运输车辆安全驾驶行为的分析

摘要

车联网是物联网技术在交通领域的典型应用，随着市场成熟的提高，其内涵也被进一步拓展。本文通过对某运输企业车辆自动采取的当前驾驶行为下的车辆状态信息以及环境信息进行处理并建立数学模型进行分析，对车辆形势过程中的驾驶行为对行车安全、运输效率与节能情况进行了评估。

对问题一，对数据进行预处理，对相同经纬及里程数的数据进行去重，将日期变量转换成 Unix 时间，将经纬度变量投影成平面坐标，使用卡尔曼滤波对轨迹数据进行纠偏去除噪声，然后使用时空密度聚类对轨迹进行分段，并去除冗余值，最后通过调用谷歌地图 API 刻画了路线图。

对问题二，首先在问题 1 的基础上对数据进行进一步的预处理，对 450 辆车的数据进行清洗，删除无效的表格。然后通过查阅国家相关法律法规，设立公式，对不良驾驶行为进行加权求和，并通过调参给出最终的安全等级，将驾驶情况分为四个等级。

对问题三，使用神经网络三层全连接模型在问题二的基础上加入天气特征，重新建立模型，对行车安全情况进行预测和验证，取得了较好的结果，准确率在 95%以上，能较为准确地预测车辆在该驾驶行为下在实时的气象环境中的安全评分。

关键词：驾驶行为，轨迹数据挖掘，时空密度聚类，神经网络

Abstract

The Internet of Vehicles is a typical application of the Internet of Things technology in the field of transportation. As the market matures, its connotation is further expanded. This paper analyzes the vehicle status information and environmental information under the current driving behavior automatically taken by a transportation enterprise vehicle and establishes a mathematical model to analyze the driving behavior, transportation efficiency and energy saving situation in the vehicle situation. .

For question one, preprocessing the data, de-duplicate the same latitude and longitude and mileage data, convert the date variable to Unix time, project the latitude and longitude variable into plane coordinates, use Kalman filter to correct the speed, and then use time and space. Density clustering segments the trajectory and removes redundant values.

For question two, by consulting the relevant national laws and regulations, formulating a formula, weighting and summing the bad driving behavior, and giving the final safety level by adjusting the parameters.

For question three, a neural network three-layer full-connection model is used to re-establish the model by adding weather features to the second problem.

Keywords:driving behavior, trajectory data mining,density clustering, neural network

目录

1. 挖掘目标

1.1 挖掘背景

随着中国经济实力的不断提高，国民对生活质量的要求也越来越高，汽车在国民生活中扮演着不可或缺的角色。据公安部统计，截止 2018 年 12 月份，全国机动车保有量达 3.25 亿辆，与 2017 年底相比增加 1556 万辆；机动车驾驶人达 4.07 亿人，与 2017 年底相比增加 2236 万人。同时私家车也增速加快。以个人名义登记的小型和微型载客汽车达 1.87 亿辆，每百户家庭私家车拥有量已超过 40 辆。从城市情况看，全国有 61 个城市的汽车保有量超过百万辆，26 个城市超 200 万辆，8 个城市超 300 万辆。

车辆数量的骤增引发了一系列社会问题，包括交通事故频发、城市交通堵塞和停车困难等现象，车辆的安全驾驶、交通管理和车辆间的信息数据交换等日益引起人们的关注。其中，道路交通安全直接关系这人们的生命财产安全，是重中之重。在车辆驾驶的过程中，影响驾驶安全的三个主要因素分别是：机动车辆，驾驶环境和驾驶员。数据表明，超过 85%的道路交通事故是由驾驶员操作失误而造成的。深入了解发现，这部分交通事故原因中出现概率较高的主要是酒后驾驶、超速行驶、疲劳驾驶，以及急加速、急减速、急刹车和急转弯等不良操作，这些驾驶员的危险驾驶行为成为了事故的主要原因。多数驾驶员仅以交通法规为标准，不重视自己的危险驾驶行为，这对道路交通安全造成了隐患。

1.2 挖掘目标

车辆行驶数据，主要涵盖车主的出行行为及驾驶行为两方面。出行行为，包括出行时间、路线、里程等信息的记录，通过这些信息，能够很容易的分析出车主的出行目的及平时的出行习惯。而车主的驾驶行为，以车辆为研究对象来看，主要体现在车辆行驶状态信息和对车辆的驾驶操作信息，如驾驶中的均速、加油频率、转向次数、碰撞等数据，以及车主的四急操控等驾驶行为，这些数据能比较客观地反映出车主的日常驾驶习惯。

本文的挖掘目标主要包含以下三部分：

一、根据已知数据，提取并分析车辆的运输路线以及其在运输过程中的速度、加速度等行车状态。

二、挖掘每辆运输车辆的不良驾驶行为，建立行车安全的评价模型，并给出评价结果。

三、综合考虑运输车辆的安全、效率和节能，并结合自然气象条件与道路状况等情况，为运输车辆管理部门建立行车安全的综合评价指标体系与综合评价模型。

1.3 研究现状

为了帮助驾驶员安全出行，减少道路交通事故的发生，国内外一些学者对驾驶过程中的驾驶行为进行了大量的研究和分析。研究发现，两种方式能够比较有效地降低使驾驶者在驾驶过程中危险性驾驶行为的出现概率：一是让驾驶员在危险驾驶行为出现时得到及时提醒；二是记录下行车过程中的不良驾驶行为并将数据反馈给驾驶员。

前者的有效方案是通过提供车载驾驶辅助系统，对驾驶员状态、驾驶行为以及道路环境等安全驾驶因素进行实时监控，感知车辆行驶状态、获取驾驶员操作行为数据，对其具有危险性的驾驶行为进行识别，并及时向驾驶员发出预警。与车载辅助驾驶系统的研究和开发相比，通过行车数据采集，车辆数据处理和驾驶行为分析建立的安全驾驶评价模型在现阶段有着更广泛的应用。这种驾驶行为的评价和反馈的机制在可用性、系统复杂度、性能要求以及研发成本等多方面有着明显优势。

目前，为个人车主提供车辆和安全驾驶信息服务的车联网产品主要有两种模式，一种是基于智能手机的行车助驾应用，利用智能手机中的传感器和 GPS 收集汽车的里程、速度、加速等信息，对车主的出行和驾驶行为进行分析，其中比较有代表性的产品是小牛助驾。另一类是以汽车 OBD 盒子为车载感知终端的信息服务产品，相比智能手机，通过车辆 OBD 直接获取数据具有更高的可靠性，可实现对车辆油耗、车辆性能、行车记录和驾驶行为等多方面的实时监控，除了辅助安全出行外，还能提供包括驾驶信息、安全信息、商业信息以及管理信息的综合信息服务。

2. 分析过程与方法

2.2 问题一：行车状况可视化

2.2.1 处理流程

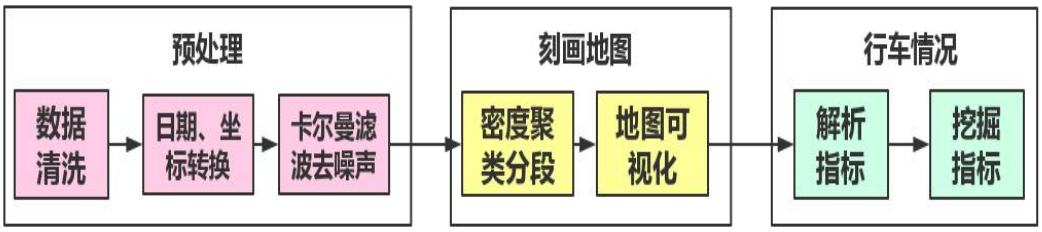

图 1 问题一流程图

2.2.2 数据预处理

1）对相同经纬度和里程数的数据进行去重

对于附表中的十辆车的行车数据中，首先使用 python pandas 函数对数据进行缺失值的检查，发现转向灯列全为空值，对其进行删除，然后使用 duplicated() 函数检查重复值，考虑到刻画地图所需要的信息，将参数变量设为 lng,lat,mileage，发现每辆车的行车数据中都出现了很多的重复轨迹数据，即某一时间驾驶车辆的位置与行车里程数相同，在刻画地图时增加了额外的数据开销，因此对重复值进行删除。

2）将日期变量转为 Unix 时间

如下表，比赛所给日期变量的格式为格式化字符串时间，为建模方便，定义了 datatime2unixtime()函数将所有日期变量转化成了整数形式的 Unix 时间，即 1970 年 1 月 1 日（UTC/GMT 的午夜）开始所经过的秒数，不考虑闰秒。

表 1 日期格式转换表

车辆	原始时间	转换后时间
AA00001	2018/8/4 1:22:02	1533316922
AA00001	2018/8/4 1:23:53	1533317033
AA00001	2018/8/4 1:23:54	1533317034
AA00001	2018/8/4 1:23:55	1533317035
AA00001	2018/8/4 1:23:56	1533317036
…	…	…

3）将经纬度坐标投影为平面坐标

常用的地球经纬度与平面坐标的转换方法有米勒投影、墨卡托投影、横轴墨卡托投影（也叫 UTM 投影，百度地图 api）、高斯-克吕格投影、Lambert 等角正割圆锥投影等，由于题目需求是地球经纬度坐标转为平面笛卡尔坐标，所以这里选和墨卡托投影方式类似的米勒投影。假设有一个和赤道垂直的圆柱套在地球上，然后在地心点亮一盏灯，灯光将地球各个点投影在圆柱上，在把圆柱展开，得到地球的平面投影，示意图如下：

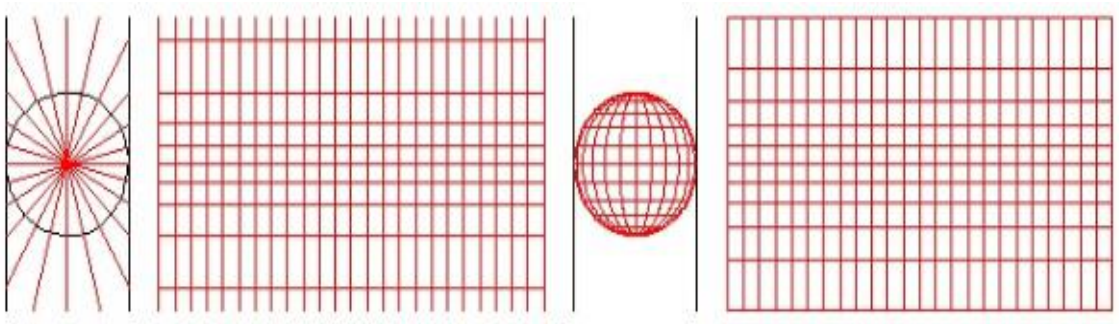

图 2 投影方法示意图

使用这种方式得到的投影地图在两极会拉长，如图所示：

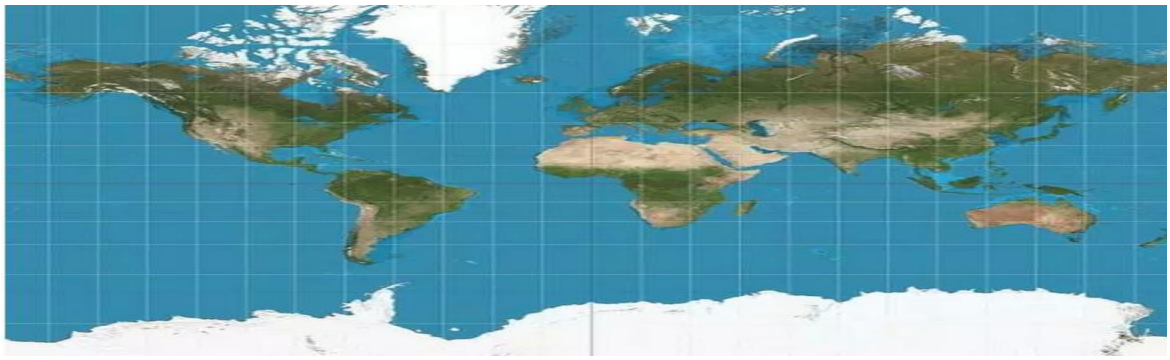

图 3 投影效果示意图

坐标转化结果如下：

表 2 平面坐标转换图

车辆	原 lng	原 lat	转化后:lng_x	转化后:lat_y
AA00001	115.944523	28.651165	12906885.260 6289	3331321.6335 1903
AA00001	115.944523	28.651151	12906885.260 6289	3331319.8575 9331
AA00001	115.944571	28.65114	12906890.603 9645	3331318.4622 2327
AA00001	115.944603	28.651126	12906894.166 1882	3331316.6862 9798
AA00001	115.944643	28.651118	12906898.618 9678	3331315.6714 8363
...	...	...	...	...

4）使用卡尔曼滤波对速度进行纠偏

由于传感器噪声和其他因素，如在城市峡谷中收到较差的定位信号，空间轨迹永远不会完全准确。 有时，错误是可接受的（例如，车辆的几个 GPS 点落在实际驾驶车辆的道路之外）在其他情况下，如下图所示，像这样的噪声点的误差

太大（例如距离其真实位置几百米），以得出诸如行进速度等有用的信息。因此，在开始刻画路线点图之前，我们需要从轨迹中滤除这些噪点。本小组主要采用了卡尔曼滤波对坐标和速度值进行纠偏。粒子滤波的初始化步骤是从初始分布生成 P 粒子$x_i^{(j)}$，j =1, 2, . . . , P.例如，这些粒子将具有零速度并且 Guass 分布的初始位置测量周围聚集。第二步是“重要性抽样”，它使用动态模型 $P(x_i|x_{i-1})$概率地模拟粒子在一个时间步长上的变化。第三步使用测量模型$w_i^{(j)} = P\left(z_i|\hat{x}_i^{(j)}\right)$计算所有粒子的“重要性权重”。更重要的权重对应于更好地被测量支持的粒子。然后重要的权重被归一化，所以它们相加到一个。当从与归一化重要性权重$w_i^{(j)}$成正比的$\hat{x}_i^{(j)}$中选择一组新的 P 粒子$x_i^{(j)}$时，循环中的最后一步是“选择步骤”。最后，我们可以通过$\hat{x}_i = \sum_{i=1}^{P} w_i^{(j)}\hat{x}_i^{(j)}$来计算权重和。纠正后的速度更加平滑，求得的加速度更准确。

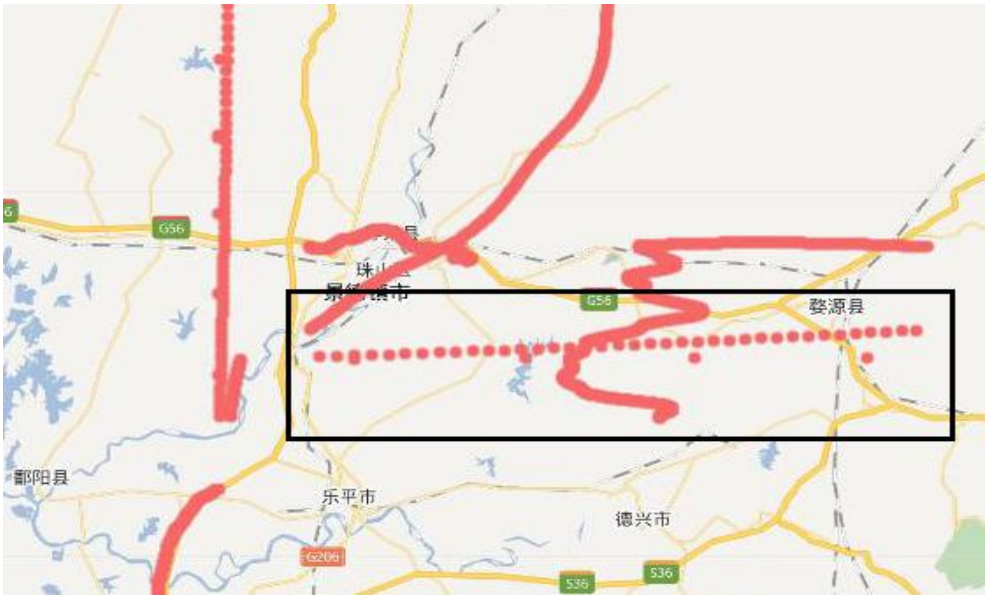

图 4 路线中的噪声图

2.2.3 刻画路线

1）基于密度的轨迹时空聚类

由于车辆行驶路线坐标存在噪声大，位置偏移等问题，本小组分别采用了基于密度的轨迹时间聚类和时间空间聚类对路段信息进行聚类，并生成路线。

聚类分析是按照相似程度划分数据，并保持类间距离最大，类内距离最小。轨迹聚类则是聚类分析在时空轨迹上的扩展，其目的是基于空间或时间相似性，把具有相似行为的时空对象划分为一类，通过聚类可以发现物体的移动模式，分析移动规律，甚至预测未来的运动行为。具体过程如下图：

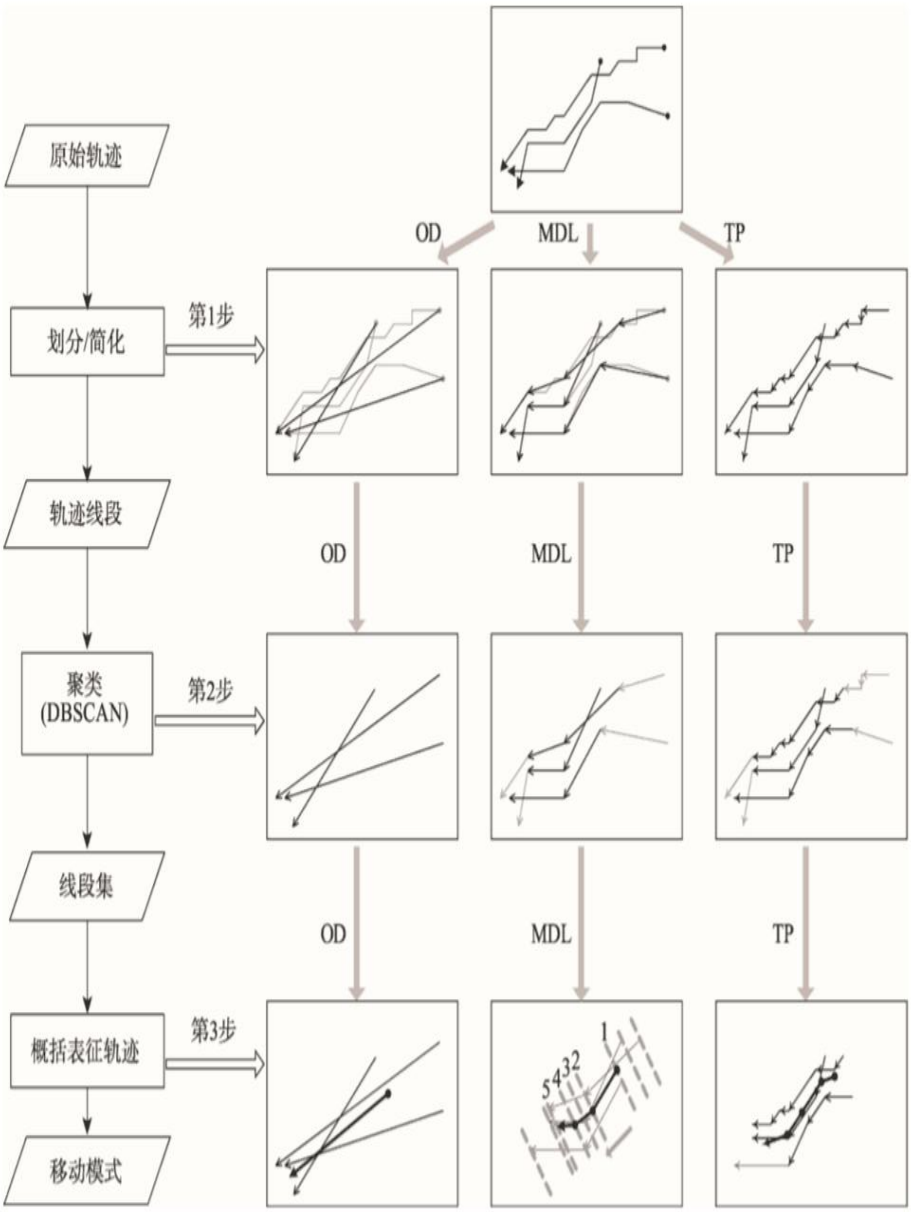

图 5 时空聚类方法图

本小组采用基于密度的聚类方法对轨迹进行分析，综合考虑车辆移动的时空信息，在现有密度的部分轨迹线段聚类方法的基础上进行改进，定义线段时间距离的度量方法，通过调参，确定其阈值和空间邻域一同构建线段的时空邻域。基于时空邻域内的轨迹密度进行聚类，得到每段路线的行驶路线，聚类方法如下图：

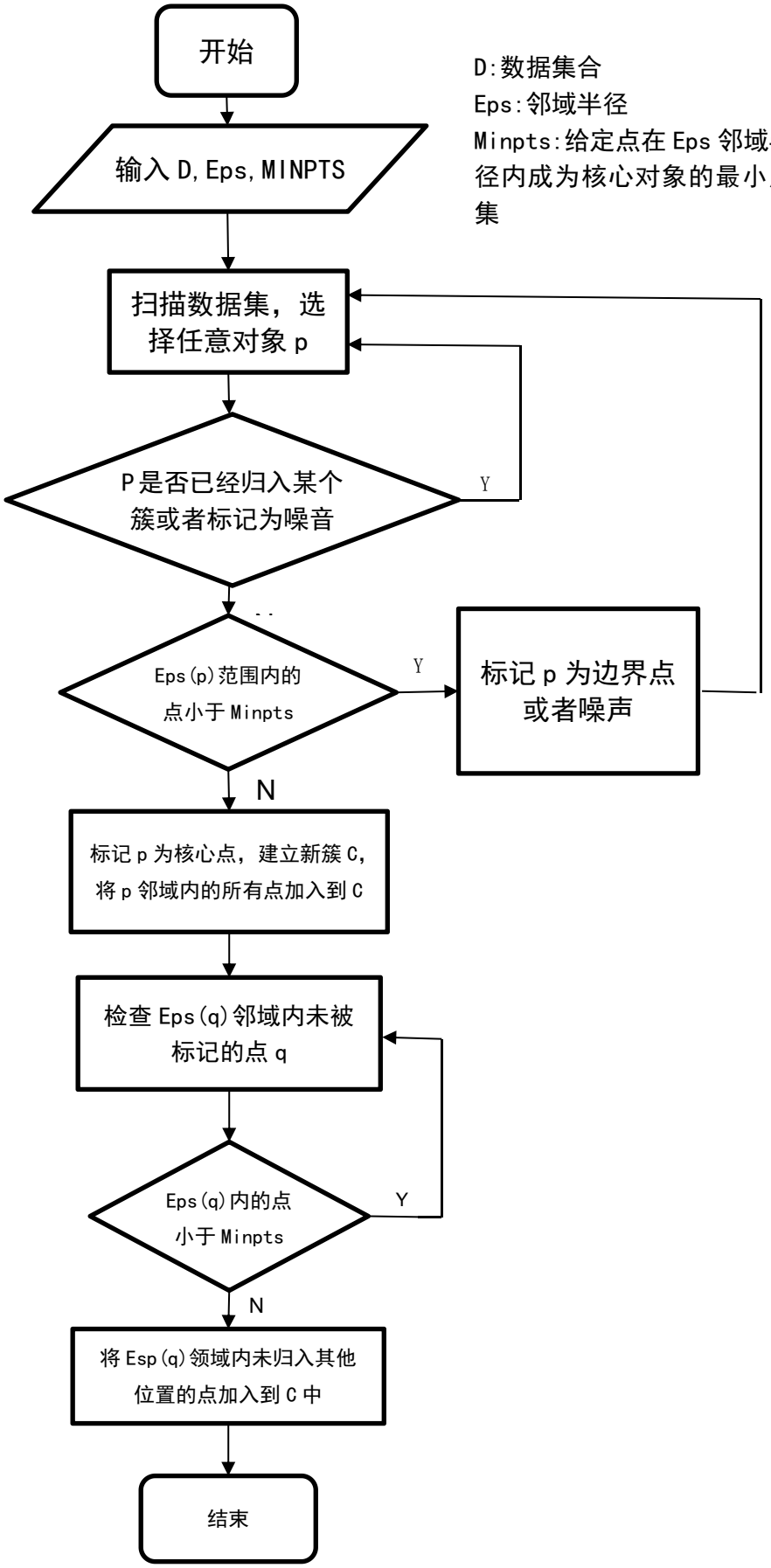
开始
输入 D,Eps,MINPTS
D:数据集合
Eps:邻域半径
Minpts:给定点在 Eps 邻域半径内成为核心对象的最小点集
扫描数据集，选择任意对象 p
P 是否已经归入某个簇或者标记为噪音
Y
Eps(p)范围内的点小于 Minpts
Y
标记 p 为边界点或者噪声
N
标记 p 为核心点，建立新簇 C，将 p 邻域内的所有点加入到 C
检查 Eps(q)邻域内未被标记的点 q
Eps(q)内的点小于 Minpts
Y
N
将 Esp(q)领域内未归入其他位置的点加入到 C 中
结束

图 6 密度聚类流程图

根据上述聚类流程，将其算法步骤归纳如下：

Step1.获得数据的行和列(一共有 n 条数据)。

Step2.计算时间距离矩阵和空间距离矩阵。

Step3.将矩阵的中小于 minPts 的数赋予 1，大于 minPts 的数赋予零，然后 1 代表对每一行求和,然后求核心点坐标的索引。

Step4.初始化类别，-1 代表未分类，遍历所有的核心点。

Step5.寻找种子点的 eps 邻域且没有被分类的点，将其放入种子集合。

Step6.通过种子点，开始生长，寻找密度可达的数据点，一直到种子集合为空，一个簇集寻找完毕。

Step7.将邻域内且没有被分类的点压入种子集合。

Step8.簇集生长完毕，寻找到一个类别。

首先使用已转化为 Unix(连续整数值)的时间变量进行聚类，通过调参将时间参数设为 7200 秒，默认值邻域值为 30。得到聚类结果如下图，其中横坐标为时间，纵坐标为路线类别。可见聚类效果较好，路段的类间距离较大。

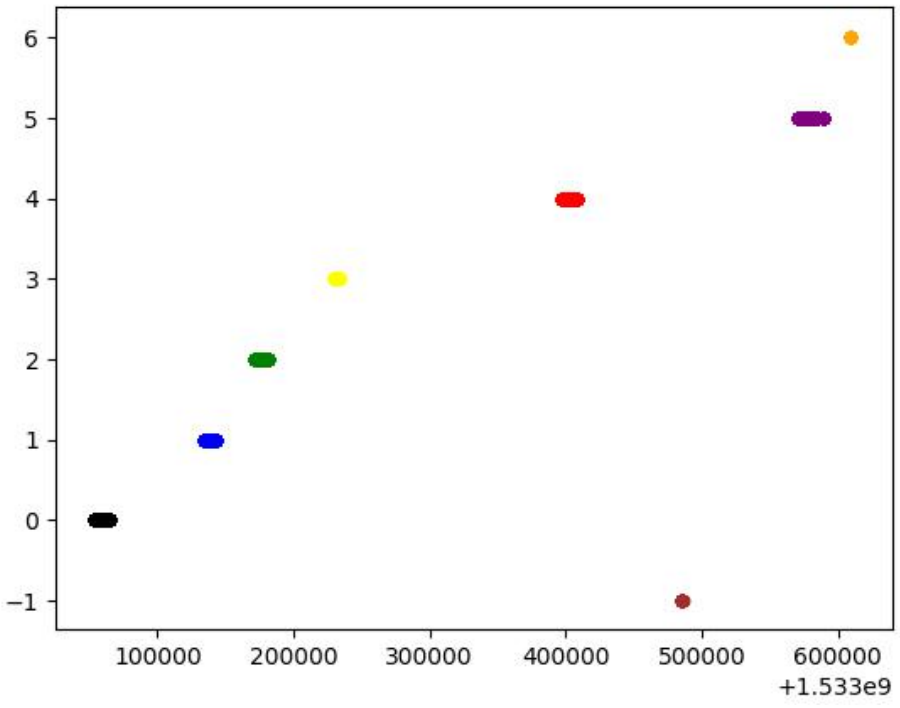

图 7 时间聚类效果图

然后在时间参数的基础上加入空间参数，设置为 500 米，默认领域值仍为 30，得到聚类效果如下图，其中横坐标为时间，纵坐标为路线类别，可见聚类效果更加清晰，并能够筛选出更多类别为-1 的离群点。至此，每段行车路线已能较好的分类出，如图 ，其得到了 10 辆车共 81 条行车路线（删除异常路线）。

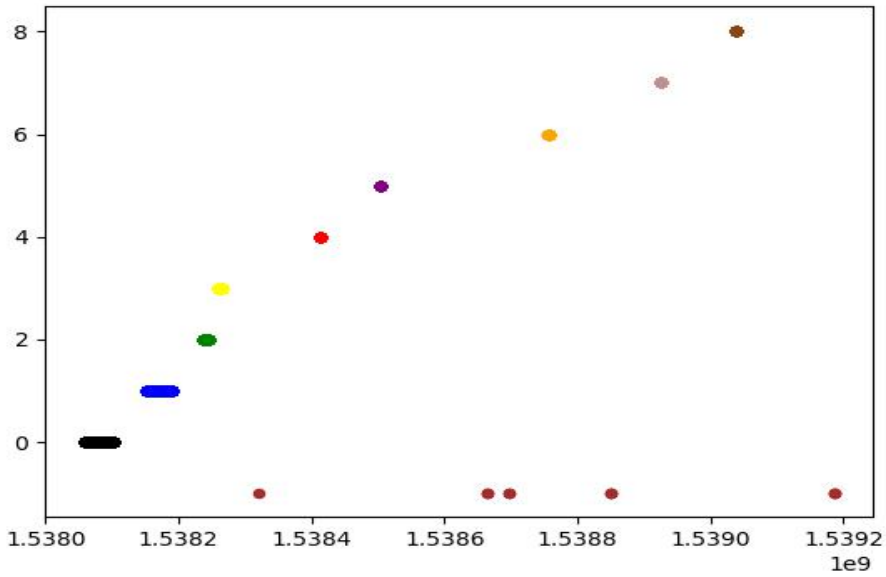

图 8 时空聚类效果图

2）地图可视化

使用谷歌地图的 API 借口将分段后的路线可视化，对不同路线标不同颜色，然后绘制在谷歌地图上，保存为 HTML 文件。部分路线图效果如下图，其中不同颜色代表不同的路线，图 9、图 10 展示了运输车辆在城市中的运输路线，轨迹较为复杂，图 11 展示了车辆在高速公路上的轨迹，较为简单。

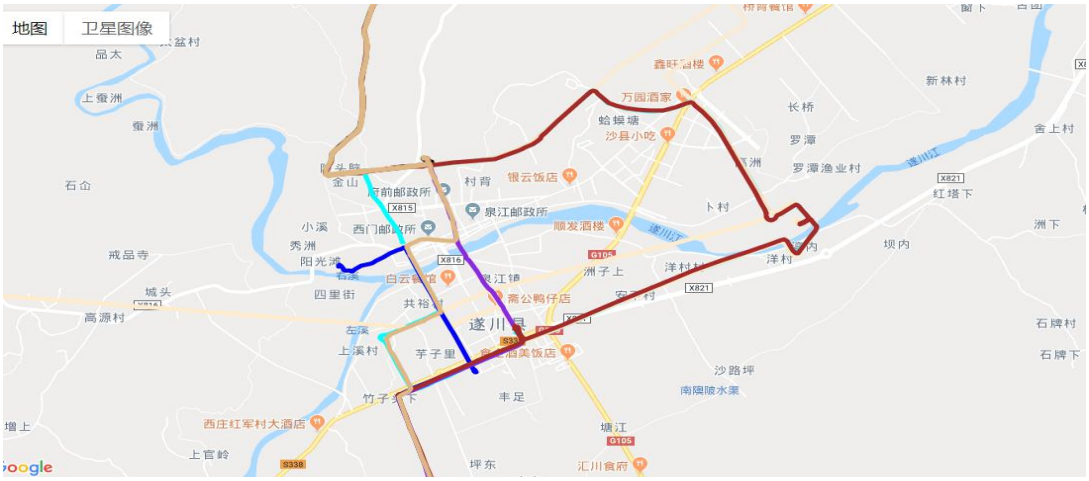

图 9 路线图 a 展示

图 10 路线图 b 展示

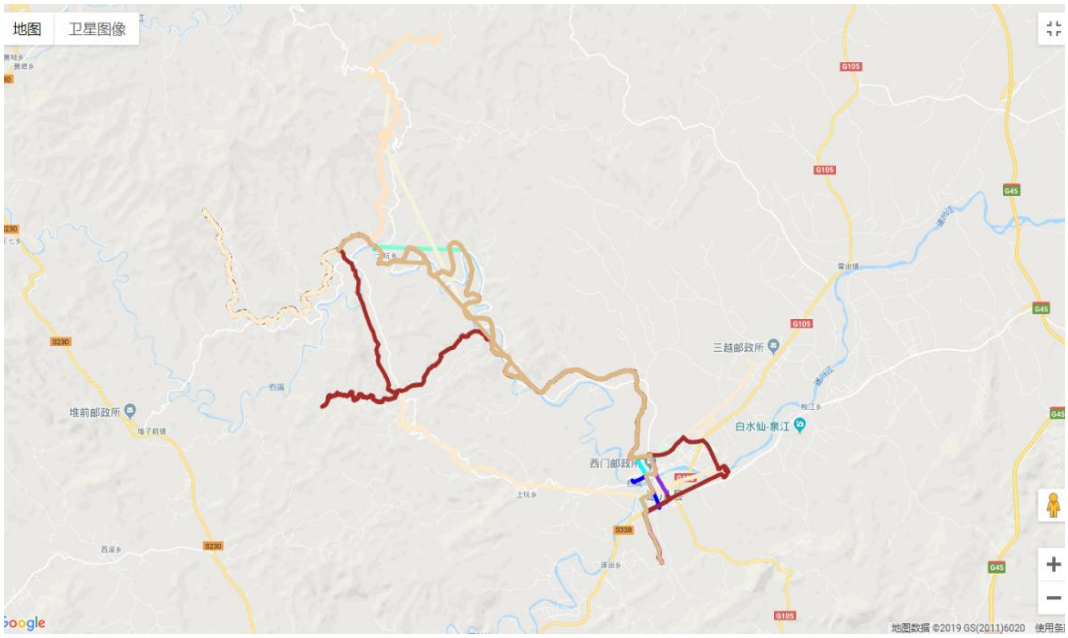

图 11 路线图 c 展示

注：根据《中华人民共和国测绘法》，我国在绘制地图时采用 GCJ-02 编码体系，该坐标系会将真实的 GPS 坐标经过算法偏移真实距离，这也造成谷歌地图与真实情况存在偏移。因此在绘制车辆行驶路线图时，路线图不会完全重合在公路上。

2.2.4 挖掘行车情况

1）指标说明

A：行车里程

计算规则：以每段路线的最终里程减去起始里程。

单位：公里

B:平均行车速度

计算规则：计算规则统计每辆车每段路径的速度求和，除以速度个数，得到平均速度。

单位：公里每小时

C：急加速急减速情况

计算规则：参考多家运输货车公司的三急标准（急加速、急刹车、急转弯），将行车加速度绝对值大于阈值（2.2m/s^2）时定为急加速急减速。然后计算货车每一时刻的加速情况，累加得到急加速急减速次数。

单位：次数

2）指标生成

通过指标说明，建立数学模型，将每段路线的运输情况输出为表格，如下表。

表 3 运输情况表

车牌号	路线	行车里程	平均行车速度	急加速急减速
AA00001	Route1	210	76.02097278	2
AA00001	Route2	6	30.48606811	0
AA00001	Route3	0	0	0
AA00001	Route4	3	26.13076923	0
AA00001	Route5	250	71.70554177	1
AA00001	Route6	482	68.67209776	9
AA00001	Route7	128	53.64674278	1
…	…	…	…	…

2.3 问题二：不良驾驶行为分析

2.3.1 处理流程

图 12 问题二流程图

2.3.2 预处理

首先使用问题一的预处理方法对问题二的数据进行预处理，在使用 python pandas 对所给的 450 辆车的行车数据进行检查的过程中，发现存在两个表格数值全为空值（AF00105、AD00112）,将其删除。

2.3.3 特征说明

1）疲劳驾驶

说明：根据国家法律规定，驾驶机动车有下列行为：连续驾驶机动车超过 4 小时未停车休息或者停车休息时间少于 20 分钟，记为疲劳驾驶。

计算规则：首先用速度变量对每辆车进行停留点检测，将车辆启动的时间初始化为 0，若行驶时间大于 4 小时，到达停留点时将时间再次初始化，车辆再次启动时，求得停留时间，若在停留点时间不到 45 分钟，则标注为疲劳驾驶。

2）急加速

说明：车辆的瞬时加速度值大于给定阈值

计算规则：通过对每列相邻数据的 GPS 速度值进行求差，得到加速度值并进行存储，通过查阅国家交通相关法律法规与多家运输公司交通规则，将阈值定为

2.22m/s^2。

3）急减速

说明：车辆的瞬时加速度值大于给定阈值

计算规则：通过对每列相邻数据的 GPS 速度值进行求差，得到加速度值并进行存储，通过查阅国家交通相关法律法规与多家运输公司交通规则，将阈值定为-2.22m/s^2。

4）怠速预热

说明：原地怠速热车，就是挂空挡，发动发动机，进行热车。

计算规则：取数据中 GPS 速度值为 0 但 ACC 状态为 1（发动机处于工作状态）的连续数据，连续数据的时间跨度在三至五分钟的定义为怠速预热。

5）超长怠速

说明：怠速预热时间过长，超过阈值

计算规则：计算方法如（4）怠速预热，将阈值改为五分钟以上及四十五钟以下。

5）熄火滑行

说明：在机动车行驶中，驾驶员把变速杆置于空档位置，利用车辆惯性行驶。

计算规则： 取数据中 ACC 状态为 0，且 GPS 速度不为 0 的行，标注为熄火滑行。

7）超速

说明：车辆行驶速度过快，超过指定阈值。

计算规则：假设车辆都为大型运输火车，行驶道路为高速公路，查阅国家交通相关法律法规，将阈值定为 100km/h。

8）急变道

说明：车辆在行驶过程中，在极短的时间内改变行车方向角进行变道。

计算规则：对每列相邻数据的方向角变量进行求差，得到方向角变化值量并进行存储，通过查阅国家交通相关法律法规，将方向角变化值大于 10°，且行驶速度大于 50km/h 的时刻定义为急变道。

2.3.4 模型建立

根据相关资料，将车辆的某一时刻的 8 个不良驾驶行为的存在情况（存在则为 1，不存在则为 0，如表），加权建立安全评价模型。

表 4 不良驾驶行为表

车辆	时间	急加速	急减速	疲劳驾驶	怠速预热	超长预热	熄火滑行	超速	急变道
AA00001	2018/8/4/1:28:50	0	1	0	0	0	0	0	0
AA00001	2018/8/4/1:24:29	0	0	0	1	0	0	1	0
AA00001	2018/8/4/1:26:59	0	0	0	0	0	0	1	1
AA00001	2018/9/9/8:06:29	0	1	1	0	0	0	0	0
AA00001	2018/8/5/9:56:12	0	0	1	0	0	0	1	0
…	…	…	…	…	…	…	…	…	…

建立安全评价模型：

$$Y=6n_1+3n_2+2n_3+n_4+n_5+n_6+0.5*(n_7+n_8)$$

其中n_1为疲劳驾驶，n_2为超速，n_3为急变道，n_4为急加速、n_5为急减速、n_6为熄火滑行、n_7为怠速预热、n_8为超长怠速，Y 为评分。

危险标准：

一级危险：评价模型评分>=12

二级危险：评价模型评分>= 8 and 评价模型评分<12

三级危险：评价模型评分>=4 and 评价模型评分<8

四级危险：评价模型评分>0 and 评价模型评分<4

2.3.5 评价结果

将 450 辆车的运输数据加入到安全评价模型，得到评分，然后对评分进行划分，得到最终的评价结果，如下表。

表 5 不良驾驶行为评分表

车辆	时间	急加速	急减速	疲劳驾驶	怠速预热	超长预热	熄火滑行	超速	急变道	评分	安全等级
AA00001	2018/8/4/1:28:50	0	1	0	0	0	0	0	0	1	四级
AA00001	2018/8/4/1:24:29	0	0	0	1	0	0	1	0	3	四级
AA00001	2018/8/4/1:26:59	0	0	0	0	0	0	1	1	5	三级
AA00001	2018/9/9/8:06:29	0	1	1	0	0	0	0	0	7	三级
AA00001	2018/8/5/9:56:12	0	0	1	0	0	0	1	0	9	二级
AA00001	2018/9/9/8:34:22	0	0	1	0	0	0	1	1	11	二级
…	…	…	…	…	…	…	…	..	…	…	…

2.4 问题三：综合评价

2.4.1 处理流程

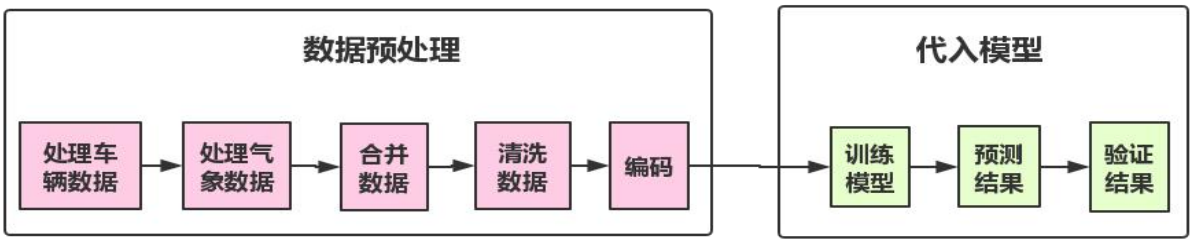

图 13 问题三流程图

2.4.2 数据预处理

1）日期格式转换

由于谷歌地图无法精确地将经纬度地位到中国区域的城市及区县，这里重新采用百度地图的 API 来定位车辆并将车辆连接到气象数据中。

将数据所给的日期变量格式转换成百度地图 API 所需要的日期格式，同时，为提高车辆与气象的连接准确性，提取了更多特征，在日期信息中额外提取了工作日信息，如下表

表 6 日期转换表

车辆	原日期	转化后日期	工作日
AA00001	2018/8/4 1:22:02	4/8/2018	星期六
AA00002	2018/8/5 3:30:01	5/8/2018	星期日
AA00003	2018/8/6 4:12:09	6/8/2018	星期一
AA00004	2018/8/7 5:45:54	7/8/2018	星期二
AA00005	2018/8/8 2:12:43	8/8/2018	星期三
AA00006	2018/8/9 7:21:23	9/8/2018	星期四
…	…	…	…

2）坐标转换

调用百度地图 API 需要重新将原始经纬度信息（国际标准）转化为百度坐标偏移标准格式，具体方法为先将国际坐标格式转为火星坐标格式，再转为百度坐标。然后就可以调用百度 API 获取每个经纬度的 POI（区域）信息，精确到第三级（省、地级市、县级市），如下表。

表 7 坐标定位表

车辆	经度	纬度	省	地级市	县级市
AA00001	115.944523	28.651165	江西	南昌	南昌
AA00001	116.28789	28.69091	江西	南昌	南昌县
AA00001	116.42503	28.717805	江西	南昌	进贤
AA00001	116.424698	28.717458	江西	南昌	进贤
AA00001	117.64985	29.242711	江西	上饶	余干
AA00001	117.64985	29.242711	江西	上饶	婺源
AA00001	116.99973	28.68104	江西	上饶	万年
AA00001	116.9973	28.68096	江西	上饶	万年
…	…	…	…	…	…

3）连接车辆数据与气象数据

将上述处理过后的车辆信息数据的 POI 信息作为外键与气象数据连接，对数据进行清洗、去重，得到了合并的数据集，如表。

表 8 车辆天气数据结合表

车辆	省	地级市	县级市	…	天气状况	风向	风力
AA00001	江西	南昌	南昌	…	雷阵雨转多云	西南风	1-2 级
AA00001	江西	南昌	南昌县	…	雷阵雨转多云	西南风	1-2 级

AA00001	江西	南昌	进贤	…	雷阵雨转阵雨	南风	1-2 级
AA00001	江西	南昌	进贤	…	雷阵雨转阵雨	南风	1-2 级
AA00001	江西	上饶	余干	…	阵雨	南风	1-2 级
AA00001	江西	上饶	婺源	…	阵雨	南风	1-2 级
AA00001	江西	上饶	万年	…	阵雨	南风	1-2 级
AA00001	江西	上饶	万年	…	阵雨	南风	1-2 级
…	…	…	…	…	…	…	…

4）对类别特征进行 one-hot 处理

数据集中存在若干不为连续值的的类别变量，如风向、天气状态。在机器学习任务中，对于这样的特征，通常我们需要对其进行特征数字化。

本文对这些特征采用 one-hot 编码，one-hot 编码，又称独热编码、一位有效编码。其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。

2.4.3 模型建立

1）神经网络模型

神经网络(Neural Network, NN)，在机器学习和认知科学领域，是一种模仿生物神经网络（动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统，通俗的讲就是具备学习功能。现代神经网络是一种非线性统计性数据建模工神经网络通常是通过一个基于数学统计学类型的学习方法（Learning Method）得以优化，所以也是数学统计学方法的一种实际应用，通过统计学的标准数学方法我们能够得到大量的可以用函数来表达的局部结构空间，另一方面在

人工智能学的人工感知领域，我们通过数学统计学的应用可以来做人工感知方面的决定问题（也就是说通过统计学的方法，人工神经网络能够类似人一样具有简单的决定能力和简单的判断能力），这种方法比起正式的逻辑学推理演算更具有优势。

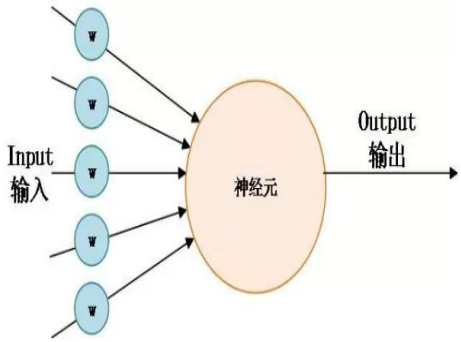

图 14 神经元结构

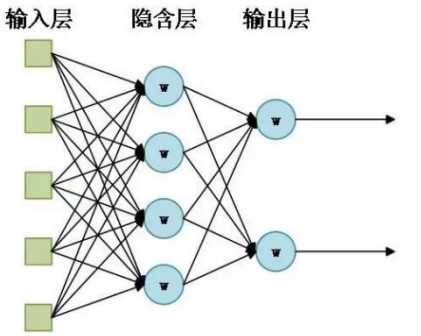

图 15 三层神经网络结构

2）架构图

模型采用三层全连接神经网络，对合并后的车辆特征和天气特征作为 input（输入），逐层传播并将损失（由 loss_function 损失函数计算）后向传播，输出误差最小的评价结果。

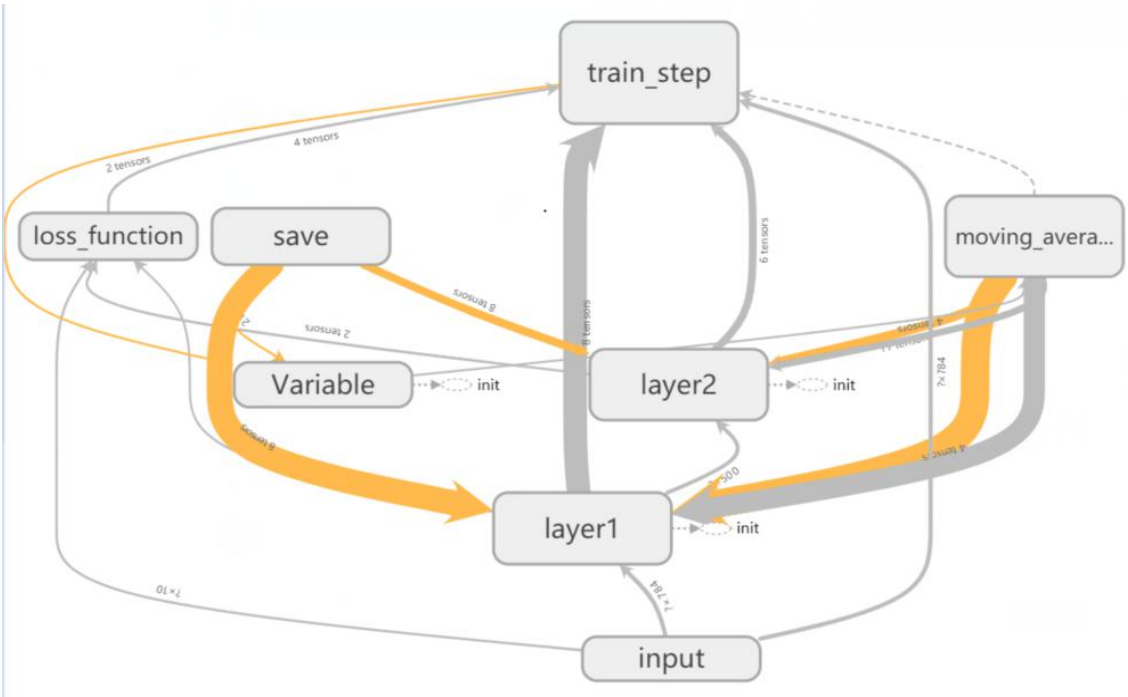

图 16 模型架构图

3）算法流程

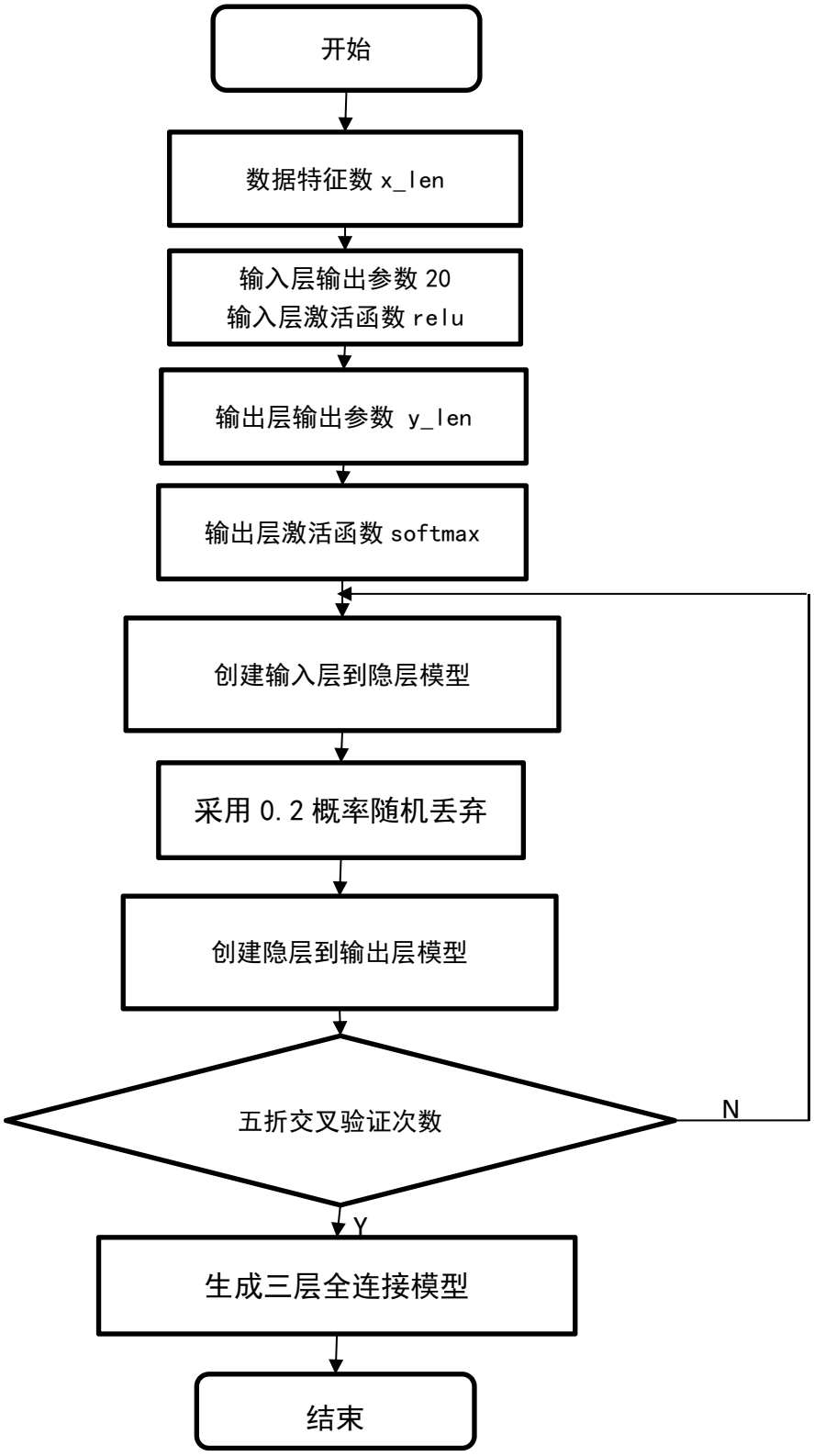

图 17 三层全连接流程图

2.4.4 评价结果

1）评价结果

最终输出每辆车的驾驶行为安全评分与结合实时环境的安全评分，结果如下表。

表 9 综合评价结果

车辆	时间	…	驾驶行为评分	…	综合危险评分	综合危险等级
AA00001	2018/8/4	…	0	…	0	安全
AA00001	2018/8/4	…	0	…	0	安全
AA00001	2018/8/4	…	2	…	3	一级
AA00001	2018/8/4	…	3	…	3	一级
AA00001	2018/8/4	…	5	…	5	二级
…	…	…	…	…	…	…

2）准确率与误差

通过 7:3 将 30 万条数据集划分成了训练集和测试集，并留出 10 万条数据作为验证集，在测试集上的正确率 95。57%，在验证集的正确率 95.53，测试效果如下图，图 18 横坐标为神经网络迭代次数，纵坐标为准确率，图 19 横坐标同为神经网络迭代次数，纵坐标为结果损失。

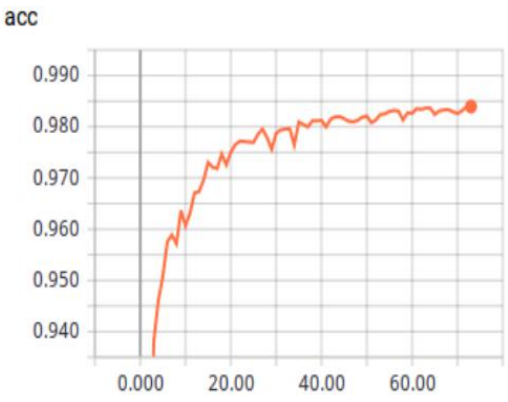

图 18 准确率图

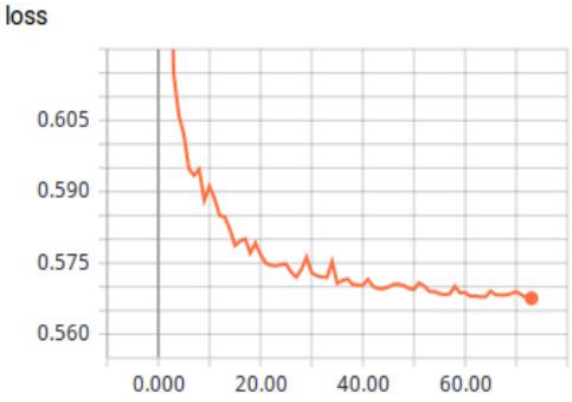

图 19 误差损失图

3）F1 评分

F1 值是精确率（P）和召回率（R）的调和均值，即 F1=2PR/(P+R)，相当于精确率和召回率的综合评价指标。下图横坐标和神经网络层数，纵坐标为 F1 值，可见神经网络层数越高得到的评分效果越好，本小组采用三层全连接网络，在 1 层 f1-score=0.9，2 层 f1-score=0.85，3 层 f1-score=0.8，已能较好的预测安全等级评分。

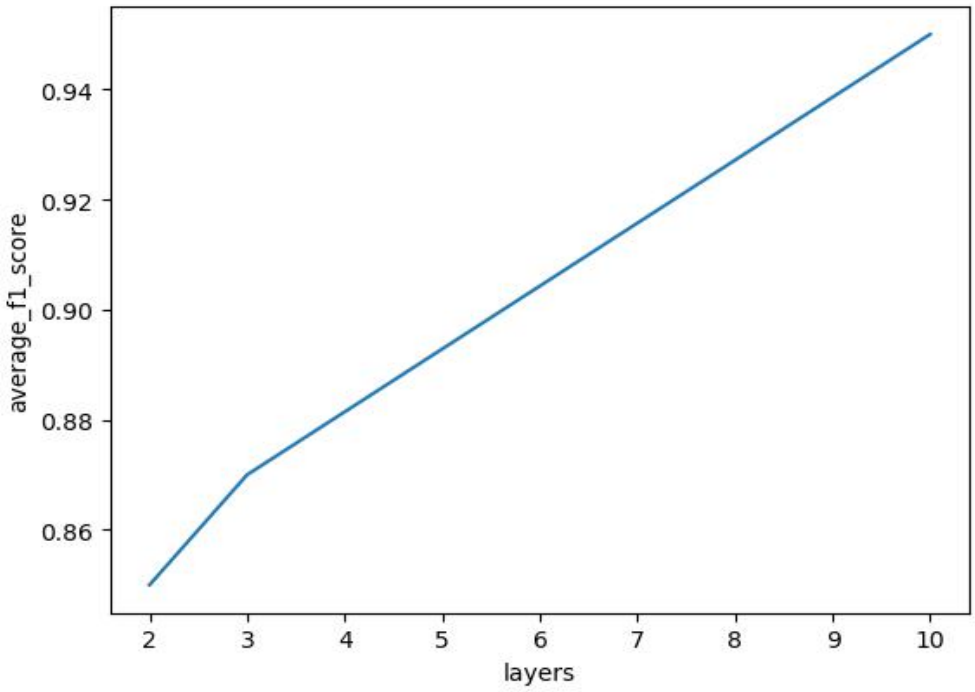

图 20 F1 值效果图

3. 结论

由于数据量较大无法上传全部过程数据，本题所使用的过程数据与三个问题的结果，存放在了源程序压缩包的 result 目录下。

题目数据量较大，存在比较多的数据噪声，于是本文通过预处理后的轨迹数据，利用了卡尔曼滤波纠正偏移，去除了大部分的坐标偏移。通过简单的时间邻域对道路进行分段，得到效果并不理想，于是采用了基于密度聚类的时空聚类方法来对坐标进行聚类，分段效果较好。由于国家测绘法规定，我国境内经纬度与真实位置存在一定偏差，误差在数据分析可以容忍的范围内，可以忽略。

指标提取的工作大部分为设立标准，然后利用简单的表格处理技术得到特征，工作重心在对交通规则的查阅与对危险的定义。

最后采用机器学习方法来结合环境数据和行车记录仪中的数据，利用三层全连接神经网络模型来训练，得到最终的综合安全评分，分析结果来看，综合安全评分基本与仅仅通过行车记录仪记录的数据得到的评分匹配度在 95%以上，较为吻合。

由于缺乏对人体行为与环境影响领域的深层研究，本文给出的模型没要考虑到特征变量之间的深层关联关系，使用了机器学习的模型来进行分析，显然，还有很多工作需要研究。

4. 参考文献

[1] 牟乃夏，张恒才，陈洁,等. 轨迹数据挖掘城市应用研究综述[J]. 地球信息科学学报, 2015, 17(10):1136-1142.

[2] 张光亚. 多目标无源定位数据的轨迹提取与模式挖掘算法研究[D].

[3] 移动对象轨迹数据挖掘方法研究[D]. 中国矿业大学, 2012.

[4] 程歆. 出租车 GPS 数据的地图匹配算法研究[J]. 中国战略新兴产业, 2017(24):106.

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

[6] Hagan M T, Beale M, Beale M. Neural network design[M]. 2002.

[7] 崔金红，王旭. Google 地图算法研究及实现[J]. 计算机科学, 2007, 34(11):193-195.

[8] 基于密度的轨迹时空聚类分析[J]. 地球信息科学学报, 2015, 17(10):1162-1171.