武漢理工大學

（申请工学硕士学位论文）

基于海量物流轨迹数据的分析挖掘系统

培养单位：信息工程学院

学科专业：通信与信息系统

研 究 生：甘波

指导教师：周云耀教授

2014 年 6 月

分类号 _____ 密　　级 公开

UDC _____ 学校代码 10497

武漢理工大學

学　位　论　文

题　目 基于海量物流轨迹数据的分析挖掘系统

英　文

题　目 Analysis and Mining System Based on Massive Data

研究生姓名 甘波

指导教师 姓名 周云耀 职称 教授 学位 博士

单位名称 武汉理工大学信息工程学院 邮编 430070

申请学位级别 硕士 学科专业名称 通信与信息系统

论文提交日期 _____ 论文答辩日期 _____

学位授予单位 武汉理工大学 学位授予日期 _____

答辩委员会主席 刘泉 评阅人 刘泉

刘可文

2014　年　6　月

独 创 性 声 明

本人声明，所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽这里所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得武汉理工大学或其他教育机构的学位或证书而使用过的材料。与这里一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名：_____日 期：_____

学位论文使用授权书

本人完全了解武汉理工大学有关保留、使用学位论文的规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权武汉理工大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存或汇编本学位论文。同时授权经武汉理工大学认可的国家有关机构或论文数据库使用或收录本学位论文，并向社会公众提供信息服务。

（保密的论文在解密后应遵守此规定）

研究生（签名）：_____ 导师（签名）：_____日期：_____

摘 要

电子商务盛行的今天，物流行业空前繁荣，物流车辆的海量 GPS 数据量也越来越多，这些数据包含很多关于交通路况、车辆甚至社会经济发展等信息。轨迹数据挖掘主要通过统计和分析车辆行驶距离、停车时间、地理位置信息、车辆特征等发现货运线路特征，为物流公司提供基于时间、成本等车辆调度方案以及衍生出来的一系列 LBS[1]应用提供服务。

本文以海量 GPS 数据作为数据源，利用海量轨迹数据挖掘和道路推荐相关理论，通过建立聚类模型和分析海量 GPS 数据来了解物流车辆行驶规律，提出针对物流车辆货运线路推荐系统的设计框架并实现。其中重点就是数据预处理方法，停车点侦测和路径分割方法，相似货运轨迹聚类和货运线路推荐四个方面进行了深入研究。具体工作如下：

（1）作为轨迹数据挖掘的必要工作，研究了预处理方法，包括数据清洗，数据中的异常进行侦查和排除，并针对本系统所有的 GPS 数据进行了特征分析和提出了一种基于历史轨迹数据的异常点检测算法。本文提出的算法在处理海量轨迹数据时具有时间复杂度低的特点。

（2）停车点侦测和路径分割可以发现物流车辆的上下货的模式，本文依据朴素贝叶斯算法提出一种新的基于历史数据的路径分割算法，根据物流车辆在上下货时的停车和普通停车在时空属性上的不同，将轨迹进行分割。

（3）相似货运轨迹聚类将相同起始点和终点的轨迹规则化后投射到同一纬度然后分析轨迹特征，采用 K 均值聚类算法将这些规则化后的轨迹聚类，聚类后的结果中可以发现物流车辆频繁的行驶轨迹。

（4）货运线路推荐方面，设计了基于历史轨迹数据在时间，距离以及成本的不同，得出相应的推荐线路指导物流司机采用合理的行驶方案。

经测试表明，论文中使用的轨迹预处理方法与传统预处理方法相比速度更快、效率更高，但是牺牲了一些准确度，停车点侦测和轨迹分割达到了良好的效果。研究成果对于缺失车辆卸货点的轨迹分析有十分重要的理论意义。

关键词：贝叶斯分类器，轨迹数据挖掘，路径分割，异常点过滤，K 均值聚类

Abstract

In the age of the electronic commerce,the electronic commerce, logistics industry, logistics vehicles are more prosperity than ever.More and more logistics vehicles GPS data are produced.These data contain a lot of traffic information such as road conditions,vehicles,and even social and economic development.Through statistics and analysis of vehicle driving distance, time, location, vehicle parking characteristics, trajectory data mining can find shipping line characteristics,provide logistics company based on vehicle scheduling schemes such as time, cost, and derived a series of LBS application.

Taking massive GPS data as the data source, using massive trajectory data mining and related theory of road recommending, composed by online and offline system, this paper proposes and realizes a designframework of route recommending system for logistics vehicle through establishing clustering model and analyzing massive GPS data to understand the driving rule of logistics vehicles. The key approach is deep studying on data preprocessing, stops detecting, route segmenting, similar freight trajectory clustering and freight lines recommending. The specific work is as follows:

As a necessary work in trajectory data mining，i study the pretreatment method, including data cleaning, data of abnormal detection and exclusion, and with the characteristics of this system all the GPS data in analysis and put forward a kind of anomaly detection algorithm based on the historical trajectory data. The algorithm for massive data processing has low time complexity.

Parking points detection and path integral can find that the pattern of logistics vehicles and goods.In this paper, on the basis of naive bayes algorithm, i put forward a new way for trajectory segmentation,according to the logistics vehicle parking and ordinary ponits using the different attributes of time and space between them when goods are loaded and unloaded.

I regulate freight trajectories clustering similar to the same starting point and end point of the trajectory . then project them on the same latitude.After that using k-means algorithm, the characteristic of trajectory is analyzed, and finally get

logistics vehicle general movement tracks.

In the recommendation of shipping lines, based on the difference of historical trajectory data in time, distance and cost ,i design and draw the corresponding recommended route guidance which logistics driver adopts through reasonable driving scheme.

Compared with traditional pretreatment method.These tests show that the method of pretreatment trajectory is faster, more efficient, but sacrificing some precision. Detecting parking spots and track segmentation achieve good effects. In the cases of missing stops of vehicle trajectory analysis,the research results have very important theoretical significance.

Keywords: bayesian classifier, trajectory data mining, carving, abnormal point filter, k-means clustering

目 录

第 1 章　绪论

1.1 课题来源

中科院深圳先进技术研究所和深圳市宇易通科技有限公司合作开发设计的一个易流云平台系统，该平台旨在为物流企业提供真实物流信息服务的真实运力服务。

1.2 研究的背景和意义

当今社会属于互联网高速发展的时代，许多传统行业都受到了剧烈的冲击，其中电子商务逐渐兴起，商业活动的网上交易呈爆发式增长，伴随来的是物流行业的蓬勃发展，为物流车辆轨迹数据挖掘提供了海量的数据源。轨迹数据挖掘是数据挖掘在轨迹上面的新的应用，它包括了轨迹数据存储、轨迹数据预处理、轨迹数据获取和挖掘以及应用。尽管有关于传统的数据挖掘系统已经有了多年很成熟的理论和方法，但是由于轨迹数据的特殊性，依然有许多轨迹问题需要深入研究，例如轨迹索引、查询、模式挖掘、不确定性和隐私保护等。其具体特点如下：

（1）数据海量性：物流车一般以 30s 的间隔向数据中心发送当前位置，这些移动在全国各地路网中的物流车辆每天生成的 GPS 数据都达到了 GB,甚至 TB 规模，并且还在不断增长中。这既是发展数据挖掘的驱动力，也是对数据挖掘的面临的难题。

（2）数据稀疏性：虽然这些轨迹数据规模庞大，但是由于地理因素（如车辆行驶在山区、雨雪天气）、设备故障等原因，并不能保证每一个路段都有完整的 GPS 信息，甚至会有一些是错误的 GPS 数据。

（3）数据复杂性：物流车辆在实际行驶过程中受到各方面主客观等因素难以简单通过某个模型或者理论进行评估和预测。主要有下列因素，每个司机都有自己的驾驶习惯，即使同一个司机在驾驶过程中也会针对不同客观条件改变自己的驾驶行为，例如天气、实时路况，这些不确定性无疑增加了轨迹数据挖掘的复杂性。

（4）数据丰富性，在海量的轨迹数据背后隐藏着全国实时路况信息、物流运输状况信息和我国不同区域经济发展水平。对于我国道路基础建设、交通路

径规划、物流车辆调度等提高我国物流行业水平具有重大意义。

车辆轨迹是车辆的位置和时间的记录序列，可以很容易的使用小型 GPS 记录仪、车内导航设备、甚至手机获取，作为轨迹数据挖掘中的重要研究对象，分析和挖掘这种数据类型可以应用于城市热点区域分析、智慧物流和交通规划等多个方面[2]。不同的物流车辆轨迹通过分解在不同时间、空间等很多维度上以后，既有相同或者类似的部分，也有不同的地方，通过分析和统计其相似性和相异特征可以挖掘出轨迹数据背后包含的很多知识。在我国建设信息网络技术，城市，交通，物流的背景下，这些知识作为宝贵的财富可以不断推动我国向信息化、智慧化城市、物流积极发展，降低运输成本，提高经济效益，最终实现物流业智能化、智慧化。

1.3 国内外研究现状

轨迹数据中的知识模式发现和处理有很多不同的方式，可以首先通过不同概念模型描述轨迹，然后在不同维度上的特征将轨迹分组，接着分析这些经过分组后的轨迹组内和组间相似性和相异性，发现偏离正常数据的异常轨迹，针对不同应用场景调整轨迹分类策略等，最终达到发现获取轨迹背后的知识。在这一个过程中，经常会遇到各种各样难以克服的困难，譬如数据量巨大、数据维数灾难、数据受到主客观因素污染、数据不确定、知识发现角度多种、知识表示困难等技术难点[3]。

轨迹数据挖掘发现的知识类型和所使用的方法密切相关。所发现的知识的价值受到数据挖掘算法的影响，常用的轨迹数据挖掘技术有规则归纳、概念簇集、关联发现等。在实际轨迹数据挖掘应用中，应当根据不同的需求采用不同的工具、方法以及理论。

目前的轨迹数据挖掘研究工作中主要为轨迹聚类、轨迹分类、离群点检测、兴趣区域、隐私保护、位置推荐等方面。作为轨迹挖掘重要的一部分：异常轨迹检测中，也已经提出了许多算法。传统的轨迹异常检测中，通常是提取轨迹某些特征，计算这些特征间的差值再进行加权得到轨迹间的距离。克诺尔[4]通过将轨迹分解、降低维数得到若干个包含主要轨迹有用信息并且相互独立的特征,如轨迹所包含的 GPS 点的数量、轨迹运动快慢、轨迹起始点的坐标位置、轨迹运动趋势等。通过检测的异常和正常轨迹数据路径的距离不同，以确定其缺陷的异常信息，它的缺陷在于:由于轨迹内部不同局部区域也存在特征上的差异,因

此上述方法只适用于特征单一或者长度较短的轨迹。伊利诺伊大学的 Li[5]建议构建了一种轨迹异样检查框架 ROAM。该框架将首先通过轨迹离散化分成一个个独立的名为 Motif 的片段，该片段提取轨迹的某些特征信息构成 Motif 特征空间，利用构建的 Motif 中的属性信息，这个分类器最终用于将不同轨迹数据分类从而获取轨迹背后蕴藏的知识。

为了克服传统轨迹中无法针对轨迹较长或者特征较复杂进行有效的检测。刘良序[6]首先通过不同轨迹间的相似程度的不同提出了部分相似、完全相似和离群轨迹的模型，将一段较长的轨迹分为若干独立无关的轨迹段，利用之前定义的模型和概念，然后比较每一个分段之间的匹配程度，设定不同阈值来确定这些较长的轨迹是否相似，并且使用了 R 树来来克服计算量过大的问题。也有一些科研人员通过数据挖掘中的密度聚类思想，密度越大的地方，轨迹越趋向于正常轨迹，密度越小的地方轨迹越有可能为异常轨迹，譬如 Liu[7]。

轨迹聚类是在相似的轨迹中找到不相似部分的过程。轨迹特征空间中不同密度代表不同轨迹在该属性上相似程度的不同，并且特征空间中不同的属性对轨迹相似程度的影响也各不相同[8]。不同轨迹从时间区间这个角度来看，其相似性也各有不同，本文从时间间隔出发，将不同聚类方法分为如下几类，如表 1-1 所示。这些方法逐渐在相似的时间间隔从时间上的要求相似性，本地时间间隔相似，最后到没有时间对应的相似性下降。它反映了人们在时间和空间探索时空轨迹和轨迹相似性度量的多样性。

表 1-1 轨迹聚类方法分类列表

相似性度量	代表聚类方法
全区间时间相似	欧式距离[9]，最小外接矩形距离[10]
全区间变换对应相似	动态时间规整[11]
多子区间对应相似	最长公共子序列距离[12]
单子区间对应相似	子轨迹聚类[13]
单点对应相似	历史最近距离[14]
无时间区间对应相似	单向距离[15]

目前有关轨迹数据挖掘的研究主要关注在轨迹的时空特性上，已经建立了一些关于轨迹数据建模、数据存储、轨迹索引、轨迹查询、轨迹挖掘方面，但

是有关轨迹的语义信息的理论研究却并不多，Yan[16]等人提出了面向轨迹数据语义信息分析与挖掘，获取物体有关运动的未知知识，这些知识是轨迹挖掘更深层次的应用。

在目前云计算和大数据的时代下，只有对轨迹数据挖掘进行更深入分析和挖掘，研究物流系统模仿或者实现人类的在不同诸如天气、实时路况等诸多客观因素下的行为，具备模仿人的智能，学习推断并自适应解决出现在物流运输、存储的问题的能力，也就是当商品从出库、车辆中转调度、行驶路线和时间一系列问题作出合理正确的规划，最终达到物流的智慧化。

1.4 论文内容和组织结构

本文依据轨迹数据挖掘的一般流程，首先分析 GPS 数据特征，并提出针对海量轨迹数据的预处理方法，接着提出采用贝叶斯分类器算法，并将此算法应用到的轨迹分割处理中，将不同车的不同轨迹提取出来，接着研究了轨迹聚类的相关算法，提出将 K 均值的聚类算法运用到轨迹聚类中。最后针对以上算法和系统进行了实验，将以上结果应用带货运线路推荐系统中。

第一章中本文系统阐述了轨迹数据挖掘产生的原因和意义和一些已有的方法和理论研究的发展和现状，以及存在的问题，最后是论文的结构。

第二章主要阐述了轨迹挖掘概念、一般过程和方法，提出了一种基于历史轨迹数据的货车运送线路推荐系统。

第三章介绍了轨迹计算的一般流程，详细分析了数据预处理、停车点识别和轨迹分割的流程，提出了轨迹数据异常检测算法，基于朴素贝叶斯分类器的轨迹分割算法，完成了货运车辆的起止点识别，从而为轨迹分割提供了依据。

第四章详细阐释了货运车辆轨迹聚类意义和存在的问题，分析了轨迹聚类流程，首先将轨迹规则化，然后通过 K-均值算法将轨迹聚类获取货运线路常用行驶路线，并提出了基于 GPS 数据的线路推荐方法。

第五章对整个系统进行实现和验证，利用 matlab 绘图总结分析得出结论。

第六章总结与展望，对于本文所做工作不足之处进行了总结和对未来轨迹数据挖掘的展望。

第 2 章　轨迹数据挖掘技术研究

2.1 轨迹数据挖掘介绍

2.1.1 轨迹数据挖掘概念

轨迹数据挖掘（Trajectory Data Mining）是数据挖掘技术中的一个重要的新兴领域，它的研究对象来源于越来越多可移动装置上装有 GPS 等定位设备并不断记录人类或者车辆的运行轨迹，在传统的数据挖掘过程和算法基础上针对移动轨迹数据特征，重点研究轨迹数据预处理、轨迹数据中的不确定性研究、轨迹数据索引与存储、轨迹模式发现、轨迹隐私保护以及基于位置信息的社会化服务。是计算机技术、存储技术、统计学、地理信息学和新技术等多学科的整合。轨迹数据本身的海量性、复杂性也对传统的数据挖掘算法提出了很多新的挑战，原有的数据挖掘对象往往数据量比起轨迹数据而言不大，为此很多新兴的数据库存储和索引技术、大数据处理解决方案也层出不穷不断，例如空间数据库、内存数据库、批处理数据处理框架、实时流计算框架等[17]。

一般而言，轨迹数据挖掘，是指从大量轨迹数据的集合 C 中发现隐含模式 m 和知识 n 的结果 S。因此，轨迹数据挖掘的过程可以看作为一个函数：

$$£ : C \rightarrow S(m, n)^{[18]}, \qquad (2\text{-}1)$$

输入是轨迹数据，输出是隐含模式 m 和知识 n。通过使用某些技术、理论，从大量的轨迹数据提取模式、发现庞大知识的一个过程。

2.1.2 轨迹数据挖掘内容

轨迹数据挖掘目前的研究热点集中于轨迹聚类、异常点检测、轨迹分类、位置推荐等方面。如图 2-1 所示。

（1）轨迹聚类。通过轨迹聚类的方式可以发现轨迹数据中的相似性和异常特征，从而得到对于轨迹应用中有益的模式，例如发现热点区、通过大量物流车辆的历史轨迹可以找到收益最高的行驶路线、监控物流货运车司机驾驶行为等。通过研究不同轨迹数据在时空等方面的特征，定义设计不同的准则去度量轨迹的相似性，利用相似性的不同将轨迹区分开来。

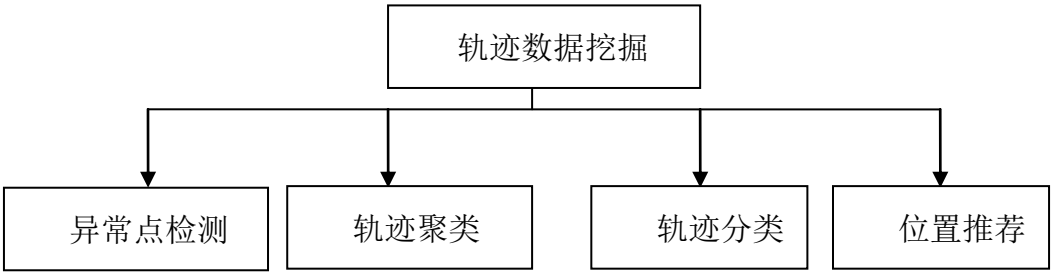

图 2-1 轨迹数据挖掘热点分类图

（2）异常点检测。在数据挖掘领域，异常数据的识别是其中比较重要的一部分，所谓的异常数据指的并非由随机误差造成的偏离大部分数据的特征的那部分数据。异常数据就识别出数据集中的异常点。同时混杂在异常数据中的正确数据也需要识别。它涉及怎么样定义异常数据和寻求有效的算法来识别并剔除掉这部分数据。

（3）轨迹分类。轨迹分类与轨迹聚类的目的相反，指的是通过统计和分析不同轨迹间的时空特征，抽象出轨迹模型，并以此模型作为分类器，对目标轨迹进行分类，这是一个不断迭代的过程，并且出于不同的分类目的定义目标轨迹模式，通过轨迹历史模型，从而智能的将轨迹数据分类。

（4）位置服务。车辆的行驶轨迹不仅仅是车辆行驶路线的一种记录，更是反映了驾驶人员针对驾驶活动期间客观因素，例如天气、实时路况、加油站等地理位置信息的智能反应，通过搜集、统计、分析这些历史轨迹可以极大提高车辆管理、监控、调度、路径规划效率，在当今越来越开放的网络条件下，实时共享位置信息，可以为公司和客户提供精准且高效的物流配送服务。随着这些位置信息的不断挖掘和共享，极大降低了物流行业中的沟通成本，降低中间环节消耗。

2.2 轨迹数据挖掘流程

2.2.1 数据来源和预处理

轨迹数据来源是移动设备所发出的位置信息，对于物流轨迹数据而言，通常是 GPS 信息。预处理的过程首先是分析所获取数据的总体特征，再依据数据的特征采用不同过滤算法。

（1）挖掘数据来源

轨迹数据挖掘来源通常是终端设备上产生的位置记录，然后位置信息传回数据中心以日志文件形式存放。本文采用的是数据中心的 GPS 日志记录。表 2-1 是 GPS 日志的记录表结构。

表 2-1GPS 日志记录表结构

属性域	描　述
车辆编号	车辆的唯一编号
经　度	以度为单位的维度值乘以10的6次方
纬　度	以度为单位的维度值乘以10的6次方
里　程	0.1KM
速　度	KM/H
方　向	0-359，正北为0，顺时针
高　度	海拔高度，单位米
GPS时间	接收到GPS的时间
状　态	0位：0：未定位　1：3D定位

一条典型的 GPS 日志记录如下，其中各个字段之间用逗号隔开,车牌号码已经被加密。XI081FB4GU,115045136,30511584,412698,62,75,0,13-12-1 20:5:48,1。

（2）轨迹数据特征分析

轨迹数据特征分析是指观测轨迹记录中所具有的包括空间属性、数字特征、分布结构等在内的特征，是数据应用的基础，它包括很多方面的统计和分析，例如离散轨迹点随着时间增长时候的方向信息、起止点最远距离信息，最大时间间距信息等一维数据信息。在这些一维数据的基础上分析轨迹点的密度的稀疏性、分布结构的信息有利于发现热点区域等，通过线性回归的方式有利于常见轨迹模式的发现和提取，不断研究轨迹时间与时间、时间和空间、空间与空间之间的关系。

（3）轨迹数据异常检测

轨迹数据挖掘领域异常检测一直是研究热点，通过数据挖掘中的常规模式析取可以发现大量轨迹数据时空特征以及背后的语义信息，但是有时候异常的出现也包含很多重要信息，例如可以修正之前轨迹特征空间划分的方式，发现

隐藏或者突发的信息等。所谓异常，有很多各种各样的概念描述，概括而言指的是这样的一些数据具有整个数据集合中不同寻常的属性和特征，异常与错误不同，异常的产生并非人为原因造成，也并不是随机因素影响。轨迹异常检测可以分为两个过程实现，首先在轨迹数据特征分析的基础之上发现和定义异常特征，然后利用这些特征空间与轨迹数据集比对，找到符合这些异常特征的数据集，不断迭代修正异常特征空间，最终检测出异常轨迹数据。目前关于异常检测有很多种方法：例如使用统计学知识对轨迹的时空分布特征统进行分析，找到不符合常规分布特征的数据集合。通过定义轨迹间距离的方式来发现异常轨迹，关于距离的定义也有很多种，欧式距离、曼哈顿距离、汉明距离、切比雪夫距离等。定义轨迹在时空属性的密度特征也可以发现异常轨迹，密度大的区域的轨迹趋向于正常轨迹，密度小的区域的轨迹趋向于异常轨[19]

2.2.2 轨迹数据路径分割和聚类分析

通过 GPS 设备获取的轨迹数据只包含每一轨迹点的经纬度和对应的时刻信息，通过这些数据无法直接得到活动行为的特征信息，如停车时间、是否卸货、物流的目的地、以及其他信息。要想获取这些信息首先就必须进行停车点侦测，判断是那些时刻是真实停车，还是 GPS 产生了漂移误差，或是由于红绿灯、交通拥堵造成的停车，或者是停车卸货等一系列重要信息和知识。图 2-2 为停留点侦测示意图。

轨迹分割是轨迹数据挖掘的基础和前提，目前，轨迹分割算法有探索和机器学习方法两大类[20]。探索性方法考虑移动对象停留和移动时的时空特征或定位设备的特征，作为已知经验，设计算法对轨迹原始数据进行处理和分析。机器学习是人工智能的一个分支，目的是构建一个可以从数据集中学习的系统，机器学习主要解决数据表示和泛化的问题，数据实例的表示和评估是所有机器学习系统中的重要部分，机器学习系统对数据集泛化的能力是对未知数据分类和计算的核心关键部分。通过机器学习的方式可以使轨迹数据挖掘更加智能，不仅能够发现轨迹已有的特征，还能针对未知轨迹数据学习发现新的特征。机器学习方法有很多种，决策树学习，关联规则学习，人工神经网络，支持向量机学习，聚类，贝叶斯网络等[21]。

聚类的目的是尝试将具有相似特征的轨迹划分开来，凸显不同轨迹间的相似性，为更深层次的研究打下基础。可以将轨迹特征空间中不同属性相互独立

抽取出来，找到每一个被划分的属性与整体分布特征之间的关系[22]。根据相似度量的不同，常见的基于距离轨迹聚类方法有基于欧氏距离、最长公共子序列距离、时间聚焦距离、历史最近距离等。

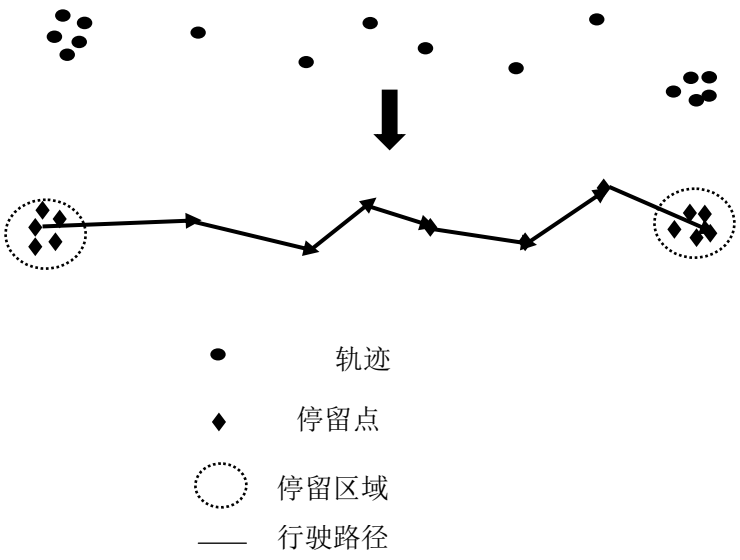

图 2-2 停留点侦测示意图

2.3 基于历史轨迹的线路推荐服务

目前基于位置的服务（LBS）已经在市面上取得了大量应用，例如旅游线路分享、车辆调度和安保等。简单的将轨迹展现在地图或者其它的媒介上，统计其行驶距离、时间、频率、停车位置等难以发现轨迹中包含的司机驾驶习惯，交通道路信息，热点区域以及路径信息等。其实，轨迹记录了驾驶人员在真实世界的活动，而这些活动将在一定程度上体现了驾驶时的各种环境因素，比如交通路况、天气、经济成本等。而传统的路径推荐系统主要是通过基于地图的最短路径算法生成的推荐线路，通过这种的算法规划得到的路径由于没有考虑到实际中地理位置信息，例如停车场，加油站，货运园区工厂以及道路限行、限速路段，监控区域等，并且这些地理位置信息经常出现新增，变更，删除，不准的状况。基于历史轨迹的路线推荐系统由于是考虑了实际需求，由于是司机驾驶的实际路径结果，包含了司机对真实地理天气和道路综合考虑的结果，因此推荐的线路也更合理、更多样，可以基于最少时间、最少费用，也可以最

短路径等。

基于历史轨迹的线路推荐服务的一般结构如图 2-3 所示。它是依赖聚类结果，包含基础的数据处理并挖掘相关知识。包括数据采集、数据预处理、数据分析。

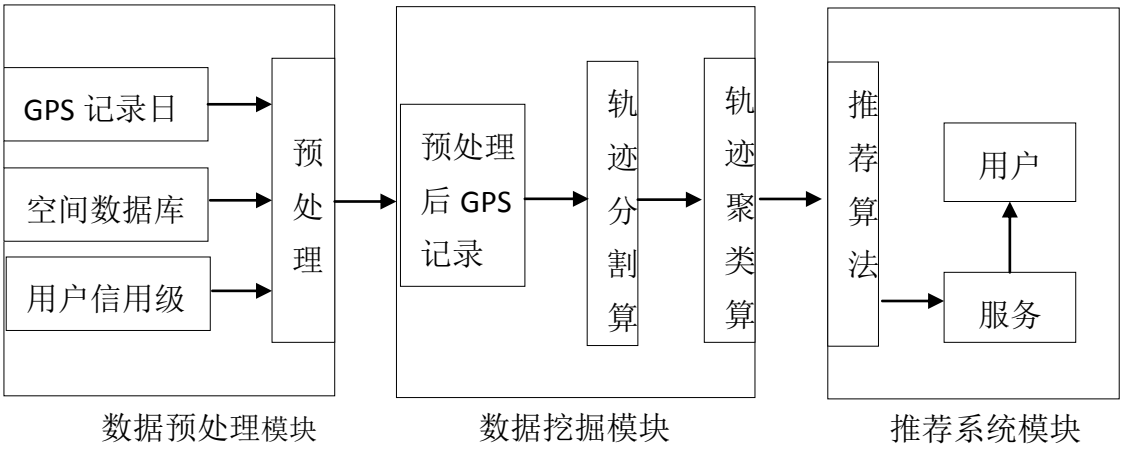

图 2-3 基于历史轨迹的线路推荐服务一般结构

2.4 本章小结

本章细述了轨迹数据挖掘的基本流程：数据源的获取，数据预处理，轨迹分割、轨迹聚类，推荐服务。对基于历史轨迹的线路推荐服务系统做了分析，说明了使用的聚类算法和推荐算法，根据数据挖掘的知识完成货运线路推荐功能。

第 3 章　轨迹数据预处理和轨迹分割方法研究

3.1 轨迹计算流程

通过定位技术采集到的原始轨迹数据只是一系列的经纬度、时间、速度等信息，通过这些信息无法直接得到物流货运车的活动行为的特征信息，例如运送货物的起始点、途经哪些城市信息，以及更深层次的活动规律等。这些原始的 GPS 数据必须经过一系列的处理步骤，才能获取到物流货运车的送货规律等特征信息[23]。如图 3-1 为轨迹计算的流程图。其中轨迹预处理包括数据规范化、异常点去除等，原始的 GPS 位置信息并不包含停车点信息，停车点识别指的是从轨迹点中提取停车位置信息，上步中识别出来的停车点信息由于并不包含货运车辆的上下货的位置信息，轨迹分割便是从轨迹中识别出一趟完整的货运信息，包括货运的起点和终点等信息。 这是本文的核心部分，最终得到的轨迹输出结果可以用于物流货运车辆的轨迹推荐。

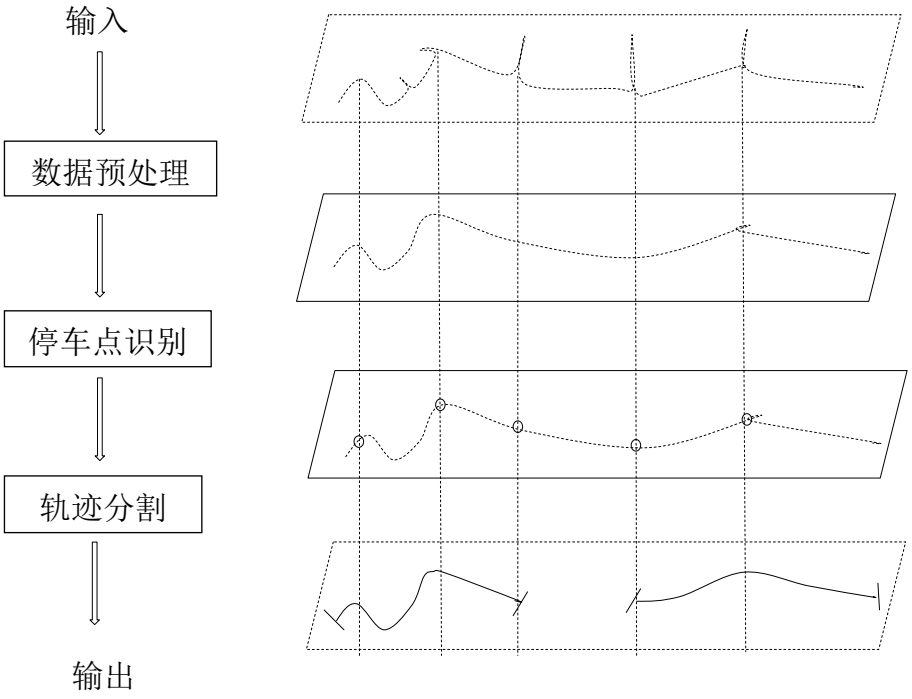

图 3-1 轨迹计算流程图

3.2 轨迹数据预处理

3.2.1 轨迹数据特征

使用车辆的海量 GPS 数据要面临的首要问题，是如何发现和处理大量数据中的异常元素。有很多客观因素诸如天气、实时路况、设备异常会使轨迹数据发生异常。本文基于某物流公司每天实际的 GPS 数据进行了空间分析，包括时间、速度和位置信息，得出的空间特征和规律如下。

（1）数据量大。本文中货运车辆每周产生的数据记录数约为 4 千万条。详尽的数据对分析和挖掘有利。但对后续的数据的分析过程和算法提出了高的要求。针对此问题，本文基于 Hadoop 分布式[24]平台以提升 GPS 并行处理的能力。

（2）数据重复。造成数据重复的原因是多种多样的,例如当车辆处于信号差的区域，山区，隧道等，或者设备本身异常或者故障导致重复发送相同 GPS 数据，车辆停车时也可能会造成 GPS 数据发送重复，所以，因这对这些数据记录进行标注并删除。这样便可以有效的压缩减少无效的数据。

（3）数据缺失。物流货运车辆在其运行期间 GPS 接收机设定的接收时间间隔一般为 30 秒到 1 分钟之间, 但是由于地理因素（如车辆行驶在山区、雨雪天气）、设备故障等原因，并不能保证每一个路段都有完整的 GPS 信息，甚至会有一些是错误的 GPS 数据。这些缺失的数据对于获取和分析轨迹行驶信息造成的严重的影响，这些缺失的数据可以通过借助一些地理信息补回。

（4）GPS 漂移。在 GPS 设备定位过程中，所标识的位置和用户实际位置有一定的出入，常见的现象是实际轨迹和先漂移轨迹混杂在一起。车辆即使实际原地位置不动的时候产生的经纬度信息也是不断变换的。有很多客观原因会造成这种现象，例如 GPS 设备在长期使用过程中并没有初始化或者调校，造成实际位置和显示的位置之间有一定的误差，设备实时搜到的卫星数量、卫星本身的位置分布等，由于 CPU 处理速度或者算法不够好，使得车辆在以较快速度行驶时的 GPS 信号与车辆静止时候相比较，经纬度偏移。目前 GPS 设备在城市中定位精度在 10 米左右，偏移在 50 米左右，由于城市道路密集和复杂，这些偏移足能够影响轨迹数据分析结果。

使用包含重复、异常和错误的 GPS 数据会影响后续轨迹数据挖掘的结果和效率，因此对海量 GPS 数据中的异常元素进行排除具有很重要的意义。

3.2.2 轨迹数据异常点检测

目前关于轨迹异常点排除算法大多通过基于划分[25]、统计[26]、密度[27]等方法。基于统计的方法通常是使用一些数据在统计学上的分布特征，例如正态分布异常点检测，如果某个数据对象偏离数据集均值到阈值则被归为异常点。该方法依赖于数据的分布、异常点类型等，该方法有坚实的数理统计理论支撑，然而当缺少数据分布特征的参数时，通过一些方法确定分布来拟合也是十分复杂低效的。使用划分的方法是一种常见的聚类分析手段。它通过将所有数据聚类成不同的簇，然后没有归类到任何簇的数据点则为异常点，该方法的时间和空间复杂度低，发现的异常点可靠性高，例如常使用 K-均值聚类算法将轨迹聚类。该方法存在的缺点是需要预先知道聚类数 K 值，聚类中心选取不准可能导致无法得到正确的分类结果。密度检测方式是数据挖掘中的常用方法，轨迹数据集中每一条轨迹被投射到不同维度上，然后比较每一块区域内的密度以及相邻区域大小，密度大的地方轨迹数据越趋向正常，密度越小的区域轨迹为异常数据可能性较大。它存在一个很大的问题就是计算量大，每次必须计算每一点的邻域，造成速度慢 。此外还有利用路网等 GIS 信息进行道路匹配的思想进行的异常点检测，该方法由于需要精确的知道路网信息，此外算法时间和空间复杂度高，当面临本文所遇到的海量轨迹数据处理的时候难以适应在实际生产中需要快速得到分析结果的应用中。

为了快速高效的检测轨迹中异常点，本文采用了基于网格划分的思想来对异常点进行检测的思想。其算法过程如下：

（1）将地图区域按网格划分。本文将 GPS 数据定义为一个二维平面中的一个点 $p_i=\{lon_i,lat_i\}$，表示经度，lat_i 表示为纬度。以地球纬度和经度作为坐标轴将包含地图区域可以简单映射为一个二维平面 $R^2\{(lon,lat)\mid lon\in R^*, lat\in R^*\}$。然后使用平行于坐标轴的直线把地图划分大小相等的网格 $R_i=\{lon_{\max},lon_{\min},lat_{\max},lat_{\min}\}$，其中 $lon_{\max}$、$lon_{\min}$、$lat_{\max}$、$lat_{\min}$ 分别为格子的上边界、下边界、右边界、左边界，所划分得到格子的集合 $S=\{R_1,R_2,R_3...R_{i-1},R_i\}$。定义一个映射关系 $F:R^2\to S$，通过分别对经纬度上下取整得 $R_i=\{ceil(lon_i),floor(lon_i),ceil(lat_i),floor(lat_i)\}$,其中 lon_i 和 lat_i 的精度的不同将会影响格子的大小。

（2）计算映射到网格轨迹数据点数量 P_{cnt}。判断 P_{cnt} 是否小于异常点阈值 C_{thre}，如果成立，则该网格内的所有数据点判为异常点，否则非异常点。

（3）对于非异常点的网格R_i，找到网格内数量最大的网格作为第一个类G_1，计算网格内部的中心点$R_c=(R_1+...+R_n)/n$，然后找到欧式距离它最远的网格点作为第二个类G_2。

（4）计算其它网格到初始两个类的网格的欧式距离d_{ij}。如果网格距离小于D_{thre}，则将该网格归到此类，否则，将其定义为新类G_i。

（5）获取没有被归类到任何类中的网格，该网格内的点既可以被认为异常点。

本算法相比较传统的划分方法的异常点检测有如下优点:（1）不必事先指定分类数 K，在不断的分类迭代中获取分类数。（2）初始分类中心并不是随机生成的，有效避免了只能收敛到局部最优的情况。（3）可以有效且方便的检测出异常点，由于道路一般比较分散，尤其是货运车辆所行驶的线路，不在网格内的数据点既可以判为异常点。（4）时间复杂度低。由于对整个区域按照网格划分计算，所有其时间复杂度接近于线性复杂度，当分类数越小时，数据越集中时时间复杂度越低。

本算法的流程图如图 3-2 所示。

3.3 停车点识别方法研究

轨迹分割就是首先将空间上离散的轨迹点划分成停车点和移动点两大类，并且停车点可以划分为三种类型，第一种主要为在城市道路中由于红绿灯等待造成的停车，一般这种停车时间较短，对于轨迹分割没有意义。第二种由于司机加油、吃饭、交通拥堵造成的停车，这种停车时间一般较长，对于本文的轨迹分割有较大影响。第三种为停车卸货，通过这个停车点可以识别出一趟货运线路的起始点位置等信息。在充分考虑了 GPS 数据特征的情况下，我们可以依货运车的速度信息来判断是否停车，本文采用了基于速度的停车点算法。该算法分为 3 个步骤，计算 GPS 点的速度，判断疑似停车点，停车点点识别[28]。

（1）计算 GPS 点的速度

由于 GPS 的定位精度较准，误差一般在 10m 到 50m 之间，时间间隔一般为 30s 到 60s 之间。GPS 点的速度可以由相邻前后 GPS 点连成直线的平均速度来替代。如图 3-3GPS 点速度计算示意图。计算P_4点的速度公式为：

$$v_{p_4}=\frac{d_{3,4}+d_{4}}{\Delta_{t_{3,4}}+\Delta_{t_{4}}} \tag{3-1}$$

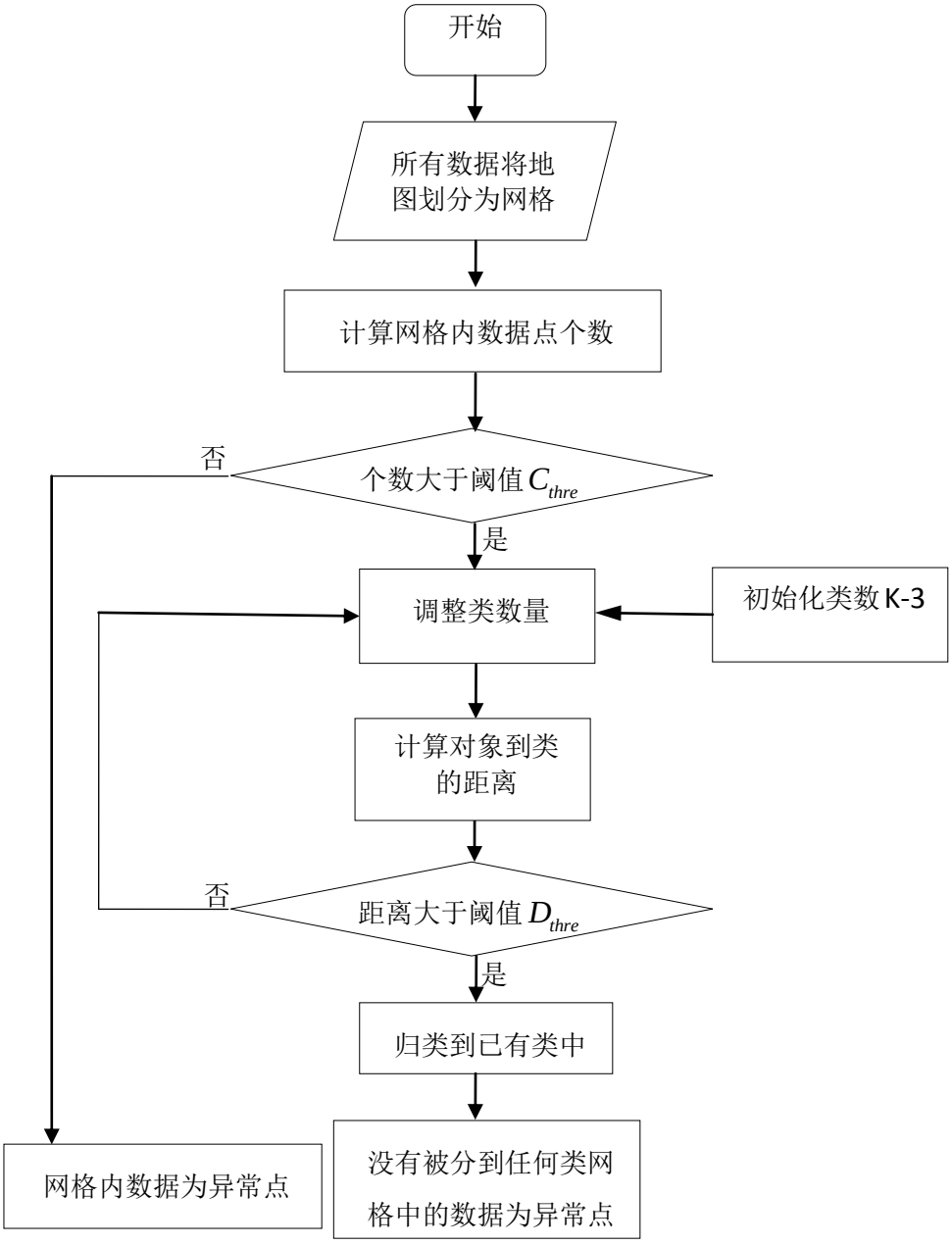

图 3-2 轨迹异常检测算法流程图

式 3-1 中 $d_{i,j}$ 表示 GPS 点 p_i 和 p_j 之间的位置差 $\Delta_{t_{i,j}}$ 为 p_i 和 p_j 的时间间隔，v_{pi} 为 p_i 时刻的速度。

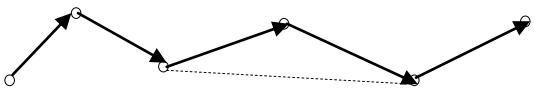

图 3-3GPS 点速度计算示意图

（2）判断疑似停车点

停车点的判断需要靠速度阈值来判断，可以设置一个速度上限 v_{max} 来判断是否停车，首先依据 v_{max} 将所有的 GPS 点划分为两类，疑似停车点和行驶点，并且由于停车的时候不应该只有一个点的速度小于 v_{max}，至少应当有若干个。形成一个停车点候选区域。停车上限 v_{max} 可以设置为人步行的速度约为 1m/s。该过程如示意图 3-4 所示。其中速度的单位 m/s。

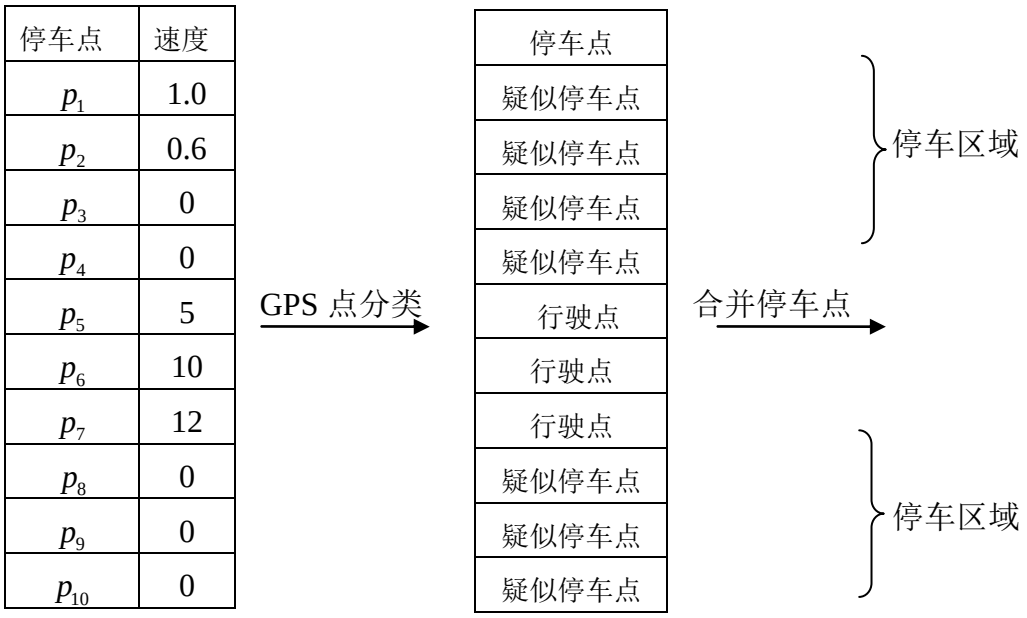

图 3-4 停车区域判断示意图

（3）停车点识别

通过前面计算得到的一系列疑似停车点，可以结合停车距离的阈值范围，选择停车时间最长的疑似停车点作为最终的停车位置。具体算法步骤如下：

I. 读取轨迹中的第一个疑似停车点 s_1，将其放入停车点序列 *Seq* 中。

II. 判断是否还有疑似停车点 $s_i(i=2,3,4,5...)$，如果有则计算这个点与上一个停车点 s_{i-1} 的距离间隔 $d_{i-1,i}$，如果 $d_{i-1,i}$ 小于距离阈值 d_{max}，则将该点放入停车序列 *Seq* 中，并重复步骤II，否则进入步骤III。

III. 计算 *Seq* 中的停留起始时间 *Seq.start* 和结束时间 *Seq.end*，如果停留时间小于时间阈值 $\Delta_{t_{min}}$，则清空 *Seq*，如果大于 $\Delta_{t_{min}}$，则保留此次停车记录。

通过停车点识别，可以找到物流货运车的简单行驶规律，例如停车地点，逗留时间，但是对于深层次的轨迹信息挖掘这是远远不够的，只能通过对停车点更深层次的挖掘才能获取货运车辆一趟完整的货运信息。

3.4 轨迹分割方法研究

3.4.1 贝叶斯分类器概述

轨迹数据挖掘中，如何对轨迹数据集合分类是一个重要问题，准确的分类是后续轨迹分析的基础。分类是从数据集合中提取描述数据类中重要属性的过程，通过分类可以更深入了解数据特征，机器学习中已经有很多种分类方法，例如模式识别和统计学方法，传统的分类方法中都是基于较小的数据规模，近年来随着大数据越来越受到科研界的重视，逐渐发展出来一些针对海量数据分类的技术。分类器可以用来判定一个未知的数据归于哪一类，例如在轨迹数据挖掘中，当进行轨迹分割的时候就需要判断哪些点属于停车点，哪些点属于卸货点，这就是分类问题的一个典型应用。数据分类过程可以分为两步，第一步是学习，也就是构造分类模型，第二部是分类，就是利用第一步产生的分类器对未知数据进行分类。

第一步中，通过描述已经带有类型信息的数据集合来构造分类器，这被称为训练阶段，例如，一个数据由向量 X 组成，其有 n 维属性组成 $X=\{x_1,x_2,x_3...x_n\}$，整个数据集由 $C=\{x_1,x_2,x_3...x_n\}$ 组成。每一个数据 X 都已经归属于特定的类属性 $A_1,A_2,A_3,...A_n$ 其中每一类属性 A_i 都是离散值并且无序。由数据 X_i 组成的集合被称为训练集，并且每一份数据都是随机抽样生成。由于数据类型已知，这一步被称为有监督学习，与此对应的是无监督学习，数据集合中的数据类型未知。这一步分类过程可以视为学习一个这样的映射或函数关系 $y=f(x)$，其中，x 代表属性，y 代表类型，从这个角度来看我们希望获取实际的这个映射关系 f .通常，这种映射关系以分类规则、决策树或者数学方程的形式展现。在停车点类型判断中，这个映射关系就是一个可以判断某一个停车点为卸货点还是普通停车点，并且被用于对未知数据进行判断。在第二步中利用第一步生成的分类器对数据进行分类，首先分类器的准确性是需要考虑的问题，如果使用训练集来评估准确性，显然可以得到非常理想的结果,因为分类器倾向于拟合这些数据，因此必须考虑使用测试数据集来评估分类器的准确性，这些测试数据集和训练集之间是相互独立的。分类器的准确性由测试数据集中有多大比例数据被正确分类决定，如果一个分类器的正确性在可以接受的范围，该分类器便可以对未知数据集进行分类，否则得迭代生成新的分类器。因此分类器的质量受到了训练集数量和质量、数据特征空间选取等多方面影响。

贝叶斯分类器是一种统计学分类器，它能预测给定数据属于哪一种类型的概率，贝叶斯理论的基础是贝叶斯公式。在该公式中 X 是数据本身，通常假设其有 n 维属性，将 H 设为 X 属于某一特定类型 C 的假设，在分类问题中，我们想要求出的是 $P(H|X)$,也就是已知 X 的情况下，它属于类型 C 的概率大小，这种概率被称为后验概率，与此对应的是先验概率 $P(H)$,在轨迹数据停车点类型判断中该概率即为所有停车点集合中卸货点的概率。与此类似 $P(H|X)$ 是已知假设 H 的情况下是 X 的概率。综上贝叶斯公式即为

$$P(X|H)=\frac{P(X|H)P(H)}{P(X)}\ [29] \qquad (3\text{-}2)$$

朴素贝叶斯分类[30-31]是一种十分简单的分类算法，在不考虑类型之间的关系时候，它是一种最小错误率分类器。朴素贝叶斯分类器的基本思想是在待分类项已知的情况下，求解其归属每一种类型的后验概率，具有最大后验概率的类型时便可以将待分类项判定为该类，图 3-5 即为朴素贝叶斯分类器流程图。其基本流程：

（1）由训练数据 $X=\{x_1,x_2,x_3...x_n\}$ 组成的训练集，并且 X 的类型已知，x_i 为 X 一个特征属性。

（2）设有 n 种类别的集合 $C=\{C_1,C_2,C_3...C_n\}$。分类器判断具有最大后验概率的 X 归属类型，也就是 X 归属为 C_i 的条件为 $P(C_i|X)>P(C_j|X)$，其中 $1\le j\le m$。

3.4.2 构造停车点训练集

货运车辆停车有多种原因，例如红绿灯、交通拥堵、司机休息交接、停车装运货物、停车卸载货物等一系列人为和非人为因素。其中装货点和卸货点是各家物流公司及司机综合考虑了时间、距离、成本等要素的结果。对于后面的货运线路优化、推荐有重要指导作用。本文所使用的轨迹数据中不包含装卸货信息，因此缺乏足够的装卸货的先验知识和训练集。通过计算停车时间等或者行驶距离等方法来从一系列的停车点中识别出装卸货点，忽略了这些连续停车点之间相互的关联关系。例如，假如有在一条历史轨迹中存在相邻两个停车点。并且停车时间也相近，距离上一个停车点也经历了较长的距离和时间。分类器必须对这两个停车点识别分类的时候，如果通过上述间隔时间和距离的判断方法，则会将第一个停车点识别为装卸货点。

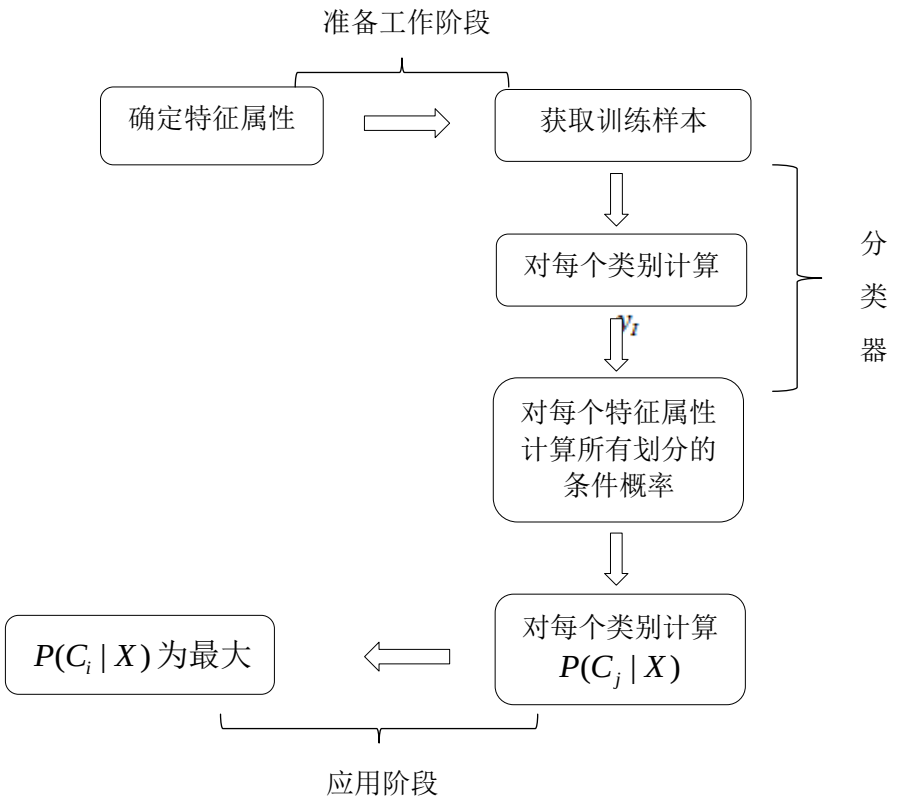

图 3-5 朴素贝叶斯分类器流程图

为了解决上述问题，本文采用了数据挖掘和机器学习的方法来完成了算法设计。具体流程是从历史轨迹中发现货运车辆装卸货的时空特征，获取车辆停车时间长度和停车点相邻距离等先验知识，有了足够的先验知识以后，就可以依据这些先验知识预测车辆下一次停车装卸货时间和地点。先验知识的准确度将会影响分类器的效果，具体到本文的轨迹数据该先验知识可以从物流货运车里程和停车时间观测获得。一般而言，同一类型物流货运车装卸货一般所消耗的时间τ和一次运输所行驶的里程v一般为一个趋向于一个常数，这是因为物流货运车司机驾驶习惯、所行驶的线路相对固定，有的车辆偏向于短途、有的偏向于长途等。

本文首先使用停车点时间间隔和空间间隔的特征，找到确信的一趟货运线路的起点或者终点，依据物流行业的经验，一辆车停车时间在 2 个小时以上基本可以判断为装卸货，由于起点和终点在停车时间角度上基本上是一样的，可以将这两种停车时间看作一种类型。基于此本文提出了简单停车点识别算法。具体算法步骤为依次读取通过上文中停车点识别获取的停车点信息，如果相邻

停车点的距离大于$v_{\max}$，停车时间阈值$\tau_{\max}$,则该停车点为起始点。算法描述如下，式中 GPS 数据点$r_i=\{t_i,lon_i,lat_i,odm_i\}$,$t_i$表示 GPS 发射时刻，$lon_i$,$lat_i$表示此时刻的经纬度，$odm_i$表示此时刻物流车的总的行驶里程。停车点$vs_i=\{lon_i,lat_i,odm_i,t_{si},t_{ii},prevdis\}$,其 $prevdis$ 相邻两次停车点的里程间隔，t_{si}为停车起始时刻，t_{ii}为停车时长。rs_i代表停车点，cs_i代表普通停车点。本算法的本质是将每一个货运车辆的停车点分为起止停车点或者普通停车点。本文提出的简单停车点识别方法如算法 3-1 所示。

停车时间设置足够长的时候上述算法可以正确识别出起止停车点和普通停车点。然而对于相邻距离较近的停车点或停车时间不够长的点却无法区分。对于剩余难以分类的停车点，本文采用了梯度下降法的算法来识别它们在停车时长的区别。经过了简单距离和时间的识别算法后，剩余的起止点构成集合RS_{remaii}，其余的普通停车点构成CS_{remain}。由于每辆车行驶的路径，司机驾驶习惯相对固定，所有最终真实的起止点时间$\tau_{\max}$会收敛到τ，所以当计算RS_{remain}中的每次起止点停车时间与τ的时候,如果ε足够小，那么τ就是该辆车的起止停车点的停车时长。然而τ显然无法从原始轨迹中直接获取，本文采用梯度下降法逐步将ε减小，最终将起止点停车时间收敛至τ。

$$\varepsilon=|RS_{remain}t_{ii}-\tau| \tag{3-3}$$

本文提出算法如下算法 3-2 所示。该算法从函数 $Interval$ 的初值τ_{init}出发，并考虑如下序列τ_1，τ_2，τ_3，$_o\cdot_p.\tau_t$使得$\tau_i=\tau_{i-1}\pm\nabla_\tau$，可以得到$Interval(\tau_1)>Interval(\tau_2)>Interval(\tau_3)\geq...$最终可以顺利收敛到$\tau_{optimum}$。上述算法中$Interval(\{VS\},\tau)$的具体实现如算法 3-3 所示下。

从算法 3-3 易看出当τ_{init}初始值很大时，这样起止点集合RS_i将为空，ε为所有停车点时间长度之和，随着τ_i沿着∇_τ逐渐减小，其ε也会随之减小。然后下降到期望值附近以后可能会'之字型'下降，最终收敛到$\tau_{optimum}$。

综上，本文首先通过简单判断方法依据停车时间长度$\tau_{\max}$和停车点之间的间隔$v_{\max}$初步获取了最可靠的起止停车点集合RS_i，然后使用梯度下降法完成了对起止停车最优时长$\tau_{optimum}$的计算，然后依据该最优值去分析停车点类型，构造训练集。

3.4.3 基于朴素贝叶斯分类器的停车点分类

朴素贝叶斯分类器是基于贝叶斯定理的一类分类器，贝叶斯定理如下

$$P(Y \mid X) = \frac{P(Y)\prod_1^d P(X_i \mid Y)}{P(X)} \quad (3\text{-}4)$$

式 3-4 中 Y 是类别变量， $X = \{x_1, x_2\}$ 是待分类的特征向量。具体到本文中；$X = \{x_1, x_2\}$ 中的 $x_1 = vs_i.prevdis$,即停车点之间的距离， $x_2 = vs_i t_{ii}$ ，即停车时长。$Y = \{y_1, y_2\}$ 有两种类别， $y_1 = 0$ 代表普通停车， $y_2 = 0$ 代表起止点停车。停车点分类问题可以描述为对某一类别 X 求其 $\arg\max(P(Y = y_i \mid X))$ ，然后依据其概率值判断 X 的类别。

Input

$VS = \{vs_1, vs_2, vs_3 ... vs_n\}$,一条轨迹中的停车点集合

$v_{\max}$,相邻停车点的距离阈值

$\tau_{\max}$,停车时间阈值

Output

$RS = \{rs_1, rs_2, rs_3 ... rs_{n1}\}$ 起止点集合

$CS = \{cs_1, cs_2, cs_3 ... cs_{n2}\}$ 普通停车点集合

Pr*ocedure*

$prevdis =, n1 = 0, n2 = 0, RS \leftarrow null, CS \leftarrow null$

for $i = 1; i = n; i++$ *do*

$vs_i.prevdis = vs_i.odm_i - prevdis$;

if $vs_i.prevdis > v_{\max}$ *and* $vs_i t_{ii} > \tau_{\max}$ *then*

$previds = vs_i.odm$

$rs_{n1} = vs_i$

$RS : add(rs_{n1})$

$n1 = n1 + 1$

else

$cs_{n2} = vs_i$

$CS : add(cs_{n2})$

$n2 = n2 + 1$

end if

end for

return $\{VS, RS, CS\}$

算法 3-1 简单停车点识别算法

Input

$VS=\{VS_1,...VS_n\}$多条轨迹中集合,τ_{init},停车时间长度∇_τ， 为迭代步长

Output

$\tau_{optimum}$ 最优起止点停车时长,$CS=\{cs_1,cs_2,cs_3,...,cs_{n2}\}$普通停车点集

Pr*ocedure*

while $\tau_{optimum}\neq\tau_{init}$

$\tau_{cur}\leftarrow\tau_{optimum}$

$\varepsilon_1\leftarrow Interval(\{VS\},\tau_{cur})$

$\varepsilon_2\leftarrow Interval(\{VS\},\tau_{cur}+\nabla_\tau)$

$\varepsilon_3\leftarrow Interval(\{VS\},\tau_{cur}-\nabla_\tau)$

$\tau_{optimun}\leftarrow\arg\min\{\varepsilon_1,\varepsilon_2,\varepsilon_3\}$

end while

return $\tau_{optimum}$

算法 3-2 复杂停车点识别算法

Input

$VS=\{VS_1,VS_2,VS_3...VS_n\}$多条轨迹中集合,$\tau$ 预测起止点停车时间长度

Output

ε 起止点停车时间与预测时间的差值

Pr*ocedure*

$\varepsilon\leftarrow null$

for $i=1;i=n;i++$ *do*

$\{VS_i,RS_i,CS_i\leftarrow NaiveMethod(VS_i,v_{\max},\tau_{\max})\}$

if $RS_i=null$ *then*

$\varepsilon=|VS_i.allt_i-\tau|$

else

$\{RS_{remain},CS_{remain}\leftarrow remove(VS_i,RS_i,CS_i)\}$

$\varepsilon=|RS_{remain}t_{ii}-\tau|$

end if

return ε

算法 3-3 函数 $Interval(\{VS\},\tau)$ 具体实现

在一个给定的训练集中，显然 $P(Y)$ 可以通过观察或者计算得出，然而训练集的构造需要一些简单办法难以区分的停车点，假设有 n_p 个起止停车点，n_c 个普通停车点，n_h 个简单办法难以区分的停车点，然后通过朴素贝叶斯方法识别出来的 n_r 个起止停车点，普通停车点的概率为式 3-5 所示，起止停车点概率如 3-6 所示

$$P(Y=y_1)=\frac{n_c+n_h-n_r}{n_p+n_c+n_h} \tag{3-5}$$

$$P(Y=y_2)=\frac{n_p+n_r}{n_p+n_c+n_h} \tag{3-6}$$

由于 x_1 和 x_2 是连续型变量，可以设其为正态分布，因此

$$P(X_i=x_i\,|\,Y=y_i)=\frac{1}{\sqrt{2\pi\sigma_{ij}}}\exp^{\frac{-(x_i-\mu_{ij})^2}{2\sigma^2_{ij}}} \tag{3-7}$$

其中 μ_{ij} 为 X_i 的均值，σ_{ij} 为训练集中的方差，$X_i\in y_j$。

朴素贝叶斯是一种简单的统计学概率的分类器，它的基础理论是贝叶斯定理，假定类之间属性是相互独立的，分类效果极其稳定，在本文的停车点判断，停车时间长度和车辆行驶距离是相互独立的。由于类条件密度的估计是基于所有的数据集，所以有很强的抗孤立噪声的能力，若干个停车点的噪声错误都最终被整个数据集平均化了。当一条轨迹中的所有起止点都被识别出来以后，以这些停车点位分割点，将轨迹进行分割，则就获取了每辆车所有的运输起始点和终点，为后续的轨迹聚类提供了数据支持。

3.5 本章小结

本章主要介绍了轨迹计算的流程，一般而言其需要经过数据预处理、停车点识别、轨迹分割，数据预处理是之间依据轨迹数据特征，对数据进行异常点检测、去除重复、数据规范化等操作，重点介绍了一种基于网格的轨迹异常点检测算法，该算法通过将轨迹点映射到不同区域中，然后网格内数据点个数和对网格进行分类从而发现异常点。然后通过车辆速度进行停车点识别，得到了一系列可靠的停车点，然后基于该停车点，利用朴素贝叶斯分类器算法获取货运车辆起止卸货停车点，完成了轨迹分割。

第 4 章 海量轨迹数据聚类算法研究

4.1 轨迹聚类的意义和问题

海量轨迹数据中包含了经纬度信息、时间标签、速度、方向、海拔、里程等时空信息，有的还有参考停车启动信息，以及隐藏在背后关于交通道路，司机对于实际地理位置信息等实际状况判断信息,深入挖掘这些信息对于提供LBS服务将会有广阔的应用前景[32]。例如在新浪微博、腾讯微博、微信等一系列 LBS 应用中在百度地图、新浪微博、微信、公交软件等一系列 LBS 应用中，用户可以不断的实时在地图上显示自己的位置和轨迹，朋友之间聚会时开启实时位置共享便可以看到好友的实时位置信息，司机在行驶途中也可以广播实时路况，这些信息都不仅生动展现了人们所处环境信息也帮助人们直观了解了未知区域信息，更为重要的是这些信息是实时的。但是若只是简单的分享位置信息、在地图上展现信息并不能充分的挖掘这些背后隐藏的知识[33]。通过轨迹聚类的方式可以挖掘出轨迹中的频繁模式，可以从海量的轨迹中蕴含的知识中发现有用的知识，这对于物流公司货运线路推荐、车辆调度和管理十分有意义。

目前，海量轨迹聚类研究的问题包括两类。

（1）怎样表示轨迹之间的相异度以及相异度的大小。用通俗的话说，相异度就是两个东西差别有多大，当车辆轨迹展示在地图上的时候，我们可以通过直观感受大体得到轨迹之间的相似程度，但是计算机没有这种直观能力，必须通过一些数学手段将我们人体直观感受到的区别大小定量的表示出来。

（2）依据轨迹数据的特征和轨迹分析的需求，提出面向应用的轨迹特征概念和聚类方法。即在时空轨迹的定义、模型和表达方式的基础之上，针对不同的分析需求采取不同的聚类算法和策略[34]。

综上，可以将轨迹聚类方法划分为轨迹表达、相似性度量和聚类算法三个步骤[35]。本文提出了如图 4-1 所示的轨迹聚类算法流程图。轨迹表达包括轨迹定义和线性插值两部分。车辆轨迹数据是一系列的离散点序列连成的折线[36]，可以看作为一个映射 $f:t\to R^d$,其中将物体在时刻 $t\in R^+$ 的位置映射到一个 d 维的空间，一般是二维或者三维空间。车辆在实际运行中其实际轨迹是连续的，但是采集并存储在计算中的时候是离散化的，并且往往依据 GPS 采样频率的不同对轨迹进行线性插值，扩充轨迹语意表达。通过相似性度量可以确定轨迹之间的时空关系，设分别有两条轨迹 $A=\{a_1,a_2..a_n$ $B=b_1,b_2,b_n..$，可以使用

$d(A,B)=f(A,B)\to R$表示轨迹A和B的距离，也即是相异度，其中R为实数域。聚类与分类不同，分类问题中类型信息是已知的，而在聚类问题中，类型信息是未知的，聚类问题将一个完成的数据集合分为若干组，组内有的数据尽可能的相似，组间的数据尽可能相异度越大。其中相异度由距离来表征。聚类是研究领域内的一个重难点问题，他需要以下若干个条件（1）可扩展性（2）兼容各种数据格式（3）能够容忍噪声。其中每个组叫做一个簇。轨迹聚类依据分析挖掘目的的不同，基于不同的相似性度量，选择不同基于密度、距离、统计等的聚类算法完成轨迹聚类从而发现常用轨迹模式。

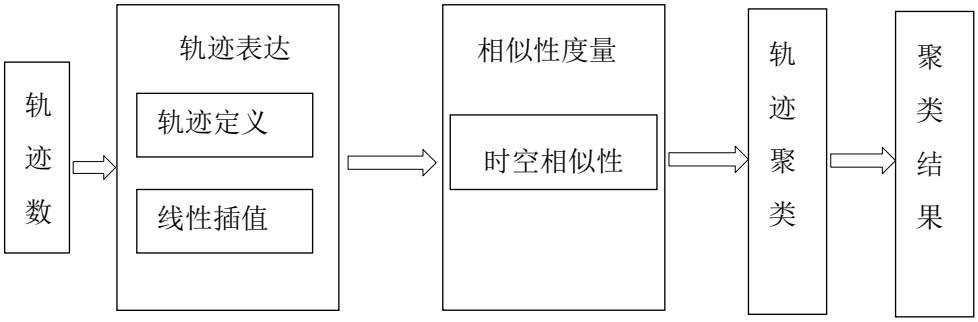

图 4-1 轨迹聚类算法流程图

4.2 轨迹表达和相似性度量

4.2.1 轨迹规则化

通常，在轨迹挖掘方法中，例如聚类算法中需要直接对数据进行加法、减法或矩阵运算等。然后轨迹数据是由一系列离散点组成，不同的轨迹有不同数量的轨迹点数据，因此直接对这些轨迹进行上述运算并不可行，一般而言，在轨迹挖掘方法中若需要对轨迹进行直接运算需要支持如下两种操作。

$$dist(X,Y)=dist(Y,X)=|X-Y| \quad (4\text{-}1)$$

$$mean(X,Y)=mean(Y,X)=|X-Y|/2 \quad (4\text{-}2)$$

其中 X 和 Y 分别代表任意两条轨迹。轨迹规则化可以将任意原始轨迹投影到统一的多维度空间以支持上述两种运算。

在前面的章节中，已经提到过一条包含 n 个 GPS 点的轨迹可以描述为 4-3 所示，式中$n\in Z^{+}$和$x_i,y_i\in[0,360]$

$$R=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\} \quad (4\text{-}3)$$

x_i 表示纬度，y_i 表示纬度轨迹起始点为 (x_1, y_1)，终点为 (x_n, y_n)，若不考虑轨迹方向，可以将轨迹表述如下：

$$y = \begin{cases} a_1 x + b_1 & x \in [x_1, x_2) \\ a_2 x + b_2 & x \in [x_2, x_3) \\ a_3 x + b_3 & x \in [x_{n-1}, x_n) \end{cases} \tag{4-4}$$

其中

$$\begin{cases} a_i = \dfrac{y_{i+1} - y_i}{x_{i+1} - x_i} \\ b_i = y_i - a_i x_i \\ a_i, b_i \in R \\ n \in Z^+ \\ x_i, y_i \in [0, 360] \\ i \in [1, n] \end{cases} \tag{4-5}$$

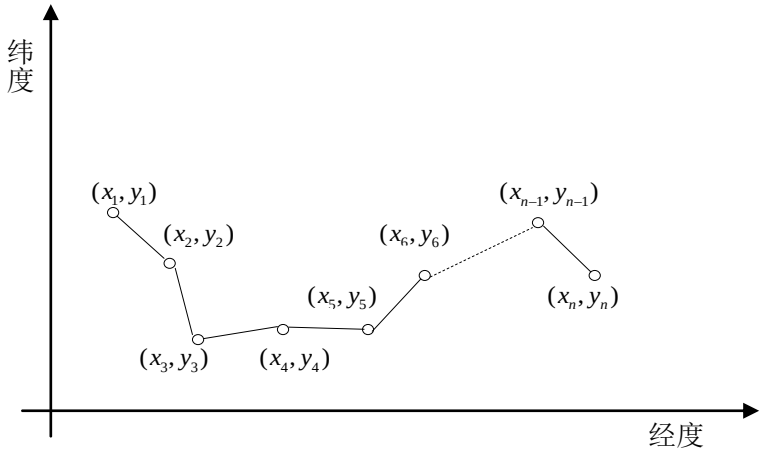

图 4-2 为式 4-4 描述的轨迹图

该轨迹可以通过算法 x 将其投影到一个固定维数的特征空间中去。该算法的定义如下，S：GPS 点所在平面。$S^\alpha : S$ 平面逆时针旋转 α 度。$f^\alpha(x) : S^\alpha$ 平面内的轨迹表达式。算法 x 的过程如下：（1）将平面坐标系统 S 分别旋转 α, β 构造两个平面坐标系统：S^α 和 S^β。（2）将平面 S 内的轨迹分别投射到 S^α 和 S^β 中。（3）在 S^α 中获取连续的 j 个抽样点 $[x_1, x_2, ..., x_n]$，计算 $y_i = f^\alpha(x_i)$；同理在 S^β 中获取连续的 j 个抽样点 $[x'_1, x'_2, ..., x'_n]$，计算得到 $y'_i = f^\alpha(x'_i)$。（4）经过上述步

骤后得到轨迹。

$$R=\{y_1, y_2, \ldots y_j, y'_1, y'_2, \ldots y'_n\}. \quad (4\text{-}6)$$

轨迹 R 可以看作为一个二维空间里面的曲线，属于一个参数集合 $\{\alpha, \beta, j\}$，α和β可以视为在二维空间内从两个角度去观测轨迹，j 为轨迹的分辨率，j 值越大，轨迹的表达越精确。图 4-3 为通过算法 4-1 得到的轨迹图。

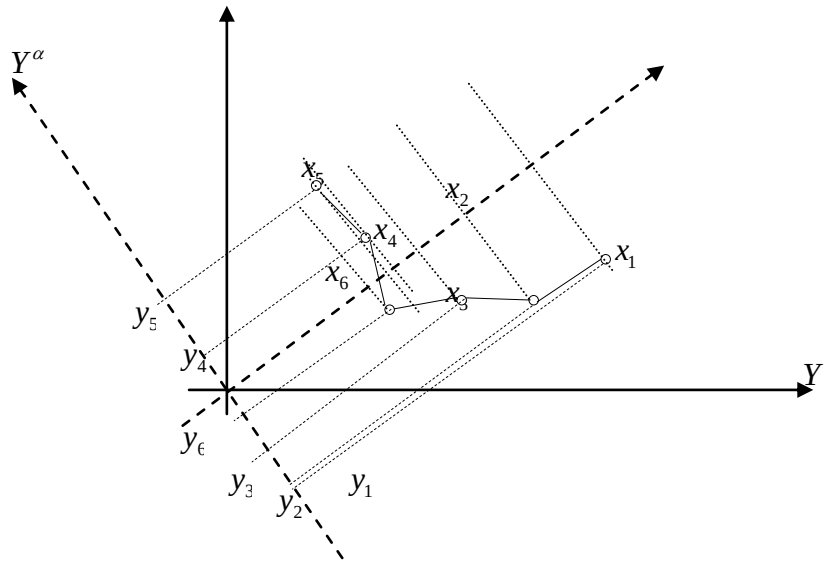

图 4-3 算法 4-1 得到的轨迹图

轨迹规则化预处理算法如算法 4-1 所示。本文提出的算法 4-2 如下所示。

通过规则化后，轨迹 R 依然保留了轨迹的走向趋势特征，这是因在上述的算法第三步过程中将原来处于平面 S 中的轨迹 R 转换为 $[(x_1, y_1), (x_2, y_2), \ldots, (x_j, y_j)]$，其中 $y^i_j = f^\alpha(x^i_j)$，因此轨迹 A 和轨 B 的距离可定义为：

$$dist(A,B)=\sum_{i=1}^{2i} f((x_i, y^A{}_i),(x_i, y^A{}_i))=\sum_{i=1}^{2i} f(g(x_i, y^A{}_i), g(x_i, y^A{}_i)) \quad (4\text{-}7)$$

通过式 4-7，显然可以看出 $g(x_i, x_i)$ 是一个常数，因此在计算轨迹 A 和 B 之间距离的时候可是简化为：

$$dist(A,B)=\sum_{i=1}^{2i} f(g(y^A{}_i, y^B{}_i))=\sum_{i=1}^{2i} d(y^A{}_i, y^B{}_i) \quad (4\text{-}8)$$

其中 $d = f * g$，从式 4-8 中可以看出轨迹 A 和轨迹 B 的区别在于 $y^A{}_i$和$y^B{}_i$，其中 $0 \le i \le 2j$，因此式 4-6 的轨迹保留了其轨迹走向特征。由于轨迹所在平面是二维空间，所以当需要描述二维空间中的一个物体形状的时候需要从两个角度去观测，所以本文将二维平面S分别旋转 α, β 构造两个平面坐标系统 S^α 和 S^β。

此外α,β大小的选取也十分重要，在轨迹规则化过程中，必须使规则化后的轨迹信息中包含全部原始轨迹中形状和趋势信息。如图 4-3 中所示，一条完整的轨迹是一系列单向的 GPS 点组成，而轨迹的走向也反映到这些 GPS 点序列中，因此线性拟合这些折线可以获取得到轨迹的走向趋势信息。假若在原始平面中S中线性拟合的直线与 x 轴之间的夹角为θ，则α,β分别为$\theta+45^{\circ}$和θ+135°。

Input:

D，原始轨迹数据

N，轨迹数据集中轨迹的条数

M,单条轨迹数组

Output：

L，预处理后的轨迹数组

Procedure:

while $i < N$ *do*

while $j < M[i]$ *do*

if $(x_2 - x_1) \neq 0$ *then*

$a \leftarrow (y_2\text{-}y_1)/(x_2\text{-}x_1)$

$b \leftarrow x_1 * a + y_1$

else

$a \leftarrow 0$

$b \leftarrow 0$

end if

add $\{a, b, x_1, x_2\}$ *into* T

$j \leftarrow j+1$

end while

add T into L

$i \leftarrow i+1$

end while

算法 4-1 轨迹规则化预处理算法

通过上述将轨迹 R 规则化以后，所有的轨迹可以看作为被投影到一个维度为$2j$的空间中，因此它们可以像二维平面中的 GPS 点数据一样运算。在现实世

界中，当轨迹没有受到噪声等干扰时，大多数轨迹都是以一种或者多种模式反复出现， 并且这些轨迹之间十分紧密。图 4-4 为通过 PCA 方法将轨迹数据降到

Input:

D，原始轨迹数据，L，轨迹数据集，xR,经度的右边界

N，轨迹数据集中轨迹的条数M,单条轨迹数组，xL,经度的左边界

Output：

R，规则化后的轨迹数组

Procedure：

$N, M \in Z^{+}$

$T \leftarrow null\ list$

$R \leftarrow null\ list$

$interval \leftarrow (xR - xL)$

while $i < N$ *do*

add $xL + interval * i$ into T

end while

$i \leftarrow 0$

$j \leftarrow 0$

$z \leftarrow 0$

while $i < N$ *do*

$currentroute \leftarrow ith\ route\ as\ a\ list\ \ from\ L$

while $z < N_s$ *do*

if $currentsample \le currentroute[j][3]$ *then*

add $(currentroute[0] * currentsample + currentroute[1])$ *into* R

else

$j \leftarrow j + 1$

end if

end while

$z \leftarrow z + 1$

end while

$j \leftarrow j + 1$

end while

算法 4-2 轨迹规则化算法

三维后的 GPS 数据分布图，该图上有 2 个区域的点紧密聚集在一起，这说明有两地之间有两条常行驶路线。

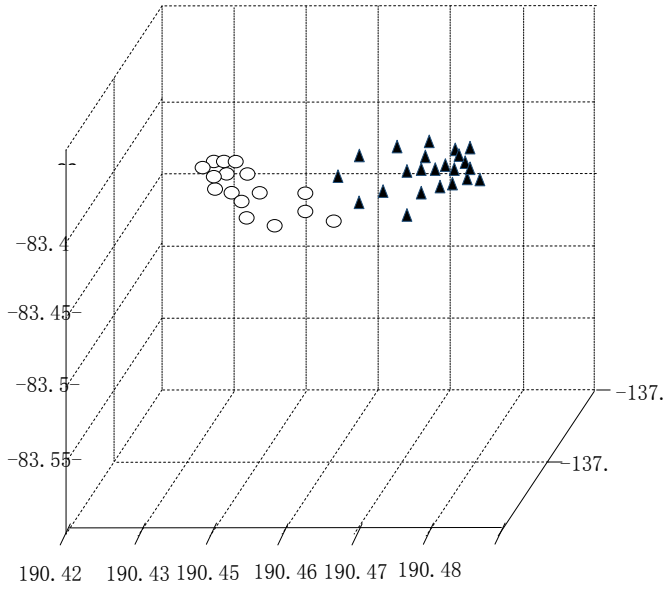

图 4-4 轨迹数据降到三维后的 GPS 数据分布图

4.2.2 轨迹相似性度量

在分类的过程中，同样的输入轨迹样本采取不同的相似性度量准则的时候表征出来的相异度往往也不一样，因此合理采用相异度计算准则非常重要，否则即使本应该归为一类的轨迹也会因为距离计算方法不当而导致巨大差异，直到影响后续的聚类效果。通常，在空间计算中，距离可以依据不同数据集的特征来选择不同，但是不论何种计算方法，该方法必须满足如下三种条件：（1）$dist(a,b)=dist(b,a)$；（2）$dist(a,b)+dist(b,c)\geq dist(a,c)$；（3）$dist(a,b)=0$。本文主要讨论如下三种定义距离的方法。

假设有两条轨迹 X 和 Y 分别通过规则化后得到为$[x_1,x_2,...x_k]$和$[y_1,y_2,...y_k]$，并且定义 y_i-x_i 为轨迹第 i 个点之间的距离。

(1) 欧氏距离

$$dist(X,Y)=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \qquad (4\text{-}9)$$

多维空间中点与点之间的距离常用该方式计算，它表示其真实距离，在轨迹计算中，两个点的欧式距离即为真实地理上的距离。在二维空间中的欧氏距

离就是两点之间的直线段距离。欧式距离看作轨迹的相似程度，通常在大多数情况下，距离越近就越轨迹就越相似,但是也有不适用情况，如图 4-5 所示，如果使用欧式距离，轨迹 A 和轨迹 B 的距离显然更大，切比雪夫距离可以很好的处理这种状况，但是在大多数状况下，欧氏距离依然是一个很好的选择。

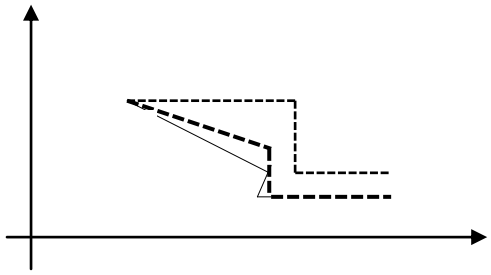

图4-5欧式距离不适用情况示意图

（2）曼哈顿距离

$$dist(X,Y)=\sum_{i=1}^{k}|x_i-y_i| \tag{4-10}$$

曼哈顿距离是一种几何学的欧几里得距离，是一种新的衡量标准两点之间的距离，它采用两点的绝对笛卡尔坐标的差异。在实际应用中，轨迹数据中通常存在噪声，而这些噪声往往属于正态分布，因此曼哈顿距离可以有效的修正轨迹中存在的正态分布的噪声所造成的误差。如图 4-6 为曼哈顿距离不适用情况示意图，显然轨迹 A 和 B 是不同的，但是它们之间的距离并不为零，可见曼哈顿距离并不使用所有情况。

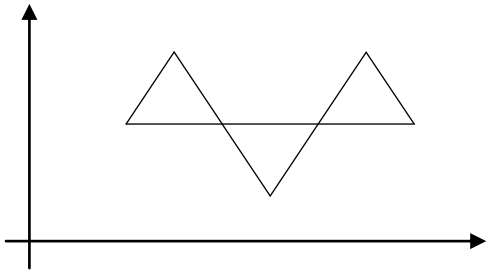

图 4-6 为曼哈顿距离不适用情况示意图

（3）切比雪夫距离

$$dist(X,Y)=\max_i(|x_i-y_i|) \tag{4-11}$$

切比雪夫距离是一种描述向量之间距离的方式，它用两个向量之间坐标差异最大值表示。这中计算距离的方式有很好的抗噪声性能，轨迹是否平行也有很好的距离表示性能。图 4-7 为切比雪夫距离示意图。

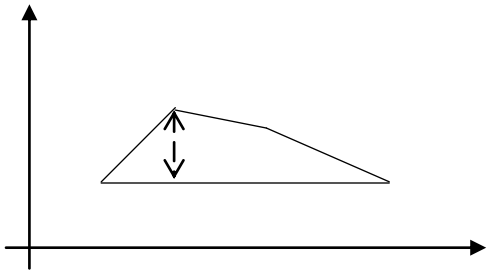

图 4-7 为切比雪夫距离示意图

以上三种距离定义方式，都能在一些场景下适用于轨迹间的距离定义，本文基于轨迹规则化的结果，采用了欧氏距离的定义距离方式，可以有效的消除类似图 4-5 所示的误差状况。

4.3 轨迹聚类

4.3.1 常见聚类算法比较

聚类是将一个集合分为若干个子集合的过程，每个子集都是一个簇，每一个簇内的对象都是相似的，簇间的数据对象并不相似。关于聚类已经有了很多不同的算法。聚类的过程是由聚类算法实现的，而不是人实现的，因此通过聚类算法可以发现未知的簇。在海量轨迹数据挖掘中，它可以发现常出现的轨迹模式。作为统计学的一个分支，聚类分析被广泛的研究和应用，聚类算法本身也必须满足一些要求才能真正达到将数据聚类的效果，这些因素包括对于不同属性数据类型必须有可扩展，可以抗噪声，能发现不同聚类形状的数据簇。有很多种聚类算法，这些方法之间很难有一个明确的界限，因为这些算法互相之间有交叠的部分，通常有如下几种聚类算法[37]。

分割方法：给定一个有 n 个数据对象的集合，将这个集合分割为 k 个部分，其中 k<n 每个部分称为一个簇.也就是将该数据集分为 k 个组，每个组内至少包含一个对象，并且每一个对象只能属于其中一个组，在模糊分类中这并不一定需要满足。大多数分区方法是基于距离的，给定 k 个初始中心，使用不断迭代的技术将一个数据对象放入另一个组中，好的分隔规则得到分类结果中同一个分区类的数据对象距离更近，不同分区类的数据对象更远。有很多种不同规则判断分区效果，通过分隔方式来实现全局最优经常需要不断迭代，大多数应用

采用启发式方法，例如 k-means 算法和 k-medoids 算法，这些贪婪算法对应中小型数据规模可以找到全局最优解，然而在处理海量数据是需要扩展此类贪婪式算法。

层次方法：该方法将给定数据集分解为一个层次，根据层次分隔方式的不同可以将其划分为凝结和分裂，凝结采取自底向上的方式执行，初始时候每一个数据对象构成一个独立的组，然后不断迭代将相邻的数据对象合并，知道所有的数据对象被合并为一个组，也即是最高层次，或者设定一个结束条件。分裂的方式与其相反，它是自顶向下执行，初始时候所有数据对象被归为一类，在每一次迭代过程中，每一个类被分解为更小的类，直到最终每一个数据对象被分为一类或者设定一个终止条件。层次方法可以是基于距离或者密度。层次方法聚类时每一步骤一旦执行，就不能回溯，这种方式可以减少计算量。

密度方法：大多数分类方法是基于数据对象之间距离，这些方法只能发现类球体形状的数据集，很难发现任意形状的数据集，然后基于密度方法的聚类则可以解决此类问题，其主要思想是只要某一个区域内部的数据点数达到了某个阈值就将其归为一类，此算法的优点在于可以剔除噪声点和发现任意形状的聚类。基于密度的方法可以将一组集合分为若干个独立或层次的簇。

网格方法：该方法将对象空间划分为有限个网格，所有的数据对象操作都是基于网格，该方法主要的优点是处理速度快，其时间复杂度不依赖于数据对象个数，而是每个方格内数据的维数，因此十分适合海量时空数据处理，网格方法可以与上述基于密度和层次方法相结合。上述其它聚类方法是数据驱动，即所有的聚类方法都是去契合数据分布特征，然而网格方法数空间驱动，即将对象空间划分为独立的网格。

4.3.2 基于k-均值算法的轨迹聚类

k-均值算法是目前使用最广泛的基于聚类的分割算法。该方法将规模为 n 的对象划分为k个聚类区域，计算每个数据点到所有聚类中心距离，将该点划归为最近的聚类中心，反复迭代这个过程，直到最终类内部尽量紧密，类间距离尽量稀松。对剩余的每个对象根据其与各个簇中心的距离，将它赋给最近的簇，然后重新计算每个簇的平均值。重复这个过程直到准则函数收敛。

$$v=\sum_{i=1}^{k}\sum_{x_j\in S_i}(x_j-\mu_i)^2 \qquad (4\text{-}12)$$

分别计算组内均方差和组间均方差，不断迭代，将所有组内方差总和最小。假设有 k 个群组 $S_i, i=1,2,...,k$。μ_i 是群组 S_i 内所有元素 x_j 的重心，或叫中心点。

传统的k-均值算法存在以下缺点：

（1）k-均值算法的聚类数k需要先验知识预估，但是实际情况是往往在轨迹聚类中并不确定实际聚类数。

（2）由于k-均值算法以选定的初始聚类中心为基础，随机选取聚类中心可能会是聚类结果不稳定，收敛慢等问题。

（3）此外k-均值算法需要不断调整分类情况，直至收敛，数据集数量很大时，若距离计算方式选取不当，将会严重增加算法的时间复杂度。

（4）聚类有效性的判别方式的不同影响聚类效果的评价。合理有效的有效性判别方式可以产生有效的聚类结果集合。

综合以上情况，本文提出相似性度量方式使用欧式距离来表示物流货运车轨迹间的距离，初始化 k 值聚类中心采用基于最大最小距离的方式。

提出该方法的基本依据是：

（1）在一般的物流货运场景中，车辆都是按照一定的路线、速度和方向行驶的，物流货运车尤其是长途货运车通常行驶在高速公路、一级公路、二级公路上。

（2）轨迹内部“高内聚”，即轨迹间距离小，方差也小，轨迹之间“低耦合”，即不同类别之间的轨迹距离大，方差也大。

（3）物流轨迹通过规则化后，可以是相同起止城市的所有 GPS 抽样点个数相同，只用比较轨迹的走向趋势，可以使用时间复杂度和空间复杂度小的距离计算方式。

具体算法流程是:（1）确定聚类数 k。（2）确定初始聚类中心。（3）不断聚类直到收敛。（4）聚类有效性的判别。

聚类数的确定方法有多种[38-39]。例如密度法，它使用的样本的统计性质特征选取凝聚点，最终的聚点数即为分类数，若密度的大小选取不当，会导致分类数的不准确。爬上法，它是一种最优聚类数的判断方式，但实际在轨迹聚类中，并不一定会出现拐点针对上文中提到的物流轨迹特点，即相同起止城市路径方式并不多，可以依据经验确定聚类数 $[k_{\min}, k_{\max}]$，依据经验[40]，通常 $k_{max} \le \sqrt{n}$，n即为聚轨迹数目。这种方法的优点在于方便、快捷和高效。$k_{\min}=1$ 代表轨迹之间无明显差异，所有轨迹被归为一类，一般从聚类数从 2 开始。

初始聚类中心的选取。若初始轨迹聚类中心随机抽取，可能会造成一系列

问题，例如聚类结果不稳定，初始中心过于临近，多个初始聚类中心在同一个类中，导致轨迹聚类时收敛速度慢，提升了聚类的复杂度。本文提出首先统计所有轨迹的总体均值，选取与总体轨迹样本均值最近的轨迹作为初始化中心，这种方式所得到的轨迹聚类结果稳定，且收敛速度优于随机抽取初始聚类中心。不断聚类直到收敛。（1）假设有一个轨迹集合 $RS=\{R_1,R_2,...R_n\}$,首先依据上述提到的方式获取初始化中心 R_i，聚类数 $k=1$，意味着所有轨迹都被归为以 R_i 为聚类中心的类中。（2）聚类数 k 为 2，基于最大距离的方式，遍历除了 R_i 以外整个轨迹集合得到距离它最远的轨迹作为第二个聚类中心 R_{i+1}，然后以最小距离的方式将集合中所有其它轨迹分类。（3）聚类数设为 3，与步骤 2 类似，计算剩余轨迹中距离已有聚类中心 R_i 和 R_{i+1} 中最远的轨迹作为第三个聚类中心，依次以最小距离的方式将剩余轨迹分类。（4）当聚类数 $k \le k_{\max}$， 在已有的 $k-1$ 个聚类中心，计算剩余的轨迹中与这些已有的聚类中心最远的点 R_{i+k} 作为新的聚类中心，重新计算整个轨迹集合得到最终的聚类结果。

聚类有效性的判别[41]。聚类就是将一组对象分为若干个簇，然后使簇内的对象差异性越小，即距离越小，簇间的对象差异性越大，即距离越大。关于聚类有效性指标有很多种，常用的有 Dunn 指标[42]，类与类之间的紧密程度使用类之间距离最近的两个点来表示，类内部的距离最远的两个点的距离表示类间的距离。CH 有效性指标[43]使用类中各点与类中心距离的平方和来表征类内部的紧密度，通过计算各类中心点与数据集中心点距离来表征类间距离。在 Silhouette 指标[44]中，通过计算不同类之间所有数据对象的距离以及类内部所有数据对象相互之间的距离来计算聚类质量。其中，在众多衡量聚类质量中 Silhouette 计算简单效率高被广泛应用，实验也证明，其对于物流货运轨迹也起到了良好的分类效果。

k -均值算法流程图 4-8 如示。

4.4 基于 GPS 数据的线路推荐方法

目前，市面上已经有物流货运推荐线路的应用，大多是直接在地图的数据上，增加了大量与货车相关的数据信息，包括货运市场、货运物流园区、停车场、加油站、工厂和其它位置的信息，同时包括一些针对货运车辆的道路信息，例如货车限行区域，道桥吨位等。但现有的物流货运线路推荐系统主要是通过基于地图的最短路径算法生成的推荐线路，这种方式由于没有考虑实时路况信

息，例如有些高速路在高峰时段是禁止货车通行，或者某个时间段禁止超过一

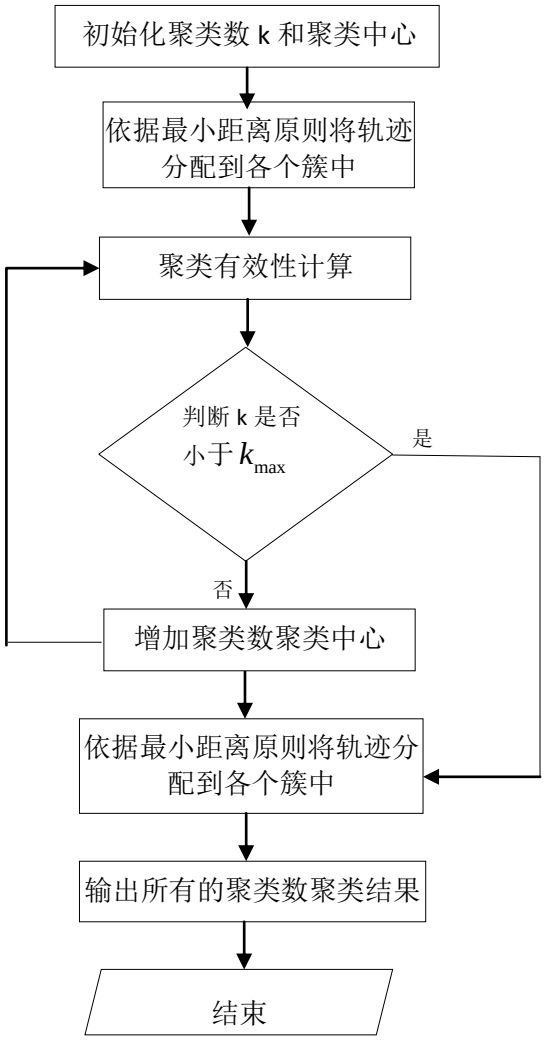

图 4-8 k -均值轨迹聚类流程图

定吨位的货车通行等，而且最短路径算法生成的货运推荐线路往往只有一条不能根据实际需求进行选择，例如货运车辆从广州运送货物到武汉，其可能更高经济效益的方式是先从广州运送货物到南昌，再从南昌运送货物到武汉，这样就提高了货运车辆的利用率，降低了空载率。

本文提出一种基于 GPS 数据的线路推荐算法通过实时发送货运车辆的 GPS 信息，通过货运车辆的实时 GPS 信息进行分析和处理，根据多个最佳推荐线路的不同属性，解决货运线路推荐实时性问题，根据实际情况选择适合司机货运航线，提高物流和汽车货运效率，并便于物流企业进行智能化的调度计划，节

省交通资源能耗，降低城市污染指数。

具体算法流程如下：

（1）GPS 终端通过无线网络技术向服务器发送自己的 GPS 信息。

（2）通过服务器接收 GPS 信息，并提取车辆唯一标识、当前信息发送的时间、当前经纬度、当前行驶里程等特征信息，并根据当前经纬度信息计算车辆所在城市。

（3）根据车辆唯一标识将每辆车的轨迹数据进行分组，并按信息发送时间升序排列，使用轨迹数据异常点检测算法对异常的轨迹数据进行过滤。

（4）将上述的轨迹数据第三章提到的轨迹分割算法将每辆车的轨迹数据记录截断，生成多条路径。

（5）将所有车辆生成的多条路径按照相同起始终止城市进行分组，对每一组的路径进行聚类分析。

（6）将聚类以后的路径分别按照不同属性（时间、距离、费用）计算生成至少两条以上的最优推荐线路。

4.5 本章小结

本章主要细述了轨迹聚类问题和意义，通过轨迹聚类的方式可以挖掘出轨迹中的频繁模式，可以从海量的轨迹中蕴含的知识中发现有用的知识。然后讲述了轨迹聚类的流程，通过轨迹表达、相似性度量、轨迹聚类最终得到聚类结果。首先通过轨迹规则化将相同起止城市的轨迹投影到相同维度空间中去，然后使用欧式距离作为相似性度量，使用基于最大最小距离的 k 均值聚类算法将轨迹聚类，得到最终结果，即为城市之间的常用行驶轨迹。

第 5 章 系统验证和结果分析

5.1 实验基础和条件

所有的设计和实验基于一个 Hadoop 集群。该集群由 10 台天阔 620 服务器组成，依次命名为 Hadoop-0、Hadoop-1、.. Hadoop-9。每台服务器均为 8 核 32G 内存，cpu 型号为 IntelXeon E5-2650 2.0G，操作系统为 Ubuntu10.04。系统使用 Java 和 php 语言编码实现，数据存储通过 Hdfs 文件系统完成，最后实验结果由 Matlab 软件和谷歌地图绘制。

实验用的物流轨迹数据来自于易流公司 2013 年 10 月到 2013 年 11 月的所有车辆 GPS 数据，总共 35G。

5.2 系统实现与验证

以下实现和验证了设计方案中的主要模块，并用 GPS 数据验证得到物流货运车的常用行驶路线，并得出推荐结果。

5.2.1 数据预处理

原始 GPS 数据由 Java 编写的 MapReduce 程序进行数据预处理，将数据时间格式规则化、数据清洗、数据过滤。

（1）原始GPS数据

GPS 数据一般是按文件日期分割，在本应用中，给出的 2013 年 10 月 31 日的 GPS 原始数据部分图，如图 5-1 所示。

（2）GPS 数据清洗

由于 Linux 服务器在接受数据的时候，每当日期变更的时候会把次日的数据放在当天的数据集合中，所有每当处理某一天的数据的时候需要将第二天的数据合并到当天的数据集中。此外由于 GPS 接收时数据误差，可能会造成某些字段为零，例如经纬度、时间等字段。日期时间由于不是标准格式，不利于对后期对 GPS 进行时间计算,将其转换为标准格式。

因此，本文对该数据的清洗操作包括

（1）提取前一天的数据集和当天的数据集进行合并

（2）对原始数据的日期格式进行转换，转换成标准日期格式

（3）去除 ID，时间，经纬度为空值的数据

（4）去重

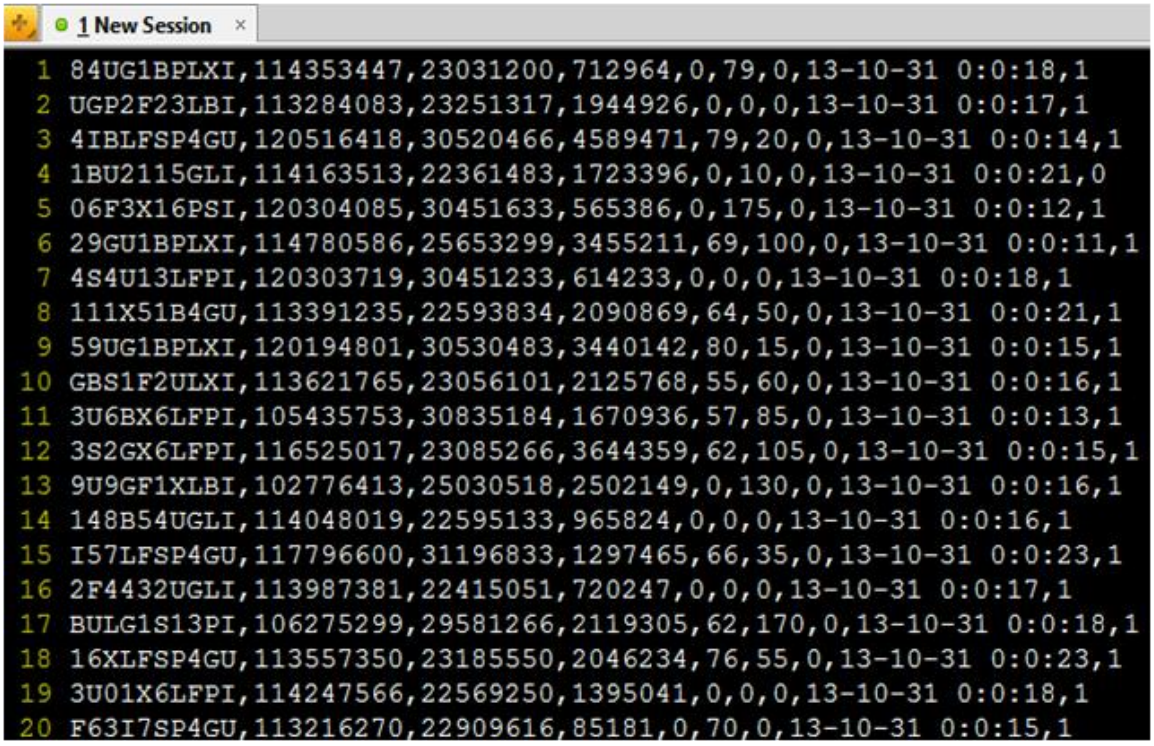

图 5-1 原始 GPS 数据图

经过清洗后的数据如图 5-2 所示。

（3）轨迹数据异常点检测

本文使用上述经过 GPS 数据清洗的轨迹数据进行异常点检测，使用了经度范围在[114.56,114.58],纬度范围在[30.456,30.466]内的一共 2410 个 GPS 数据作为测试数据。真实的地图范围如下图 5-3 所示，经过基于网格的轨迹异常点检测算法得到的轨迹数据点图图 5-4 所示。纬度 0.001 度在地球表面任意地方对应的地球表面距离都是大约 100 米稍多。经度 0.001 度在赤道上对应的地球表面距离约为 $100\cos\alpha$米。所以图 5-4 中一个网格的面积大小为 100*87m^2 其中网格内 GPS 个数阈值设为 C_{thre} 设为 15，距离阈值 D_{thre} 设为 200 米。

5.2.2 停车点识别

本文使用测试数据为 2013-10-23 06:24:31 到 2013-10-24 03:29:50 期间某辆车的 GPS 数据进行停车识别，该点的起始经纬度为 36.252865，120.438614，终点经

1 New Session

```
1 2013-10-31T00:00:00.000Z,02U1G2P1FE,116072983,26094533,1537512,79,108,0,1
2 2013-10-31T00:00:00.000Z,070XFSP4GU,126629585,45777882,634499,0,0,0,1
3 2013-10-31T00:00:00.000Z,0ISF51B4GU,112888266,33221950,660925,0,0,0,1
4 2013-10-31T00:00:00.000Z,0L151FB4GU,114149496,22961600,109939,0,0,0,1
5 2013-10-31T00:00:00.000Z,0S64P1FBUI,106345616,38097233,349329,0,0,0,1
6 2013-10-31T00:00:00.000Z,11451FB4GU,114623616,23605583,69251,74,94,0,1
7 2013-10-31T00:00:00.000Z,11IX4BP4GU,121432200,31325833,53957,0,0,0,1
8 2013-10-31T00:00:00.000Z,1202BI4UFE,115756583,29178116,2457870,64,6,0,1
9 2013-10-31T00:00:00.000Z,15101FB4GU,119010352,25245000,304973,0,30,0,1
10 2013-10-31T00:00:00.000Z,19GUX16PSI,120090034,32193218,995212,74,65,0,1
11 2013-10-31T00:00:00.000Z,1B14LXP4GU,104422050,25738199,615542,99,35,0,1
12 2013-10-31T00:00:00.000Z,1IL1S6P4GU,113900033,22479450,1069159,33,160,0,1
13 2013-10-31T00:00:00.000Z,1S51I2XLGU,112109783,32697083,1428531,0,0,0,1
14 2013-10-31T00:00:00.000Z,1UG6P2XLBI,113840752,22739828,80760,50,5,0,1
15 2013-10-31T00:00:00.000Z,1XFI66P4GU,116271484,40062016,695248,0,0,0,1
16 2013-10-31T00:00:00.000Z,212S11U2FE,112532650,26730066,7237704,79,176,0,1
17 2013-10-31T00:00:00.000Z,224P2IXLGU,87812550,47351200,869294,0,0,0,1
18 2013-10-31T00:00:00.000Z,27B47US3LI,114390251,35903400,1799007,0,0,0,1
19 2013-10-31T00:00:00.000Z,2LU3X16PSI,110676765,39592499,291751,0,0,0,1
20 2013-10-31T00:00:00.000Z,31GUX16PSI,120234848,29958384,938685,66,10,0,1
```

图 5-2 清洗后的 GPS 数据

图 5-3 测试数据真实路网图

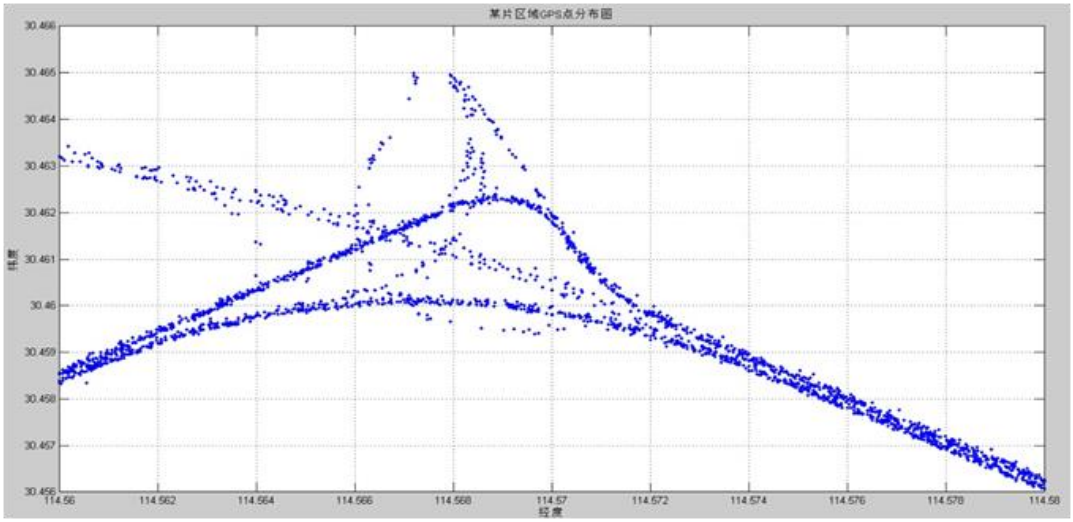

图 5-4 过滤后的 GPS 数据分布图

图 5-6 所示。从中可以看出绝大多数停车时长小于 30 分钟，在白天期间停车较频繁，晚上较少，这也基本上符合物流货运车的行驶规律。

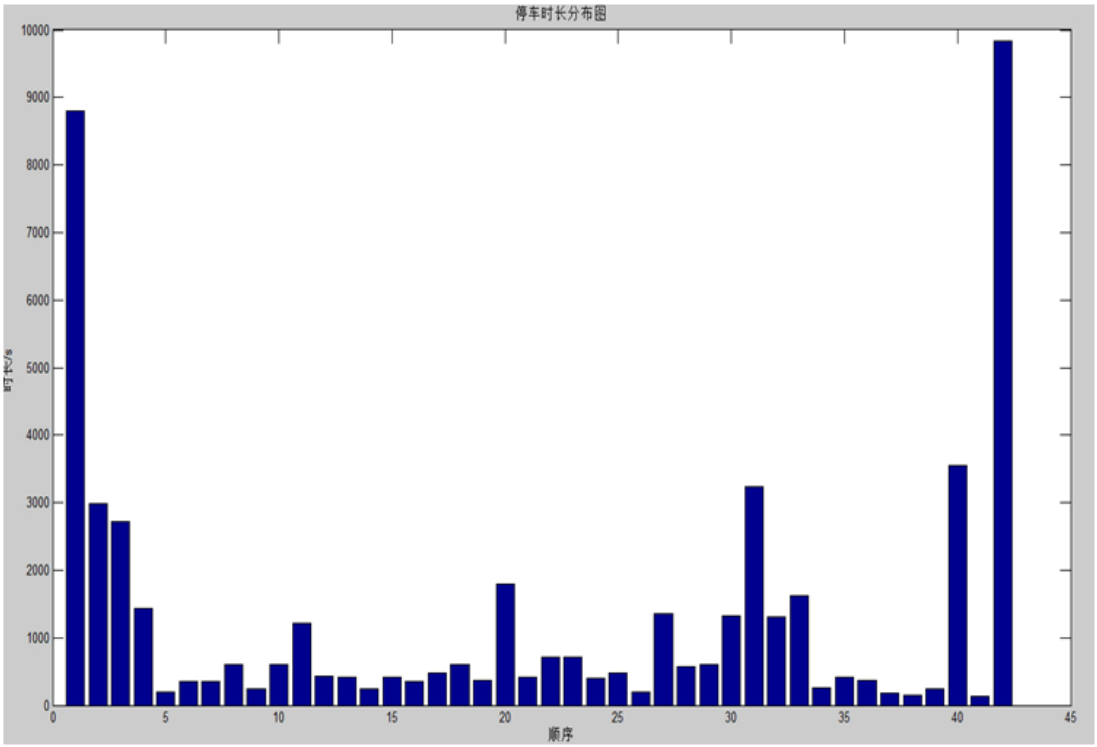

图 5-5 某辆车停车时间分布图

1 New Session

2013-10-23 06:24:31	2013-10-23 06:29:14	2013-10-23 07:33:23
2013-10-23 09:05:03	2013-10-23 09:31:03	2013-10-23 09:41:44
2013-10-23 10:01:21	2013-10-23 10:09:44	2013-10-23 10:25:34
2013-10-23 10:31:34	2013-10-23 10:43:34	2013-10-23 11:04:34
2013-10-23 11:15:44	2013-10-23 11:25:46	2013-10-23 11:33:44
2013-10-23 11:41:44	2013-10-23 11:49:46	2013-10-23 12:01:46
2013-10-23 12:13:43	2013-10-23 12:21:43	2013-10-23 12:52:55
2013-10-23 13:05:43	2013-10-23 13:21:43	2013-10-23 13:40:56
2013-10-23 13:49:43	2013-10-23 13:58:56	2013-10-23 14:09:43
2013-10-23 14:38:33	2013-10-23 14:50:04	2013-10-23 15:02:04
2013-10-23 15:28:04	2013-10-23 16:23:27	2013-10-23 17:42:56
2013-10-23 18:10:46	2013-10-24 00:20:22	2013-10-24 00:29:47
2013-10-24 00:42:14	2013-10-24 01:02:12	2013-10-24 01:05:03
2013-10-24 01:40:40	2013-10-24 03:25:50	2013-10-24 03:29:50

图 5-6 停车时间点

5.2.3 轨迹分割

本文使用测试数据为车牌号码为 2491LUX1BI（编码处理后）在时间范围为[2013-10-22, 2013-10-28]和[2013-12-11, 2013-12-17]期间的 GPS 数据。其停车点分布如图 5-7 所示。本测试数据首先用简单分类的方法，使用停车时间 $\tau_{\max}$ 为 10000s，距离 $v_{\max}$ 为 100km 将最显著的起止停车识别出来。 然后使用梯度下降法获取该车辆最佳起止提车时间 $\tau_{optimum}$ 为 7200s。图 5-8 为所划分得到的轨迹分割结果。

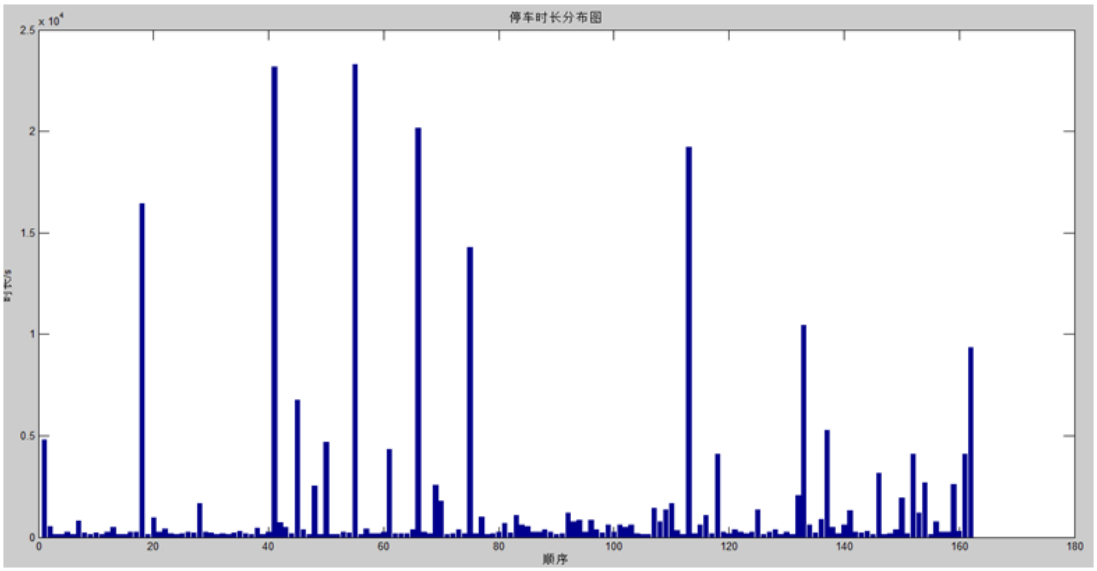

图 5-7 某辆车停车时间分布图

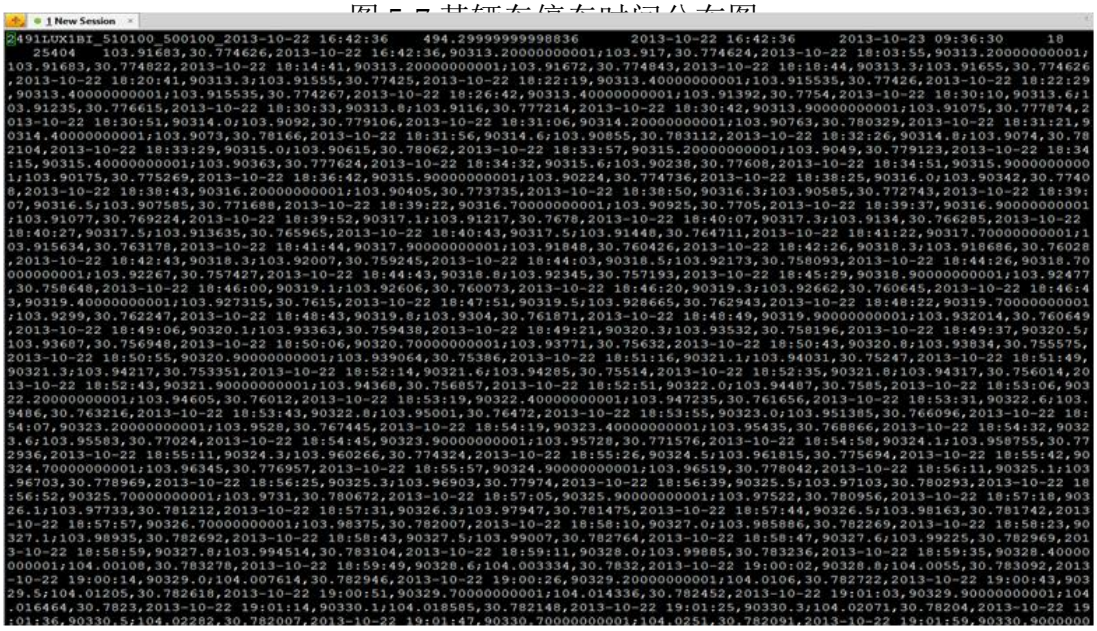

图 5-8 轨迹分割结果图

表 5-1 为轨迹分割结果数据格式。

表 5-1 轨迹分割结果数据格式

车id_起始城市id_目的城市id_本趟开始时间	2491LUX1BI_510100_500100_2013-10-22 16:42:36
里程（km）	494.299
本趟开始时间	2013-10-22 16:42:36
本趟结束时间	2013-10-23 09:36:30
中途停车次数	18
总停车时长（s）	25404
轨迹（';'分隔不同点信息）组成：经度,纬度,该点时间,该时刻里程;	103.91683,30.774626,2013-10-22 16:42:36,90313.20000000001;103.917,30.774624,2013-10-22 18:03:55,90313.20000000001……
最后一次停车的时长s	16441
所有停车点（格式同轨迹）	107.71205,29.353756,2013-10-23 04:46:05,262……

5. 2. 4轨迹聚类

本文使用基于最大最小距离的k-均值聚类算法。其结果如图 5-9 广州到深圳轨迹聚类结果。

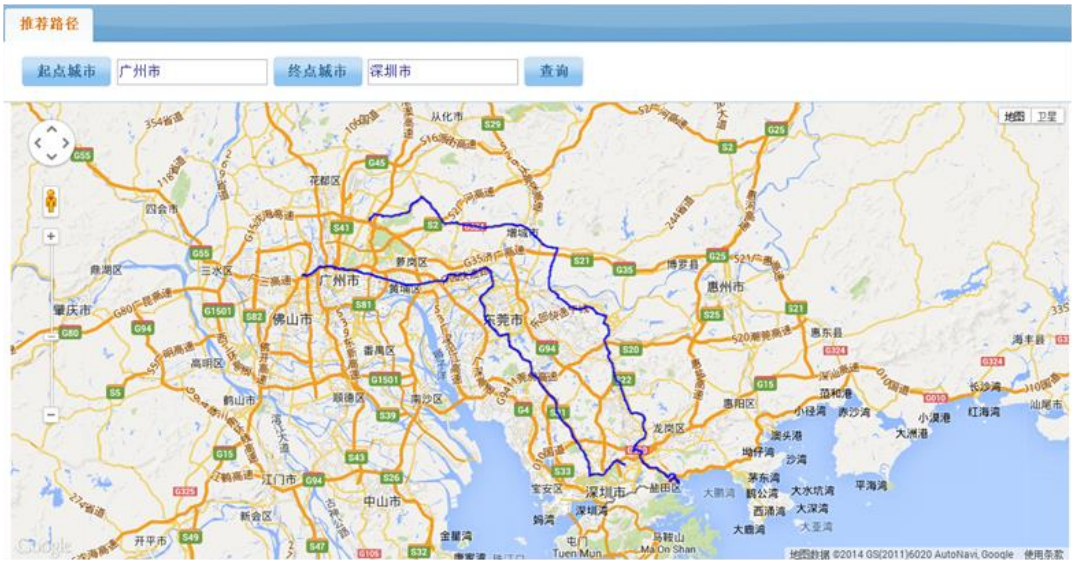

图 5-9 广州到深圳轨迹聚类结果

第 6 章 工作总结和展望

6.1 本文工作总结

本文以海量 GPS 数据作为数据源，利用海量轨迹数据挖掘和道路推荐相关理论，通过建立聚类模型和分析海量 GPS 数据来了解物流车辆行驶规律，提出使用基于海量历史线路的方法构建一个物流货运线路推荐系统，该系统最大的优势是源于真实数据，而并非简单的一些最短路径算法生成。其中重点就是数据预处理方法，停车点侦测和路径分割方法，相似货运轨迹聚类和货运线路推荐。本文重要工作如下所示：

（1）作为轨迹数据挖掘的必要工作，研究了预处理方法，包括数据清洗，数据中的异常进行侦查和排除，并针对本系统所有的 GPS 数据进行了特征分析和提出了一种基于历史轨迹数据的异常点过滤算法。

（2）首先通过停车点侦测可以将普通停车点和上下货停车点区分开来，然后通过路径分割方法可以找到物流车辆每一趟运送货物线路起止点。本文依据朴素贝叶斯算法提出一种新的基于历史数据的路径分割算法，根据物流车辆在上下货时候停车和普通停车在时空属性上的不同，将轨迹进行分割。

（3）相似货运轨迹聚类将相同起始点和终点的轨迹归一化后投射到同一纬度，然后采用 K 均值算法分析轨迹特征，最终得到物流车辆一般行驶轨迹。

（4）货运线路推荐方面，设计了基于历史轨迹数据在时间，距离以及成本的不同，得出相应的推荐线路指导物流司机采用合理的行驶方案。

（5）本文使用真实物流车辆 GPS 数据，利用上述的停车点检测方法、轨迹分割方法、路径聚类算法进行了验证和分析，得到的结果和实际物流车辆符合。

6.2 下一步工作展望

本文提出的关于轨迹挖掘的方法和算法，创新性应用在了物流货运车轨迹挖掘中，得到了良好的结果，但还有很多工作需要处理。

（1）数据预处理方面，基于网格距离的异常检测方法只能从所有车辆的所有轨迹中去检测异常点，没能够针对某辆车的单条轨迹去发现异常点，影响了后期工作的精确性。

（2）轨迹分割方面，没有考虑车辆由于司机、性能变化等原因造成了驾驶

距离和停车时间变化的问题，即只考虑到停车时间单个波峰问题，而没有考虑停车时间多波峰的影响。

（3）聚类算法方面只是沿用前人研究成果算法，没有使用到 GPS 数据中的速度、方向等特征来提供更精确的结果，速度和 CPU 性能也没有充分的利用。

（4）目前的轨迹数据挖掘中关于语义挖掘越来越受到人们的重视，轨迹中附加的语义信息可以帮助找到更加丰富的信息，研究基于语义的轨迹数据挖掘来实现更好的线路推荐。

致 谢

随着时间的流逝，一晃三年就过去了，三年前带着对未来的憧憬，我来到了武汉理工大学，开始了我的研究生的生涯，在这里我认识了许多有着同样理想的同学和朋友，也得到了周云耀教授的悉心指导，感受到了严谨的学风、认真的治学态度，使我从变为了一个实践者，踏实努力做好每一件事情，一步一个脚印。三年时间不算长，但是给我带来方方面面的影响确实十分深远的。

首先，我最想感谢周云耀教授，从研一进入武汉理工大学开始，我就感受到了周教授对学生从学业到生活方面的关怀，刚来的时候就组织我们这批刚进入研究生阶段的新生听师兄师姐们分享实习和工作的经验，让我们一开始就有了努力的方向，为我们每一个人制定一个合理可行的方向。生活方面为我们组织活动，让原本并不熟悉的我们很快的融成了一片，不分彼此。从论文的开题到写作，周教授均悉心指导，逐字逐句的对我们的论文进行推敲，在我迷茫的时候给我指明了方向，请允许我再次表示深深的感谢。

其次，我要感谢中国科学院深圳先进技术研究所数字所云计算中心的研究员张帆老师和邹玉斌、李烨、白雪、赵娟娟工程师，在先进院实习期间，他们为我们这批实习生提供了非常好的实验条件和先进的研究方向，提供了让我们深入了解国内外云计算方面最新的成果和疑难问题，并为我们指明了努力的方向，张帆老师在我开题期间为我答疑解惑，提出论文的创新点和难点，不仅在做学问方面，做人方面也令我十分钦佩，在我的研究生生涯中能遇到在工作和为人处事方面如此优秀的老师，我深感我的幸运，也预祝张帆老师早日实现自己的理想。

接着，我要感谢实验室和信研 A1101 班的同学们，在三年研究生生涯中让我感受到了关怀，积极向上的生活态度，以及严谨的学习态度，在此，我要感谢我亲爱的同学们，白小龙、朱晨华、芦祎、王波涛、王国杰、王小艳、徐光。是他们鼓舞我一步一个脚印踏踏实实努力学习和生活。特别感谢王小艳同学，是她让我在最困难的时候有机会进入到先进院实习的机会，奠定了扎实的基础。

然后，我还要感谢百度 LBS 地图基础业务部的尹颂扬、李亮、张津洁高工对我在北京实习期间工作和生活上的照顾，以及和刘振坤、徐凯伦合作时对我的帮助，另外对于已经出国的薛羽凡、陶志鹏读研我要感谢你们和祝福你们美国能够生活学习上更上一层楼。特别感谢指导人尹颂扬，是他在我工作中遇到

困难，无法解决的时候陪我到深夜查 bug，排除线上问题。为我在工作方面积累的很多经验。

最后，我想感谢我的父母，是他们在我在十九年的求学生涯中，一直默默的支持我，为我创造了良好的生活环境可以让我全身心的投入到学习和工作中去。真挚的希望母校的发展越来越好、培育更多的人才；也祝福我的老师、家人和朋友工作生蒸蒸日上、身体健康；祝愿同学们在毕业后都能找到合适的工作，理想的伴侣，工作爱情双丰收。

甘 波

2014 年 3 月

参考文献

[1] Ferraro R, Aktihanoglu M. Location-Aware Applications[M]. Manning Publications Co., 2011.

[2] 龚玺，裴韬，孙嘉，等. 时空轨迹聚类方法研究进展. 地理科学进展[J]. 2011，30(5): 522-534.

[3] 李德仁，王树良，李德毅. 空间数据挖掘理论与应用[M]. 科学出版社, 2006.

[4] Knorr EM, Ng RT, Tucakov V. Distance-Based outliers: Algorithms and applications. VLDB Journal, 2000,8(3):237−253.

[5] Li X, Han J, Kim S, et al. ROAM: Rule-and Motif-Based Anomaly Detection in Massive Moving Object Data Sets[C]//SDM. 2007, 7: 273-284.

[6] 陈锦阳，刘良旭，宋加涛，等. 基于 R-tree 的高效异常轨迹检测算法[J]. 计算机应用与软件, 2011, 28(10): 34-37.

[7] Liu Z, Pi D, Jiang J. Density-based trajectory outlier detection algorithm[J]. Journal of Systems Engineering and Electronics, 2013, 24(2): 335-340.

[8]Han J, Kamber M, Pei J. Data mining: concepts and techniques[M]. Morgan kaufmann, 2006.

[9] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets[C]//ACM SIGMOD Record. ACM, 2000, 29(2): 427-438.

[10] 雷小锋，高韬，谢昆青，等. 扩展空间对象聚类问题的研究[J]. 计算机工程与应用，2003, 39(23): 172-175.

[11] Müller M. Dynamic time warping[J]. Information Retrieval for Music and Motion, 2007: 69-84.

[12] Gorbenko A, Popov V. On the Longest Common Subsequence Problem[J]. Applied Mathematical Sciences, 2012, 6(116): 5781-5787.

[13] Lee J G, Han J, Whang K Y. Trajectory clustering: a partition-and-group framework[C]//Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, 2007: 593-604

[14] Gao Y, Zheng B, Chen G, et al. Algorithms for constrained k-nearest neighbor queries over moving object trajectories[J]. Geoinformatica, 2010, 14(2): 241-276.

[15] Lin B, Su J. One way distance: For shape based similarity search of moving object trajectories[J]. Geoinformatica, 2008, 12(2): 117-142.

[16] Yan Z, Spaccapietra S. Towards Semantic Trajectory Data Analysis: A Conceptual and Computational Approach[C]//VLDB PhD Workshop. 2009.

[17]谢远飞, 刘洋, 李海军. 空间数据挖掘方法综述[J]. 全球定位系统, 2010, 35(005): 65-68.

[18] Han J, Kamber M, Pei J. Data mining: concepts and techniques[M]. Morgan kaufmann, 2006.

[19] 姜金凤. 移动对象轨道异常检测算法的研究[D]. 南京航空航天大学, 2010.

[20] 张治华. 基于 GPS 轨迹的出行信息提取研究 [D]. 华东师范大学, 2010.

[21] Russell S. Artificial intelligence: A modern approach, 2/E[M]. Pearson Education India, 2003.

[22] Kantardzic M. Data mining: concepts, models, methods, and algorithms[M]. John Wiley & Sons, 2011.

[23] Yan Z, Spaccapietra S. Towards Semantic Trajectory Data Analysis: A Conceptual and Computational Approach[C]//VLDB PhD Workshop. 2009.

[24] White T. Hadoop: the definitive guide[M]. O'Reilly, 2012.

[25] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis[M]. Wiley. com, 2009.

[26] Barnett V, Lewis T. Outliers in statistical data[M]. New York: Wiley, 1994.

[27] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure[J]. ACM SIGMOD Record, 1999, 28(2): 49-60.

[28] 仇培元. 城市出行者轨迹数据时空挖掘方法研究[D]. 北京建筑工程学院, 2012.

[29] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine learning, 1997, 29(2-3): 131-163.

[30] Leung K M. Naive Bayesian Classifier[J]. POLYTECHNIC UNIVERSITY Department of Computer Science/Finance and Risk Engineering, 2007.

[31] Zhang H. The optimality of naive Bayes[J]. A A, 2004, 1(2): 3.

[32] Giannotti G. Mobility, data mining and privacy: Geographic knowledge discovery[M]. Springer, 2008.

[33] 郑宇, 谢幸. 基于用户轨迹挖掘的智能位置服务[J]. 中国计算机学会通讯, 2010, 6(6): 23-30.

[34] 张旭. 基于时空约束的轨迹聚类方法研究与应用[D]. 重庆邮电大学, 2010.

[35] Kharrat A, Popa I S, Zeitouni K, et al. Clustering algorithm for network constraint trajectories[M]//Headway in Spatial Data Handling. Springer Berlin Heidelberg, 2008: 631-647.

[36] Li X, Han J, Kim S, et al. ROAM: Rule-and Motif-Based Anomaly Detection in Massive Moving Object Data Sets[C]//SDM. 2007, 7: 273-284.

[37] 冯晓蒲, 张铁峰. 四种聚类方法之比较[J]. 微型机与应用, 2010, 29(16): 1-3.

[38] 张逸清，刘文才．聚类数的确定[J]．计算机与数字工程, 2007, 35(2): 42r44.

[39] 刘丹，高世臣. K-均值算法聚类数的确定[J]．硅谷, 2011 (6): 38-39.

[40] 于剑，程乾生．模糊聚类方法中的最佳聚类数的搜索范围[J]．中国科学：E 辑，2002, 32(2): 274-280.

[41] 刘燕驰，高学东，国宏伟，等．聚类有效性的组合评价方法[J]．计算机工程与应用，2011, 47(19): 15-17.

[42] DunnJ C. Well-separated clusters and optimal fuzzy partitions[J]. Journal of cybernetics, 1974, 4(1): 95-104.

[43] Caliński T, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics-theory and Methods, 1974, 3(1): 1-27.

[44] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational and applied mathematics, 1987, 20: 53-65.

作者在攻读硕士学位期间发表的专利

[1]张帆,甘波,白雪,赵娟娟,李晔,邹瑜斌,须成忠. 一种基于北斗/GPS数据的线路推荐系统及方法[P]. 广东：CN103278833A,2013-09-04.

[2]白雪,张帆,甘波,赵娟娟,须成忠. 物流公司运力分析系统及其运力分析的方法[P]. 广东：CN103295120A,2013-09-11.

基于海量物流轨迹数据的分析挖掘系统

作者： 甘波
学位授予单位： 武汉理工大学