

轨迹相似性的计算依赖于相似度函数。一般思路为将轨迹上的采样点进行匹配，通过计算各对采样点间的距离来完成轨迹相似度的计算。由于用户使用手机的习惯存在较大的差异，相较于GPS数据，信令数据的采样率较低，且采样率不稳定，因此匹配不同轨迹的采样点较为困难。

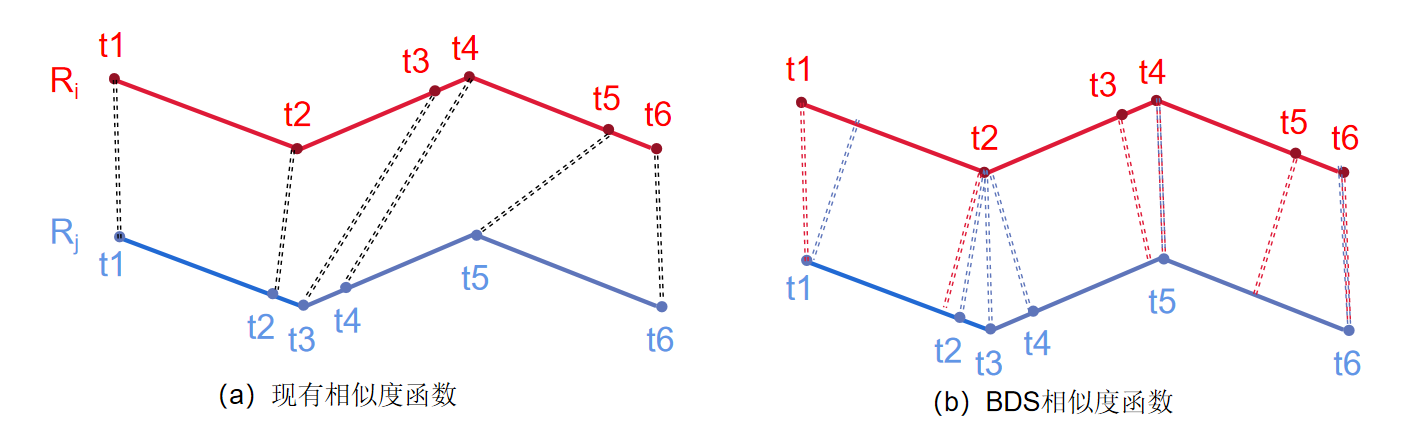


图 16 轨迹相似度对比图

图16展示了两条相似度较高的轨迹Ri与Rj，其中图16(a)采用现有的相似度函数，即将两条轨迹的采样点依据时间信息进行匹配，由于两条轨迹的采样率不同导致采样点匹配的效果不是很理想，据此函数求得的轨迹相似度较小，无法准确地衡量两条轨迹的相似程度。为此引入BDS(Bi-Directional Similarity)，即双向轨迹函数来计算两条轨迹的相似程度。如图16(b)所示，对于Ri上的任一采样点tk，BDS函数将其匹配到Rj上距离其最小的点，该点可以不是采样点；相对应地，Rj上的任一采样点也将被匹配到Ri上距离其最小的点。相较于现有的轨迹相似度函数，根据BDS函数会计算得出一个较大的相似度，从而更加准确地衡量两条轨迹的相似情况。

对于一条轨迹R，其由|R|个采样点与|R|-1个轨迹段组成，其中轨迹段由相邻的两个采样点与两个采样点所连线段上的无穷多个位置点组成。给定两条轨迹Ri与Rj，BDS函数的主要思路是将Ri上的任一采样点匹配到Rj上与距离其最近的点，相对应地，将Rj上的任一采样点匹配到Ri上距离最近的点。采样点到Rj上的最小距离可形式化表示为：

|  |  |
| --- | --- |
|  | （3.19） |

其中表示采样点到Rj中轨迹段r的最短距离，具体地，若到r的垂足落在r的内部，则的值为到r的垂线段距离；否则的值为到轨迹段r两个端点的距离较小值。

一般而言，轨迹相似度的值位于[0,1]区间内，需要对BDS函数得出的结果值进行归一化处理。为此可以设立一个距离阈值,当的值大于时，则表示采样点距离Rj距离过远，没有可以匹配的位置点，即轨迹Ri与轨迹Rj并不相似。的取值由信令数据的定位精度决定，本文中取值为300米。归一化后的结果可形式化表示为：

|  |  |
| --- | --- |
|  | （3.19） |

给定两条轨迹Ri与Rj，其BDS相似度函数可形式化定义为：

|  |  |
| --- | --- |
|  | （3.19） |

对于某个用户一周的信令数据，采用BDS函数计算其每两天轨迹的相似度，取多个轨迹相似度的均值作为该用户一周内的轨迹相似度。

1. POI特征

兴趣点（Point of Interest）在地理信息系统中表示点对象，其中包含了对象的属性，位置等信息。在本文中主要指与特殊人群聚居区域相关的地理实体，如大学、中学、小学等教育设施；居住小区以及工厂等。POI的分析对于特殊人群的个人属性预测具有较为重要的影响，因此本文选取了POI特征作为出行特征之一。

考虑到学生，工厂工人日常的活动区域，本文从互联网地图中爬取了学校、住宅区、工厂这三种类型的POI。并结合POI中的位置信息与停留点中的位置信息，提取停留点所属的POI类别。具体方法为以停留点的位置为中心点，将其扩张为边长600米的方形区域，统计该区域内各个种类的POI数目，以数目最多的种类为该停留点的POI类别。结合停留点中的POI信息与本文在用户职住地提取部分中定义的居家时间段与工作时间段，本文给出如下POI特征的定义：

1. 用户在学校区域内停留时长与居住时长的比率，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.19） |

其中home.et表示居家时间段的终止时间，home.st表示居家时间段的起始时间，sPoint.home.et表示停留点与居家时间段交集的终止时间，sPoint.home.st表示停留点与居家时间段交集的起始时间，停留点的种类为学校。

1. 用户在学校区域内停留时长与工作时长的比率，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.20） |

其中work.et表示工作时间段的终止时间，work.st表示工作时间段的起始时间，sPoint.work.et表示停留点与工作时间段交集的终止时间，sPoint.work.st表示停留点与工作时间段交集的起始时间，停留点的种类为学校。

1. 用户在工厂区域内停留时长与居住时长\工作时长的比率

其定义与用户在学校区域停留时长的比率类似，不同的是停留点的种类为工厂。

1. 特征相关性分析

本文基于用户一周5天工作日的信令数据，从均值与标准差两个层面统计上述特征，最终得到的特征如表6所示：

表 6用户出行特征说明表

|  |  |  |  |
| --- | --- | --- | --- |
| 特征编号 | 特征含义 | 特征编号 | 特征含义 |
| 1 | 离家时间均值 | 10 | 离家时间标准差 |
| 2 | 回家时间均值 | 11 | 回家时间标准差 |
| 3 | 移动距离均值 | 12 | 移动距离标准差 |
| 4 | 移动范围均值 | 13 | 移动范围标准差 |
| 5 | 出行时间均值 | 14 | 出行时间标准差 |
| 6 | 轨迹相似性均值 | 15 | 轨迹相似性标准差 |
| 7 | 出行频率均值 | 16 | 出行频率标准差 |
| 8 | 学校/工厂居住时长比率均值 | 17 | 学校/工厂居住时长比率标准差 |
| 9 | 学校/工厂工作时长比率均值 | 18 | 学校/工厂工作时长比率标准差 |

基于皮尔逊系数计算用户各个特征之间的相关性，计算公式为：

|  |  |
| --- | --- |
|  | （3.21） |

取值范围为[-1,1]，其中-1代表两个特征呈负线性相关，1代表两个特征呈正线性相关。特征相关性的热力图如图16所示：

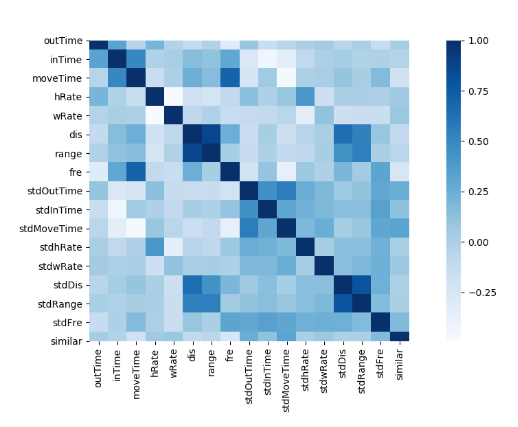


图 17 用户出行特征热力图

可以看到用户出行距离与出行范围相关性较高，因此将出行距离均值与出行距离标准差两个特征剔除，其余特征相关程度较低。

### 训练集标注

由于隐私保护的限制，手机信令数据中并没有包含有关用户个人属性的信息。通过对手机信令数据的质量分析可知，大部分用户产生的信令间隔在10分钟左右，高频的信令数据能够较为完整地反映用户的活动规律，因此结合地图数据与POI数据，能够较为可靠地根据用户的移动轨迹标识出用户的个人属性。本文以北航知春路校区与大兴安定化工厂两个典型人群聚居区域为例，介绍人群属性的标注方法，本文所标注的两类特殊属性人群的定义如表7所示：

表 7 标定人群属性定义

|  |  |
| --- | --- |
| **名称** | **定义** |
| 大学生 | 在高等院校正在接受教育，主要活动地位于院校内的人群 |
| 工人 | 从事劳动生产相关行业，主要活动地位于工厂内的人群 |

1. 北航区域学生人群标注

北京航空航天大学知春路校区占地面积1km2，是典型的学生人群聚居区域，本文根据POI信息与生活经验，将其划分为两个工作区与两个居住区，具体如图17所示：



图 18 北航区域划分图

北航区域为黑色方框内区域。其中工作区1为北航主要的教学区，工作区2为北航部分学科实验室的聚集区，居住区1与居住区2为学生主要的居住区域。本文基于5天工作日的出行链数据，首先筛选出有3天以上在北航区域停留8小时以上的用户。然后对于用户每天的出行链，计算其停留在居住区1/2的停留时长与居家时间段的交集时长，以及其停留在工作区1/2的停留时长与工作时间段的交集时长，并计算上述两个交集时长的平均值。最后根据上述两个指标，本文采用Kmeans算法对北航区域内的人群进行分类。

Kmeans算法需要预先设定需要划分的类簇数k，本文使用肘系数方法确定k的数值。肘系数的定义如下：

|  |  |
| --- | --- |
|  | （3.22） |

其中，是第i个簇，p是中的样本点，是的质心，SE是所有样本的聚类误差，代表聚类的好坏。随着K值的增大，样本分割更加精细，每个簇的聚合程度将逐渐增加。当K小于实际簇时，K的增加将大大增加每个簇的聚合度，因此SSE下降幅度较大，而当K到达实际聚类数时，通过增加K所得到的聚集程度回报将迅速减小，然后随着K值的继续增大，趋于平缓，SSE和k值的关系图是一个手肘的形状，肘部对应K值就是数据的真实聚类数。图18展示了具体的肘系数图：

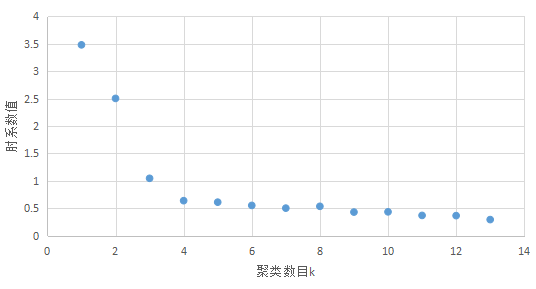


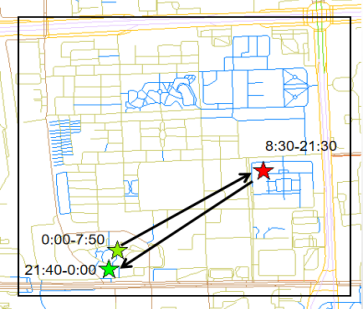
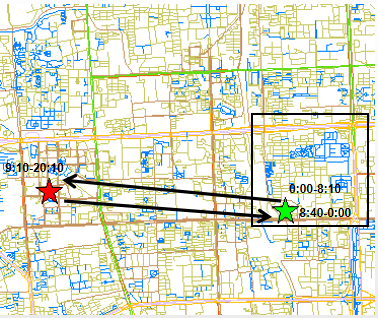
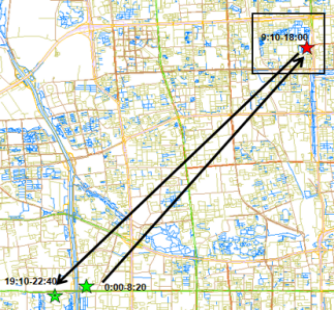
图 19 Kmeans类簇数量选择

可以看到当类簇数取4时，肘系数值的变化趋于平缓，因此k值取4。每个簇集的中心点数据如表8所示：

表 8 北航区域人群划分结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | 居住区交集时长（小时） | 工作区交集时长（小时） | 人数 | 占总人数比例 |
| 1 | 8.39 | 9.49 | 3951 | 22.15% |
| 2 | 7.89 | 0.78 | 1504 | 8.43% |
| 3 | 1.16 | 8.97 | 3406 | 19.10% |
| 4 | 0.96 | 2.05 | 8976 | 50.32% |

可以看到第1类人群日间在工作区活动时长较高，夜间在居住区活动时长较高，这类人群一般以学生为主；第二类人群夜间在居住区活动时长较高，日间几乎不在工作区活动，这类人群以外出实习的学生及居住在北航区域里的通勤人员为主；第三类人群日间在工作区活动时长较高，夜间几乎不在居住区活动，这类人群以工作地在北航区域的教职工为主；第四类人群由于基站定位精度以及区域划分偏差的影响，其活动规律表现为日间与夜间均不在划分区域内活动。图19展示了前三类人群中部分样本一天的活动轨迹：

**(a) 第1类人群样本轨迹 (b) 第2类人群样本轨迹 (c) 第3类人群样本轨迹**

图 20 北航区域人群样本轨迹图

从左至右分别为第一类人群、第二类人群及第三类人群中部分样本的活动轨迹。其中黑色框内为北航区域，绿色点为其停留时长位于居住时间段内的停留点，红色点为其停留时长位于工作时间段内的停留点。可以看到第一类样本符合学生的出行特征，第二、三类样本符合一般通勤人员的出行特征。本文将第一类人群标记为学生群体，第二、三类人群标记为非学生群体。

1. 安定化工厂人群标注

北京安定化工厂位于北京市大兴区安定镇东侧，占地面积约0.3km2 ，是较为典型的集生产、住宿及经营为一体的工厂区域。

由于工厂周边基站覆盖密度较低，且其地图信息较少，无法在工厂区域内再更加细致地划分区域。因此对于用户每天的出行链，本文只计算其停留在整个工厂区域的停留时长与居家时间段的交集时长，工作时间段交集时长的计算方法类似。安定化工厂区域人群的划分结果如表9所示：

表 9 安定化工厂区域人群划分结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | 居住时间段交集时长（小时） | 工作时间段交集时长（小时） | 人数 | 占总人数比例 |
| 1 | 9.39 | 9.21 | 307 | 42.82% |
| 2 | 8.19 | 1.46 | 137 | 19.11% |
| 3 | 0.95 | 8.97 | 119 | 16.60% |
| 4 | 3.73 | 4.76 | 154 | 21.48% |

由于表9可以看到第一类人群日间与夜间几乎都在工厂区域内活动，比较符合本文所定义工人的特征；第二类人群为居住地在工厂区域附近的通勤人员；第三类人群为工作地在工厂区域的通勤人员；第四类人群活动特征不明显。本文将第一类人群标注为工人群体，将第二类与第三类人群标注为非工人群体。

### 用户个人属性预测模型

由于涉及隐私泄露等问题，一般而言信令数据中并不包含用户的个人属性等信息。通过调查问卷与人工标注等方式获取标注信息不仅会耗费大量的资源，获得的信息量也较少。本文基于SC-Kmeans半监督聚类算法构建了用户个人属性预测模型，针对标注数据较少，信息量不足等问题，采用主动学习和基于Seed集的扩充算法来提升标注数据集的信息量；针对由用户使用手机频率不同导致的特征差异较大的问题，采用基于L2范数的欧式距离作为聚类算法中的度量函数，提升模型的预测精度。

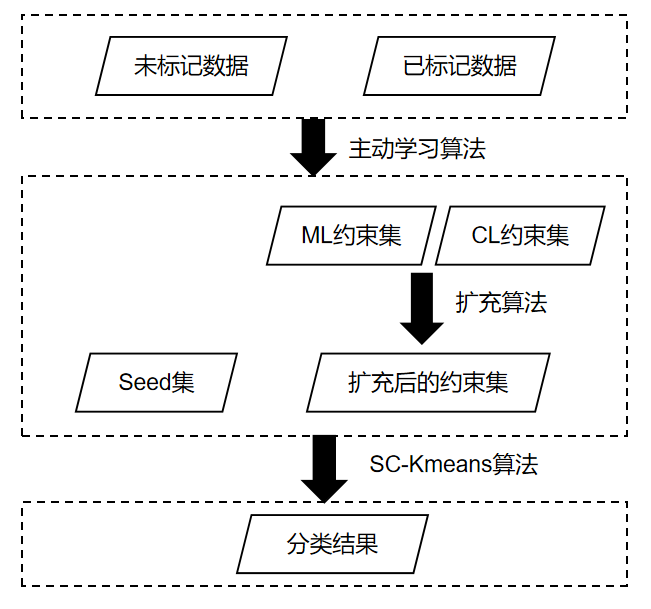


图 21 个人属性预测模型框架图

1. 模型框架

SC-Kmeans算法是一种基于距离度量的半监督聚类算法，其运用到了两种形式的监督信息：包含少量标签信息的Seed集以及must-link和cannot-link这两种约束信息。具体地，对于给定数据集，其中代表m维的特征向量。代表正约束对的集合，即集合中任一对约束对所包含的样本都属于同一类别；相反地，代表负约束集合，其中约束对所包含的样本不属于同一类别。由于约束对需满足传递性与传递性，因此上述两种约束信息还需被扩展，例如上述must-link被扩展为，cannot-link被扩展为。由数据集D与被扩展后的must-link集合可以构建邻域集合，其可以表示为以D中的一个样本为点，must-link集合中约束为边的个连通分支。同一邻域的样本均具有正约束，不同邻域中至少有一对样本具有负约束。

对于半监督学习而言，先验信息中所蕴涵的信息量越丰富，学习所得到的结果也就越精确。为此，本文采用两种方式扩充先验信息，一是基于最远距离优先策略的主动学习算法，二是利用Seed集中样本的标签信息对约束集进行扩充。同时针对部分特征差异过大的问题，引入基于L2范数的欧几里得距离作为模型中样本的距离度量函数，提升模型的预测精度。

1. 基于L2范数的欧式距离

在聚类算法中，往往需要依据距离函数来刻画样本间的相似性，通过样本间的相似性度量来确定样本所属的类别。常用的距离函数有欧氏距离、曼哈顿距离以及马氏距离等。一般地，距离函数dis(xi,xj)需要满足以下几个性质：

|  |  |
| --- | --- |
|  | （3.22） |

传统的欧式距离是衡量两个样本间的绝对距离。由于用户使用手机频率的差异较大，导致其特征向量在特征幅度上差异较大，欧氏距离不能准确地衡量样本间的相似性。为说明方便，本文采用二维向量来说明上述问题。给定3个特征向量：x1=(8,4)，x2=(4,3)，x3=(8,0)。其中向量的第一维度表示用户在学校区域的停留时长，第二维度表示用户的移动时长。x1与x2为工作在校区的教职人员，x3为校区内的学生。采用欧式距离计算可得知dis(x1,x2)>dis(x1,x3)，即相较于x2，x1与x3更加相似。欧式距离度量用户相似性的结果并不正确，其主要原因在于不同用户使用手机的频率不同，x2所代表的用户使用手机频率较低，其特征值在幅度上与其他用户的差异较大。因此采用欧式距离作为度量函数并不合理。

为较小特征幅度对用户相似性度量结果的影响，本文引入了基于L2范数的欧氏距离，给定向量与，其可形式化表示为：

|  |  |
| --- | --- |
|  | （3.22） |

其中表示向量x的L2范数。易知上式满足距离函数的四个性质。基于L2范数的欧式距离将向量缩放到单位范数，以此减少特征幅度对用户相似性度量结果的影响。以上文所述的3个向量为例，其经过缩放的结果如图22所示：

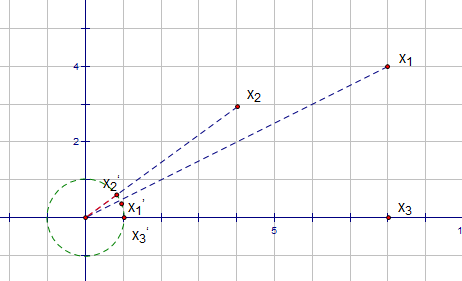


图 22 不同距离函数的相似性度量对比

采用基于L2范数的欧式距离计算可得知dis(x1,x2)<dis(x1,x3),即相较于x3，x1与x2更加相似，该结果更加符合真实情况。

1. 先验信息的扩充
2. 主动学习

半监督聚类的结果往往取决于先验数据中所蕴涵的信息量。当先验数据的数据量越多，其分布也就越接近于样本集的整体分布，聚类的效果也就越好。在实际应用场景中，获取大量的先验数据需要耗费大量的人力与物力，因此如何在拥有较少的先验数据的条件下，获取较优的聚类结果是本文需要解决的一个主要问题。主动学习是一种解决该问题的数据选择技术，其不再需要大量的先验数据，而是基于某种主动选择策略，筛选出具有信息量较多的样本交由用户去标注，并将标注好的数据扩充到先验信息的集合中。

在基于距离的Kmeans算法中，两个距离相近且类别相同的先验样本所构成的约束对所蕴含的信息量较低，其原因在于即使没有这类先验数据，算法也会将两个距离较近的样本划分到同一类别。相反地，类别相同但距离较远的样本和类别不同但距离较近的样本所蕴含的信息量就较高。获取蕴涵信息量较多的约束对是主动学习算法的第一个目标。初始聚类中心的选择会对Kmeans算法的性能有较大的影响，在半监督聚类中主要依据Seed集构建聚类中心，因此获取分布较为均匀的Seed集是算法的第二个目标。本文采用了一种基于距离的主动学习策略，其算法流程如下：

| **算法：**主动学习算法 |
| --- |
| **输入：**D={x1,x2,.....,xn}；Seed集S；约束数目N  **输出：**正约束集ML，负约束集CL，扩充后的Seed集NS |
| 1：依据类别标签将S划分为k个子集的集合，初始化ML与CL为空  2：依据初始化k个类别中心集合  2：**while** ML与CL的数目之和小于N  3： 从D中选择样本xi,使得xi距离各个类别中心的距离最远  4： **for** S中的任一子集Si  5： 从Si中选取样本点xj,组成约束对(xi，xj)  6： **if** (xi，xj)为正约束对  7： 将(xi，xj)加入ML中，令xj的类别与xi的类别相同，将xj加入 NS中  8： **else**  9： 将(xi，xj)加入CL中  10： **end if**  11： **end for**  12： **if** xi 与任意一个类别都能组成负约束关系  13： 将xi标记为噪声点  14： **end if**  15：**end while**  16：输出ML，CL，NS |

在该算法中，样本xi与各个类别中心的距离为。其中dis为基于L2范数的欧式距离。本文依照最远距离优先的策略选择样本点，这样可以获得蕴涵信息量较多的约束对以及分布均匀的Seed集。同时，若样本xi与各个类别均满足负约束关系，则可以将xi标记为噪声点。

1. 基于Seed集扩充成对约束

约束信息的数据量对聚类结果往往影响较大，先验数据中约束信息的数据量越大，聚类结果越好。Seed集中的样本带有标签信息，通过对这些标签信息进行分析可以扩充约束集的规模，进而提高聚类的效果。基于Seed集扩充成对约束的算法流程如图所示：

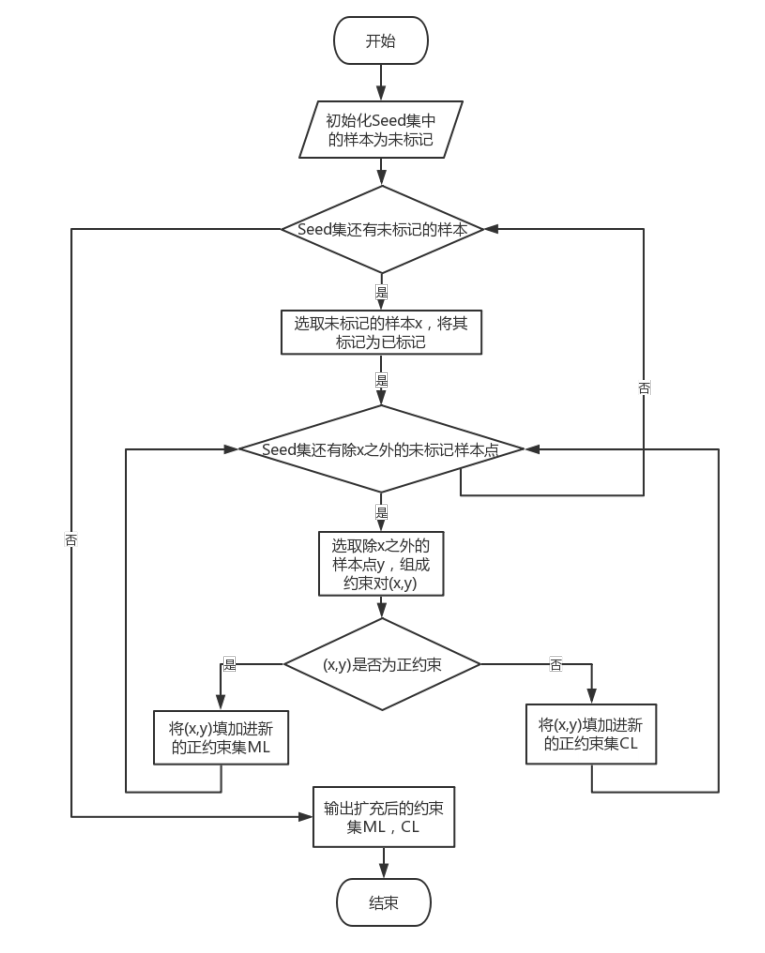


图 23 扩充约束对流程图

1. SC-Kmeans算法

SC-Kmeans算法的主要特点是能够同时应用Seed集与约束对集合这两种形式的先验信息来指导聚类。

## 区域特征分析

基于3.4节用户个人属性预测结果与3.3节的用户职住地数据，我们可以得到特殊人群的居住地位置。采用密度聚类算法可以将特殊人群的居住地聚类形成区域。公交数据与共享单车数据虽不具有手机信数据的样本代表性广泛、数据量丰富等优点，但其数据结构简单，且时空精度相较于信令数据更高，能够较为准确地反映群体的出行规律。本节的目标是基于公交刷卡数据与共享单车数据，分析区域的客流特征。

### 公交数据与共享单车数据出行链

公交数据指乘客乘坐公交车或地铁时使用一卡通刷卡所产生的数据，其中包含了乘客上下车时间、地点等时空信息。共享单车数据指用户使用单车时所产生的订单数据，其中包含了单次行程的起止时间与起止地点等信息。由于上述两类数据结构相似，本文采用相同的定义对其进行描述，相关定义如下：

* + 1. **OD（Origin-Destination）出行**

OD出行表示用户在乘坐公交或共享单车时所产生的一次行程，可形式化定义为：

|  |  |
| --- | --- |
|  | （3.23） |

其中sTime表示行程的开始时间，eTime表示行程的终止时间，sLoc表示行程的开始地点，eLoc表示行程的结束地点。

* + 1. **OD出行链**

OD出行链由用户一天内的多个行程组成，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.24） |

### 区域特征定义

1. 基于信令数据的区域特征

使用DBSCAN算法对特殊人群的居住地进行聚类操作后，可以得到表示区域的簇集dC，dC由一簇居住地构成。区域的定义可形式化表示为：

|  |  |
| --- | --- |
|  | （3.25） |

其中id表示区域的唯一标识；loc表示区域的坐标，具体由dC的中心点构成，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.26） |

radis表示区域的半径，为计算方便，具体为居住地簇集中的中心点到各个点的最远距离，可形式化表示为：

|  |  |
| --- | --- |
|  | （3.27） |

count表示区域内特殊人群的数量，具体为居住地簇集中点的数量；dense表示区域的人口密度，单位为人/m2，具体为区域内人群的数量除以区域的面积，为计算方便，本文将区域设定为正方形区域，区域的人口密度可形式化表示为：

|  |  |
| --- | --- |
|  | （3.28） |

1. 基于公交数据的区域特征
2. 区域内居住人数

对于某个区域，本文定义在该区域内居住的用户为在早高峰时间段内离开该区域，晚高峰时间段内回到该区域的用户。为确定早晚高峰时间段的范围，本文绘制了工作日期间北京市用户出行时段分布图，具体如图24所示：

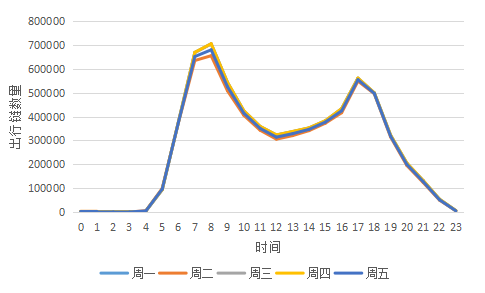


图 24 用户出行时段分布图

可以看到在6点到10点间，居民的出行达到了第一个高峰；在16点到19点间，居民的出行达到了第二个高峰。本文定义6点至10点为早高峰时间段，16点至19点为晚高峰时间段。

1. 区域内工作人数

与区域内居住人数的定义类似。对于某个区域，本文定义在该区域内工作的用户为在早高峰时间段内到达该区域，晚高峰时间段内离开该区域的用户。

1. 基于共享单车数据的区域特征
2. 区域内部骑行数量

对于某个区域，本文定义在该区域内的一次骑行行程为骑行出发点、到达点均在该区域的行程，统计一天内该类别行程的总数作为区域内部的骑行数量。

### 区域特征分析

完成了基于信令数据的区域基本特征与基于公交数据与共享单车数据区域客流特征的提取后，本节的目标是针对上述所提取的特征，采用Kmeans算法对不同的人群聚居区域进行聚类分析。聚类所需的特征如表11所示：

表 10 聚类特征说明表

|  |  |  |
| --- | --- | --- |
| 类别 | 编号 | 名称 |
| 区域基本特征 | 1 | 区域半径 |
| 2 | 区域人数 |
| 3 | 区域人口密度 |
| 区域客流特征 | 4 | 区域内居住人数 |
| 5 | 区域内工作人数 |
| 6 | 区域内骑行数量 |

由于各个特征的量纲不同，因此在进行聚类分析前需对特征进行标准化处理，标准化计算公式如下：

|  |  |
| --- | --- |
|  | （3.29） |

其中为的平均值，为的标准差。

## 本章小结

本章提出了基于多源位置数据的人员特殊聚居区域识别模型。首先对手机信令数据进行了质量分析，在了解了数据质量后，剔除了长距离基站抖动数据，并完成了优质用户的筛选。然后基于预处理后的信令数据，引入ST-DBSCAN算法识别了用户的停留态与移动态，完成了用户出行链的提取，并基于用户多天的停留点数据，提取了用户的职住地信息。为判别用户的个人属性，本文基于用户的出行链数据与职住地数据，提取相关的出行特征，使用随机森林算法完成了个体属性预测模型的构建。最后基于用户个人属性的预测结果，将特殊人群的居住地聚类形成区域，并结合公交数据与共享单车数据提取区域客流特征，完成了聚居区域的特征分析。

# **人员特殊聚居区域分析系统的设计与实现**

根据第三章描述的人员特殊聚居识别模型，本文设计并实现了一个基于多源位置数据的人员特殊聚居区域分析展示系统。

本章首先介绍了基于多源位置数据的人员特殊聚居区域分析与展示系统的需求及架构设计，然后依次从数据接口设计和关键模块设计等几个方面对系统的设计与实现进行了详细的介绍与说明。

## 系统需求分析

人员特殊聚居区域分析系统的设计目标为：基于手机信令数据提取用户的出行链与职住地信息；结合POI数据与出行链数据，计算用户的出行特征，基于人工标注的训练集完成个人属性预测模型的构建；结合公交数据与共享单车数据，分析人员特殊聚居区域的特征，并将分析结果进行可视化展示。

人员特殊聚居区域分析系统的需求如下：

1. 系统需要完成字段缺失、错误的信令数据的剔除，信令数据位置信息的补全，长距离基站抖动数据的剔除以及优质用户筛选，公交数据与共享单车数据字段约简等预处理相关的功能。
2. 系统需根据预处理后的信令数据，识别用户的停留态与移动态，完成用户出行链与职住地信息的提取。
3. 系统需结合POI数据与用户出行链、职住地数据，计算用户的出行特征。根据人工标注的训练集与用户的出行特征数据，完成用户个人属性预测模型的构建，并基于未标注的用户数据，输出个人属性的预测结果。
4. 系统需根据个人属性的预测结果，提取特殊人群的聚居区域，并结合公交数据与共享单车数据计算区域的客流特征，使用聚类算法完成区域的特征分析。
5. 系统需实现可视化展示功能，基于互联网地图完成对人群出行链与职住地信息，特殊人群的出行轨迹以及人员特殊聚居区域信息的结果展示。

## 系统架构设计

根据上述的系统需求，本文设计的人员特殊聚居区域分析系统，以手机信令数据、基站数据、POI数据、公交及共享单车数据作为输入，识别人员特殊聚居区域并分析区域特征，利用可视化工具进行展示。系统分为四层架构：原始数据层、业务处理层、业务数据层和结果展示层。系统架构图如图25所示：

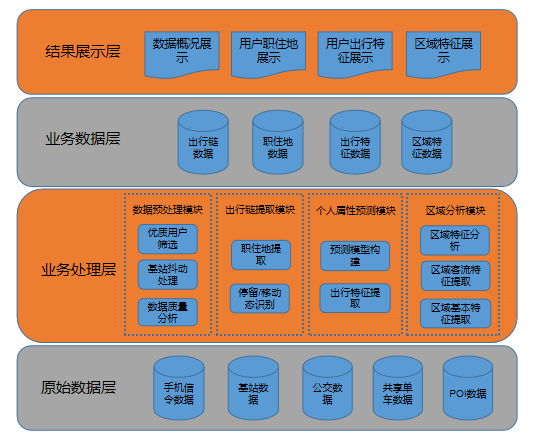


图 25 人员特殊聚居区域分析系统架构图

1. 原始数据层

本系统需要输入的数据包括：手机信令数据、基站数据、公交数据、共享单车数据以及POI数据。原始数据层为上层架构提供基本的数据源服务，包括数据的存储和管理，数据的操作等功能。

1. 业务处理层

业务处理层是人员特殊聚居区域分析系统的核心部分，主要包含了数据预处理、出行链与职住地提取、个人属性预测以及区域分析这几个模块。数据预处理模块完成了数据质量分析、基站抖动处理以及优质用户筛选等几项功能；出行链提取模块基于预处理模块生成的信令数据，完成了用户停留态/移动态的识别，提取了用户的出行链与职住地信息；个人属性预测模块结合POI数据与出行链数据，提取用户出行特征，完成了用户个人属性预测模型的构建，并输出个人属性预测结果；区域分析模块基于个体属性预测结果，提取人员特殊聚居区域的基本特征，并结合公交数据与共享单车数据提取区域的客流特征，采用聚类算法完成了区域特征的分析。

1. 业务数据层

本系统生成的中间数据包括出行链数据、职住地数据以及用户出行特征数据。结果数据为区域特征数据。业务数据层主要为业务处理层以及结果展示层提供了数据传输服务，包括中间数据在处理层各个模块间的传输等功能。

1. 结果展示层

结果展示层负责进行结果数据的图形化展示，将业务数据层中的计算结果以直观的形式将结果展示给用户。

## 数据接口设计

数据接口包括基础数据接口与中间数据接口，其中基础数据包含信令数据、公交数据、单车数据以及POI等数据，中间数据包含出行链与职住地数据、出行特征数据以及区域特征数据。基础数据的ER图如图26所示。

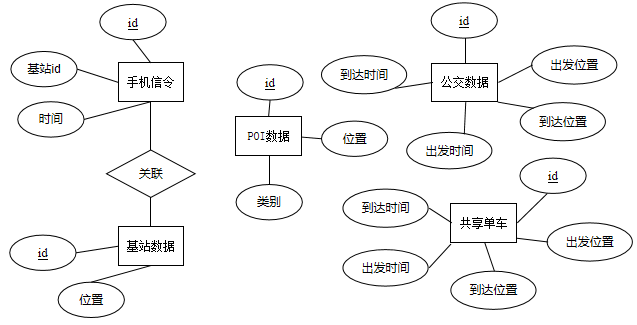


图 26 基础数据ER图

中间数据的ER图如图27所示。

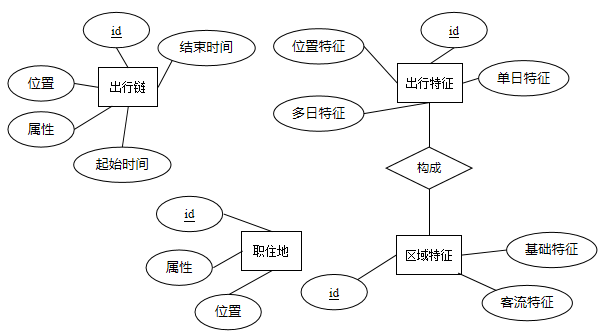


图 27 中间数据ER图

### 手机信令数据接口

1. 接口说明

手机信令数据指用户使用手机与基站通信时，运营商所记录的日志数据，其中包含了手机与基站通信发生的时间，基站的id等信息。由于原始的信令数据中只包含了基站的id信息，因此需结合基站位置数据完成信令数据位置信息的补充。

1. 接口定义

信令数据接口如表12所示，主要包含的字段有：imsi码（由用户手机号加密生成）、通信发生的时间、信令类型、基站的lac码、基站的ci码以及用户手机号前七位。

表 11 手机信令数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| imsi | 由用户手机号加密生成，可唯一标识一个用户 | String | 113F5F219A08ECDA0445D4577 |
| time | 表示用户使用手机与基站发生通信的时间 | Datetime | 20190614072000 |
| type | 信令类型，具体分为2G、3G以及4G信令 | String | 4 |
| lac | 基站的位置区编号 | Integer | 41089 |
| ci | 基站的小区编号 | Integer | 10212 |
| number | 用户手机号前7位 | String | 1851802 |

基站位置数据接口如表13所示，主要包含的字段有基站的lac码、ci码、基站的类型以及基站的经纬度。

表 12 基站位置数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| lac | 基站的位置区编号 | Integer | 41089 |
| ci | 基站的小区编号 | Integer | 10212 |
| lng | 基站位置的经度 | Double | 116.3452 |
| lat | 基站位置的纬度 | Double | 39.9762 |
| type | 基站类型，具体分为2G、3G以及4G基站 | String | 4 |

### 公交数据与共享单车数据接口

1. 接口说明

公交数据指用户乘坐公交或地铁时使用一卡通刷卡所产生的数据。共享单车数据指用户使用单车完成一次行程时，单车服务商所记录的相关行程信息。两类数据所包含的信息类似，均记录了上下车的时间信息与地点信息。

1. 接口定义

公交数据接口定义如表14所示，主要包含的字段有：一卡通号、载具类型、上车时间、上车地点经纬度、下车时间以及下车地点经纬度。

表 13 公交数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| id | 一卡通号，可唯一标识一个用户 | String | 365912222155 |
| type | 载具类型，GJ表示公交，DT表示地铁 | String | GJ |
| onTime | 上车时间 | Datetime | 20190603075001 |
| onLng | 上车位置经度 | Double | 116.33786 |
| onLat | 上车位置纬度 | Double | 40.0853 |
| offLng | 下车位置经度 | Double | 116.276531 |
| offLat | 下车位置纬度 | Double | 39.9325 |
| offTime | 下车时间 | Datetime | 20190603090431 |

共享单车数据接口定义如表15所示，主要包含的字段有：用户编号、租车位置经纬度、还车位置经纬度、租车时间以及还车时间。

表 14 共享单车数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| id | 用户编号 | String | 5510948890 |
| rentTime | 租车时间 | Datetime | 20190603175001 |
| backTime | 还车时间 | Datetime | 20190603175401 |
| rentLng | 租车位置经度 | Double | 116.3685 |
| rentLat | 租车位置纬度 | Double | 40.0023 |
| backLng | 还车位置经度 | Double | 116.3460 |
| backLat | 还车位置纬度 | Double | 39.9996 |

### POI数据接口

1. 接口说明

POI(Point of Interest)数据表示城市中某个地点的信息。一个POI可以是一栋房子、一个商铺、一个邮筒、一个公交站等。本文将POI数据用于用户出行特征分析。

1. 接口定义

POI数据接口定义如表16所示，主要包含的字段有：名称，类别，经度与纬度。

表 15 POI数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| type | POI类型 | String | 071100 |
| name | POI名称 | String | 新主楼 |
| lng | POI位置经度 | Double | 116.33832 |
| lat | POI位置纬度 | Double | 39.9771 |

### 用户出行链与职住地数据接口

1. 接口说明

出行链提取模块基于预处理模块生成的信令数据，识别用户的停留态/移动态，进而输出行链数据与职住地数据。出行链数据由停留点和移动点组成，职住地数据由用户多天的停留点生成。

1. 接口定义

出行链数据接口定义如表17所示，主要包含的字段有：用户id、停留开始时间、停留结束时间、停留结束时间、停留位置经度、停留位置纬度以及点类别。

表 16 出行链数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| uid | 用户id | String | CDA0445D457741DAC56 |
| sTime | 停留开始时间 | Datetime | 20190603175001 |
| eTime | 停留结束时间（若为移动点，则该字段为空） | Datetime | 20190603175401 |
| lng | 停留位置经度 | Double | 116.33832 |
| lat | 停留位置纬度 | Double | 39.9771 |
| type | 点类别（S表示停留点，M表示移动点） | Char | S |

职住地数据接口定义如表18所示，主要包含的字段有：用户id、位置点经度、位置点纬度以及点类别。

表 17 职住地数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| uid | 用户id | String | CDA0445D4577 |
| lng | 位置点经度 | Double | 116.33832 |
| lat | 位置点纬度 | Double | 39.9771 |
| type | 点类别（H表示居住地，W表示工作地） | Char | H |

### 用户出行特征数据接口

1. 接口说明

用户出行特征数据是个人属性预测模块结合POI数据与出行链数据以及职住地数据生成的，用于支撑个人属性预测模型的构建。

1. 接口定义

用户出行特征接口定义如表19所示，主要包含的字段有：用户id、移动范围均值与标准差、离家\回家时间均值与标准差、轨迹相似性均值与标准差、移动频率均值与标准差、移动时间均值与标准差、出行时间均值与标准差以及相关POI特征。

表 18 用户出行特征接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| uid | 用户id | String | CDA0445D457741DAC56 |
| rangeM | 移动范围均值（米） | Double | 581.9 |
| rangeStd | 移动范围标准差 | Double | 39.1 |
| outTimeM | 离家时间均值（小时） | Double | 7.6 |
| outTimeStd | 离家时间标准差 | Double | 1.6 |
| similarM | 轨迹相似性均值（取值范围为[0,1]） | Double | 0.78 |
| similarStd | 轨迹相似性标准差 | Double | 0.12 |
| inTimeM | 回家时间均值（小时） | Double | 20.2 |
| inTimeStd | 回家时间标准差 | Double | 3.1 |
| freM | 移动频率均值 | Double | 2.1 |
| frStd | 移动频率标准差 | Double | 0.3 |
| mTimeM | 移动时间均值（分钟） | Double | 30.4 |
| mTimeStd | 移动时间标准差 | Double | 3.4 |
| hRateSM | 学校区域内停留时长与居住时长的比率均值 | Double | 0.76 |
| hRateSStd | 学校区域内停留时长与居住时长的比率标准差 | Double | 0.03 |
| wRateSM | 学校区域内停留时长与工作时长的比率均值 | Double | 0.84 |
| wRateSStd | 学校区域内停留时长与工作时长的比率标准差 | Double | 0.13 |
| hRateFM | 工厂区域内停留时长与居住时长的比率均值 | Double | 0.76 |
| hRateFStd | 工厂区域内停留时长与居住时长的比率标准差 | Double | 0.03 |
| wRateFM | 工厂区域内停留时长与工作时长的比率均值 | Double | 0.84 |
| wRateFStd | 工厂区域内停留时长与工作时长的比率标准差 | Double | 0.13 |

### 区域特征数据接口

1. 接口说明

区域特征数据是区域分析模块结合个体属性预测结果与公交数据以及共享单车数据生成的，用户支撑区域的分析。

1. 接口定义

区域特征数据接口定义如表20所示，主要包含的字段有：区域id、区域类型、区域位置、区域内聚居人群数量、区域半径、区域人口密度、区域内居住\工作人数、区域内骑行次数。

表 19 区域特征数据接口

| 字段名称 | 字段含义 | 字段类型 | 数据示例 |
| --- | --- | --- | --- |
| did | 区域id | String | 3 |
| type | 区域类型（S为学校，F为工厂） | Char | S |
| count | 区域内聚居人群数量 | Integer | 357 |
| lng | 区域中心位置经度 | Double | 116.33832 |
| lat | 区域中心位置纬度 | Double | 39.9771 |
| radis | 区域半径（米） | Double | 231.4 |
| dense | 区域人口密度（人/m2） | Double | 0.0034 |
| hCount | 区域内居住人数 | Integer | 476 |
| wCount | 区域内工作人数 | Integer | 213 |
| inCount | 区域内骑行次数 | Integer | 21 |

## 关键模块设计与实现

### 数据预处理模块

数据预处理模块主要完成了手机信令数据的质量分析、错误字段数据剔除、基站抖动数据删除与优质用户筛选等功能。其模块结构图如图28所示：

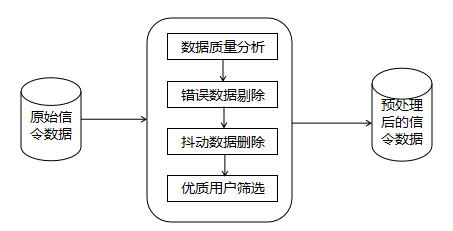


图 28 预处理模块结构图

该部分主要包含cellData类、userData类、dataProcess类以及dataAnaly类。数据预处理模块的类图如图29所示：



图 29 预处理模块类图

其中cellData类表示原始信令数据；userData类表示一个用户产生的信令数据，其由cellData类组合而成；dataAnaly类主要完成信令数据质量分析等相关功能。由于信令数据量巨大，每天的数据量可以达到110GB，因此将信令数据以文件的形式存储在HDFS文件系统中,dataAnaly类与dataProcess中方法的参数均为HDFS文件路径。dataAnaly类的主要成员如表21所示：

表 20 dataAnaly成员

| 成员变量/方法名称 | 成员变量/方法说明 |
| --- | --- |
| getTimeInterval(path:String) | 统计信令数据产生的时间间隔 |
| getDataCount(path:String) | 统计用户每天产生的信令数量 |
| getNativeCount(path:String) | 统计北京市用户数量 |
| getForeignCount(path:String) | 统计外地用户数量 |

dataProcess类主要完成数据预处理功能，主要包括超出北京市范围以及字段缺失、错误信令数据的剔除，基站抖动信令数据的删除以及优质用户的筛选。dataProcess的主要成员如表22所示：

表 21 dataProcess成员

| 成员变量/方法名称 | 成员变量/方法说明 |
| --- | --- |
| cityRlng,cityRlat | 北京市右上角坐标 |
| cityLlng,cityLlat | 北京市左下角坐标 |
| getLegalData(path:String) | 错误信令数据剔除 |
| deleteShake(path:String) | 删除基站抖动数据 |
| getGoodUser(path:String) | 优质用户筛选 |

### 出行链提取模块

出行链提取模块主要基于经过预处理后的信令数据，识别用户的停留态与移动态，提取用户出行链，并基于用户多天的停留点数据生成用户职住地信息。出行链提取模块的结构图如图30所示：

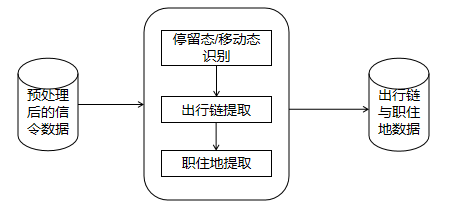


图 30 出行链提取模块结构图

该部分主要包含userData类、stopPoint类、hwPoint类、userChain类以及chainProcess类。数据预处理模块的类图如图31所示：

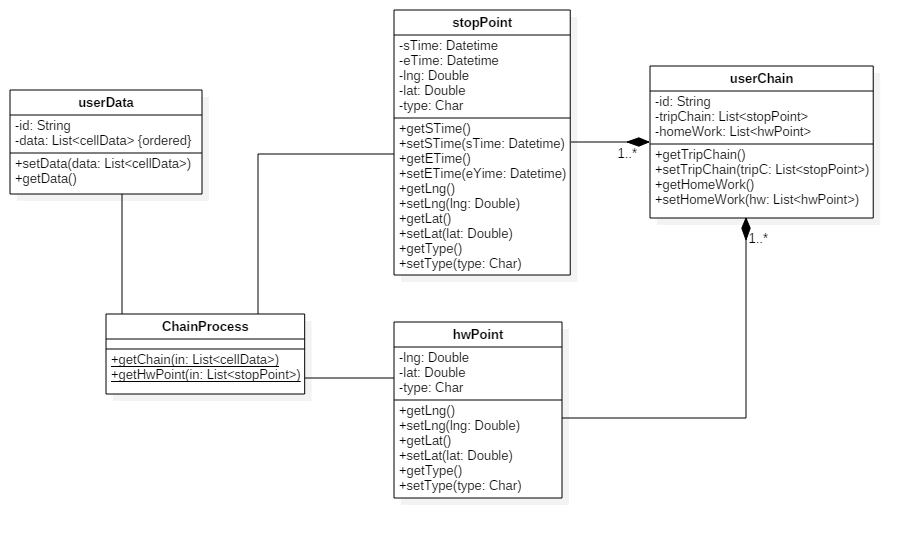


图 31 出行链提取模块类图

其中ChainProcess类中的getChain方法主要实现了ST-DBSCAN算法，用于识别用户的停留态与移动态；getHwPoint方法主要实现了基于多日停留点数据的用户职住地提取算法。userChain类表示该模块输出的中间结果数据，其主要由stopPoint类与hwPoint类组合而成。stopPoint类表示用户的停留点，其主要成员如表23所示：

表 22 stopPoint类成员

| 成员变量/方法名称 | 成员变量/方法说明 |
| --- | --- |
| Datetime sTime | 停留开始时间 |
| Datetime eTime | 停留结束时间 |
| Double lng | 停留位置经度 |
| Double lat | 停留位置纬度 |
| Char type | 点类型，S表示停留点，M表示移动点 |

hwPoint类表示用户的职住地，其主要成员如表24所示：

表 23 hwPoint类成员

| 成员变量/方法名称 | 成员变量/方法说明 |
| --- | --- |
| Double lng | 位置点经度 |
| Double lat | 位置点纬度 |
| Char type | 点类型，H表示居住地，M表示工作地 |

### 个人属性预测模块

个人属性预测模块基于用户出行链数据、职住地数据以及POI数据计算用户的出行特征，结合随机森林算法完成个人属性预测模型的构建，并对没有属性标注的人群进行预测，输出个人属性预测结果。个人属性预测模块图如图32所示：

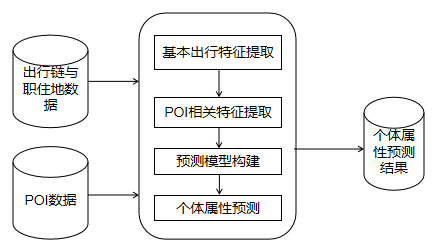


图 32 个人属性预测模块结构图

该部分主要包含userChain类、poiData类、Feature类以及predictModel类。个人属性预测模块的类图如图33所示：

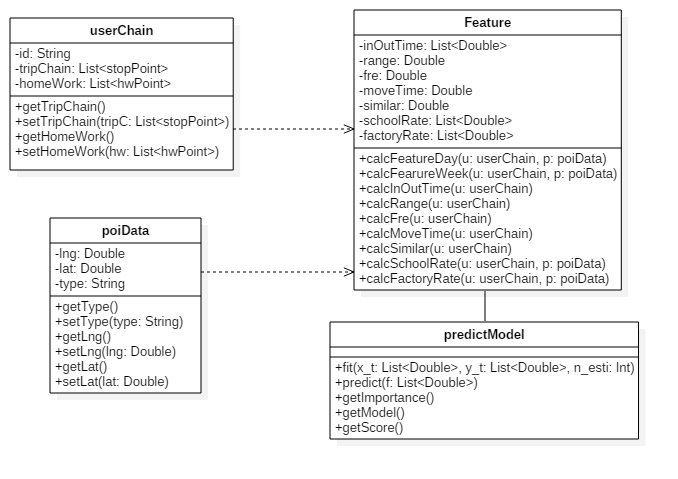


图 33 个人属性预测模块类图

其中Feature类基于用户出行链数据、职住地数据以及POI数据，完成了用户的出行特征计算。predictModel类基于随机森林算法，完成了个人属性预测模型的构建以及没有属性标注人群的预测。其中Feature类的主要成员如表25所示：

表 24 hwPoint类成员

| 成员变量/方法名称 | 成员变量/方法说明 |
| --- | --- |
| List<Double> inOutTime | 用户离家、回家时间 |
| Double range | 用户移动距离 |
| Double fre | 用户移动频率 |
| Double moveTime | 用户移动时间 |
| Double similar | 用户移动路径相似性 |
| List<Double> schoolRate | 用户停留在学校区域时长比率 |
| List<Double> factoryRate | 用户停留在工厂区域时长比率 |
| calcFeatureDay(u : userChain, p : poiData) | 用户单日出行特征计算 |
| calcFeatureWeek(u : userChain, p : poiData) | 用户多日出行特征计算 |

### 区域分析模块

基于个人属性模块输出的预测结果，区域分析模块首先将目标人群的居住地聚类形成区域，并计算区域的基本特征，如区域半径、密度、位置等信息；然后结合公交数据与共享单车数据计算区域的客流特征，如区域内工作人数、区域内骑行数量等客流特征；最后对区域特征进行聚类分析。区域分析模块的结构图如图34所示：

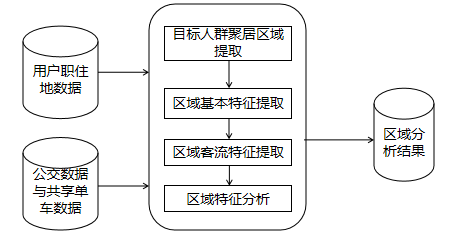


图 34 区域分析模块结构图

该部分主要包含userChain类、busData类、bikeData类、clusterUtil类以及类。区域分析模块的类图如图35所示：

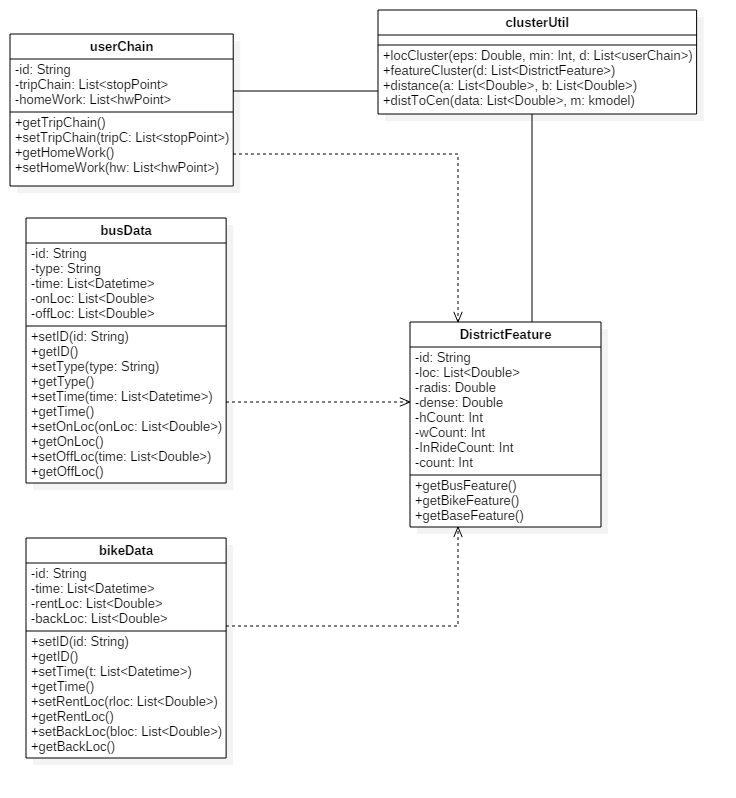


图 35 区域分析模块类图

其中clusterUtil类中locCluster方法使用了DBSCAN算法，基于userChain类中的职住地属性将目标人群的居住地聚类形成区域；DistrictFeature类基于clusterUtil类的聚类结果计算区域的基本特征，并结合公交数据与共享单车数据计算区域的客流特征。clusterUtil类中featureCluster方法使用Kmeans算法对区域特征进行聚类分析。clusterUtil类的主要成员如表26所示：

表 25 clusterUtil类成员

| 成员变量/方法名称 | 成员变量/方法说明 |
| --- | --- |
| locCluster(eps:Double,min:Int,d:List<userChain>) | 将目标人群的居住地聚类形成区域 |
| featureCluster(d:List<DistrictFeature>) | 区域特征聚类分析 |
| distance(a:List<Double>,b:List<Double>) | 计算特征向量间的距离 |
| distToCen(data:List<Double>,m:kmodel) | 计算肘系数 |

### 可视化模块

基于高德地图提供的API与地图数据，可视化模块结合Web页面对人员特殊聚居区域分析系统生成的中间数据与结果数据进行了展示，展示页面如图36所示：

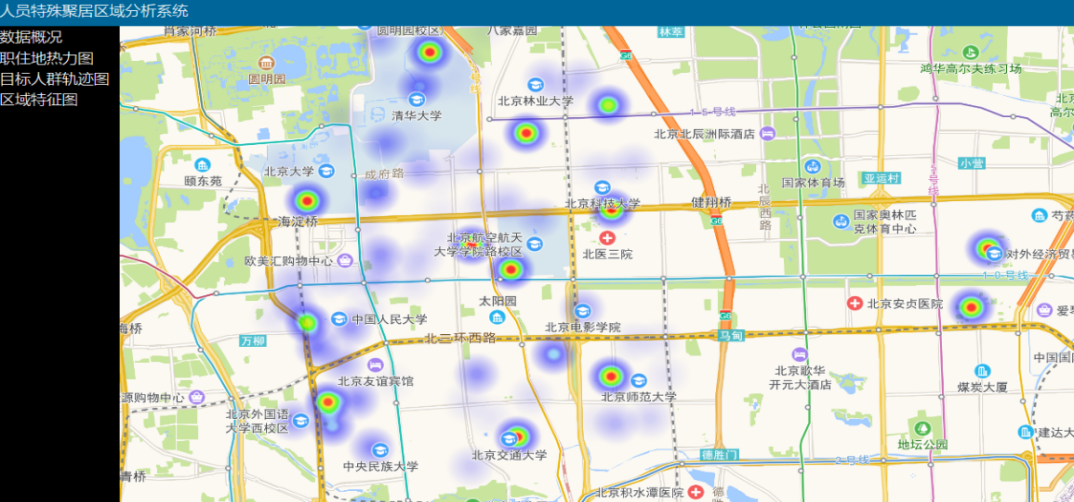


图 36 可视化页面展示

## 本章小结

本章以第三章提出的基于多源位置数据的人员特殊聚居区域识别模型，设计并实现了人员特殊聚居区域分析系统。本章首先进行了系统需求分析，明确了系统需实现哪些功能；然后对系统的架构进行了设计，确定了系统的关键模块与各层架构间的依赖关系；接着对系统的数据接口与关键模块进行了较为细致地介绍；最后设计了可视化页面，对系统计算得出的中间数据与结果数据进行了展示。

# **系统测试与分析**

在前述章节中，本文完成了基于多源位置数据的人员特殊聚居区域识别模型性的研究，并设计了人员特殊聚居区域分析及展示系统。本章将对本文提出的出行链与职住地提取、个体属性预测以及区域特征分析功能的准确性进行验证。

## 实验环境

### 测试系统环境

测试实验在两台台式机组成的集群上运行，其中每个集群节点的配置信息如表27所示：

表 26 集群节点配置信息

|  |  |
| --- | --- |
| **配件** | **配置** |
| CPU | Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz |
| 内存 | 250GB |
| 硬盘 | 30TB |
| 操作系统 | Centos 7.3 |

节点中各个服务组件的信息如表28所示：

表 27 节点组件信息

|  |  |  |  |
| --- | --- | --- | --- |
| **组件** | **版本** | **节点数** | **说明** |
| Hadoop | 2.7.5 | 2 | 分布式并行处理框架 |
| Zookeeper | 3.4.12 | 2 | 分布式协调系统 |
| Spark | 2.3.1 | 2 | 分布式内存计算框架 |

展示系统的运行环境为个人PC机，其配置信息如表29所示：

表 28 个人PC机配置

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CPU | 内存 | 硬盘 | 操作系统 | WEB浏览器 |
| 英特尔酷睿 i7-3770 @3.6GHz | 16G | 1TB | Windows 10 | Chrome 78.0.3904.97 |

### 测试数据环境

实验采用了北京联通公司提供的2019年1月1日至2019年7月31日北京市范围内的4G信令数据，具体数据概况表1所示。实验采用的公交数据为2019年6月10日2019年6月16日的公交刷卡数据，具体数据概况如表30所示：

表 29 公交数据概况

|  |  |
| --- | --- |
| **属性** | **数值** |
| 涉及公交站点数量 | 44524 |
| 涉及轨道站点数量 | 320 |
| 用户数量 | 4827180.5人/天 |
| 出行链数量 | 9171324.3条/天 |

实验采用的共享单车数据为2019年6月10日至2019年6月16日的北京市范围内的单车订单数据，数据概况如表31所示：

表 30 共享单车数据概况

|  |  |
| --- | --- |
| **属性** | **数值** |
| 涉及单车公司数量 | 4 |
| 单车数量 | 457802.3辆/天 |
| 订单数量 | 2186422.7单/天 |

## 用户出行链与职住地提取结果测试与分析

用户出行链是由若干个停留点以及移动点组成，本文引入了ST-DBSCAN算法判别用户的移动态与停留态，并基于用户多天的停留点提取其职住地信息。本节将对停留态识别与职住地提取两方面内容展开分析与验证。

### 停留点识别

图37展示了某个用户的停留点提取实例。图中展示了用户原始的信令记录，算法识别出的移动点记录以及停留点记录。由图可以看出算法准确地识别出了空间相近但时间上不连续的停留点（停留点0:00-8:10与停留点20:50-0:00）。

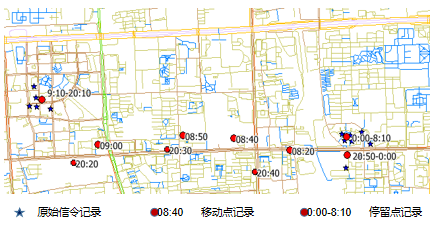


图 37 用户停留点提取结果实例图

本文统计了用户单日的停留点数量分布来从整体层面上分析停留点识别结果的准确性。停留点数量在一定程度上可以反映用户的类型：居家人员或无工作人员一天一般产生1-2个停留点；通勤人员一天会产生3-4个停留点；从事物流运输或外卖相关职业的人员一天可能会产生5个以上的停留点。图38展示了具体的停留点数量分布：

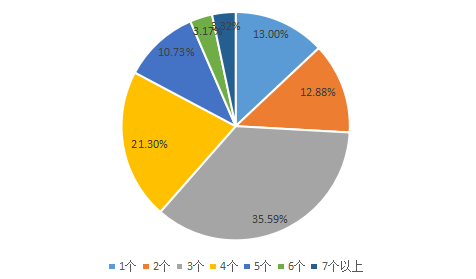


图 38 用户停留点数量分布图

可以看到拥有3-4个停留点的用户占比到达了约57%，根据交通发展年报[33]统计结果显示，居民出行中通勤活动占比约为51%。本文有关停留点的计算结果较为符合真实情况。

### 职住地识别

本文首先基于10个工作日的停留点数据，提取了用户的职住地信息。然后将北京划分为50×50个网格，每个网格的面积为1km2。最后将用户职住地信息映射到网格层级，绘制了基于网格的职住热力图。为验证职住地识别的准确性，本文将绘制的职住热力图与交通发展年报中的职住热力图进行对比分析。图39与图40展示了职住地热力图的对比：

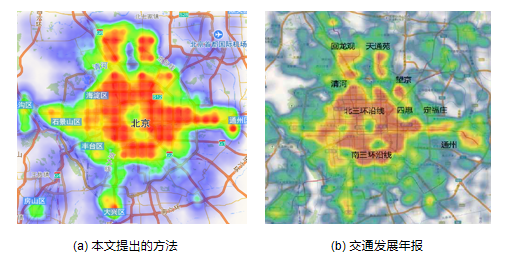


图 39 居住地热力图

由居住地热力图可以看出，北京市的居住地多集中于城市中心区，主要分布在南北三四环沿线附近。对于典型的密集居住区，比如天通苑、回龙观、望京以及通州地区，本文提出的方法都能够有效地识别。

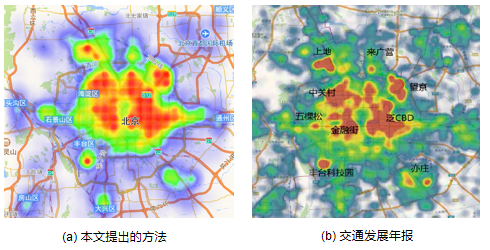


图 40 工作地热力图

由工作地热力图可以看出，北京市的工作地主要集中于CBD、望京、中关村以及上地一带。本文提出的方法能够较为准确地识别上述区域。

## 个人属性预测结果测试与分析

基于3.4.2节对于训练集标注的介绍，本文选取部分院校区域与工厂区域作为标注区域，构建训练集，并基于随机森林算法构建个人属性预测模型。其中训练集数据的概况如表32所示：

表 31 训练集概况

|  |  |  |
| --- | --- | --- |
| **名称** | **定义** | **数量** |
| 学生 | 在高等院校正在接受教育，且职住地都位于院校内的人群 | 4461 |
| 工人 | 从事劳动生产相关行业，且职住地都位于工厂内的人群 | 914 |
| 其他 |  | 10296 |

在预测性能评价方面，本文采用准确率与召回率两个指标进行评价。指标的计算公式如下所示：

|  |  |
| --- | --- |
|  | （5.1） |
|  | （5.2） |

其中P表示准确率，R表示召回率。对于二元分类问题，TP表示预测为正例，实际也为正例的样本数量；FP表示预测为正例，实际为负例的样本数量，FN表示预测为负例，实际为正例的样本数量。对于多元分类问题，可以将每个类别单独视为正，所有其他类型视为负，计算每个类别相对其他类别的准确率与召回率。

为验证随机森林算法的性能，本文选取SVM(Support Vector Machine, 支持向量机)算法进行对比分析，对比结果如表33所示，可以看到两种机器学习算法的预测精度相近，随机森林算法的预测精度相对更高。

表 32 个人属性预测结果对比分析

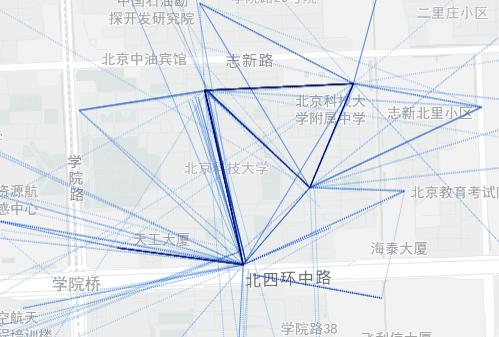
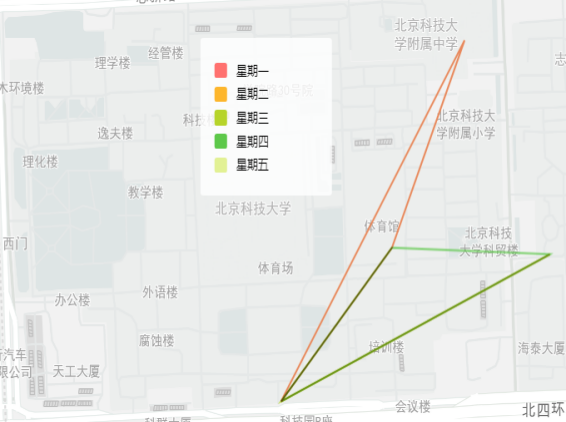
|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | 名称 | 准确率 | 召回率 |
| 随机森林 | 学生 | 77.2% | 79.9% |
| 工人 | 75.6% | 78.7% |
| 其他 | 76.3% | 83.6% |
| SVM | 学生 | 71.3% | 70.1% |
| 工人 | 72.7% | 77.7% |
| 其他 | 74.7% | 79.2% |

## 区域特征分析

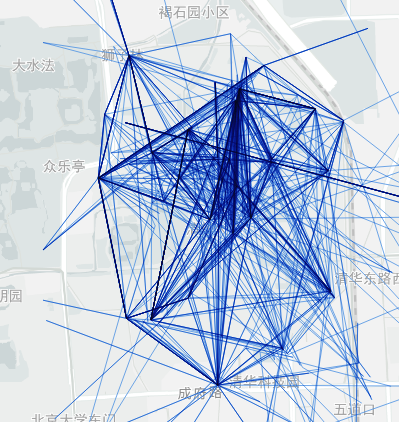
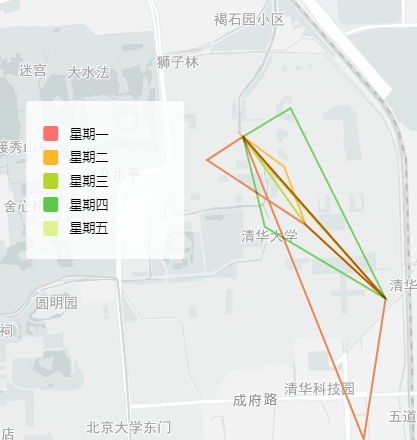
在完成个人属性预测后，本文将学生、工人这两类特殊人群的居住地聚类形成区域，并结合公交数据与共享单车数据分析区域特征。本节将对区域位置与特征进分析验证。

### 学校区域识别验证

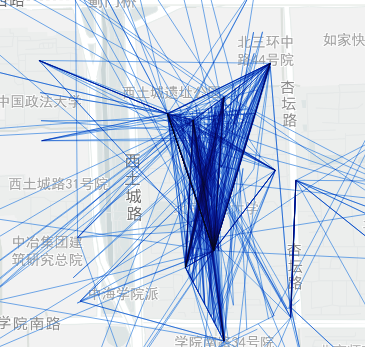
将学生群体的居住地聚类形成区域后，可以得到学校区域的位置信息。本节基于北京市范围内一周的信令数据，共识别出学生74211人，聚类形成区域84个。图41展示了学校区域识别的一个样例：

**(a) 样例1群体轨迹** (b) **样例1个体轨迹**

**(c) 样例2群体轨迹** **（d）样例2个体轨迹**

**(e) 样例3群体轨迹** **（f）样例3个体轨迹**

图 41 学校区域识别样例

其中样例1为北京科技大学附近区域，样例2为清华大学附近区域，样例3为北京邮电大学附近区域，红色方框表示本文识别出的区域。由图可以看出，样例群体的活动范围基本位于院校范围内，部分个体的活动较为较为规律，往返于院校内的教学区与工作区之间。算法较为准确地识别出了学生群体的聚居区域。学生群体的居住地热力图如图42所示：

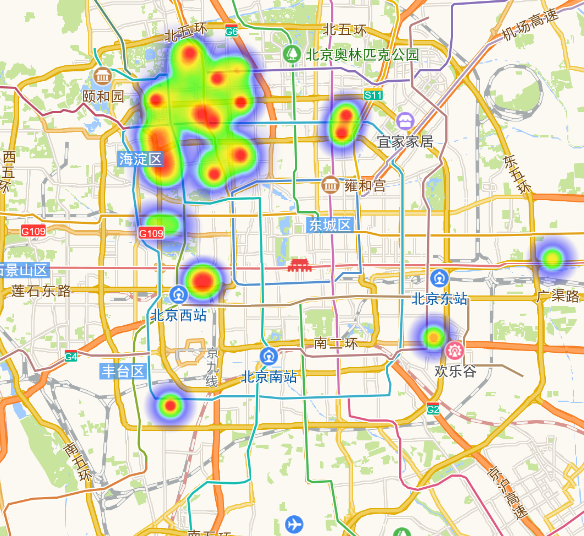


图 42 学生群体居住热力图

由于高等院校的地图信息较为丰富，因此本文结合POI信息与人工观察的方法，对北京市范围内93所高等院校进行分析验证，结果如表34所示：

表 33 学校区域识别结果

|  |  |  |
| --- | --- | --- |
| 名称 | 准确率 | 召回率 |
| 学校区域 | 80.95% | 73.11% |

可以看到学校区域识别的准确率较高，但召回率较低。部分院校如北京城市学院，其学生以走读生为主，用于构建预测模型的训练集中没有包含这类学生群体；中华女子学院，其院校范围较小，且周边基站分布较为稀疏导致定位误差较大。算法没有准确地识别出上述院校。后续应在定位精度与样本代表性方面展开更加深入的研究。

### 工厂区域识别验证

本节以大兴区为例对工厂区域的识别结果进行分析。基于大兴区范围内一周的信令数据，共识别出工人6801人，聚类形成114个区域。大兴区工人居住热力图如图43所示。

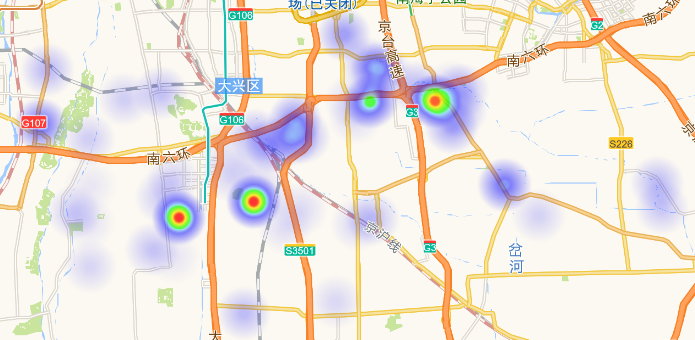


图 43 大兴区工人居住热力图

对区域半径、区域人口密度、区域内居住人数、区域内工作人数以及区域内骑行数量进行聚类分析，其中聚类中心如表35所示：

表 34 大兴区工厂区域聚类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 类别 | 占比 | 半径（米） | 人口密度 | 居住人数 | 工作人数 | 骑行数量 |
| 1 | 34% | 98.55 | 0.011 | 13 | 17 | 5 |
| 2 | 27% | 113.56 | 0.023 | 107 | 69 | 7 |
| 3 | 17% | 421.67 | 0.007 | 771 | 95 | 124 |
| 4 | 15% | 269.56 | 0.014 | 144 | 373 | 11 |
| 5 | 7% | 913.7 | 0.017 | 532 | 256 | 154 |

其中第一、二类区域半径较小，区域内客流量也较低，该类区域主要由地处偏僻的独立工厂组成；第三、四类区域内客流量较高，区域半径适中，该类区域主要由地处交通较发达地区的公司、工厂组成；第五类区域半径较高，人口密度也较高，其主要由大型工业园区中的企业组成。图44展示了各类型区域的位置分布图。

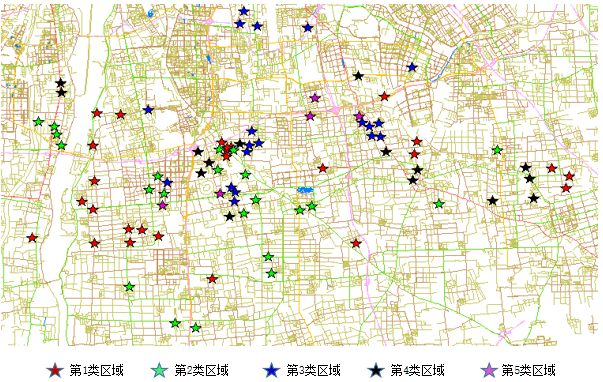


图 44 各类型区域位置分布

本文结合北京市应急管理局提供的企业台账信息对工厂区域的识别结果进行验证分析。企业台账包含了企业的行业属性、员工数量、位置坐标等信息。由于台账信息存在数据缺失、位置信息错误等问题，且只记录了部分工业园区中的企业信息。因此本文主要选择3、4、5类中的区域，结合区域内人数与区域内企业人数进行验证。考虑到联通公司的市场占有率约为50%，因此将基于信令数据计算得出区域内人数乘以2再与区域内企业人数进行对比。对比结果如图45所示：

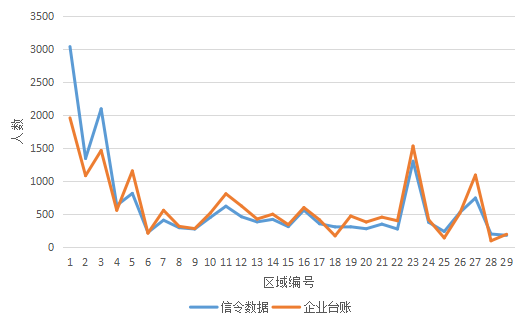


图 45 信令数据与企业台账区域人数对比

由图45可以看出，基于信令数据计算得到的区域人数与根据企业台账得到的人数存在较强的相关性。计算得到两者的皮尔逊相关系数为0.81，说明部分工业园区的工厂区域识别结果较为准确。

## 本章小结

本章对基于多源位置数据的人员特殊聚居区域识别模型进行了分析与验证。首先对实验环境与实验数据进行了简要的介绍。然后对用户停留点识别与职住地提取进行分析，将停留点数量与职住地热力图与交通发展年报统计的结果进行对比，验证了本文提出方法的有效性。接着对用户个人属性的预测结果进行分析，使用随机森林算法与SVM算法进行对比，证明了随机森林算法的准确性优于SVM算法。最后基于公交数据与共享单车数据对区域特征进行了分析，并结合地图信息与企业台账信息说明了工厂区域与学校区域识别结果的准确性。

# **结论与展望**

部分具有特殊社会属性的人群如学生、工人等，其出行规律与一般通勤人员相比具有较大的差异性，这些人群聚居在一起形成了职住功能混合的区域，该类区域具有人员流动性低，人口密集等特点，具有较为严重的安全隐患。伴随着移动通信技术的飞速发展，人群在使用手机时产生了大量时空精度较高的4G信令数据。对信令数据进行分析，挖掘人群的出行特征，利用出行特征辨别目标人群的社会属性，识别目标人群的聚居位置对相关部门进行预警防范工作具有重要的意义。

本文完成的工作如下：

1. 通过对手机信令数据进行质量分析，证明了信令数据可以对后续分析提供有效的数据支撑。对信令数据进行预处理，剔除长距离基站抖动的数据，筛选出优质用户的数据。
2. 针对信令数据具有时空信息的特点，引入ST-DBSCAN算法对用户信令数据进行聚类分析，辨别用户的停留态/移动态，完成了用户出行链的提取，然后基于用户多日的停留点数据，完成了用户的职住地提取。
3. 基于用户出行链数据与职住地数据，采用LCSS算法计算用户出行轨迹相似性、出行距离等出行特征。根据人工标注的训练集，构建基于随机森林算法的预测模型，完成了个体社会属性的估计。
4. 根据个体社会属性的预测结果，将学生、工人两类特殊人群的居住地聚类形成区域，并结合公交数据、共享单车数据计算区域的客流特征，采用K-means算法分析区域特征。

本文后续的研究方向如下：

在对个人属性进行预测时，本文采用的是一种有监督的学习方法，只针对学生、工人这两类特殊人群进行了分析。后续的研究应采用无监督的学习方法，对所有人群进行分析，挖据出不同属性人群的出行模式。

本文在分析公交数据与共享单车数据时，只考虑了其体现的客流特征。后续的研究应充分利用共享单车数据定位精度高、样本代表性广泛等优点，识别区域边界等更加细致的区域特征。

# **参考文献**

[1] Adomavicius, Gediminas, Tuzhilin, Alexander. Using Data Mining Methods to Build Customer Profiles[J]. Computer, 2001, 34(2):74-82.

[2] Sugiyama K. Adaptive web search based on user profile constructed without any effort from users[C]// International Conference on World Wide Web. ACM, 2010:675-684.

[3] Li R, Wang S, Deng H, et al. Towards social user profiling:unified and discriminative influence model for inferring home locations[C]// ACM Knowledge Discovery and Data Mining. 2012:1023-1031.

[4] Zhu Y，Zhang Y，Shang W，et al.Trajectory enabled service support platform for mobile users'behavior pattern mining[C].Mobile and Ubiquitous Systems:Networkin-g&.Services.MobiQuitous，2009.MobiQuitous 6th Annual International.IEEE.2009：1-10.

[5] Chen, Xihui, Jun Pang, and Ran Xue. Constructing and comparing user mobility profiles for location-based services. Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM, 2013.

[6] Ying, J.J.C., Lu, E. H. C., Lee, W. C., Weng, T. C., & Tseng, V. S. (2010, November). Mining user similarity from semantic trajectories. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks(pp. 19-26). ACM.

[7] 黄文彬,徐山川,吴家辉,王军.移动用户画像构建研究[J].现代情报,2016,36(10):54-61.

[8] 毕晓萤.面向城市低收入人群的交通改善策略研究[D].南京:东南大学交通学院,2013.

[9] Srinivasan S,Ｒoger P．Travel behavior of low-income residents: studying two contrasting locations in the city of Chennai,India［J］．Journal of Transport Geography，2005.

[10] Salon D，Gulyani S．Mobility，poverty，and gender: travel choices of slum residents in Nairobi，Kenya［J］． Transport Ｒeviews，2010，30( 5) :641-657．

[11] 程龙，陈学武.基于结构方程的城市低收入通勤者活动-出行行为模型[J].东南大学学报,2015,45(5):32-37

[12] 廖琨.基于手机信令数据的区域住房价格与居民出行特性相关性研究. 西南交通大学,2018

[13] 杜影.基于感知成本的低收入人群通勤交通方式选择研究. 东南大学,2015

[14] 城市老年人活动——出行行为特征及相关建成环境影响研究[J]. 宋彦李青,王竹影.西南交通大学学报(社会科学版) . 2018(06)

[15] 黎海涛,张平.蜂窝无线定位技术的发展及应用[J]. 电信科学, 2001, 17(9):20-23.

[16] 王明旭. 4G-LTE移动通信技术概述. 移动通信工程 TN929.5

[17] Information Technology - Data Streaming; Studies from Sun Yat Sen University in the Area of Data Streaming Described

[18] Calabrese F, Ferrari L, Blondel V D. Urban Sensing Using Mobile Phone Network Data [J]. Acm Computing Surveys, 2014, 47:1-20.

[19] Ahas R, Aasa A, Silm S, et al. Mobile positioning in space–time behaviour studies: social positioning method experiments in Estonia[J]. Cartography and Geographic Information Science, 2007, 34(4)

[20] Xu Y, Shaw S L, Zhao Z, et al. Another Tale of Two Cities: Understanding Human Activity Space Using Actively Tracked Cellphone Location Data[J]. Annals of the American Association of Geographers, 2016, 106(2).

[21] 杨佩瑜.基于低质量手机信令数据的出行链提取模型的研究与应用. 北京航空航天大学,2013

[22] 林楠,尹凌,赵志远.基于滑动窗口的手机定位数据个体停留区域识别算法[J].地球信息科学学报,2018

[23] Calabrese F, Colonna M, Lovisolo P, et al. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(1):141-151.

[24] Quercia D, Lathia N, Calabrese F, et al. Recommending social events from mobile phone location data[C]. Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 971-976.

[25] Horn C, Klampfl S, Cik M, et al. Detecting Outliers in Cell Phone Data: Correcting Trajectories to Improve Traffic Modeling[M]. 2014.

[26] Widhalm P, Yang Y, Ulm M, et al. Discovering urban activity patterns in cell phone data[J].Transportation, 2015, 42(4): 597-623.

[27] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. KDD. 1996, 96(34): 226-231

[28] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial–temporal data[J]. Data & Knowledge Engineering, 2007, 60(1): 208-221.

[29] 赵阳.基于手机信令数据的城域出行模式提取模型研究与实现. 北京航空航天大学,2014

[30] Yang C, Liu Y, Ma X, et al. Towards estimating urban population distributions from mobile call data[J]. Journal of Urban Technology, 2012, 19(4): 3-21.

[31] Long Y, Zhang Y, Cui C. Identifying Commuting Pattern of Beijing Using Bus Smart Card Data[J]. Acta Geographica Sinica, 2012, 67(10):1339-1352.

[32] 毛海虓. 中国城市居民出行特征研究[D]. 北京工业大学, 2005.

[33] 北京交通发展年报.北京交通发展研究中心.2018

[34] Bothe C M, Belik V, Couronné T, et al. Unravelling daily human mobility motifs[J]. Journal of The Royal Society Interface, 2013, 10(84): 20130246.

[35] 曹劲舟,涂伟,李清泉等. 基于大规模手机定位数据的群体活动时空特征分析[J].地球信息科学学报, 2017, 19(4): 467-474.

[36] 许宁,尹凌,胡金星.从大规模短期规则采样的手机定位数据中识别居民职住地[J]. 武汉大学学报(信息科学版), 2014, 39(6):750-756.

[37] Moiseeva A, Jessurun J, Timmermans H. Semiautomatic imputation of activity travel diaries: use of global positioning system traces, prompted recall, and context-sensitive learning algorithms[J].Transportation Research Record: Journal of the Transportation Research Board, 2010 (2183): 60-68.

[38] Mc Gowen P, Mc Nally M. Evaluating the potential to predict activity types from GPS and GIS data[C]. Transportation Research Board 86th Annual Meeting, Washington. 2007.

[39] Tu W, Cao J,Yue Y,et al. Coupling mobile phone and social media data:A new approach to understanding urban functions and diurnal patterns[J]. International Journal of Geographical Information Science, 2017, 31(12): 2331-2358.

[40] H.Gao and H.Liu,Mining human mobility in location-based social networks,Synthesis Lectures on Data Mining and Knowledge Discovery,vol.7,no.2,pp.1–115,2015.

[41] 林楠.基于大规模手机定位数据的居民活动链挖掘方法. 中国科学院大学,2018

[42] Zheng Y, Li Q, Chen Y, et al. Understanding mobility based on GPS data[C] Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008: 312-321.

[43] Liu L, Hou A, Biderman A, et al. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen[C] Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on. IEEE, 2009: 1-6.

[44] Kim K, Oh K, Lee Y K, et al. An analysis on movement patterns between zones using smart card data in subway networks[J]. International Journal of Geographical Information Science, 2014 (ahead-of-print): 1-21.

[45] Xu J H, Liu H. Web user clustering analysis based on KMeans algorithm[M]. IEEE, 2010.

[46] Sander J, Ester M, Kriegel H P, et al. Density-Based Clustering in Spatial Databases: The Algorithm DBSCAN and Its Applications[J]. Data Mining and Knowledge Discovery, 1998, 2(2):169-194.

[47] Breirnan, L. (1996a）.“Bagging predictors” Machine Learning, 24(2):123 140.

[48] Breirnan, L. (2001a）.“Random Forest” Machine Leαming, 40(3):113-120.

# **攻读硕士学位期间取得的学术成果**

1. 诸彤宇，**许伟**，魏翔. “一种基于手机信令数据与POI兴趣点数据的个体行为分析方法”，中国国家发明专利（已申请）

# **致 谢**

在研究生生涯即将结束之际，我衷心地感谢两年半以来所有帮助过我的老师、同学、朋友以及家人。正是由于你们无私的帮助，我才能够顺利完成充实而富有挑战的学业。

由衷地感谢我的导师诸彤宇副教授。从研究生入学至今，诸老师严谨的治学态度、深邃的洞察力，影响了我的整个研究生生涯，他饱满的工作热情和积极的生活态度对我的人生产生深远的影响。从诸老师身上，我学到了很多在书本上学不到的知识，诸老师平日工作繁忙，但从来没有忽视对学生的指导，正是由于老师耐心的帮助和指导，才使得我的研究工作可以顺利进行。

感谢郭之超、陆志龙、陈丽等师兄师姐以及魏翔、庞倩等师弟师妹对我科研工作上的帮助与指导。感谢朝夕相处的课题组同学，陆禹城、刘文倩等都给了我很多关心和帮助，我们共同学习、共同进步。感谢詹师傅和谈姐为我营造了良好的科研环境。

我由衷地感谢我的父母与我的室友陈珺、杨德剑，是他们在我研究中最紧张、最失落的时期给予我安慰与支持，让我能够战胜所有的困难，并且让我拥有了在未来继续在数据分析方向进行研究的希望与信心。

最后十分感谢各位评审老师在百忙之中评审我的论文。