基于 BP 神经网络的驾驶行为综合评价方法

摘要

车联网（Internet of Vehicles）是由车辆位置、速度和路线等信息构成的巨大交互网络。通过 GPS（Global Positioning System）、RFID（Radio Frequency Identification）、传感器、摄像头图像处理等装置，车辆可以完成自身环境和状态信息的采集；通过互联网技术，所有的车辆可以将自身的各种信息传输汇聚到中央处理器；通过计算机技术，这些大量车辆的信息可以被进行分析和处理。随着 5G 时代的到来，以及 C-V2X（Cellular Vehicle-to-Everything）技术的不断深入研究，车联网技术迎来了快速发展。本文在这种背景下，对车辆行驶数据进行深入挖掘分析，最终基于 BP 神经网络（Back Propagation Neural Network）构造出了驾驶行为的安全评分模型和综合评分模型。

针对问题一：在数据预处理部分，首先对整体数据完成了去重处理；实现了对经纬度偏移点的修正；对里程异常值进行了分情况处理；为了后续挖掘的方便，将天气数据进行了格式上的统一；最后通过将地理坐标数据作为中间件，实现了对天气数据的集成。接着，分别在 leaflet 地图库和百度地图中绘制出车辆静态轨迹与动态轨迹，并划分出车辆的行驶路段，对不同路段的急加速行为、急减速行为、行驶路程和平均行车速度进行了挖掘与展示。

针对问题二和问题三：结合天气信息，挖掘出了所有车辆涉及到安全、效率和能耗的 17 个驾驶行为指标，采用了加权的方法来分别构建安全评价模型和综合评价模型。为了避免层次分析法中判断指标重要程度过程混乱以及难以应对多指标计算和熵权法中过度依赖数据差异以及准确性不高的问题，本文选择了 BP 神经网络来计算权重。在具体实验中，将专家评分结果作为训练样本，对模型进行了训练与学习。将模型在测试集上的评分结果与专家评分进行了验证，展示出了模型的可行性。

最后，在数据分析阶段，对所有车辆的评分进行了汇总统计，找出了低分值车辆的具体原因；对所有车辆的轨迹生成热力图后发现了该运输企业的区域性特点，并针对气候特征提出了相应的注意事项；通过 K-Means 算法对驾驶员的倾向性完成了聚类，为驾驶员和企业管理部门提供了另一角度的参考。

关键词：车联网、评分模型、BP 神经网络、K-Means 算法

Abstract

Internet of Vehicles is a huge interactive network composed of information of vehicle location, speed and route. Vehicles can complete the collection of their own environment and state information through GPS, RFID, sensors, camera image processing and other devices; all vehicles can transmit their own information to the central processing unit through Internet technology; these large amounts of vehicle information can be analyzed and processed through computer technology. With the advent of 5G era and the continuous in-depth study of C-V2X technology, Internet of Vehicles has reached a rapid development. In this context, the vehicle driving data are deeply mined and analyzed. Finally, the safety scoring model and comprehensive scoring model of driving behavior are constructed based on BP neural network.

Aiming at Question 1, in the part of data preprocessing, removed duplicate of data firstly; the longitude and latitude drift points are corrected; the anomaly mileage values are processed, respectively; for the convenience of subsequent mining, the weather data is unified in format; finally, the integration of weather data is realized by using geographic coordinate data as middleware. Then, the static and dynamic trajectories of vehicles are drawn in leaflet map library and Baidu map, respectively, and the driving sections of vehicles are divided. The acceleration, deceleration, driving distance and average driving speed of different sections are mined and displayed.

Aiming at Question 2 and Question 3, 17 driving behaviors involving safety, efficiency and energy consumption of all vehicles are mined after combined with weather information, and weighted methods are used to construct safety evaluation model and comprehensive evaluation model, respectively. In order to avoid the confusion in the process of judging the importance of indicators and problem in multi-index calculation in AHP, and the difficulty in dealing with the problems of over-reliance on data differences and inaccuracy in entropy weight method, this paper chose BP neural network to calculate the weight. In the specific experiment, the model was trained and learned with the result of expert scoring sample. The scoring results of the model on the test set and the expert scoring are contrasted, and the feasibility of the model is demonstrated.

Finally, in the stage of data analysis, the scores of all vehicles are summarized, and the specific reasons of low-value vehicles are found out. After generated heat map of all vehicles'trajectories, the regional characteristics of the transportation enterprise are found, and corresponding precautions were put forward according to the climate characteristics. The driver's inclination is clustered by K-Means algorithm, which provided another angle of proposal for drivers and enterprise management departments.

Key words: Internet of Vehicles, Scoring model, BP neural network, K-Means algorithm

目录

1. 研究背景与目标

1.1 研究背景

近 20 年来，随着我国社会，经济和科学技术的全面快速发展，物流运输产业已经成为我国的支柱型产业之一。经济一体化的发展也使得企业的采购，仓储，销售，配送等协作关系日趋复杂，运输产业越来越被得到重视。而作为运输方式中的重要的一环——公路运输的管理也变得更加重要，我国公路基础设施建设迅速发展，公路运输能力大大提高，在国民经济增长和人民生活水平提高方面发挥着越来越重要的作用。

但是随着日益增长的运输需求和人们对于运输要求的变高，公路运输在信息时代迎来新的挑战。车联网技术应运而生，车联网是指借助装载在车辆上的电子标签通过无线射频等识别技术，实现在信息网络平台上对所有车辆的属性信息和静、动态信息进行提取和有效利用，并根据不同的功能需求对所有车辆的运行状态进行有效的监管和提供综合服务的系统。当前道路运输行业等相关部门已经开始利用车联网等技术，开展道路运输过程安全管理的数据分析，以提高运输安全管理水平和运输效率。

车载传感器等设备能够实时的采集到大量的车辆行驶数据，如何从大量、复杂的采集数据中提取车辆的行程、对车辆轨迹进行挖掘、对车辆行驶轨迹进行管控、分析驾驶员的不良驾驶行为、为驾驶员提出有效建议是至关重要的。并且综合考虑安全、效率和能耗三个因素进行驾驶数据分析与挖掘能够为驾驶员、运输企业和交通管理部门提供十分有价值的信息。

1.2 研究目标

1) 轨迹挖掘，通过采集到的 GPS 数据，进行轨迹的重放，并对每段路程的急加速、急减速、行驶路程和平均速度等情况进行挖掘。
2) 安全评价模型，参考不良驾驶行为的业内标准，建立行车安全的评价模型，并给出评价结果。
3) 综合评价模型，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为运输车辆管理部门建立综合评价模型。

2. 问题分析

2.1 问题一的分析

根据提供的数据在经纬度坐标系下绘制出运输路线图及对应的每一段的急加速、急减速、行车里程、平均行车速度等情况。附件 1 中给出了 449 辆行驶车辆的部分 OBD 数据，包括经纬度信息、转角方向、车辆的 gps 速度和总行驶里程等。确定要绘制车辆轨迹后，首先要对采集到的数据进行预处理，对车辆轨迹中的缺失值，异常值(漂移值)进行处理。然后分别在百度地图和 leaflet 地图库中绘制出车辆轨迹，并划分出车辆的行驶路段，对不同路段的急加速、急减速、行驶路程和平均行车速度进行挖掘与展示。

2.2 问题二的分析

利用附件 1 所给数据并结合附件 2 天气数据，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。因为采集到的都是车辆方面的数据，所以想要分析驾驶员的不良驾驶行为(如疲劳驾驶、酒驾、毒驾等)没有其他的更直观的参考依据，例如车内的视频监控，司机的实时身体特征等数据可供分析。所以需要通过分析车辆采集到的转向角、经纬度、ACC 的状态、车速、里程、时间的变化以及驾驶时天气状况等，在已有问题一的四个指标的前提下，来继续深入挖掘车辆的不良驾驶行为并统计出每辆车的各个不良驾驶行为的严重程度，基于统计出的各种不良行为的多少建立安全行车的评价模型并给出评分。

2.3 问题三的分析

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综合评价模型。在问题二的基础上，需要将运输车辆的效率和节能纳入考虑，通过分析车辆的平均运输速度和导致车辆能耗增大的驾驶行为，来对效率和节能情况进行分析，结合问题二已有模型建立综合评价模型并给出评分。

3. 数据预处理

为了提高数据挖掘的质量，数据预处理是必不可少的环节。现实中采集到的数据大多都带有噪声，数据预处理是数据挖掘过程中必不可少的环节，优秀的数据预处理是数据挖掘成功的基石。

附件 1 中给出了车辆上传感器采集的信息，每张表的格式如下表 1 所示。

表 1 附件 1 的数据说明

指标名称	指标说明	说明
vehicleplatenumber	车牌号码	
device_num	设备号	设备的 ID
direction_angle	方向角	范围：0-359（方向角指从定位点的正北方向 起，以顺时针方向至行驶方向间的水平夹角）
lng	经度	东经
lat	纬度	北纬
acc_state	ACC 状态	点火 1/熄火 0
right_turn_signals	右转向灯	灭 0/开 1
left_turn_signals	左转向灯	灭 0/开 1
hand_brake	手刹	无 0/有 1
foot_brake	脚刹	无 0/有 1
location_time	采集时间	
gps_speed	GPS 速度	单位：km/h
mileage	GPS 里程	单位：km

在现实的车辆数据采集中，往往会出现各种各样的噪声，比如会出现 GPS 定位漂移，里程数据异常等，于是对这些数据进行必要的处理是至关重要的。

3.1 数据去重

在数据预处理的最开始，应该进行去重处理，这可以避免很多麻烦，例如：可以避免用前后两条记录的经纬度距离差和时间差进行除法运算算速度时，分母为零的问题。

3.2 异常检测

3.2.1 经纬度坐标异常

1.异常原因分析[1]

1） 天气情况

下雨天，空气中水分多，影响了信号的传输。这也是为什么夏季 GPS 信号稍弱的原因，夏季雨多潮湿，再加之高温蒸发，使得空气中的水分变多，从而影响 GPS 信号的传输。

2） 高楼因素影响 GPS 信号

在一些高层建筑物的低层或者地下建筑，如车辆行驶在隧道内，由于受到墙体的遮挡，室内信号衰减非常大，就形成了信号覆盖弱点，所以造成定位不精准，误差大等情况。

3） 卫星数量

农村及偏僻地区上空安置卫星数量少，造成位置定位偏差大。

2.异常经纬度检测

在这里，使用基于经纬度距离的异常点检测算法来筛选出经纬度异常，并随后进行处理。这种检测算法是基于这样的思想：通过当前节点的经纬度、当前节点的记录时间以及下一节点的记录时间，来判断下一节点的经纬度是否异常。简而言之，就是通过判断间隔时间和位移距离与阈值的大小关系来筛选出异常点。图 1 展示了该思想的流程。

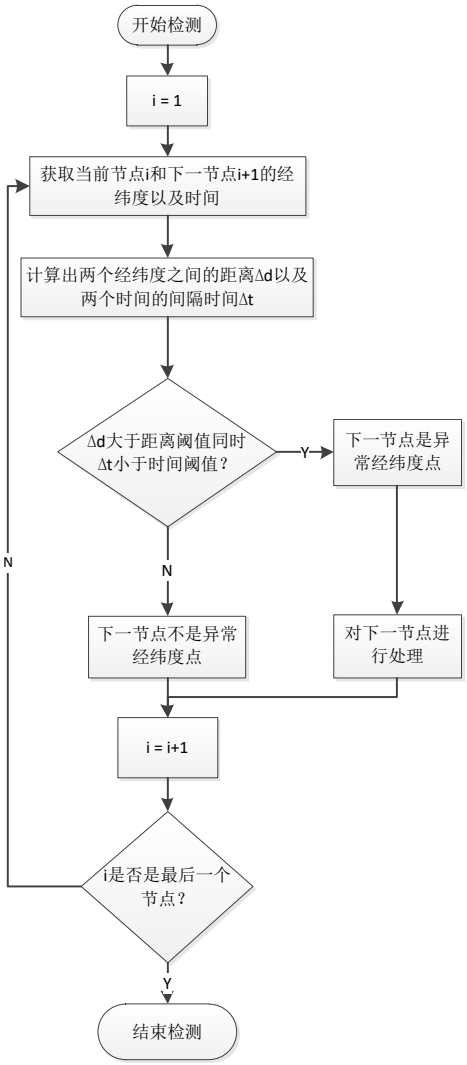


图 1 经纬度异常点的检测流程图

至于检测出的异常经纬度点要进行怎样的处理，将在后续详细阐述。

3.2.2 里程数据异常

1.异常原因分析[2]

里程是用来显示车辆累计行驶距离。在车辆行驶过程中，通过网络将采集到的车辆实时数据返回给后台，在传输过程中可能出现错误导致里程数的异常。也有可能人为的缘故导致里程数异常。

2.异常里程检测

对于异常里程的定义：第一，某一里程大于其后面数据的里程。正常行驶的车辆里程数是随着行车时间增加的，若前面采集到数据的里程值大于后面数据的里程值，则视为异常里程。第二，里程缺失情况，由于采集设备的异常或人为的

原因，会丢失一部分的数据，这部分丢失的数据时间短则几秒钟或者几分钟，长则几天甚至几个月，这导致了里程数的不一致，产生缺失的情况。

- 里程数值异常

举例说明，我们通过将车辆 AD00003 的里程值绘制成折线图进行观察，如图 2 所示。

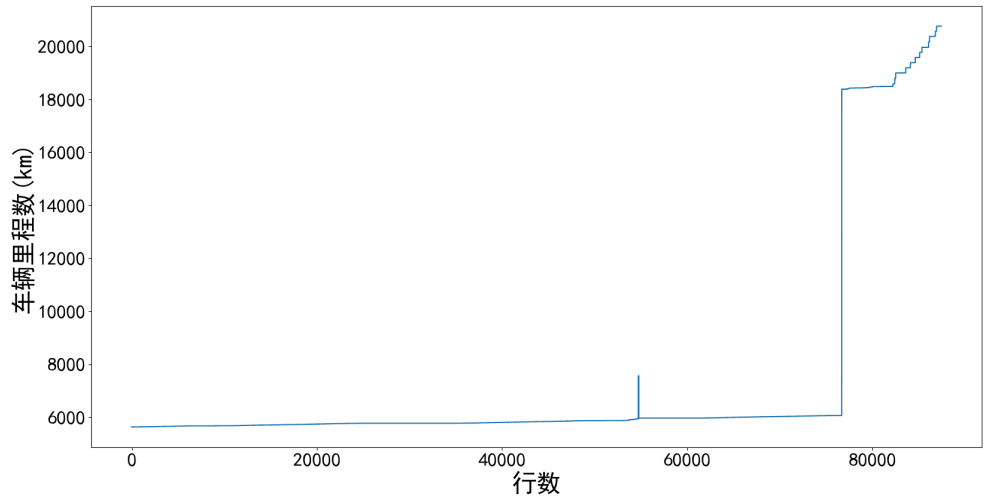


图 2 车辆 AD00003 的里程图

从图中我们可以发现 AD00003 车辆的数据是由 80000 多行的记录构成，其中观察发现在 0 至 40000 多行，里程是处于平稳增长，在约 54000 行的位置发生了突增和突减，由此断定出现了里程异常点。进一步观察原始数据表，如表 2 所示:

表 2 AD00003 里程数值异常处数据

row_num	Vehicleplate number	device_num	direction_angle	lng	lat	…	location_time	gps_speed	mileage
54746	AD00003	AAA7109003	347	114.9773	27.62996	…	2018/8/7 1:41	62	5942
54747	AD00003	AAA7109003	347	114.9773	27.62996	…	2018/8/7 1:41	62	5942
54748	AD00003	AAA7109003	347	114.9772	27.63012	…	2018/8/7 1:41	63	7567
54749	AD00003	AAA7109003	347	114.9772	27.63012	…	2018/8/7 1:41	63	7567
54750	AD00003	AAA7109003	357	114.9343	27.81196	…	2018/8/7 4:38	0	5967
54751	AD00003	AAA7109003	357	114.9343	27.81196	…	2018/8/7 4:38	0	5967

由原始数据表清楚的发现，在第 54748 行和 54749 行，里程跳跃至 7567km，之后时间从 2018/8/7 1:41 跳至 2018/8/7 4:38 后里程又减到 5967km，之后仍按正常情况增多。由此可以说明此处的 7567km 里程为异常里程值。

- 里程缺失

由于采集设备的异常或者人为的原因，会丢失一部分的数据，这部分丢失的数据时间短则几秒钟或者几分钟，长则几天甚至几个月。我们还是以 AD00003 车为例。观察图 2 可以发现里程有一截突然从 6000km 左右突增为 18000km 左右，再进一步观察原始数据，如表 3 所示：

表 3 AD00003 里程缺失处数据

row_num	Vehicleplate number	device_num	direction_angle	lng	lat	…	location_time	gps_speed	mileage
76690	AD00003	AAA7109003	70	115.1331	27.20776	…	2018/8/7 7:43	0	6064
76691	AD00003	AAA7109003	70	115.1331	27.20776	…	2018/8/7 7:43	0	6064
76692	AD00003	AAA7109003	261	115.1333	27.20772	…	2018/10/7 23:54	0	18386
76693	AD00003	AAA7109003	261	115.1333	27.20772	…	2018/10/7 23:54	0	18386

观察原始数据发现数据在 2018/8/7 7:43 和 2018/10/7 23:54 之间是缺失的，里程也从 6064km 变为 18386km,时间缺失两个月，里程跳跃了 12322km。

无论是里程数值异常还是里程缺失，都可以借助检测异常经纬度的思想来筛选出。

3.3 异常处理

在上一节中，已经检测出了经纬度异常和里程值异常，接下来将对这些异常点进行必要的处理。对于经纬度异常，需要根据两个节点间的经纬度间隔（位移距离）和时间间隔来决定怎样处理。图 3 展示这一过程的流程。

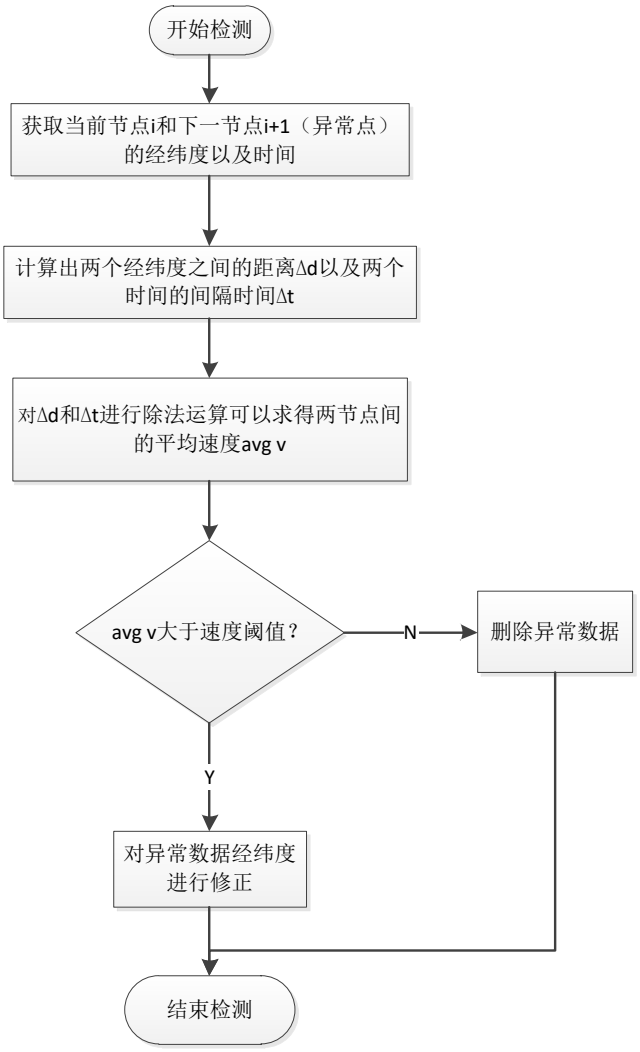


图 3 经纬度异常点的判断处理方式流程图

这种思想就是通过判断两节点间的平均速度与速度阈值的大小关系来确定异常点能否进行修正，简而言之，就是根据车辆能否在间隔时间内行进间隔的距离来决定该异常记录是否是必须删除的脏数据。对于脏数据，直接删除，而如果是能够修正的异常点，则需要结合汽车方向角来进行修正。

至于里程数据异常的情况，同样可以使用两种处理方式对其进行处理。

3.3.1 经纬度异常点的修正

对经纬度异常点进行修正，其实就是一个知道当前节点经纬度求下一节点经纬度的问题。如图 4 所示，待修正的经纬度（lng_1，lat_1）可以通过当前经纬度（lng_0，lat_0）、两节点间隔时间 Δt、当前节点速度 v_0 以及当前车辆方向角 θ 求出。

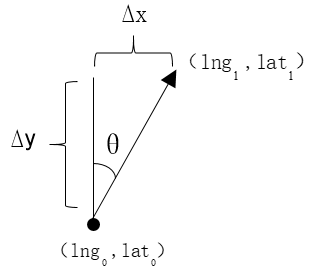


图 4 经纬度异常修正示意图

其中，通过Δt 与 v0 相乘可以得到间隔点间行驶的距离。然后将该距离进行三角函数运算，可以得到相对应的Δx 和Δy，它们分别也就是在经度上和在纬度上的变化距离，然后将该距离换算成当地对应的经度变化值和纬度变化值，将该变化值与当前节点坐标（lng0，lat0）进行相加减，则可以计算出下一异常节点的正确坐标。

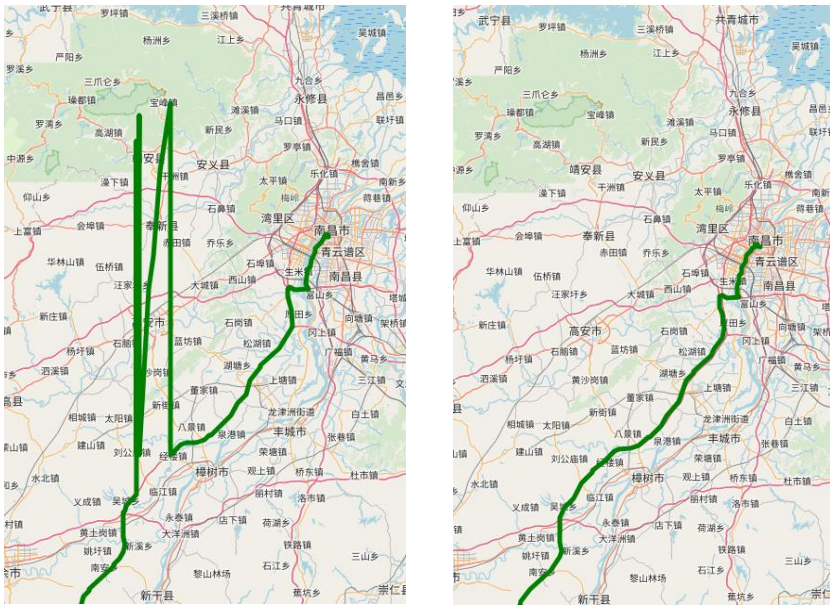
例如在$\theta \in [0,90]$时，可以用公式（1）和公式（2）来进行修正下一节点坐标。

$$lng_1 = lng_0 + f(\sin(\theta) * v_0 * \triangle t) \quad (1)$$

$$lat_1 = lat_0 + g(\cos(\theta) * v_0 * \triangle t) \quad (2)$$

其中函数 $f()$ 是用来将距离长度换算成当地的经度变化，函数 $g()$ 是用来将距离长度换算成纬度变化，之所以经度变化要换算成当地的经度变化，这是因为不同坐标下的同样大小的经度所代表的实际距离会有所不同。

图 5（a）展示了一个典型的可修正异常点的情况，图 5（b）是使用上述算法进行修正后的轨迹情况。



（a）原始轨迹　　（b）修正异常数据后的轨迹

图 5 异常经纬度记录修正前后对比图

图 5 直观的展示了该算法的可行性，然而，还有一些异常数据，并不能使用该算法进行修正，这种脏数据只能对它们进行删除。

3.3.2 经纬度异常点的删除

对于那些速度为零并且时间超过设定阈值并且距离变化也超过距离阈值的点，认定它们为脏数据，直接采取删除处理。

下图以 AA00251 为例，展示了这种异常点在经纬度坐标系中的情况。

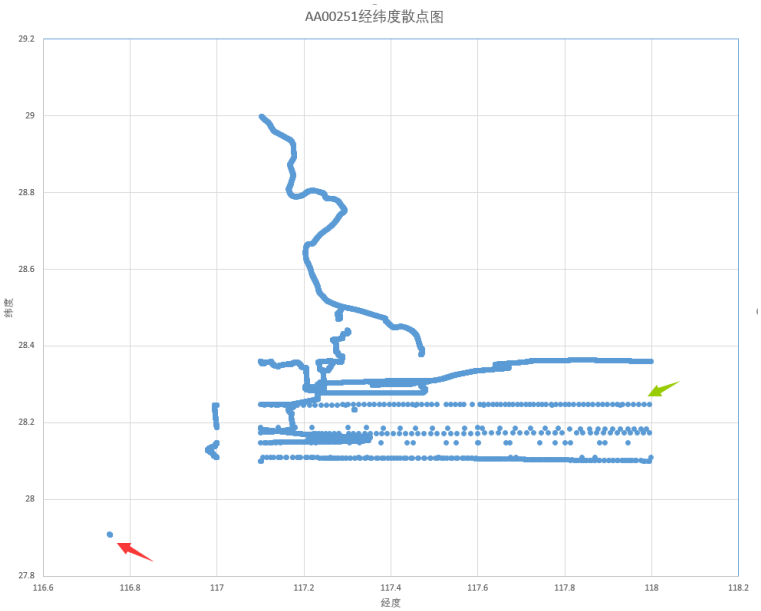


图 6 AA00251 经纬度散点图

在图中，绿色箭头指向的一连串的异常值是属于能够通过上一部分叙述的算法进行修正的异常记录。而红色箭头指向的异常点，往往是速度为零并且与上一记录时间间隔非常大而且距离变化也非常大的脏数据，由于没有道路轨迹坐标库，而这样的道路轨迹又毫无规律可言，对于这样的脏数据，采用删除的处理方式对其处理。

3.3.3 里程数据异常的处理

在 3.2.2 节中，详细描述了里程异常的两种情况，和经纬度异常类似，不同的异常情况采用不同的处理方式。

对于图 2 和表 2 中的里程同时突增突减的情况，认定这种记录为脏数据，删除记录行。而对于表 3 中时间与里程缺失的情况，认定为是新的一段行程的开始，通过添加零数据行对其进行分段处理。

3.4 数据格式统一化

在附件 2 所给的天气表中，有些列的格式比较混乱，故对该表的某些列进行了格式统一化处理。类似于附件 1 的汽车车头方向角，将天气表中的风向由标注性数据转为数值型的角度数据；风级用当日最高风级代替；最高最低温度去掉单位用数值取代。

3.5 数据集成

要用到附件 2 中的天气信息，就必须要将附件 1 中的每条记录与附件 2 的天气情况进行关联。

在这里，使用了额外的包含各省市县的地理坐标信息的 json 文件作为中间件来对附件 1 和附件 2 进行数据集成。

要将汽车行驶的大量经纬度数据分别与地理位置相关联，在这里采用的是基于位移距离的判断方法实现的。其实质就是寻找与当前经纬度坐标最近的省市县坐标，然后将当前记录划分在这个省市县的辖区内。虽然从行政划分的角度上讲，这样的处理没法保证准确性，但因为关联到省市县信息的目的是为了和天气信息相关联，位移距离越接近的天气状况更相似。从这个角度上考虑，这种基于位移距离的判断方法对当前的任务研究将更准确有效。

确定了经纬度和地理信息相关联的方式后，问题就转换成了三个数据表的关联问题，在不将数据导入数据库构建索引的前提下，这样的三表关联往往需要花费大量时间，其时间复杂度将达到$O(m*n*r)$，其中 m 为驾驶数据表的行数，n 为带有坐标信息的省市县数据表的行数，r 为天气数据表的行数。

为了解决表关联的时间开销太大的问题，在这里提出了一种基于判断坐标距离与日期变化的多表关联算法，其思想如下图所示。

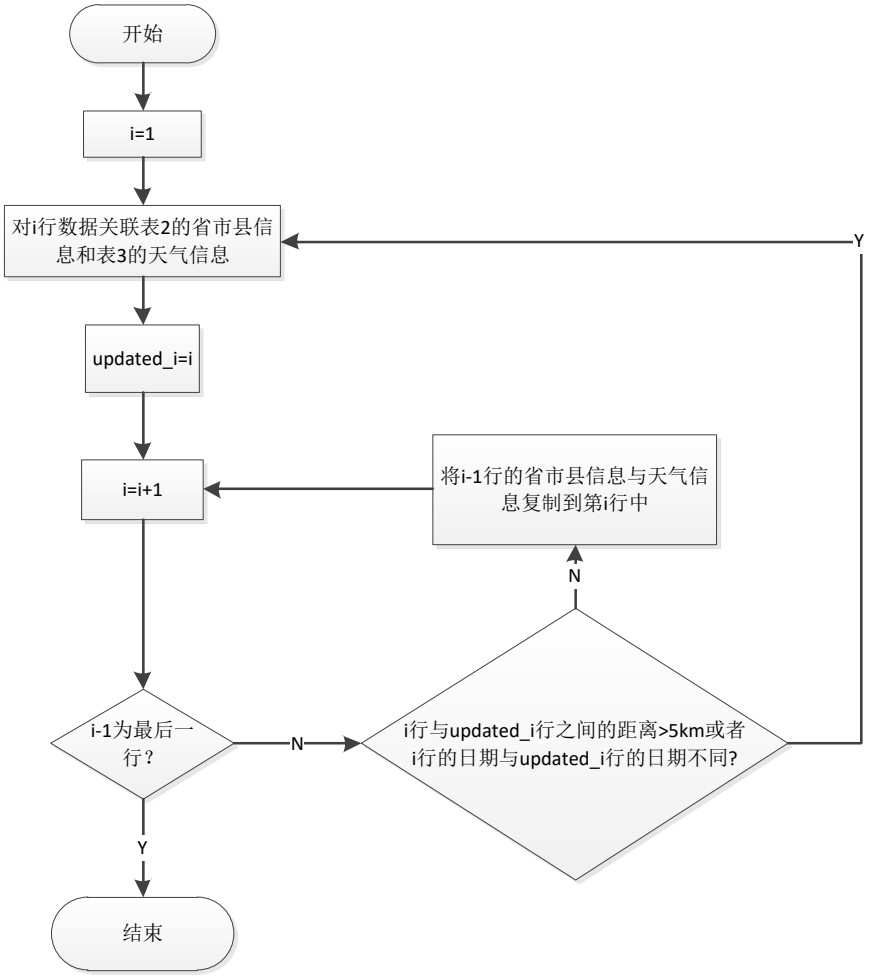


图 7 三表关联流程图

简而言之，就是以 5 公里为界，如果和上次更新地理位置和天气信息的记录相隔了这个界限或者日期有变化了，则去检索更新天气信息，并标记为更新记录。否则与上一条记录保持一致。这样的关联数据的算法既保证了结果的准确性，又大大的节省了时间上的开销。

表 4 对数据集成后的数据进行了部分行的展示。

表 4　AA00251 数据集成后的数据格式

vehicleplate number	…	mileage	province	prefecture_ city	county	wind_ direction	wind_ power	high_ temp	low_ temp	conditions	relative_ humidity	precipitation
AA00251	…	23755	江西省	鹰潭市	贵溪市	45	2	28	20	多云	77%	0
AA00251	…	23755	江西省	鹰潭市	贵溪市	45	2	28	20	多云	77%	0

4. 驾驶行为指标判断标准

通过结合天气信息，本文提出了计算共计 17 个指标的评判标准，它们可以严格对应到安全、效率和能耗的影响因素中去。

这 17 个指标分别是：急加速次数、急减速次数、疲劳驾驶次数、怠速预热次数、超长怠速次数、熄火滑行次数、超速次数、急转弯次数、汽车是否在报废里程内、汽车运行时的平均速度、行驶时的速度稳定性、低能见度时超出限速次数、侧面大风时高速行驶时长、八级及以上大风时驾驶时长、恶劣天气驾驶速度过高次数、逆风时高速驾驶时长、非经济车速比例。分别计算出各个指标的具体情况将为后续的驾驶行为评分模型提供数据来源。

4.1 急加速

急加速描述的是车辆起步阶段或行驶过程中猛踩加速踏板的过程,急加速行为不但容易造成追尾事故的发生，而且在能耗、乘客舒适度和货物完整程度上考虑都是不利的。急加速的判别标准依照的是车辆加速度的大小，当车辆的加速度超过阈值 A 时则认定为一次急加速行为。根据文献[3]，A 的取值为：$A=3\mathrm{m/s^2}$。并且，通过分析本次研究中 449 辆车辆的行驶数据，能够发现大量的记录时间跳跃的情况。为了确保计算结果的准确性，在这里，对前后两条记录间的时间间隔做出限制：$0<\triangle T\leq 3\mathrm{s}$。急加速行为识别与计算的流程图如下图所示。

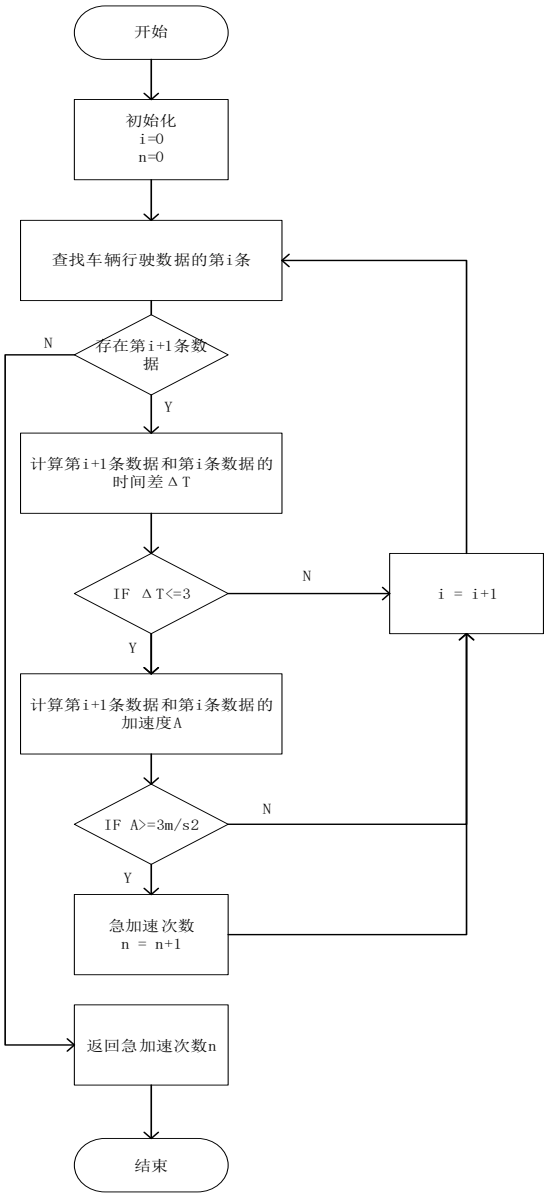


图 8 计算急加速行为算法流程图

急加速行为判别算法主要流程步骤如下：

1) 获取须处理车辆设备采集的所有数据，初始化索引 i=0，急加速次数 n=0；

2) 取第 i 行和第 i+1 行的数据，且判断是否存在第 i+1 行的数据(即是否是数据的最后一行)，若否，则执行步骤 3；若是，则执行步骤 6。

3) 对第 i 行数据和第 i+1 行的数据时间间隔进行计算，计算两条数据时间间隔ΔT，若时间间隔$0 < \Delta T \leq 3s$，若是，则执行步骤 4；若否，则执行 i 自增 1，并返回执行步骤 2，选取下一行的数据。

4) 计算第 i 行和第 i+1 行数据的加速度，用两条数据的速度差除以ΔT，计算出加速度 A,,若加速度 A>=$3m/s^2$,若是，则执行步骤 5，若否，则执行 i 自增 1，并返回执行步骤 2，选取下一行的数据。

5) 满足了时间阈值的条件和加速度条件，则急加速次数增加一次，n=n+1。然后返回执行步骤 2，选取下一行的数据。

6) 若程序执行到最后一行，已经处理完整个文件，将急加速次数 n 输出。结束急加速行为的判断。

4.2 急减速

与急加速相反，急减速描述的是车辆行驶过程中猛踩刹车制动的行为,急减速的判别标准同样是依照车辆加速度的大小，当车辆的加速度小于阈值 A′时则认定为一次急减速行为。根据文献[3]， $A'=-3m/s^2$ 。

除了加速度阈值的设定外，急减速行为的计算与急加速行为的计算流程相一致。

4.3 疲劳驾驶

疲劳驾驶是一种及其严重的危险驾驶行为，对人身和财产安全带来极大的危害，特别容易发生在运输企业之中。根据《道路交通安全法实施条例》第六十二条相关规定和文献[4]中给出的定义，在这里将疲劳驾驶的评判标准设置为：在 24 小时内车辆连续驾驶时长超过 8 小时，或者是连续驾驶车辆超过 4 小时而休息时间少于 20 分钟。

4.4 怠速预热

怠速预热即为车辆在每次行驶前，司机发动汽车后以怠速在原地热车，等上几分钟再开车上路，原地热车会导致发动机积碳，影响发动机性能，进而危害到驾驶安全。根据文献[3]和[5]，判别一辆车是否怠速预热，我们通过设置时间阈值 T=120s，当车辆在发动后超过时间阈值 T 仍未起步，则判别为一次怠速预热。

其主要流程如下：

1) 获取须处理车辆设备采集的所有数据，初始化索引 i=0，时间 T 为空，怠速预热次数 n=0。

2) 查询车辆数据的第 i 条，进行下一步判断。

3) 判断是否为最后一条数据，若是，执行步骤 12；若否，则执行步骤 4。
4) 判读是否存在第 i-1 条数据，即考虑是否为第一条数据，若是，执行步骤 5；若否，执行步骤 9。
5) 读取第 i 条数据，判断当前的 acc_state 状态是否为 1，且当且速度是否为 0。若是，执行步骤 6；若否，执行步骤 7。
6) 将第 i 条数据的时间追加到时间列表 T 中，然后对 i 自增 1，返回执行步骤 5。
7) 当第 i 条数据不满足 acc_state 为 0 且速度为 0，则计算时间列表 T 中的各个时间，计算出累计时间差 ΔT。
8) 判断计算出的 ΔT 是否大于 120s,若是，对怠速预热次数 n=n+1，然后清空时间列表，对 i 自增 1，执行步骤 2；若否，将时间列表清空，对 i 自增 1，返回执行步骤 2。
9) 判断第 i-1 条数据的 acc_state 状态是否为 0。若是，则执行步骤 10；若否，则清空时间列表 T，对 i 自增 1，返回执行步骤 2。
10) 读取第 i 条数据,判断第 i 条数据 acc_state 状态是否为 1 且速度为 0，若是，执行步骤 11；若否，则执行步骤 8。
11) 将第 i 条数据的时间追加到时间列表 T 中，并对 i 自增 1，返回执行步骤 10。
12) 若第 i 条数据为最后一条数据,于是将怠速预热次数 n 输出。程序结束。

4.5 超长怠速

超长怠速和怠速预热在概念上很容易相混淆，怠速预热是发生在车辆的启动阶段，而超长怠速是车辆处于行驶过程中产生的怠速，即 acc 状态不发生改变。根据文献[3]和[5]，判别一辆车是否超长怠速，我们同样可设置时间阈值 T=120s 来判断。

主要判断流程如下：

1) 获取须处理车辆设备采集的所有数据，初始化索引 i=0，时间 T 为空，超长怠速次数 n=0。
2) 查询第 i 条数据,进行下一步的判断。
3) 判断是否为最后一条数据，若否，则执行步骤 4；若是，执行步骤 12。

4) 判断是否存在第 i-1 条数据，若否，执行步骤 5；若是，执行步骤 6。

5) 读取第 i 条数据，判断 acc_state 是否为 1 且速度为 0，若是，对 i 自增 1，继续执行读取第 i 条数据进行判断；若否，则对 i 自增 1，返回执行步骤 2。

6) 判断第 i-1 条数据 acc_state=0，若是，则执行步骤 7；若否，执行步骤 8。

7) 读取第 i 条数据，判断第 i 条数据 acc_state=1 且速度为 0，若是，则对 i 自增 1，继续读取第 i 条数据进行判断；若否，返回执行步骤 11。

8) 进行下一个判断，判断第 i 条数据 acc_state=1 且速度为 0，若是，执行步骤 9；若否，执行步骤 11。

9) 读取第 i 条数据，判断第 i 条数据 acc_state=1 且速度为 0，若是，则对 i 自增 1，继续读取第 i 条数据进行判断；若否，则执行步骤 10。

10) 计算时间列表中的累计时间差 ΔT。若 ΔT>120,则超长怠速次数 n=n+1；若否，则执行步骤 11。

11) 对 i 自增 1，并将时间列表 T 清空，返回执行步骤 2。

12) 第 i 行为最后一行数据，则将超长怠速的次数 n 输出。程序结束。

4.6 熄火滑行

熄火滑行即为在行驶过程中，车辆熄火后仍按照惯性向前滑行一段距离。对于采用了助力转向系统的车辆而言，熄火滑行会使转向盘变重，难以控制；对于采用真空助力刹车系统的车辆而言，熄火滑行会使刹车系统失效。

该算法流程图如下图所示：

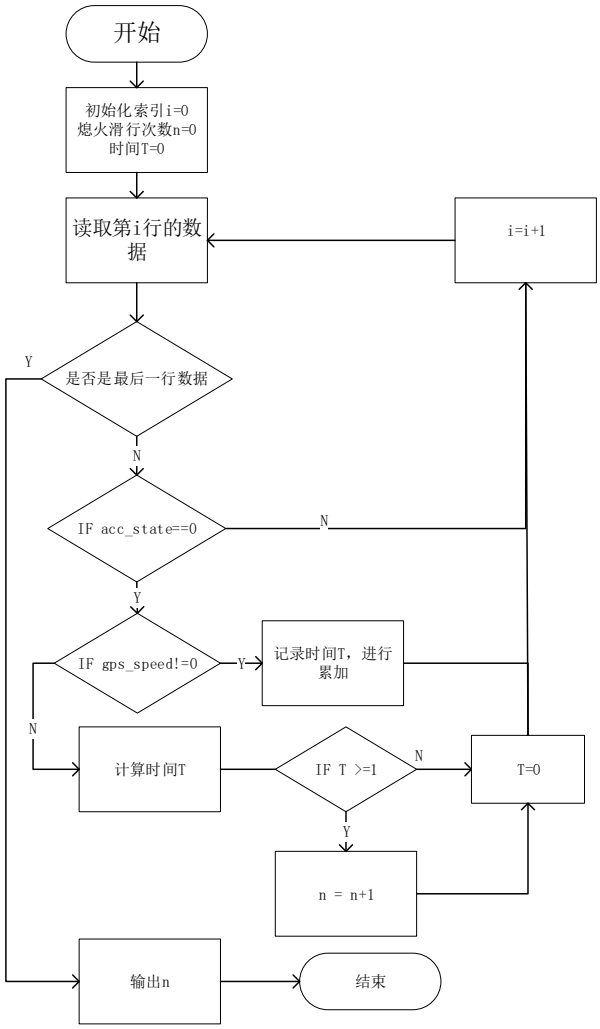


图 9 计算熄火滑行行为算法流程图

4.7 超速

超速即为车辆超过当前行驶路段的最高时速限制。根据文献[4]对 2007 年至 2010 年间发生的货车重特大交通事故分析，因超速而引起的重特大交通事故占所有因素中的 17.65%，位居超载、故障车上路之后的第三影响因素。不过由于缺少车辆行驶路段的具体信息，在这里根据《道路交通安全法》中关于大客车在高速上的限速规定设置了一个阈值 V=100km/h。为了避免卫星定位漂移带来的不必要误判，还需要设置一个时间阈值 T=3s，只有当速度超过 V 且持续时间超过 T 时才会计算一次超速行为。

判别超速行为的流程如下图所示。

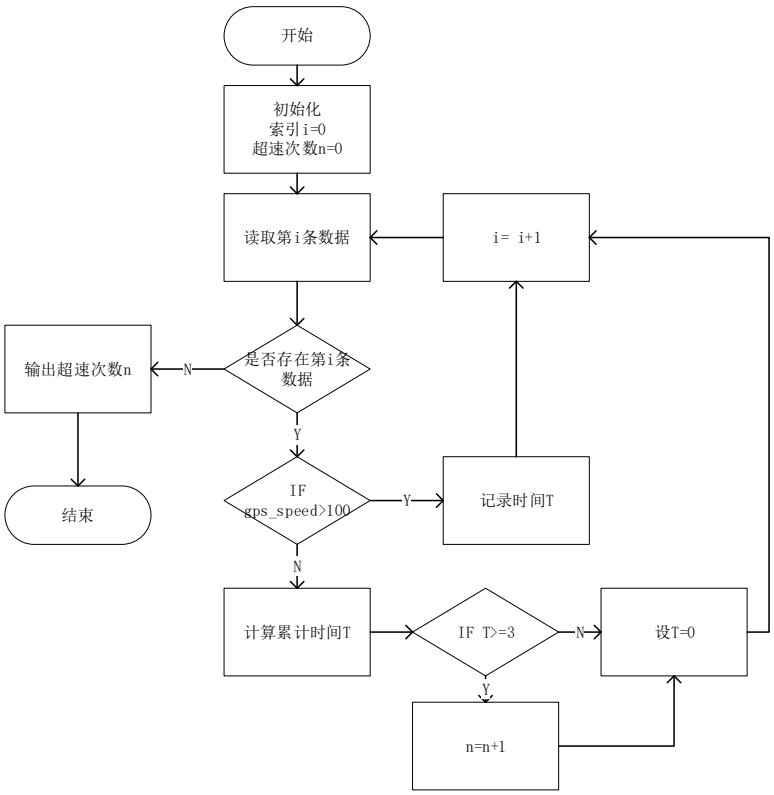


图 10 计算超速行为算法流程图

主要流程算法如下：

1) 获取须处理车辆设备采集的所有数据，初始化索引 i=0，超速次数 n=0。

2) 读取第 i 条数据，并进行下一步判断。

3) 判断是否存在第 i 条数据，即判断数据是否来到数据末尾。若否，执行步骤 4；若是，执行步骤 6。

4) 判断第 i 条的速度是否大于 100km/h,若是，则记录时间 T，对 i 自增 1，返回执行步骤 2；若否，执行步骤 5。

5) 判断累计时间是否超过 3s,若是，超速次数 n=n+1,然后将时间 T 清零，对 i 自增 1，返回执行步骤 2；若否，将时间清零，对 i 自增 1，返回执行步骤 2。

6) 输出超速次数 n,程序结束。

4.8 急转弯

急转弯也是一种危险驾驶行为，特别是对于运输车辆，容易造成车辆侧翻或是发生侧面碰撞。根据文献[6]，本文将急转弯的速度阈值设为 20km/h，行车方向改变的角度阈值根据常规的十字交叉路口设为 90° 。

其计算流程如下：

1) 获取须处理车辆设备采集到的所有数据,初始化索引 i=0,急转弯次数 n=0。
2) 读取第 i 行数据，进行下一步判断。
3) 判断是否存在第 i+1 行数据，若是，则执行步骤 4；若否，则执行步骤 6；
4) 计算第 i+1 行和第 i 行的时间差ΔT 和方向角差ΔD。
5) 判断是否同时满足时间间隔ΔT 等于 1s，方向角变化ΔD 大于 90 度，车辆速度大于 20km/h。若是，则急转弯次数 n=n+1,然后对 i 自增 1，返回执行步骤 2；若否，对 i 自增 1，返回执行步骤 2。
6) 读取到数据最后一行，则将急转弯次数 n 输出，结束程序。

4.9 超出报废里程

当车辆行驶总里程超过一定的范围,其车辆各方面的稳定性都会大大不如以前，若再继续行驶，会有较大的安全隐患。根据《机动车强制报废标准规定》，将本文研究的运输车辆行驶总里程限制为 40 万公里。

4.10 运行时的平均速度

本文将车辆运行状态下的平均速度也纳入了考虑,其值能够很好的说明效率与能耗方面的情况。

4.11 车速稳定性

车速稳定性是指车辆在行驶过程中的速度稳定程度,用于刻画行驶时速度的波动情况。当车速稳定性过低时，不但会影响到乘客的舒适性与货物的完整性还会增加行驶油耗。本文借助了标准差公式来体现出车速的稳定性,具体公式如下：

$$s = 1 - f\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(v_i - \bar{v})^2}\right) \tag{3}$$

其中，s 为计算出的稳定性，$f()$ 为归一化函数，N 为记录的行驶速度的个数，v_i 为记录的第 i 个速度，$\bar{v}$ 为所有行驶速度的平均数。

4.12 低能见度时超出限速

在低能度低时，车辆速度过快会因为视力受限而极容易发生追尾事故。根据文献[7]中提出的公式（4），可以通过降水量大致计算出能见度。

$$VBS = 13410r^{-0.66} \quad (4)$$

其中，VBS为能见度（单位：m），r为降水量（单位：mm/h）。

根据《中华人民共和国道路交通安全法实施条例》第 81 条规定，在这里将速度阈值按照如下要求设置：

1） 能见度小于 200 米时，车速不得超过每小时 60 公里；

2） 能见度小于 100 米时，车速不得超过每小时 40 公里；

3） 能见度小于 50 米时，车速不得超过每小时 20 公里。

4.13 侧风高速

在侧向风力过大时，从安全的角度上考虑，应该降低车速，保持车身稳定驾驶。在这里对侧风高速行驶做出如下定义：在风力大于等于 6 级且风向与汽车行驶方向垂直的情况下，汽车速度大于等于 70km/h。

4.14 大风行驶

当风力级别过大时，驾驶员应当将车就近停至室内停车场，不应该继续驾驶机动车。对于大风行驶的定义是：风力级别大于等于 8 级的情况下继续驾驶。

4.15 恶劣天气驾驶速度过高

在恶劣的雨雪天气下驾驶需要格外小心。这种天气下往往会视力有限、路面湿滑，同时路上行人与非机动车视力同样不佳，所以降低车速是格外重要的。本文针对不同的雨雪天气，设置了不同的速度阈值：大暴雪及暴雪天的速度阈值为 30km/h;大雪天的速度阈值为 40km/h;中雪天的速度阈值为 50km/h;暴雨天的速度阈值为 50km/h;大雨天的速度阈值为 60km/h;中雨天的速度阈值为 80km/h。

4.16 逆风高速

当风速级别过大时，车辆逆风高速行驶会因为车速过高而导致风的阻力加大，消耗的燃油大都用来克服风的阻力了，这样会大大地增加油耗。本文给出的逆风高速的标准是：在风力大于等于 6 级且风向与行驶方向相反的情况下，车速大于等于 70km/h。

4.17 非经济车速比例

当汽车的机械运转状况最好、各种部件配合最默契，发动机、变速箱以及各方面的运转最和谐的时候，汽车的油耗也是最经济的，这个时速就是“经济时速”[8]。参考多篇文献，发现经济车速的区间并没有一个统一的标准，这是因为车型不同，所对应的“经济时速”也有所不同。本文将车速在[70,90]之间的速度设置为经济车速。

5. 驾驶行为评分模型

根据驾驶车辆的 OBD 数据和天气数据结合后进行的挖掘计算，能够得到共计 17 个能反映驾驶员行为优劣的指标：急加速次数、急减速次数、疲劳驾驶次数、怠速预热次数、超长怠速次数、熄火滑行次数、超速次数、急转弯次数、汽车是否在报废里程内、汽车运行时的平均速度、行驶时的速度稳定性、低能见度时超出限速次数、侧面大风时高速行驶时长、八级及以上大风时驾驶时长、恶劣天气驾驶速度过高次数、逆风时高速驾驶时长、非经济车速比例。

对十七个指标进行总结分析后，可以严格的对应到安全、效率和能耗三类中。其中：

1) 安全：急加速次数、急减速次数、疲劳驾驶次数、怠速预热次数、超长怠速次数、熄火滑行次数、超速次数、急转弯次数、汽车是否在报废里程内、行驶时的速度稳定性、低能见度时超出限速次数、侧面大风时高速行驶时长、八级及以上大风时驾驶时长、恶劣天气驾驶速度过高次数。
2) 效率：超长怠速次数、汽车运行时的平均速度。
3) 能耗：急加速次数、急减速次数、怠速预热次数、超长怠速次数、行驶时的速度稳定性、逆风时高速驾驶时长、非经济车速比例。

以上一章的各个指标判断标准为依据，现在要依照各辆车的各项指标数值来构建安全评价模型和综合评价模型。基于多参数理论，采用加权的方法来构建驾驶行为安全评价模型：

$$z = f(x,p) = 100 - \sum_{1}^{14}(x_i \times p_i) \tag{5}$$

其中，z 为计算出的安全评分，x_i 为涉及到安全评价的第 i 个指标的值，p_i 为涉及到安全评价的第 i 个指标所对应的权重。

同样地，可以构建驾驶行为综合评价模型：

$$h = f(y,p) = 100 - \sum_{1}^{17}(y_i \times p_i) \quad (6)$$

其中，h为计算出的综合评分，y_i为所有评价指标中的第 i 个指标的值，p_i为所有评价指标中第 i 个指标所对应的权重。

至此，问题的关键就是权值的选取了。权重选取的合适与否将直接影响着最终结果的正确性。

目前,权重确定方法主要分为定量和定性两大类，其中，定量法有熵权法、灰色关联度法、人工神经网络定权法、因子分析法、回归分析法和路径分析法，定性法有德尔菲法、层次分析法、模糊聚类法和比重法[9]。目前使用最多的是层次分析法和熵权法来确认各个指标的权重,例如文献[10]将层次分析法运用到城市电网评估中权重的确定上，文献[11]通过熵权法完成了对水质的综合评价。

层次分析法,是指将与决策总是有关的元素分解成目标、准则、方案等层次，在此基础之上进行定性和定量分析的决策方法。该方法是美国运筹学家匹茨堡大学教授萨蒂于 20 世纪 70 年代初，在为美国国防部研究"根据各个工业部门对国家福利的贡献大小而进行电力分配"课题时，应用网络系统理论和多目标综合评价方法，提出的一种层次权重决策分析方法。它的主要原理是把问题分成若干层次，在每一层次上将人的主观思想数量化，通过加权和的方法计算出各个指标的权重[12]。

虽然层次分析法中将人的主观思想数量化了,但其实质仍是一个依据人的主观判断的定性方法，并且当待确定权重的指标数量过多时，必然会出现各指标间的重要程度判断混乱，矩阵尺寸庞大，计算过程复杂，时间花销大。

熵权法是利用指标的熵值,也就是信息量来计算指标权重的一种客观赋权方法。指标间的差异程度越大，其权重越大。熵权法的客观性和适应性强，能很好的解释权重结果[13]。

熵权法在体现了它的客观性的同时，却牺牲了结果的准确性，这是因为它的结果主要表现和突出局部差异，这使得某些非重要指标的权重过大。

针对本次任务研究数据指标特点，同时要确保评价结果的准确性，本文采用BP(back propagation)神经网络为权重分配模型来确定权重。

它既能克服层次分析法中判断指标重要程度过程混乱和难以应对多指标计算的问题又能解决熵权法的过度依赖数据差异和准确性不高的缺点。

5.1 权重分配的 BP 神经网络模型

首先，需对指标列进行归一化处理：

$$X=[x_1,x_2,\cdots,x_n],x_i\in[0,1],i=1,2,\cdots,n \tag{7}$$

它们对应的决策结果为：

$$Y=[0,100] \tag{8}$$

设 X 和 Y 之间存在一种映射关系 G，则有：

$$Y=G(X) \tag{9}$$

将各指标列 $[X_1,X_2,\cdots,X_n]$ 作为 BP 网络的输入，决策结果 Y' 作为 BP 网络的输出，利用 BP 网络的学习算法对 X 和 Y 组成的样本集进行训练，当网络收敛后，可实现 X 和 Y 之间的映射关系 H，且满足：

$$|H(X)-G(X)|<\varepsilon \tag{10}$$

其中，ε 为任意小的正数。考虑到每个神经元的输入对该神经元的作用体现在连接权的大小上，因此，可以预见 X 对 Y 的影响作用也必然体现在连接权的大小上。于是可以建立如下的三层 BP 网络模型，其拓扑图如图 11 所示[14]。

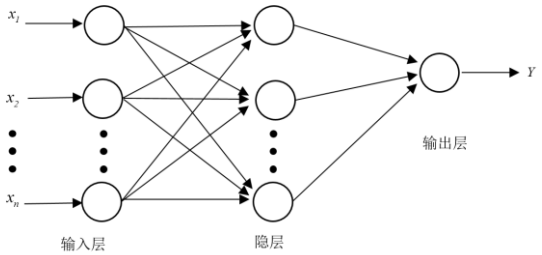


图 11 三层 BP 神经网络拓扑结构

5.2 权重分配的 BP 神经网络原理

输入层中的 n 个单元对应了 n 个指标，输出层的 Y 则代表了最终的决策结果，而隐层包含了 r 个单元，用 w_{ih} 表示输入层单元 i 与隐层单元 h 之间的连接

权，用V_{hj}表示隐层单元 h 与输出层单元 j 之间的连接权，那么隐层单元 h 的输出b_h为：

$$b_h = f(\sum_{i=1}^{n} W_{ih} x_i + \theta_h), h = 1,2,\cdots,r \tag{11}$$

式中， $f()$ 为 Sigmoid 函数，θ_h 为第 h 单元的阈值。

同样地，输出层中第 j 单元的输出 y_j 为：

$$y_j = f(\sum_{h=1}^{r} V_{hj} b_h + e_j), j = 1,2,\cdots,m \tag{12}$$

其中， e_j 为输出层中第 j 单元的阈值。

由公式（11）和（12）可以得到 x_i 对 y_j 的灵敏度为：

$$\frac{\partial y_j}{\partial x_i} = \frac{\partial y_j}{\partial b_1} \bullet \frac{\partial b_1}{\partial x_i} + \frac{\partial y_j}{\partial b_2} \bullet \frac{\partial b_2}{\partial x_i} + \cdots + \frac{\partial y_j}{\partial b_r} \bullet \frac{\partial b_r}{\partial x_i} \tag{13}$$

令

$$Ob_h = \sum_{i=1}^{n} W_{ih} x_i + \theta_h \tag{14}$$

$$Oy_j = \sum_{h=1}^{r} V_{hj} b_h + e_j \tag{15}$$

由公式（12）可得：

$$\frac{\partial y_j}{\partial b_i} = f^1(Oy_j) \mathrm{V}_{1j} \tag{16}$$

$$\frac{\partial y_j}{\partial b_2} = f^1(Oy_j) \mathrm{V}_{2j} \tag{17}$$

$$\frac{\partial y_j}{\partial b_r} = f^1(Oy_j) \mathrm{V}_{rj} \tag{18}$$

$$\frac{\partial b_1}{\partial x_i} = f^1(Ob_1) \mathrm{W}_{i1} \tag{19}$$

$$\frac{\partial b_2}{\partial x_i} = f^1(Ob_2) \mathrm{W}_{i2} \tag{20}$$

$$\frac{\partial b_r}{\partial x_i} = f^1(Ob_r) \mathrm{W}_{ir} \tag{21}$$

其中，对 Sigmoid 函数求导可以得到：

$$f^1(u)=\frac{e^{-u}}{(1+\mathrm{e}^{-u})^2} \tag{22}$$

由公式（13）和（16）~（22）可得：

$$\frac{\partial y_j}{\partial x_i}=f^1(Oy_j)[f^1(Ob_1)\mathrm{V}_{1j}\mathrm{W}_{i1}+f^1(Ob_2)\mathrm{V}_{2j}\mathrm{W}_{i2}+\cdots+f^1(Ob_r)\mathrm{V}_{rj}\mathrm{W}_{ir}] \tag{23}$$

同样地，可以得到 x_k 对 y_j 的灵敏度为：

$$\frac{\partial y_j}{\partial x_k}=f^1(Oy_j)[f^1(Ob_1)\mathrm{V}_{1j}\mathrm{W}_{k1}+f^1(Ob_2)\mathrm{V}_{2j}\mathrm{W}_{k2}+\cdots+f^1(Ob_r)\mathrm{V}_{rj}\mathrm{W}_{kr}] \tag{24}$$

由公式（23）和（24）可以得出：

$$\left|\frac{\partial y_j}{\partial x_i}\right|-\left|\frac{\partial y_j}{\partial x_k}\right|=f^1(Oy_j)[f^1(Ob_1)\left|V_{1j}\right|(|W_{i1}|-|W_{k1}|)+f^1(Ob_2)\left|V_{2j}\right|(|W_{i2}|-|W_{k2}|)+\cdots+f^1(Ob_r)\left|V_{rj}\right|(|W_{ir}|-|W_{kr}|)] \tag{25}$$

将 X 和 Y 组成的训练集作为 BP 神经网络的学习样本进行训练，设 W_{ih} 和 W_{kh} 分别为 x_i 和 x_k 对应的输入单元 i 和 k 与隐层单元 h 之间的连接权系数，如果 $|W_{i1}|>|W_{k1}|,|W_{i2}|>|W_{k2}|,\cdots,|W_{ir}|>|W_{kr}|$，则系数 x_i 的灵敏度 S_i 比系数 x_k 的灵敏度 S_k 要大。

这是因为当网络收敛后，一般有：

$$f^1(Ob_1)\left|V_{1j}\right|\approx f^1(Ob_2)\left|V_{2j}\right|\approx\cdots f^1(Ob_r)\left|V_{rj}\right| \tag{26}$$

依照这种判断灵敏度大小的方法，可以对各个指标的权重进行求解。

令

$$|W_{si}|=|W_{i1}|+|W_{i2}|+\cdots+|W_{ir}| \tag{27}$$

假定 n 个指标的权重分配为 $\lambda_1,\lambda_2,\cdots,\lambda_n$，并且 $\sum_{i=1}^{n}\lambda_i=1$，可以得到：

$$\lambda_i=\frac{|W_{si}|}{\sum_{j=1}^{n}|W_{sj}|},i=1,2,\cdots,n \tag{28}$$

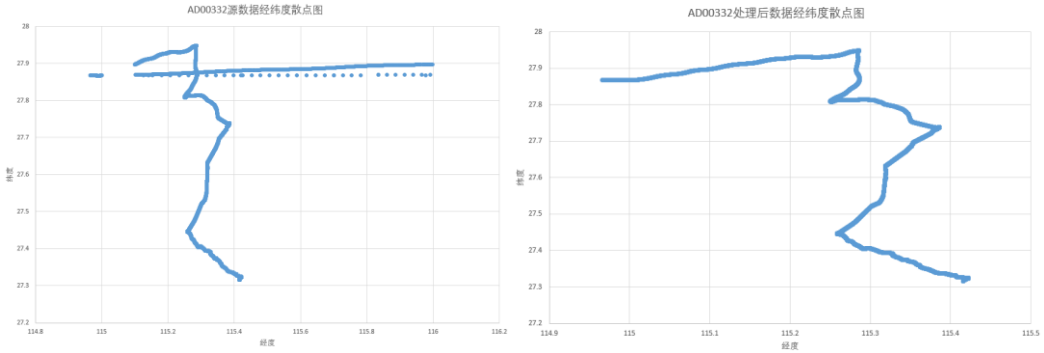
其中，$|W_{si}|=|W_{i1}|+|W_{i2}|+\cdots+|W_{ir}|$ 为网络收敛时的系数 x_i 所对应的输入单元 i 和所有隐层单元之间连接权值的绝对值之和[15]。

6. 实验与结果

在整体流程及其具体步骤确认了后，接下来将对手上的数据进行实际的实验操作。

6.1 数据预处理

首先，根据第 3 章的设计思想，对所有车辆的驾驶数据进行预处理，完成数据去重、异常值修正与删除、数据格式统一化，并且通过坐标地理数据实现对天气信息的集成。在第 3 章中表 4 展示了预处理过后的数据格式，图 12 将使用经纬度散点图模拟汽车轨迹，展示预处理前后经纬度坐标的变化。



（a）源数据散点图　　（b）预处理后数据散点图

图 12 AD00332 数据预处理前后散点图

6.2 车辆轨迹绘制

下一步，使用 folium 地理可视化库实现了对车辆轨迹的静态绘制，并划分出了车辆的行驶路段，对行驶过程中的急加速、急减速、行驶路程和平均行车速度进行了展示。

然而，静态轨迹并不能完全的体现汽车行进动向，无法获取车辆行进的实时动态。于是，进一步地，通过将经纬度坐标转至 BD-09 标准并调用百度地图 API 实现了车辆轨迹的动态绘制。图 13 和图 14 分别为动态轨迹页面与静态轨迹页面中的部分区域截图。

图 13 AB00006 动态轨迹页面截图

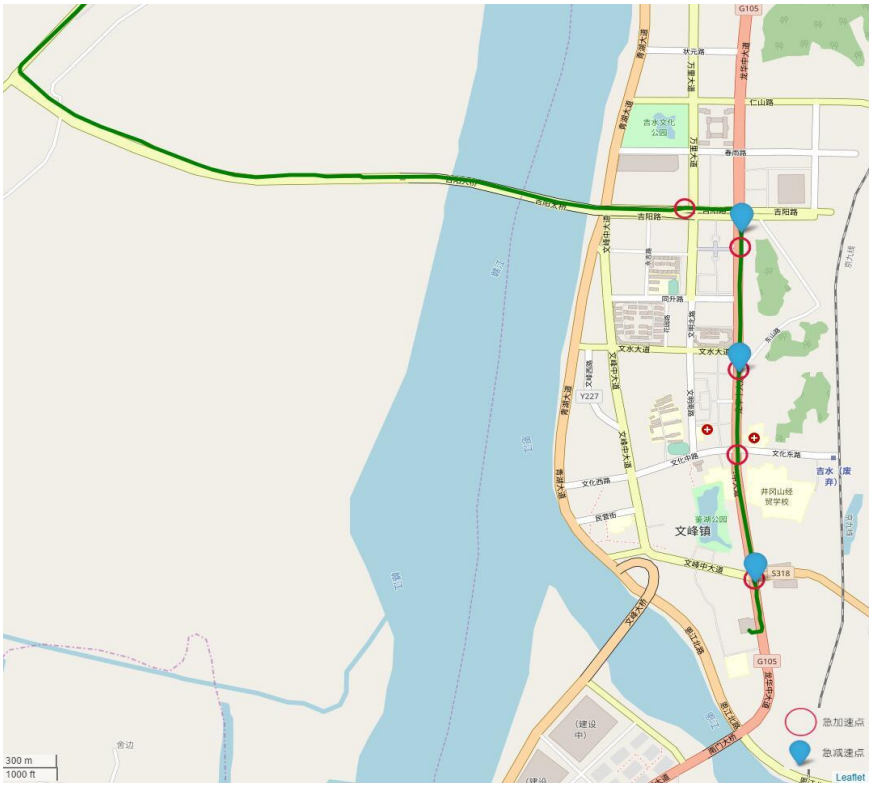


图 14 AD00003 静态轨迹页面截图

仔细观察轨迹图，容易发现急加急减速大多是相伴出现的，并且绝大多数都出现在路口处。从安全、乘客的舒适性、货物的完整性以及节约能耗的角度上考虑，这种行为是需要引起驾驶员警惕的。

6.3 驾驶行为挖掘

接下来，根据第 4 章的驾驶行为标准，对所有车辆的 17 个指标进行了统计汇总，最终的汇总文件格式如下表所示。

表 5 驾驶行为汇总表部分行

Vehicleplate number	rapid_ times	deceleration_ times	fatigueDriving_ times	idlePreheating_ times	overlongIdle_ times	coastingEngineoff_ times	speeding_ times	…	diseconomicSpeed_ rate
AA00001	408	437	1	0	77	0	0	…	0.832243
AA00002	188	190	1	0	30	0	90	…	0.771821
AA00004	235	232	0	7	15	0	0	…	0.687519
AA00036	181	172	0	1	26	0	0	…	0.756099

17 个驾驶行为指标分别为：急加速次数、急减速次数、疲劳驾驶次数、怠速预热次数、超长怠速次数、熄火滑行次数、超速次数、急转弯次数、汽车是否在报废里程内、汽车运行时的平均速度、行驶时的速度稳定性、低能见度时超出限速次数、侧面大风时高速行驶时长、八级及以上大风时驾驶时长、恶劣天气驾驶速度过高次数、逆风时高速驾驶时长、非经济车速比例。

6.4 构造 BP 神经网络进行评分

在一开始，需要对驾驶行为汇总表进行数据划分。由于采用专家打分作为训练集，如果打分记录过多会导致评价标准混乱，所以将 449 条数据分成了 9 份（每份约 50 条记录），按照 2:7 的比例划分为了两个部分：训练集 99 条，测试集 350 条。专家打分后其格式如下表所示：

表 6 专家打分后的数据格式

Vehicleplate number	rapid_ times	deceleration_ times	fatigueDriving_ times	…	diseconomic Speed_rate	safe score	total score
AA00001	408	437	1	…	0.832243	71	67.5
AA00002	188	190	1	…	0.771821	65	64.25
AA00004	235	232	0	…	0.687519	77	76
AA00036	181	172	0	…	0.756099	79	75.5

依照第 5 章 BP 网络分配权重的原理，通过训练 3 层 BP 网络实现对各个驾驶行为确定权重并对每辆车辆完成评分。

在安全评分中，共涉及 14 项指标：急加速次数、急减速次数、疲劳驾驶次数、怠速预热次数、超长怠速次数、熄火滑行次数、超速次数、急转弯次数、汽车是否在报废里程内、行驶时的速度稳定性、低能见度时超出限速次数、侧面大风时高速行驶时长、八级及以上大风时驾驶时长、恶劣天气驾驶速度过高次数。综合评分时要涉及到所有的 17 个指标，除了安全指标已经列出的 14 个之外，还有：汽车运行时的平均速度、逆风时高速驾驶时长、非经济车速比例。

因此在训练安全评分模型时，输入层神经元个数为 14。训练综合评分模型时，输入层神经元个数为 17。隐层神经元个数 m 按照经验公式（29）选取。

$$m = n + 0.618(n - t) \tag{29}$$

其中，n 为输入层神经元的数目，t 为输出层神经元的数目。

网络模型的其他参数分别如下：网络学习效率为 0.1，最大迭代次数为 10000 次，最大期望误差为 0.1。

需要强调的是，在输入训练集时，需要将各个指标进行涉及训练集与测试集所有数据的归一化处理，使他们限制在[0,1]的区间中，这是为了使这些指标在求权重之前能够保持同等重要性，并且加快收敛速度。在这里使用的是最大最小值归一化法，其核心思想是公式（30）。

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{30}$$

其中，x' 为归一化后的值，x 为待归一化的值，$x_{\max}$ 为当前指标列中的最大值，$x_{\min}$ 为当前指标列中的最小值。

在训练完成后的网络模型中输入测试集数据，得到每辆车相应的评分，评价计算结果如表 7 所示。

表 7 评价计算结果部分展示

Vehicleplate number	rapid_ times	deceleration_ times	fatigueDriving_ times	idlePreheating_ times	…	diseconomic Speed_rate	safescore byPredict	totalscore byPredict
AB00365	91	94	0	2	…	0.476636	79.57843	78.69113
AB00370	98	110	0	1	…	0.999777	78.97718	73.55906
AB00380	110	102	0	0	…	1	79.42409	74.04497
AB00386	157	164	3	6	…	0.779366	41.18823	33.49892

最后，选取了 20 条模型打分结果与专家打分结果进行了对比。结果如图 15 所示。

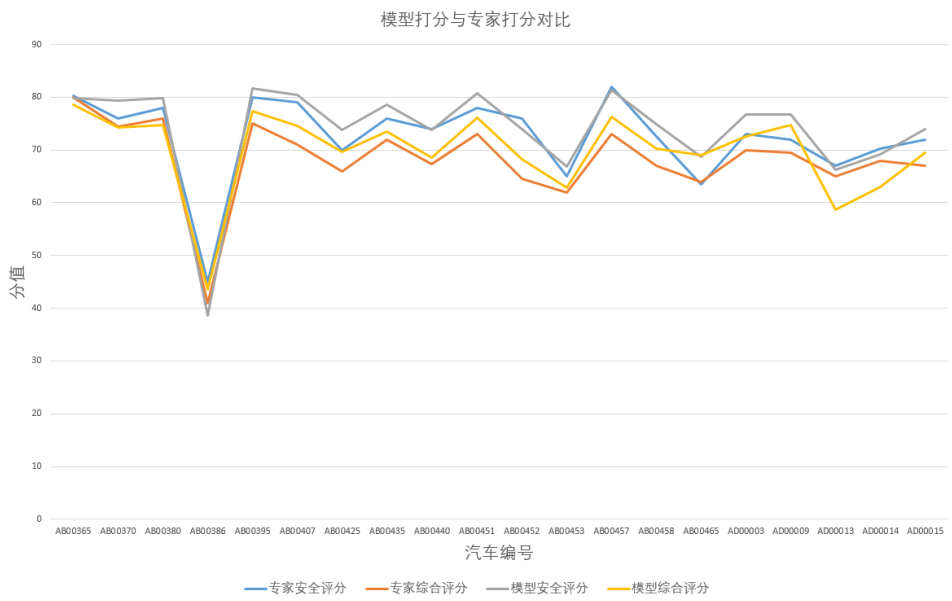


图 15 模型打分结果与专家打分结果对比

结果显示，模型打分与专家评价的结果高度一致，通过验证计算得到本模型安全评分的平均偏差为2.235271035分，综合评分的平均偏差为2.754255789分。

7. 数据可视化与分析

7.1 得分统计与分析

将449辆车驾驶行为导入评分模型中，得出的各个车辆对应的驾驶员的得分汇总情况如图16所示。

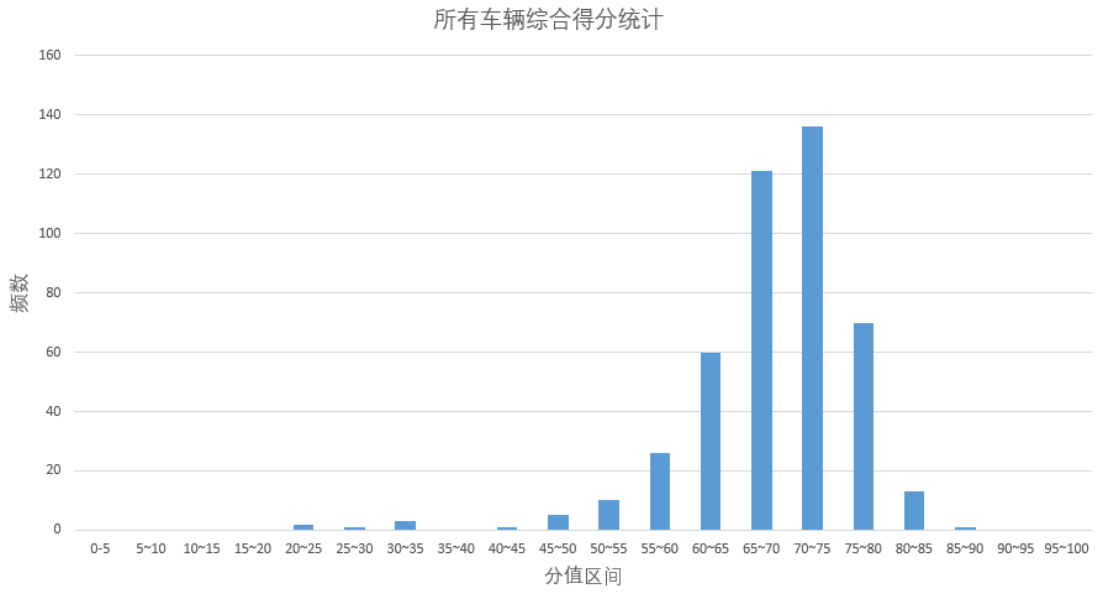


图 16 所有车辆综合得分统计

通过观察，发现得分情况大致符合负偏态分布，大部分驾驶员得分集中在区间[60,80]内。在有频数的最高区间[85,90]内只有 1 辆车，但却有 2 辆车驾驶行为评分在[20,25]区间内，1 辆车驾驶行为评分在[25,30]内，3 辆车驾驶评分在[30,35]区间内。

为了对评分最低的六辆车的最终评分进行验证，仔细分析它们的行为统计结果，如下表 8 所示。

表 8 低分车辆驾驶行为

vehicleplate number	rapid_ times	deceleration_ times	fatigueDriving_ times	idlePreheating_ times	overlongIdle_ times	speeding_ times	…	average_ speed	diseconomic Speed_rate	total score
AD00050	252	255	5	6	23	549	…	76.43174	0.624188	20.3271
AD00292	289	295	4	9	21	501	…	69.84067	0.711437	21.28982
AD00320	96	93	0	10	10	458	…	77.23121	0.730658	25.99487
AD00369	81	80	0	9	6	367	…	77.67321	0.901801	33.17816
AB00386	157	164	3	6	15	319	…	70.03252	0.779366	33.49892
AD00290	36	35	0	4	3	361	…	51.69119	0.963675	34.71924

仔细观察，不难发现这些车辆的驾驶员在行驶过程中都有大量的超速行为，部分还有多次疲劳驾驶的危险行为。结果也在某个角度验证了模型的合理性。

7.2 驾驶轨迹区域性分析与建议

为了探索这些车辆的行驶地域特征，进而为驾驶员和企业提供更有效的建议与指导，在这里对车辆行驶的经纬度用热力图在地图上做出展示，以刻画驾驶的区域密度与轨迹的地域划分特点，具体如下图所示。

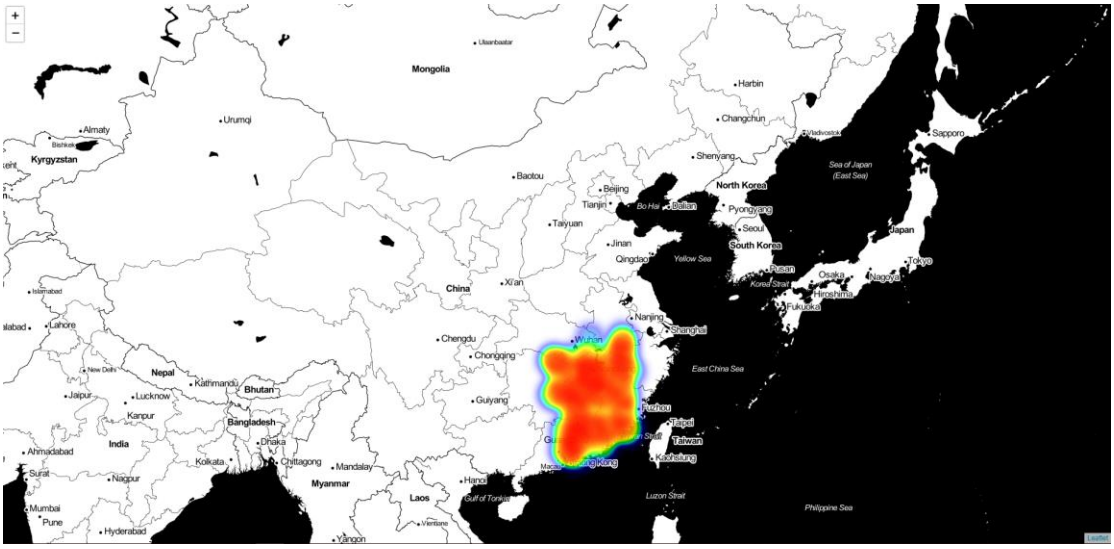


图 17 车辆行驶轨迹热力图

通过观察，可以发现，这些车辆的行驶主要集中在江西省、福建省、广东省和湖南省内。这几个南方地区属于典型的亚热带季风气候，冬季降水降雪较少，但夏季梅雨期将持续 20-30 天，并且在九、十月，福建和广东的沿海地区还会遭遇较多的台风入侵[16]。

针对这些气候特点，驾驶员和企业管理部门需要警惕夏季的恶劣天气对驾驶带来的负面影响，在下大雨和刮大风时需要注意合理控制车速、开启行车灯和示宽灯、避免急加急减速、小幅度打方向盘、注意路面积水情况。情况恶劣时，需要将车停至室内停车场，不要侥幸驾驶。

7.3 驾驶行为倾向性

驾驶行为倾向涉及的因素很多，此处主要涉及到和安全息息相关的驾驶员激进程度倾向性，按照驾驶员个体特性可分为：激进型、温和型、保守型三类，激进型司机容易冲动，相比于其他类型更倾向于超车与超速，更有可能出现路怒症；温和型司机一般自制力强，不易冲动；保守型司机倾向于开慢车，速度稳定，但遇到紧急情况容易慌张[17]。

为了展示结果便于理解，并且避免指标参数过多导致的计算量和复杂性剧增，本文使用 K-Means 算法对倾向性区分最为显著的两个特征：驾驶速度稳定性和平均速度进行了聚类研究。

K-Means 均值聚类算法是先随机选取 K 个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚

类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小[18]。

使用 K-Means 对驾驶行为聚类的结果如下图所示。

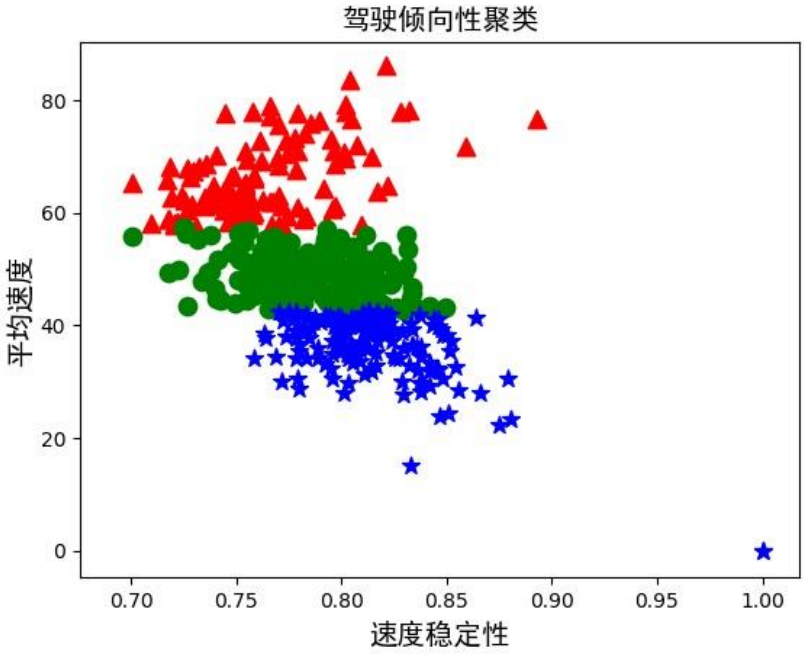


图 18 驾驶倾向性聚类

可以看到，图中将数据分为了红、绿、蓝三类。对照其速度稳定性和平均速度，发现红色数据点大都是速度稳定性低而平均速度高，参考驾驶行为表中的记录，它们都是急加速、急减速、超速次数较多的数据记录，由此认定红色数据代表着激进型的驾驶员。类似的，可以发现绿色数据点代表着温和型的驾驶员，蓝色数据点代表着保守型驾驶员。通过计算可以得到激进型驾驶员共 108 名，占比 24.05%，温和型驾驶员共 181 名，占比 40.31%，保守型驾驶员共 160 名，占比 35.64%。

针对聚类出来的结果，运输企业管理部门可以重点关注激进型和保守型的驾驶员。对于激进型的驾驶员，应当控制减少受情绪影响的驾驶行为，注意对速度的把控，为了乘客的舒适性与货物的完整性应当避免过多的急加急减速行为。对于保守型驾驶员，从效率和油耗的角度上考虑，应当在保证安全的前提下适当提高行驶速度。

8. 总结和未来展望

本文针对 449 辆运输车辆的部分 OBD 数据，结合天气数据，进行了深入挖掘与分析。在数据预处理部分，首先对整体数据完成了去重处理；实现了对经纬度偏移点的修正；对里程异常值进行了分情况处理；为了后续挖掘的方便，将天气数据进行了格式上的统一；最后通过将地理坐标数据作为中间件，实现了对天气数据的集成。接着，分别提出了 17 个驾驶行为指标的判断标准，并将它们对应到了安全、效率和能耗三类中去。在绘制轨迹图时，分别就静态展示和动态展示，绘制了静态轨迹图和基于 BD-09 坐标系的动态轨迹图。在评价模型的构建中，采用了加权的方法来构建模型，将问题转换成了权值选取的问题，本文选择了 BP 神经网络来计算权重。在具体实验中，将专家评分结果作为训练样本，对模型进行了训练与学习。将模型在测试集上的评分结果与专家评分进行了验证，展示出了模型的可行性。并且在数据分析阶段，对所有车辆的评分进行了汇总统计，找出了低分值车辆的具体原因；对所有车辆的轨迹生成热力图后发现了该运输企业的区域性特点，并针对气候特征提出了相应的注意事项；通过 K-Means 算法对驾驶员的倾向性完成了聚类，为驾驶员和企业管理部门提供了另一角度的参考。

由于作者学识水平和准备时间有限，本文仍存在一些能够进一步完善优化的不足之处，具体体现为：

1) 在预处理阶段，数据缺失部分尚未提供补齐方法。在下一步中，可以考虑通过构建历史轨迹库，对缺失部分采用在轨迹库中就近检索补齐的处理方式。
2) 由于在车辆行进过程中，除了天气，还受环境、路况、车内状态等多方面因素的影响，本文只涉及到了天气上的考虑是不够全面的。未来，除了车辆行驶数据和天气数据，还可以结合路况与车内摄像头等信息进行挖掘分析。
3) 本文选用 BP 神经网络构建评价模型，在有着大量优点的同时，也存在着如下的不足之处：参数过多，大小设置尚未存在科学定论；容易陷入局部最优；收敛速度较慢。下一步，可以考虑设计科学调参方法，对参数进行自动选取。

参考文献

[1] Yulong W , Kun Q , Yixiang C , et al. Detecting Anomalous Trajectories and Behavior Patterns Using Hierarchical Clustering from Taxi GPS Data[J]. ISPRS International Journal of Geo-Information, 2018, 7(1):25.

[2] Mining freight truck's trip patterns from GPS data[C]// IEEE International Conference on Intelligent Transportation Systems. 2014.

[3] 张志鸿. 基于OBD数据分析的驾驶行为研究[D].2017.

[4] 许书权. 基于车辆运行监控系统的驾驶行为安全与节能评价方法研究[D]. 2015.

[5] 蔡凤田, 曾诚, 殷国祥, et al. 交通运输行业标准JT/T 807-2011《汽车驾驶节能操作规范》释义[M].

[6] 任慧君, 许涛, 李响. 利用车载GPS轨迹数据实现公交车驾驶安全性分析[J]. 武汉大学学报·信息科学版, 2014, 39(6): 739-744. REN Huijun, XU Tao, LI Xiang. Driving Behavior Analysis Based on Trajectory Data Collected withVehicle-mounted GPS Receivers. GEOMATICS AND INFORMATION SCIENCE OF WUHAN UNIVERS, 2014, 39(6):739-744.

[7] 佘星源, 聂颖, 张杰. 南京机场低能见度天气变化规律及拟合预测模型[C]// 第35届中国气象学会年会. 2018.

[8] 俞倩雯. 基于车联网的汽车行驶经济车速控制方法[D].2014.

[9] 夏杰. 基于道路运输企业安全生产管理数据的驾驶行为安全与节能评价方法[D]. 2016.

[10] 李晓辉, 张来, 李小宇, et al. 基于层次分析法的现状电网评估方法研究[J]. 电力系统保护与控制, 2008, 36(14):57-61.

[11] Zou Z H , Yun Y , Sun J N . Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment[J]. Journal of Environmental Sciences, 2006, 18(5):1020-1023.

[12] 常建娥, 蒋太立. 层次分析法确定权重的研究[J]. 武汉理工大学学报(信息与管理工程版), 2007, 29(1):153-156.

[13] 程启月. 评测指标权重确定的结构熵权法[J]. 系统工程理论与实践, 2010, 30(7):1225-1228.

[14] 周志华. 机器学习[M]. 清华大学出版社, 2016.

[15] 李聪颖, 王肇飞. 基于BP神经网络的高速公路交通安全评价系统设计与实现[J]. 武汉理工大学学报(交通科学与工程版), 2010, 34(3):476-479.

[16] 赵济. 新编中国自然地理[M]. 高等教育出版社, 2015.

[17] 王晓原, 杨新月. 基于决策树的驾驶行为决策机制研究[J]. 系统仿真学报, 2008, 20(2).

[18] 张良均. Python数据分析与挖掘实战[M]. 机械工业出版社, 2016.