行车安全性多属性综合评价方法

摘　要

评价车辆行驶安全性的方式多种多样，其中，通过对行车数据进行采集并做出分析处理是一种常见且有效的手段。

通过各类电子传感器采集到的车辆数据，蕴含着大量值得挖掘的信息，能够对人们分析判断车辆是否在行驶过程中存在安全隐患，提供有力的数据支撑。但是，由于电子设备本身的局限性，采集到的数据往往存在着价值密度低、蕴含大量噪声和异常的问题，要想提取出有价值的信息，需要我们对数据进行严谨的处理和科学的分析。

针对题目中的具体问题，我们从以下 4 个方面构建行车安全性多属性综合评价方法：

1. 多角度对车辆的行驶数据进行分析与预处理（包括属性缺失值的处理以及异常值的处理），并从处理后的数据中提取出有用的信息，包括车辆的行驶速度和加速度、行驶里程、行驶路线等，同时将规定的十辆数据的每条路线图、平均速度以及急加速和急减速情况汇总到了数据的附件 1。

2. 针对问题二，我们首先在充分分析附件 1 数据的基础上，对每种不良驾驶行为首先构建判断模型，其次，对提取的结果分析，最后，与诸如百度地图等比较成熟的软件结果对比分析来验证判断方法的有效性，总共提出了劳驾驶、超速、急加速、急减速、怠速预热、超长怠速、熄火滑行、急变道等 7 种不良驾驶行为特征，在对 7 种评价指标进行了相关性分析的基础之上，最终形成了反映车辆行车安全性的 7 种评价指标，详细数据我们汇总到了数据的附件 2.

3.针对问题二，评价模型的构建，首先，分别采用主客观的方法为评价指标赋予权重，在此基础之上，通过最小二乘法对两种权重进行融合，形成兼顾专家意见与客观数据因素的评价指标权重；其次，根据理想逼近解法 TOPSIS 法对评价模型进行构建 ，并在真实数据上对研究对象进行安全等级排序；最后，我们通过 K-means 对数据进行粗略地二分类，将 TOPSIS 标记与粗略二分类的标记进行对比，选出标记相同的数据作为标注数据，训练二叉决策树评价模型，利用训练好的模型再对不确定数据进行评价，最终以投票的方式将不确定的数据进行安全性评价，详细评价结果数据具体请参看我们数据的附件 3.

4.针对问题三，在我们已经构建好了描述车辆不良驾驶的七种行为特征。在此基础之上，我们综合分析附件 2 给出的自然气象数据以及自己额外获取的关于车辆路线的道路状况数据，同时构建了行车天气环境指标以及道路状况指标，一共 12 个特征，从行车安全、节能、效率的角度分析，综合考虑行车安全、效率和节能，给出了一个行车安全的综合评价指标体系。为了反映各种指标之间相互影响影响，我们采用网络层次分析的结果，由指标的全局权重可以看出，疲劳驾驶、超速、急变速、急变道、天气类别这五个指标是评价指标体系中占权重最大的五个指标，这也与我们的主观感觉相符。而对于安全、节能、效率这三个方面，根据指标的综合权重结果可以看出，我们给出的这个综合评价指标体系，关注的程度最大的还是行车安全，然后是节能，最后才是效率。

关键词：特征提取、AHP、ANP、熵权法、TOPSIS、决策树、k-means.

Abstract

There are many ways to evaluate vehicle driving safety, among which, collecting and analyzing traffic data is a common and effective means.

Vehicle data collected by various electronic sensors contain a lot of information worth mining, which can provide powerful data support for people to analyze and judge whether there are potential safety hazards in the driving process. However, due to the limitations of electronic equipment itself, the collected data often have the problems of low value density, containing a lot of noise and anomalies. In order to extract valuable information, we need to deal with the data rigorously and scientifically.

Aiming at the specific problems in the topic, we construct a multi-attribute comprehensive evaluation method of driving safety from the following four aspects:

1. Multi-angle analysis and pre-processing of vehicle driving data (including attribute missing value processing and abnormal value processing), and extract useful information from the processed data, including vehicle speed and acceleration, driving mileage, driving route, etc. At the same time, each road map, average speed, acceleration and deceleration situation of 10 vehicle data will be specified. The data are summarized in Annex 1.

2. Aiming at the second problem, we first construct a judgment model for each bad driving behavior on the basis of full analysis of the data in Annex 1. Secondly, we analyze the extracted results. Finally, we verify the validity of the judgment method by comparing and analyzing the results with the mature software results such as Baidu Map. In total, we put forward labor driving, overspeed, acceleration, rapid deceleration, idle preheating, etc. Seven kinds of bad driving behavior characteristics, such as super-long idling, extinguishing taxiing and sudden change lane, are analyzed. Based on the correlation analysis of seven evaluation indexes, seven evaluation indexes reflecting vehicle driving safety are finally formed. Detailed data are summarized in Annex 2 of the data.

3. Aiming at the second problem, the evaluation model is constructed. Firstly, the subjective and objective methods are used to assign weights to the evaluation indexes. On this basis, the least square method is used to fuse the two weights to form the evaluation index weights that take into account both expert opinions and objective data factors. Secondly, the evaluation model is constructed according to the TOPSIS method of ideal approximation solution and on the real data. Finally, we use K-means to classify the data roughly, compare the TOPSIS markers with the rough binary markers, select the data with the same markers as the marking data, train the binary decision tree evaluation model, use the trained model to evaluate the uncertain data, and finally vote the uncertain number. According to the safety evaluation, the detailed evaluation result data refer to Annex 3 of our data.

4. To solve the third problem, we have constructed seven behavioral characteristics describing bad driving of vehicles. On this basis, we synthetically analyze the natural meteorological data given in Annex 2 and the road condition data acquired by ourselves. At the

same time, we construct the weather and environment indicators and road condition indicators, which have 12 characteristics. From the point of view of driving safety, energy saving and efficiency, we give a driving safety considering driving safety, efficiency and energy saving comprehensively. Comprehensive evaluation index system. In order to reflect the interaction among various indicators, we use the results of network analytic hierarchy process. From the overall weight of the indicators, we can see that the five indicators of fatigue driving, overspeed, rapid speed change, sudden change lane and weather category are the five most important indicators in the evaluation index system, which are also consistent with our subjective feelings. For the three aspects of safety, energy saving and efficiency, according to the results of the comprehensive weight of the indicators, we can see that the comprehensive evaluation index system given by us pays the greatest attention to driving safety, then energy saving, and finally efficiency.

Key words: feature extraction, AHP, ANP, entropy weight method, TOPSIS, decision tree, k-means.

目　录

图 录

表　录

第一章　问题描述

1.1　问题描述

随着道路交通的发展，运输行业车辆在数量增长的同时，事故发生率也呈现出多发的态势，道路运输行业的监管问题逐渐成为人们关注的焦点。但是，由于缺乏专业的技术手段，监管部门对驾驶人员的各类不良驾驶行为，如超速、疲劳驾驶等作出有效监督和及时干预，始终是行业面临的一大难题。

近年来，随着物联网和大数据技术的兴起，通过车联网采集运输车辆在行驶过程中的各类数据，通过实时或非实时的分析，可以实现对车辆静态和动态的监管，这逐渐成为了有关部门进行安全管理、提高运输效率水平的有效手段。

在车联网系统中，通过安装在车辆上的电子标签，信息平台可以对车辆各个属性进行分析，得到车辆的运行状态，而我们的任务，就是从采集到的数据中分析有价值的车辆行驶信息，并对行车安全性进行全面的分析。

首先，我们需要从附件所给的数据中，提取车辆在各个行程中有价值的行驶信息，包括车辆的行驶速度和加速度、行驶里程、行驶路线等信息；之后，我们需要提取车辆行驶过程中的不良驾驶行为，包括疲劳驾驶、超速、急加速、急减速、怠速预热、超长怠速、熄火滑行、急变道等，建立安全评价模型，判断数据集中所有车辆的行驶安全性；最后我们需要综合考虑车辆的安全、节能、效率等因素，建立行车安全的综合评价指标模型和体系。

1.2　论文的结构安排

本文共分为五章，各章内容安排如下：

第一章，对论文需要解决的问题进行描述，并简单介绍整篇论文的结构安排。

第二章，多角度对车辆的行驶数据进行分析，对数据进行预处理，并从处理后的数据中提取出有用的信息，包括车辆的行驶速度和加速度、行驶里程、行驶路线等。

第三章，全面分析车辆的不良驾驶行为，对每种不良驾驶行为构建判断模型，并通过结果分析验证判断方法的有效性。

第四章，运用层次分析法、熵权法构建不良驾驶行为的评价模型，并使用理想点逼近法对每辆车的行驶安全性进行判断，最后使用决策树算法构建安全性评判模型的评价准则。

第五章，综合考虑安全、效率、节能的因素，通过网络层次分析法建立建立行车安全的综合评价指标体系与综合评价模型。

第二章 车辆数据探索分析

在分析车辆的驾驶行为时，一些信息的提取是必要的，其中包括车辆的行程划分、行驶里程、行驶路线、瞬时速度、平均速度等，本章将主要就上述几个方面的问题，对车辆数据进行初步的探索分析。

2.1 车辆数据预处理

在数据挖掘过程中，数据预处理是第一步，同时也是很重要的一步，数据预处理的好坏直接决定着之后特征提取、分类预测等步骤能否顺利进行。在对数据进行分析之前和分析的过程中，我们逐渐对题目所给的数的认识逐步加深，并最终得出了一套较为完整的数据预处理流程。

2.1.1 缺失属性的剔除

通过对数据集进行观察，我们发现数据集并不是每个属性都是被正常记录的。其中，有4个属性：右转向灯（right_turn_signals）、左转向灯（right_turn_signals）、手刹（hand_brake）、脚刹（foot_brake）在整个数据集中，始终保持为0不变，因此，这4个属性的记录是缺失的，不能为我们提供有价值的信息。因此，我们选择将这4个属性从数据集中剔除。

2.1.2 缺失时间的记录

我们发现，在一段连续且完整的行车记录中，采集时间（location_time）是逐秒递增的。但是，数据集中时间记录的缺失十分常见，不仅体现在行车过程中某1秒的时间缺失，更体现在某一连续行车时间的缺失，如图 1。

50855	AA00001	AAA9102C	107	116.615	27.57887	1	2018/8/6 9:37	0	5456
50856	AA00001	AAA9102C	107	116.615	27.57887	1	2018/8/6 9:37	0	5456
50857	AA00001	AAA9102C	107	116.615	27.57887	1	2018/8/6 9:37	0	5456
50858	AA00001	AAA9102C	107	116.615	27.57887	1	2018/8/6 9:37	0	5456
50859	AA00001	AAA9102C	107	116.615	27.57887	1	2018/8/6 9:37	0	5456
50860	AA00001	AAA9102C	107	116.615	27.57887	1	2018/8/6 9:37	0	5456
50861	AA00001	AAA9102C	107	116.615	27.57886	1	2018/8/6 9:37	0	5456
50862	AA00001	AAA9102C	107	116.615	27.57886	1	2018/8/6 9:37	0	5456
50863	AA00001	AAA9102C	107	116.615	27.57886	1	2018/8/6 9:37	0	5456
50864	AA00001	AAA9102C	289	116.6189	27.5777	1	2018/8/6 10:00	34	5456
50865	AA00001	AAA9102C	290	116.619	27.57768	1	2018/8/6 10:00	34	5456
50866	AA00001	AAA9102C	288	116.6186	27.57775	1	2018/8/6 10:00	33	5456
50867	AA00001	AAA9102C	287	116.6187	27.57773	1	2018/8/6 10:00	33	5456
50868	AA00001	AAA9102C	288	116.6184	27.5778	1	2018/8/6 10:00	33	5456
50869	AA00001	AAA9102C	288	116.6184	27.57782	1	2018/8/6 10:00	35	5456
50870	AA00001	AAA9102C	288	116.6183	27.57784	1	2018/8/6 10:00	34	5456
50871	AA00001	AAA9102C	289	116.6182	27.57787	1	2018/8/6 10:00	34	5456
50872	AA00001	AAA9102C	288	116.6181	27.57789	1	2018/8/6 10:00	34	5456

图 1　原始数据行驶过程中的时间缺失

对于单独 1 秒时间的缺失，需要我们在分析数据时留意，每条数据之间的间隔不一定是 1 秒。在判定两条数据之间的间隔时间时，需要用时间序列进行计算，而不是用索引值计算。

对于连续时间的缺失，需要我们判断是正常缺失还是异常缺失。对于正常缺失，应该满足在一次较大的时间间隔前后，速度到 0 结束再从 0 开始，里程数保持连续，经纬度基本保持不变；对于异常缺失，速度、里程、经纬度都可能发生突变，需要我们进行记录。

2.1.3　异常里程的处理

对于正常的数据集，车辆里程记录应当连续递增，但是在处理数据的过程中，我们发现，车辆里程会发生突变，有时还会伴随着设备号的改变，如图 2，图 3。

270	AD00017	AAA9101017	170	114.6056	27.36416	1	0	0	0	0	2018/8/7 4:40	50	7176
271	AD00017	AAA9101017	170	114.6056	27.36404	1	0	0	0	0	2018/8/7 4:40	50	7176
272	AD00017	AAA9101017	172	114.6056	27.36391	1	0	0	0	0	2018/8/7 4:40	50	7176
273	AD00017	AAA9101017	172	114.6056	27.36379	1	0	0	1	0	2018/8/7 4:40	51	8108
274	AD00017	AAA9101017	206	114.5716	27.22248	1	0	0	0	0	2018/8/7 5:00	56	7193
275	AD00017	AAA9101017	205	114.5715	27.22236	1	0	0	0	0	2018/8/7 5:00	55	7193
276	AD00017	AAA9101017	204	114.5715	27.22224	1	0	0	0	0	2018/8/7 5:00	54	7193

图 2　原始数据里程数的异常突变

18545	AB00465	AAA2005280	135	115.6497	24.95989	1	0	0	0	0	2018/9/27 7:39	0	16856
18546	AB00465	AAA2005280	135	115.6497	24.95989	1	0	0	0	0	2018/9/27 7:39	0	16856
18547	AB00465	AAA2005280	135	115.6497	24.95989	1	0	0	0	0	2018/9/27 7:39	0	16856
18548	AB00465	AAA2005280	135	115.6497	24.95989	1	0	0	0	0	2018/9/27 7:39	0	16856
18549	AB00465	AAA8107465	0	114.9526	25.84488	1	0	0	0	0	2018/9/27 23:42	0	40810
18550	AB00465	AAA8107465	0	114.9526	25.84488	1	0	0	0	0	2018/9/27 23:42	0	40810
18551	AB00465	AAA8107465	0	114.9526	25.84488	1	0	0	0	0	2018/9/27 23:42	0	40810
18552	AB00465	AAA8107465	0	114.9526	25.84488	1	0	0	0	0	2018/9/27 23:42	0	40810

图 3　原始数据设备号的改变

这对于我们根据行车里程求平均速度以及判断疲劳驾驶会产生很大影响，因此需要我们在突变前后进行记录，在计算里程时忽略突变段，具体处理方法见平均速度的计算一节。

2.1.4 异常经纬度的处理

分析数据集后，我们发现，在经纬度方面，数据的记录并不十分精准。对于一段完整行程记录，车辆路线应当是连续且行驶在正常道路上的，但是，数据集显示的路线结果并不理想。

我们任意挑出车辆 AA00001 的一段行程，对路线进行了绘制，结果如图 4。

图 4 原始数据的行车路线

从图中可以看出，车辆行程路线不仅存在中断，某些时间的经纬度变化甚至不符合正常的行车水平（图中竖直虚线显示的路线，在一分钟内行驶超过 80km），因此，需要我们对经纬度数据重新进行修改，具体修改过程见经纬度修正一节。

2.1.5 异常速度的处理

我们在分析数据集中速度的异常时发现，速度不仅存在单个异常值的突变，还有一个普遍问题：

速度传感器在车辆低速行驶的状态下（速度小于等于 10km/h）是无法准确识别车辆速度的，因此判定小于 10km/h 的速度全部为 0，很多现象都印证了这一规律（速度为 0 时，方向角和经纬度还在改变；统计各速度的频数，0 和 10 之间的速度频数为 0），如图 5。

图 5　原始数据的速度频数统计

这一规律对于我们分析车辆的急加急减速造成了很大影响，不论是正常起步还是制动，数据集中的速度都会从 0 突变到 10km/h 以上，或是从 10km/h 以上突变到 0，如果不对速度进行处理，按照 $3m/s^2$ 的加速度阈值，正常的起步和制动都可能会被判定为急加速和急减速。对此，我们对速度为 0 的情况进行了填充，具体填充过程见零速度填充一节。

2.1.6　异常方向角的处理

在分析车辆方向角时，我们发现，有些时候，行车记录仪的各传感器会同时失灵：方向角归零、经纬度保持不变、速度归零，如图 6。

62026	AA00001	AAA91020	89	115.9126	28.69144	1	2018/8/6 23:01	53	5638
62027	AA00001	AAA91020	89	115.9128	28.69144	1	2018/8/6 23:01	53	5638
62028	AA00001	AAA91020	89	115.9129	28.69144	1	2018/8/6 23:01	53	5638
62029	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62030	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62031	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62032	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62033	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62034	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62035	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62036	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62037	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62038	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62039	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62040	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62041	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62042	AA00001	AAA91020	0	115.9129	28.69144	1	2018/8/6 23:01	0	5638
62043	AA00001	AAA91020	87	115.9147	28.69183	1	2018/8/6 23:01	47	5639
62044	AA00001	AAA91020	87	115.9152	28.69156	1	2018/8/6 23:01	48	5639

图 6　原始数据方向角的异常变化

事实上，对于实际情况，方向角可能仍然保持着之前的状态，对于经纬度和速度，我们有具体的修正和填充方法，但是对于方向角的突变，由于只对判断车辆的急变道有影响，

我们在分析车辆急变道行为时，只需将这一特殊情况考虑进去即可。

2.2 零速度填充

车辆的行驶速度，包括瞬时速度和平均速度，瞬时速度通过数据集本身的记录（gps_speed）得到，平均速度需要根据瞬时速度间接计算得到。本节主要针对瞬时速度进行处理。

在数据预处理一节中，我们分析得出了速度记录中存在的，低速度无法识别的问题。对于速度记录，可以分为 gps_speed = 0 和 gps_speed ≠ 0 这两种情况。

在车辆正常行驶（gps_speed ≠ 0）的过程中，需要进行修正的只有异常值突变这一种情况；对于速度大量缺失的情况，我们无法根据数据集包含的信息对缺失部分的数据进行填充。

在车辆静止（gps_speed = 0）的过程中，由于传感器不灵敏、失灵等因素，导致原本应当正确显示的速度被记录为 0，这在数据集中是十分常见的，因此需要我们进行分析和修正。

我们的任务，就是挖掘零速度和异常速度背后隐藏的正确的速度，对传感器的错误记录进行判别和修改。由于需要修改的数据大多数都是零速度记录，因此，我们将本节命名为“零速度填充”。我们采取的填充过程如下：

2.2.1 记录零速度区间

我们遍历数据，得到所有的连续的 gps_speed = 0 的区间，这样可以方便我们针对每个区间进行填充，对零速度区间我们采取元组的形式表示，元组包含元素：起始索引（start_i），终止索引（end_i），区间内记录数（num）。

根据零速度区间的定义，对于任意元组（start_i, end_i, num）属于零速度区间，有如下性质：

① speed[i] = 0（start_i <= i <= end_i）

② speed[i] > 0（i 不属于任意零速度区间）

③ speed[start_i – 1] > 0，speed[end_i + 1] > 0

但是，这样统计得到的零速度区间存在再划分的问题。

比如，对于一次车辆行程的切换，速度变化的大多数的情况为：减小到 0，并经过一段 0 速度时间（在此过程中，行车过程已经发生了切换，时间很可能发生突变，前一次行车过程结束，后一次行车过程开始），重新从 0 开始增加。如图 7。

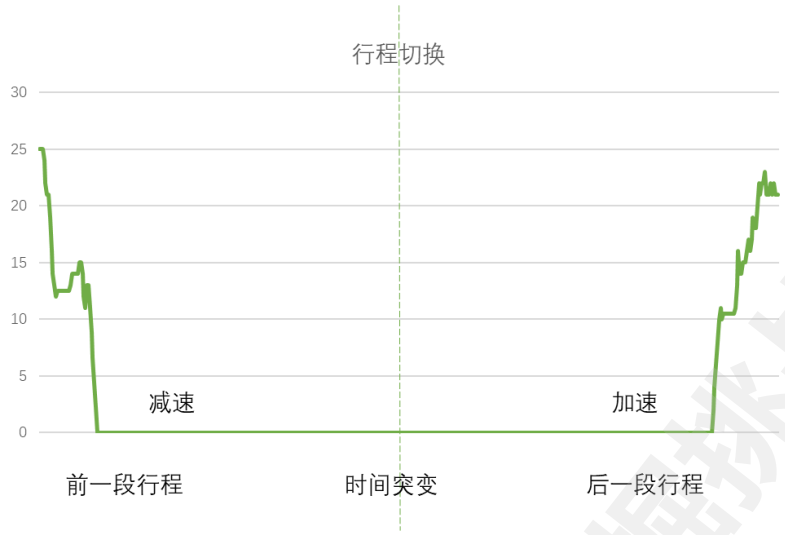

图 7　车辆行程切换和速度的变化

根据上述情况，这段零速度区间包含了 2 个不同的行车过程的加速和减速行为，而我们的策略是，根据零速度区间相邻的非 0 速度信息进行 0 速度填充。正常情况下，相邻两个行车过程的零速度区间是较长的，这样在进行填充时，一个行车过程的速度信息不会对相邻行车过程的 0 速度填充产生影响。如果该零速度区间长度较短，就可能出现前后信息的交叉，导致速度填充出现问题，如图 8。

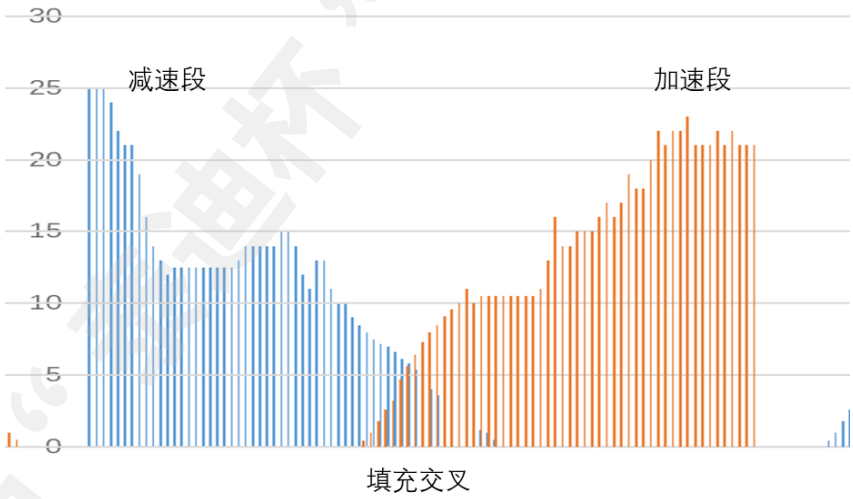

图 8　速度填充过程中的交叉现象

因此，对零速度区间进行再划分是必要的。

我们分析数据发现，相邻两个行车过程之间，会发生时间的突变，且行车过程的切换往往伴随着车辆的关闭和重启（acc_state 从 1 变为 0 再变为 1），根据这一特征，我们就可以很方便地对零速度区间进行再划分。这样经过再划分的零速度区间，将不再满足性质③，例如，对于发生再划分的零速度区间，前一区间的 speed[end_i + 1] = 0，后一区间的 speed[start_i – 1] = 0。

2.2.2 填充策略：

首先为每一零速度区间定义两个标志：发生移动标志 move_onetime 和始终移动标志 move_alltime。具体含义如下：

$$\begin{cases} move\_onetime = 0，表示在该零速度区间车辆始终未移动 \\ move\_onetime = 1，表示在该零速度区间车辆发生过移动 \end{cases}$$

$$\begin{cases} move\_alltime = 0，表示在该零速度区间车辆存在静止的情况 \\ move\_alltime = 1，表示在该零速度区间车辆始终在移动 \end{cases}$$

填充策略如下：

遍历零速度区间，对于区间 i：

①如果区间长度小于等于 5，且区间两侧速度都不为 0，则视作传感器短暂失灵，使用区间两端非 0 速度的平均值填充，填充完成后转到区间 i + 1。否则，转入步骤②；

②如果 move_onetime = 0，说明车辆没有发生移动，如果区间两端非 0 速度小于阈值 S_1，视作车辆从低速转为静止，则保持速度为 0 不变，填充完成后转到区间 i + 1。如果大于该阈值 S_1，转入步骤⑦；如果 move_onetime = 1，说明车辆发生过移动，转入步骤③；

③如果 move_alltime = 1，说明车辆始终在移动，但是没有被记录仪记录，很可能是在发生 10km/h 以下的低速移动，例如堵车等。此时用区间两端的非 0 速度的平均值填充（若该段区间发生时间突变，则视为数据缺失，仍按照平均速度进行填充），填充完成后转到区间 i + 1；否则转入步骤④；

④如果 move_alltime = 0，说明车辆发生过移动，但没有始终在移动，此时应当对区间两端的速度进行匀加速匀减速填充。在填充之前，需要判断区间长度是否小于 5，如果是，转入步骤⑤；否则，转入步骤⑥；

⑤在进行匀加速匀减速填充时，我们采取的策略是每次填充 4 条数据，即用速度差除以 5 条数据的时间得到平均加速度。如果区间本身长度就小于规定的默认填充长度，则使用该区间对应的时间差计算出平均加速度，进而计算出每条数据的速度，进行填充。填充完成后转到区间 i + 1；

⑥如果区间本身长度大于规定的默认填充长度，则使用该默认长度对应的时间差计算出平均加速度，进而计算出每条数据的速度，进行填充。填充完成后转到区间 i + 1；

⑦如果区间两端非 0 速度大于阈值 S_2，表示车辆在高速行驶状态下突然转为静止，一段时间后又突然恢复成高速行驶状态，这种情况可以确定代表着异常，发生的原因可能是车辆在高速行驶状态时，速度记录传感器失灵。因此，使用区间两端非 0 速度的平均值填充。填充完成后转到区间 i + 1；

⑧如果当前区间 i 为最后一个零速度区间，填充完成后算法结束。

算法流程如图 9。

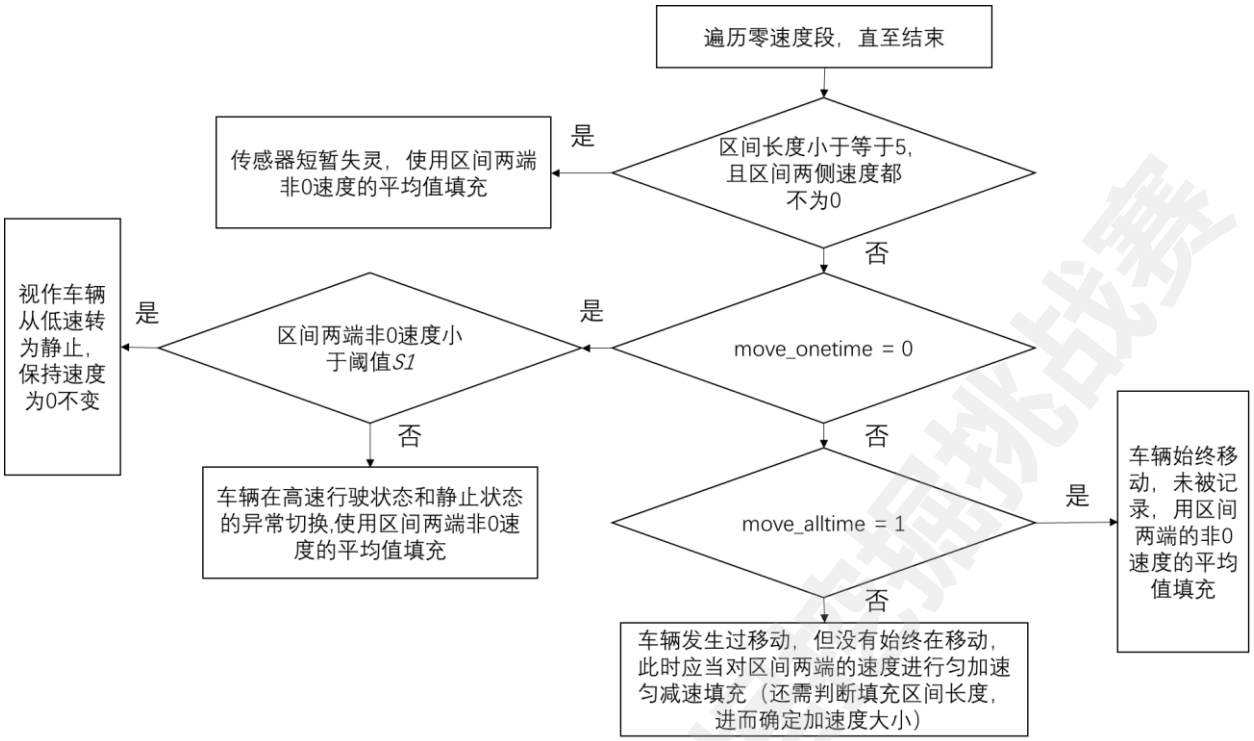

图 9　零速度填充流程图

2.3　车辆行程路线分析

通过分析数据，我们发现给出的数据中，一段时间内经纬度可能存在很大的偏差，上一时刻还在正常行驶，下一时刻记录下的经纬度就与之前的位置相差几公里，这明显是不合理的错误数据，如图 10，图 11。所以我们设计判别规则，筛选出异常经纬度的数据项，并删除部分数据或利用经纬度与速度、方向角等属性之间的关系进行修正。

117.60723	29.100486	1	2018/8/5 6:27	92
117.60738	29.100256	1	2018/8/5 6:27	92
117.60755	29.100025	1	2018/8/5 6:27	93
117.60773	29.99795	1	2018/8/5 6:27	92
117.60795	29.99563	1	2018/8/5 6:27	93
117.60818	29.99331	1	2018/8/5 6:27	93
117.6084	29.99101	1	2018/8/5 6:27	93
117.60868	29.9887	1	2018/8/5 6:27	93
117.60901	29.98638	1	2018/8/5 6:27	94
117.60936	29.98408	1	2018/8/5 6:27	94

图 10　经纬度的异常突变

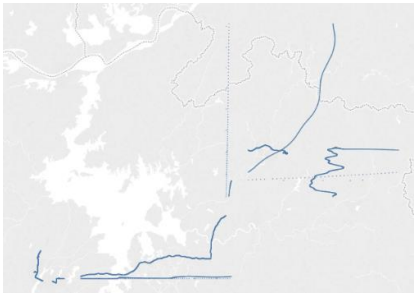

图 11　原始数据绘制出的异常路线

2.3.1　经纬度异常点发现

通过时间差将表中所有数据划分了一个个行程。通过分析，我们发现对每个行程，可以通过相邻数据项的经纬度计算出其距离（记为经纬度距离），再与速度及数据项之间的时间差计算出实际跑的距离（记为速度距离）进行比较，如图 12。如果这两种距离之差的绝对值大于某个阈值，则认为数据发生了偏移，否则，认为数据未发生偏移。

```
location_time ('两种距离之差的绝对值', '经纬度距离', '速度距离')
2018-08-05 06:27:01 (3.0837731386807654, 28.361550916458544, 25.27777777777778)
2018-08-05 06:27:02 (3.8227100380078696, 29.378265593563423, 25.555555555555554)
2018-08-05 06:27:03 (4.930396419984209, 30.485951975539763, 25.555555555555554)
2018-08-05 06:27:04 (99504.56482743126, 99530.39816076458, 25.833333333333332)
2018-08-05 06:27:05 (232.49657365554043, 258.052129211096, 25.555555555555554)
2018-08-05 06:27:06 (232.29991433590104, 258.1332476692344, 25.833333333333332)
```

图 12　经纬度距离和速度距离的比较

如图表 2 所示，从 2018-08-05 06:27:04 到 2018-08-05 06:27:05 车辆的经纬度距离很大，而速度距离很小，二者之差的绝对值大约为 10km。对于这种情况，我们就认为数据发生了跳跃。

当然，这种跳跃有三种情况：①从正确的数据跳为异常的数据，②从异常的数据跳回到正常的数据，③从异常数据跳到异常数据。如果确定该行程跳跃前的部分数据正常，则跳跃后数据为异常数据。否则，跳跃后可能为正常数据，再判断跳转后的前部分数据的两种距离差的绝对值是否小于某一阈值，如果连续 20 条数据均满足距离差绝对值小于某一阈值，则认为跳跃后数据为正常数据。否则，认为跳跃后的数据仍为异常数据。

如图 13 所示，从 2018-08-05 06:33:52 到 2018-08-05 06:33:53 车辆发生跳转，但由于 06:33:53 后的数据计算出的经纬度距离比速度时间距离仍然大得多，所以认为跳转后的数据为异常数据。

```
location_time ('两种距离之差的绝对值', '经纬度距离', '速度距离')
2018-08-05 06:33:50 (230.17781612244792, 256.28892723355904, 26.11111111111111)
2018-08-05 06:33:51 (231.85172810671085, 257.6850614400442, 25.833333333333332)
2018-08-05 06:33:52 (85950.32545522341, 85975.88101077898, 25.555555555555554)
2018-08-05 06:33:53 (2137.54597510636, 2163.1015306619156, 25.555555555555554)
2018-08-05 06:33:54 (2117.599915726623, 2143.1554712821785, 25.555555555555554)
2018-08-05 06:33:55 (2117.7637245613437, 2143.3192801168993, 25.555555555555554)
2018-08-05 06:33:56 (2117.7890206244256, 2143.3445761799812, 25.555555555555554)
```

图 13　跳转后的异常数据

如图 14 所示，从 2018-08-05 08:15:48 到 2018-08-05 08:15:49 车辆发生跳转，并且跳转后的部分数据计算出的经纬度距离与速度时间距离的差值不大，所以认为跳转后的数据为正常数据。

```
location_time ('两种距离之差的绝对值', '经纬度距离', '速度距离')
2018-08-05 08:15:47 (48271.86845272806, 48286.590674950276, 14.722222222222221)
2018-08-05 08:15:48 (69391.50710388337, 69405.95154832781, 14.444444444444445)
2018-08-05 08:15:49 (0.8862463835174132, 13.66402416129519, 12.777777777777777)
2018-08-05 08:15:50 (2.150276017862325, 15.483609351195657, 13.333333333333332)
2018-08-05 08:15:51 (0.023702618425842203, 15.254075159351935, 15.277777777777777)
```

图 14　跳转后为正常数据

对每一行程的第一个数据项，计算其与已确定的最后一个正确的数据项之间的经纬度距离以及里程之差。如果行程第一条数据为正常数据，则计算出的经纬度距离应小于等于里程之差。如果不满足此条件，则认为该行程第一个数据项就是跳转点且是跳跃为异常数据。

2.3.2 异常经纬度区间发现

对给出的区间（left,right），先检测其是否为一个完整的行程。若是则判断行程开始部分是否为正常数据，若是，则正常判断后续是否存在跳转点，否则认为行程开始即是异常数据，判断是否存在跳回点，若无跳回点，则整个区间（left，right）都是异常数据区间，否则，（left,跳回点）为异常数据区间。具体流程，如图 15：

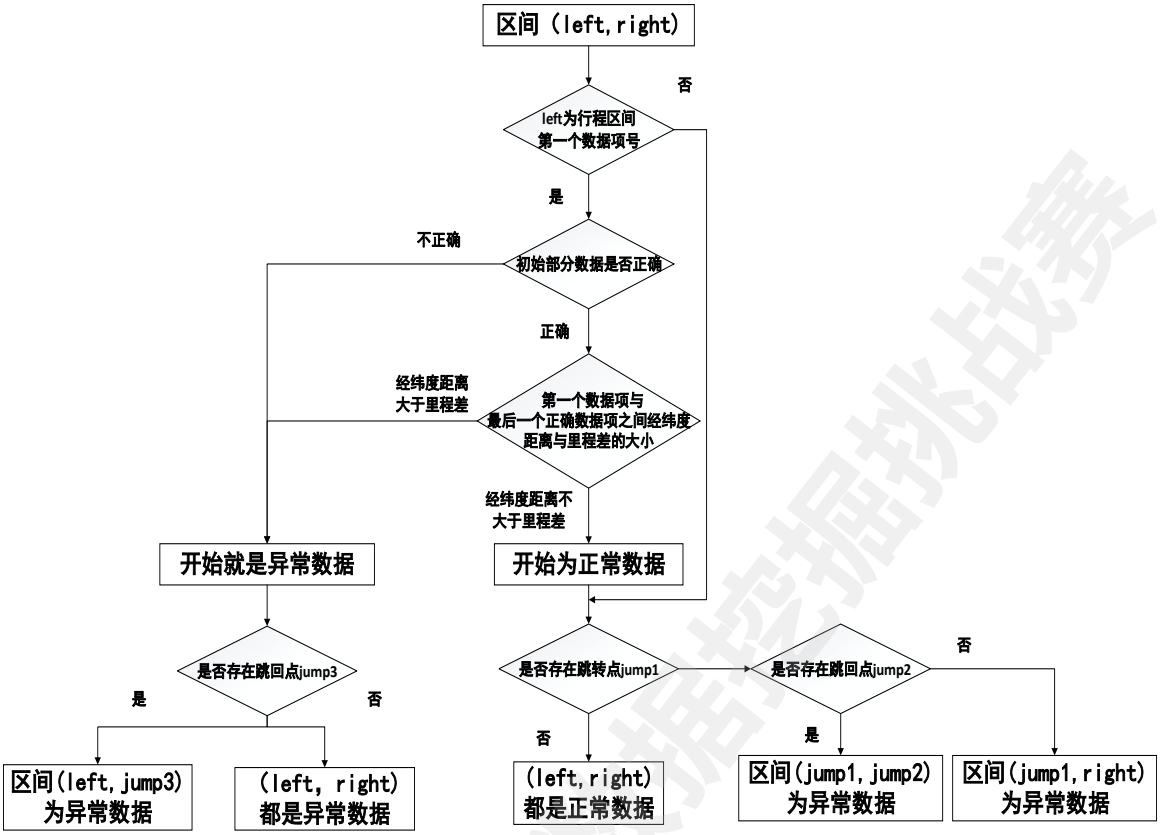

图 15　异常经纬度区间的挖掘流程图

i．如果已知一个点 A 的经纬度，从该点运行到下一点 B 的时间以及速度，以及点 B 相对于点 A 的方向角，我们可以计算出这两点之间的距离，再通过相关计算，推导出点 B 的经纬度。

ii．对于行程中的错误数据项的部分，我们可以根据跳转前最后一个正确的数据项或跳回后第一个数据项以及其经纬度、速度、时间、方向角来推出下一数据项或上一数据项的正确的经纬度值，再根据推出的值递归推出所有错误的数据项对应的正确经纬度值。

iii.如果某一行程判断出来整个过程的数据项都是错误的，则删掉该部分数据，不进行修正。（给出删除部分占数据总数的比值）如果某一错误数据项区间的左端点为某一行程开始的第一个数据项，则用该区间右端点的下一数据项逆向递推出该区间数据项对应的正确经纬度。

2.3.3 异常经纬度的纠偏

如果已知一个点 A 的经纬度，点 A 与点 B 之间的距离，以及点 B 相对于点 A 的方向角，可以推导出点 B 的经纬度[1]。而两点之间的距离我们可以根据给出的速度以及时间差求得，如图 16，图 17。

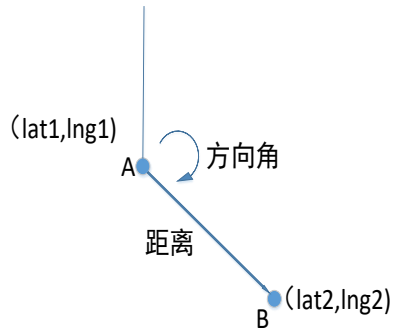

图 16　经纬度修正

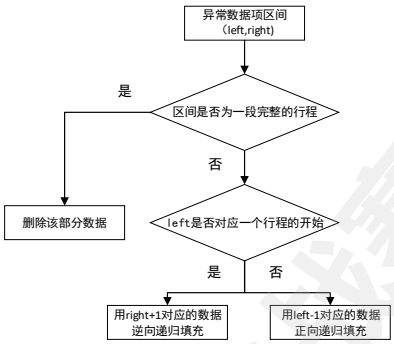

图 17　经纬度异常区间的处理流程图

对于行程中的错误数据项的部分，我们可以根据跳转前最后一个正确的数据项或跳回后第一个数据项以及其经纬度、速度、时间、方向角来推出下一数据项或上一数据项的正确的经纬度值，再根据推出的值递归推出所有异常的数据项对应的正确经纬度值。

如果某一行程判断出来整个过程的数据项都是异常的，则直接删掉该部分数据，不进行修正。如果某一异常数据项区间的左端点为某一行程开始的第一个数据项，则用该区间右端点的下一数据项逆向递推出该区间数据项对应的正确经纬度。否则，用该区间左端的上一数据项正向递归填充数据。

2.3.4 纠偏前后路线图对比

修正前后的结果的展示如图 17 图 18，图 19。

图 18　修正前路线图

图 19　修正后路线图

通过调用百度地图 API 传入修正后的车牌号为 AA00001 在 2018 年 8 月 5 日的相关数据，画出其行驶的路线图，如图 20。通过观察修正后的路线是真实存在的，说明通过我们的方法能够很好地实现轨迹的纠偏。

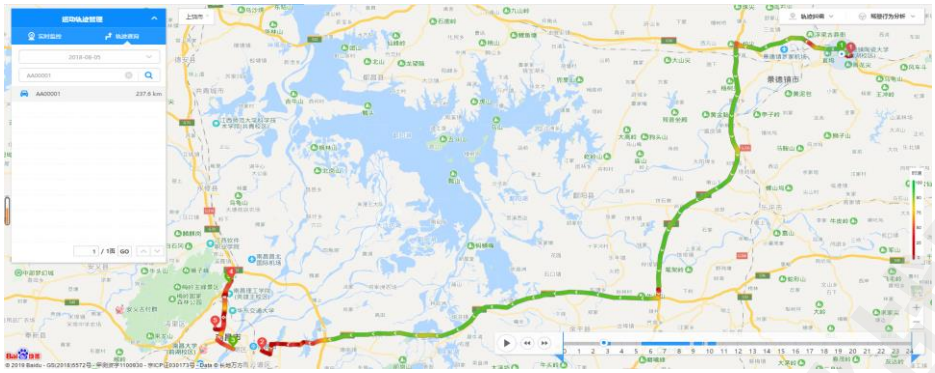

图 20　百度地图 API 检测修正后路线图

图 21 - 图 30 为十辆目标车辆的路线图的结果展示

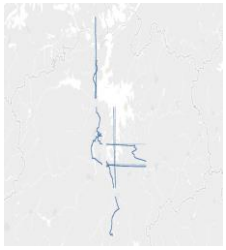

图 21　AA00002 修正前后对比图

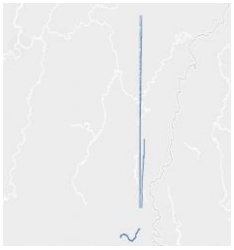

图 22　AB00006 修正前后对比图

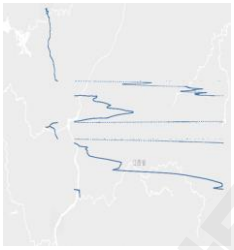

图 23　AD00003 修正前后对比图

图 24　AD00013 修正前后对比图

图 25　AD00053 修正前后对比图

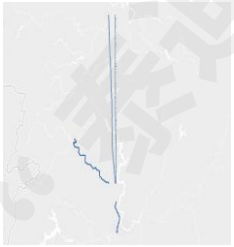

图 26 AD00083 修正前后对比图

图 27 AD00419 修正前后对比图

图 28 AF00098 修正前后对比图

图 29 AF00131 修正前后对比图

图 30 AF00373 修正前后对比图

2.4 行驶里程计算

2.4.1 计算策略

我们定义当前阶段里程数 mileage_current，总里程数 mileage_total，当前阶段里程数计算起点 mileage_start，当前样本的里程值 mileage_now。

每遍历一个样本，如果设备号没有变化，更新当前 mileage_current = mileage_now - mileage_start；如果设备号变化，将 mileage_total 增加 mileage_current，mileage_current 赋值为 0，mileage_start = mileage_now。

2.4.2 异常里程数处理

在数据预处理一节，我们展示了里程数存在的两种异常，分别为记录的异常和记录的缺失。

对于记录的异常，我们将突变的数据项直接删除，这样做并不影响我们对里程数进行累加；对于记录的缺失，我们选择不作处理，因为车辆在这段缺失的数据中同样发生了移动，行驶了对应长度的里程。

在完成了每辆车的里程统计后，我们发现，存在车辆里程数在整个数据集中保持不变的情况，且对应的速度记录也全部为 0。但是，经过分析后发现，这些车辆的某些属性依然发生了变化，可以判断是传感器的记录异常。对此，我们采取的策略是，将其他属性全部设置为保持初始数值不变，这样才能保持数据的一致性。

2.5 平均速度计算

车辆平均行驶速度的计算有三种策略：根据传感器记录的速度（gps_speed）计算、里程数（mileage）除以时间、经纬度变化推算的距离变化值除以时间。在数据预处理部分，我们知道，经纬度存在较多异常值和缺失值，通过该属性计算出的速度包含错误较多，不适合我们用来记录和分析。因此，我们主要采用前两种方法，分析车辆的平均速度。

2.5.1 瞬时速度求平均速度

通过瞬时速度求平均速度是很方便的，只需要得到每个非零速度记录时刻，在时间跳

变时速度保持为 0 的条件下，根据每条数据的时间戳变化，和速度求积后累加，即可得到行驶的总里程，再除以时间戳变化的累积量，即得到平均速度，公式如下：

$$总里程S = \sum_{i\in非零速度时刻} v_i \cdot \Delta t_i$$

$$总时间T = \sum_{i\in非零速度时刻} \Delta t_i$$

$$平均速度V = \frac{S}{T}$$

2.5.2 里程数求平均速度

本题的数据存在停车休息、数据缺失、超长怠速等情况（问题分析中提到的几种情况），会影响总里程和总时间的计算。为了避免这个问题，我们就根据这些情况，将行驶的过程分段，求每一段的平均速度。

引入变量 V_current，表示每段的平均速度；mileage_current,表示每一段的行驶里程；time_current,表示每一段的行驶时间，则

$$V\_current = \frac{mileage\_current}{time\_current}$$

再引入列表变量 speed[],存储每一段的平均速度; n=len(speed),为列表长度，即整个行驶过程分成的段数，则

$$V\_average = \frac{\sum_{0}^{n} speed[i]}{n}$$

1. 问题分析

初步分析：根据里程数求平均速度，主要得到过程中的里程数和行驶时间。我们发现数据中有五个情况会影响我们得到这两个值：

（1）同一辆车，在行驶的整个过程中，记录行驶数据的的设备号发生变化，导致里程数的记录发生巨大跳跃。这种情况需要对里程数的计算做一定处理

（2）相邻两个样本的时间差小于 20 秒，但里程数变化大于 2 公里，这种情况速度至少为 100 米/秒，显然是错误数据，也要判断出来并丢弃这部分数据。

（3）相邻两个样本时间差大于三个小时，且里程数也有变化（查阅资料，显示堵车时间平均为 2 个多小时，考虑到特殊情况，可能会有三个小时。所以时间差小于三个小时时，只要里程数在变，就有可能是堵车，这时数据在求平均速度时需要保留。），这时根据里程数的变化，求这几个小时内的平均速度，如果太小（给定一个值，9 米/秒，根据国家法律规定的机动车行驶速度，最低就是 30 公里/小时），则认为其在这段时间不是一直在行驶，我们也无法确定其行驶时间，就丢弃这部分数据。如果大于这个值，我们就认为

在正常行驶，算平均速度时考虑这部分数据。

（4）相邻两个样本时间差大于两分钟（查阅资料，一般等红灯时间不超过两分钟），且经纬度都没有变化，则认为在这段时间间隔内车辆没有在行驶。（为什么此处用经纬度来判断，而不用里程。因为里程是以公里为单位的，如果里程的变化小于 1 公里，则里程的值不会改变，无法判断是否在行驶。而经纬度的变化则比较准确，及时）

（5）速度为零情况下，可能停车，也可能堵车慢速行驶，根据速度为零时经纬度是否变化判断时慢速行驶还是停止的。

2. 相关变量的符号描述以及算法流程

本节给出里程数求所涉及到的相关概念及形式化描述。本节所用到的符号在表 1 中给出。

表 1　计算平均速度所用到的符号

符号	意义
count	标记当前为第几个样本
time_current	当前阶段的行驶时间
time_start	计算当前阶段行驶时间的起始时间
mileage_current	当前阶段的行驶里程
mileage_start	计算当前阶段行驶里程的起始里程
device_num_now	当前样本的设备编号
last_device_num	上一个样本的设备编号
time_now	当前样本的时间
timestamp	当前样本时间的时间戳
last_time	上一个样本的时间（时间戳）
mileage_now	当前样本的里程数
speed_now	当前样本的速度
last_speed	上一个样本的速度
lng_now	当前样本的经度
last_lng	上一样本经度
lat_now	当前样本纬度
last_lat	上一样本纬度
speed = [ ]	#存放每一个阶段的平均速度

当前阶段，是指初步分析中几种情况以及停车这几种状态中，任意两种状态之间的阶段。因为每遇到这几种情况都要更新一下变量值，所以每个阶段求一个平均速度，最后求整个过程的平均速度。

算法流程如图 31（注：虽然经纬度在整个数据集上来看，存在突变偏离的情况，但如果考虑的是局部两个样本间的经纬度的变化，则偏离对此影响并不大，因为我们此时关注的的是一个相对的情况，而不是一个绝对的情况。）

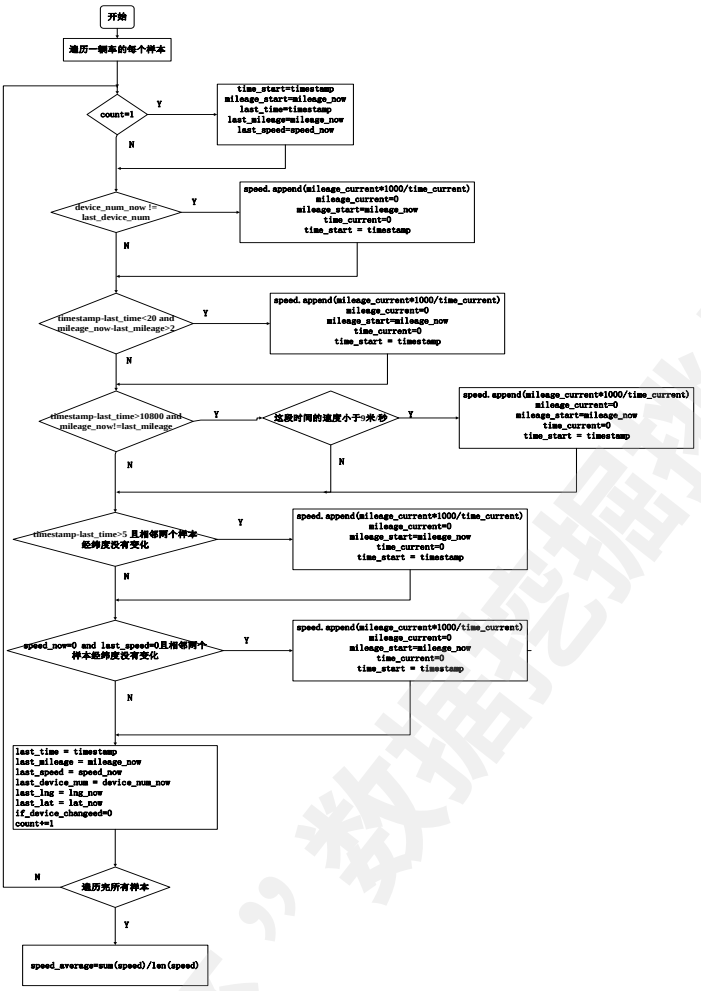

图 31　里程数求平均速度流程图

（1）对原数据按行进行遍历，即遍历每一个样本，当 count=1 时，time_start，mileage_start，last_time ，last_mileage，last_speed 等值置为第一个样本对应的值，以下便是判断各种异常情况。

（2）如果 device_num_now != last_device_num: 即设备号发生变化，则要先计算之前这一阶段的平均速度，再将当前阶段的行驶时间和行驶里程置为 0，更新时间记录的起始值和里程记录的起始值为当前样本的起始值。

（3）如果 timestamp-last_time<20 且 mileage_now-last_mileage>2，即相邻两个样本的时间差小于 20 秒，但里程数变化大于 2 公里,显然是个异常情况，速度不可能达到 100 米/秒这么大。于是停止行驶时间和行驶里程的更新，先计算之前这一阶段的平均速度，再将当前阶段的行驶时间和行驶里程置为 0，更新时间记录的起始值和里程记录的起始值为当前样本对应的值。

（4）如果 timestamp-last_time>10800 且 mileage_now!=last_mileage，即两个样本的时间差大于 3 个小时且当前样本的里程数的值和上一个样本的里程数的值不一样，再断一下这段时间内的平均速度。

如果这段时间的速度小于 9 米/秒， 认为这期间有很长一段时间都是停止的状态，计

算之前这一阶段的平均速度，再将当前阶段的行驶时间和行驶里程置为 0，更新时间记录的起始值和里程记录的起始值为当前样本对应的值。

（5）如果 timestamp-last_time>5 且 lng_now==last_lng 和 lat_now==last_lat，即相邻两个样本的时间差大于 2 分钟，且这段时间内经纬度没有变化。 将计算之前这一阶段的平均速度，再将当前阶段的行驶时间和行驶里程置为 0，更新时间记录的起始值和里程记录的起始值为当前样本对应的值。

（6）如果相邻两个样本的速度均为零，且经纬度的值也没有变化，则认为速度真的为零，车子停止，不再行驶。计算平均速度时就不考虑这部分时间。

于是计算之前这一阶段的平均速度，再将当前阶段的行驶时间和行驶里程置为 0，更新时间记录的起始值和里程记录的起始值为当前样本对应的值。

3. 平均速度结果分析

有问题背景内容知该运输企业所辖各车辆均存在常规运输路线与驾驶人员，则每辆车求得的平均速度，可以反映出该辆车对应的运输路线的路况（主要是限速情况，平均速度大的可能运输路线中以高速为主，而平均速度小的可能以低速道路为主。）以及对应驾驶员驾驶的快慢等。先通过箱尾图对平均速度的整体情况做大致分析，如图 32：

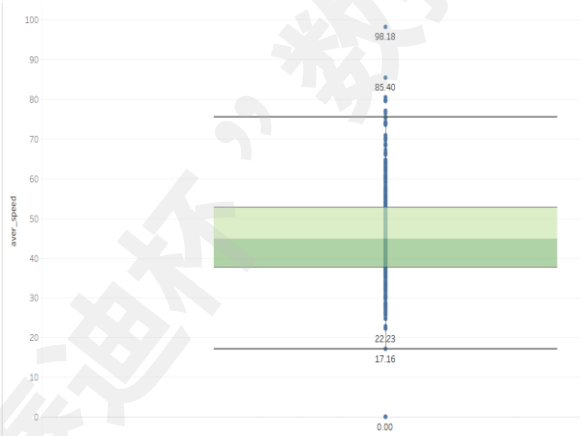

图 32　平均速度的箱尾图分析

由箱尾图可知平均速度的中位数，上下四分位数，都集中在 40-50km/h,整体合理的一个平均速度范围 20-75 km/h 之间。对图中显示的异常点进行分析，编号为 AD00112 的车辆记录的数据中速度 gps_speed 都是零，且里程数 mileage 也始终没有变化，这应该是数据记录过程出现的错误。平均速度过大的也是异常值，对应的车辆编号为 AD00117，对该辆车记录的数据进行分析，发现存在两个样本间时间间隔很小，车载记录终端的设备号也没有变化，但里程数变化特别大的情况。这种情况根据里程数计算出的平均速度就会偏大，导致最后得出的平均速度也过大。

下面再以 10km/h 为间隔，对所有车辆平均速度的分布情况做分析，如图 33：

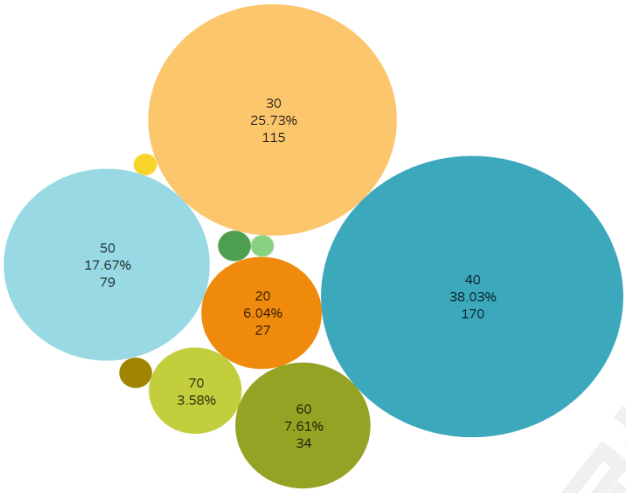

图 33　平均速度的气泡图分析

由该图很清晰的看出平均速度主要集中在 30,40,50km/h 这一个范围之内，说明该运输企业的所有路线图中还是以正常速度限制的道路为主，可能是省道、县道等，也有可能是因为这些线路上的车辆的载重量很大，导致行驶的速度不能太快。

平均速度达到 70km/h 及以上的情况比较少，说明运输线路中以高速为主的情况比较少，可能是因为中短途运输路线比较多，对运输时间要求不高，而且走高速还需要交过道费，支出与时间上的收益的比例不合适时，就不会选择高速了。

2.5.3　两种方法的合并

我们通过两种方法得到了车辆在行驶过程中的平均速度，两种平均速度分别依赖于速度记录（gps_speed）的准确性和里程（mileage）的准确性，在信息置信度水平上各有侧重，因此这两种方法计算出的平均速度均存在信息的不完整性。

对此，我们采取的策略是，将两种方法计算出的平均速度再求一次平均，得到的即为最终的平均速度。

2.6　车辆行程划分

在分析车辆驾驶行为时，有两种分析思路：既可以对一辆车的整体数据进行分析，也可以将整体数据划分行程，分析每段行程的行为。对车辆划分行程后再进行分析，得到的信息会更加具体和细致。

下面给出了划分车辆行程的具体方法。

2.6.1　行程切换的标志

在数据集中，属性 acc_state 表示车辆的点火熄火状态，具体含义如下：

$$\begin{cases} acc\_state = 1, & \text{点火} \\ acc\_state = 0, & \text{熄火} \end{cases}$$

该属性有助于我们划分车辆的行程，一般认为，两端行程之间一定存在一个过程，那就是：车辆结束上一次行程，停车熄火停车；车辆开始新的行程，点火启动。以熄火和点火的角度分析，acc_state 是车辆行程切换的一个标志之一。

除此之外，我们认为，时间发生突变也是行程切换的标志之一，因为对于驾驶员，在两次行程之间，总会有一段休息时间。长则数天，短则若干分钟。因此，我们也可以对时间的变化规定一个阈值T_1，如果相邻两条数据的间隔时间大于该阈值，则认为已经发生了一次行程切换。判断公式如下，对于两个相邻记录的时间t_1, t_2：

$$\begin{cases} |t_1 - t_2| > T_1, & \text{行程发生切换} \\ |t_1 - t_2| \le T_1, & \text{行程未切换} \end{cases}$$

2.6.2 缺失和异常的处理

对于理想的数据记录，单独分析 acc_state 的变化即可对车辆行程进行切分，但是对于本数据集，其中包含的异常是需要我们特别关注的。

首先是记录的缺失，对于数据集，记录的缺失是很普遍的，如果缺失的恰好为行程切换的数据（事实上，多数缺失都发生在行程切换时），就会对行程划分造成影响，如图 34。

20978	AB00006	AAA7109C	148	115.6464	24.96664	1	2018/8/1 0:19:56	35	4634
20979	AB00006	AAA7109C	150	115.6465	24.96656	1	2018/8/1 0:19:57	35	4634
20980	AB00006	AAA7109C	152	115.6465	24.96648	1	2018/8/1 0:19:58	35	4634
20981	AB00006	AAA7109C	154	115.6465	24.96641	1	2018/8/1 0:19:59	35	4634
20982	AB00006	AAA7109C	158	115.6466	24.96633	1	2018/8/1 0:20:00	34	4634
20983	AB00006	AAA7109C	156	115.6466	24.96625	1	2018/8/1 0:20:01	35	4634
20984	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:32	0	5030
20985	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:35	0	5030
20986	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:36	0	5030
20987	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:37	0	5030
20988	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:39	0	5030
20989	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:40	0	5030
20990	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:41	0	5030
20991	AB00006	AAA7109C	0	115.649	24.95945	1	2018/8/3 3:30:43	0	5030

图 34　行程切换的数据缺失

图中，acc_state 始终保持为 1，但是时间直接突变到了两天后，说明行程的改变并没有被记录。

此外，还有 acc_state 异常变化的情况，尽管记录发生了改变，但实际上行程并没有改变，如图 35。

27813	AA00001	AAA9102C	30	115.8607	28.81639	0	2018/8/5 9:29:41	0	5131
27814	AA00001	AAA9102C	30	115.8607	28.81639	0	2018/8/5 9:29:42	0	5131
27815	AA00001	AAA9102C	30	115.8607	28.81639	0	2018/8/5 9:29:43	0	5131
27816	AA00001	AAA9102C	30	115.8607	28.81639	0	2018/8/5 9:29:44	0	5131
27817	AA00001	AAA9102C	30	115.8608	28.81643	1	2018/8/5 9:29:45	0	5131
27818	AA00001	AAA9102C	30	115.8608	28.81643	1	2018/8/5 9:29:46	0	5131
27819	AA00001	AAA9102C	30	115.8608	28.81643	1	2018/8/5 9:29:47	0	5131
27820	AA00001	AAA9102C	30	115.8608	28.81643	1	2018/8/5 9:29:48	0	5131
27821	AA00001	AAA9102C	30	115.8608	28.81643	1	2018/8/5 9:29:49	0	5131
27822	AA00001	AAA9102C	34	115.8608	28.81644	1	2018/8/5 9:29:50	0	5131
27823	AA00001	AAA9102C	34	115.8608	28.81645	1	2018/8/5 9:29:51	0	5131
27824	AA00001	AAA9102C	34	115.8608	28.81645	1	2018/8/5 9:29:52	0	5131
27825	AA00001	AAA9102C	34	115.8608	28.81644	1	2018/8/5 9:29:53	0	5131
27826	AA00001	AAA9102C	34	115.8608	28.81644	1	2018/8/5 9:29:54	0	5131
27827	AA00001	AAA9102C	34	115.8608	28.81644	1	2018/8/5 9:29:55	0	5131
27828	AA00001	AAA9102C	34	115.8608	28.81644	0	2018/8/5 9:29:56	0	5131
27829	AA00001	AAA9102C	34	115.8608	28.81644	0	2018/8/5 9:29:57	0	5131
27830	AA00001	AAA9102C	34	115.8608	28.81644	0	2018/8/5 9:29:58	0	5131

图 35　acc_state 的异常变化

图中，时间始终保持连续增长并没有改变，但是车辆却经历了从熄火到点火再到熄火的转换，根据时间判断，此段记录应当属于同一行程。

针对上述的两种异常情况，我们采取的处理措施是，通过 acc_state 和 location_time 两个属性来划分行程，并定义了 2 个阈值，分别是行程最短间隔时长T_1和熄火最长等待时长T_2，二者含义如下：

如果两条相邻数据时间间隔$\Delta t > T_1$，即便 acc_state 没有发生变化，我们也认为车辆已经发生了行程切换；

如果两条相邻数据时间间隔$\Delta t \le T_2$，即便 acc_state 发生了变化，我们也认为车辆没有发生行程切换。

我们规定，$T_1 = 20\,\text{min}$，为了和疲劳驾驶的最小休息间隔保持一致；$T_2 = 1\text{min}$，含义是，一分钟内的车辆重启均不视作行程的改变。

通过以上判断，我们就可以消除上述异常对行程划分造成的影响。

2.7 结果的展示

本章首先对给定的车辆数据进行了预处理包括（去重、0 值填充等）；其次，根据车辆的状态属性以及经纬度的变化等，对每辆车的行程路线的起始点进行划分，并对异常经纬度数据进行纠偏操作，最后，根据纠偏后的每辆车的每段行程分析其总的里程数、平均速度以及急加速和急减速情况（急加速和急减速情况在我们第三章特征提取的时候进行详细地描述）。图 36，图 37 展示了车辆 AA00002 的路线图以及不同的行程的里程数、平均速度以及急加速和急减速情况，其他的车辆详细数据请查看我们的数据附件 1。同时如表 2，我们也展示了每辆车总的行车里程以及平均速度和急加速和急减速情况。

表 2　车辆平均速度、里程数、急加速急减速次数

	AA00002	AB00006	AD00003	AD00013	AD00053	AD00083	AD00419	AF00098	AF00131	AF00373
aver_speed(m/s)	14.6	11.33	13.64	7.92	9.8	13.63	12.76	12.93	10.69	10.84
mileage(km)	749	12200	15138	7565	779	814	4197	1032	703	2587

acc_time	188	26	168	457	47	538	9	11	0	12
dec_time	190	29	183	550	38	598	14	53	5	9

图 36　AA00002 的路线图

file_name	trip_num	aver_speed(m/s)	total_mileage(km)	acc_times	dec_times
AA00002	1	15.65	98	23	24
AA00002	2	11.71	21	14	13
AA00002	3	18.48	333	25	26
AA00002	4	15.94	229	32	34
AA00002	5	8.5	10	10	12
AA00002	6	2.85	1	1	1
AA00002	7	9.16	10	15	11
AA00002	8	8.61	20	33	35
AA00002	9	6.07	2	2	2
AA00002	10	5.77	5	9	12
AA00002	11	10.01	20	24	20

图 37　AA00002 不同的行程的里程数、平均速度以及急加速和急减速情况

第三章　不良驾驶行为的特征提取

在车辆运输过程中，不良驾驶行为主要包括疲劳驾驶、急加速、急减速、怠速预 热、超长怠速、熄火滑行、超速、急变道等[2]；对于时间的计算处理，在 Python 中都转换成时间戳，且时间戳变化 1，表示时间变化 1 秒。没有特别说明的时间都为时间戳格式。

3.1　疲劳驾驶

3.1.1　定义简述

根据国家相关交通法规规定：机动车驾驶人在 24 小时内累计驾驶时间不得超过 8 小时，连续驾驶时间不得超过 4 小时，每次停车休息时间不少于 20 分钟。

又根据相关论文，驾驶员在午饭后连续行驶 30 分钟之后就容易疲劳驾驶，所以我们也将从午饭后开始连续行驶 30 分钟认为开始疲劳[3]。

我们根据上述内容来定义疲劳驾驶，单次连续驾驶超过 4 小时，期间单次休息时间小于 20 分钟，则为单次连续疲劳驾驶。午饭后（根据本题数据无法判断驾驶员何时吃的午饭，我们将 11 点多和 12 点多开始的行驶过程认定为午饭后开始的行程。）连续驾驶 30 分钟，也认为单次连续疲劳驾驶。单日累计驾驶时间超过 8 小时，则为日累计疲劳驾驶。如下所示，

为判断单次连续疲劳驾驶，引入标记 single_warned 其定义如下：

$$single\_warned=\begin{cases}0，\text{未单次疲劳驾驶}\\1，\text{单次疲劳驾驶}\end{cases}$$

为了记录单次连续驾驶时间，引入单次连续驾驶时间 T_run，如果 T_run>14400,则令 single_warned=1。

为判断日累计疲劳驾驶，引入标记 day_warned 其定义如下：

$$day\_warned=\begin{cases}0，\text{未日累计疲劳驾驶}\\1，\text{日累计疲劳驾驶}\end{cases}$$

为了记录单次连续驾驶时间，引入单次连续驾驶时间 T_total，如果当日内行驶过程中休息时间达到了 20 分钟，则 T_total=T_total+T_run。

如果 T_tootal>28800,则令 day_warned=1。

3.1.2　问题分析

考虑疲劳驾驶最主要的还是要考虑休息时间入手，休息时间有三种情况：

(1)相邻两个样本的 local_time 时间差大于 20 分钟。这里面又分两种情况：

a)这两个相邻样本的 mileage 值相同，说明在这一段时间间隔内，车辆没有在运动，而是在休息

b)这两个相邻样本的 mileage 值不同，说明在这一段时间间隔内，车辆还在行驶，并没有完全停下来，所以并不能算在休息。

(2)gps_speed 为 0 的状态持续 20 分钟

只要速度为零，就说明车子没有在行驶，如果连续 20 分钟，虽然记录的速度为零也有可能是因为速度太慢小于 10，设备由于精度无法记录。但以这种极慢的速度持续行驶超过 20 分钟，现实中不太可能存在，就算存在，以这种速度行驶，我们也认为休息达到了 20 分钟。

(3)acc_state 为 0 的状态持续 20 分钟

对数据简单分析后，发现这种情况不存在，因为如果有记录的样本的 acc 为熄火状态，则其 gps_speed 一定为 0，所以这种情况已经在情况二中讨论过了。

3.1.3 相关变量的符号描述以及算法流程

本节给出疲劳驾驶问题所涉及到的相关概念及形式化描述。本节所用到的符号在表 3 中给出。

表 3 描述疲劳驾驶所用符号

符号	意义
count	标记当前为第几个样本
T_total	当天累计驾驶时间
T_run	当前单次驾驶时间（从上一次休息 20 分钟开始，还没到下一次休息）
T_start	单次连续驾驶时间时的起始时间。（休息时间满足 20 分钟后又开始行驶的开始时间或为数据集的第一条数据）
speed_zero_start	记录速度开始为零的时刻
speed_zero_time	记录速度为零的持续时间，如果速度为零时间超过 20 分钟，也认为其休息了 20 分钟
last_time	上一个样本时间的时间戳
last_day	上一个样本时间的日期
last_mileage	上一个样本的里程
last_speed	上一个样本的速度
time_now	当前样本的时间（local_time）
timestamp	当前样本时间的时间戳
hour_start	记录单次连续驾驶的起始时间为几时
speed_now	当前样本的速度
this_day	当前样本的日期，比如 this_day=2018-08-06
single_warned	标记单次连续疲劳驾驶
day_warned	标记当日累计疲劳驾驶

算法流程如图 38：

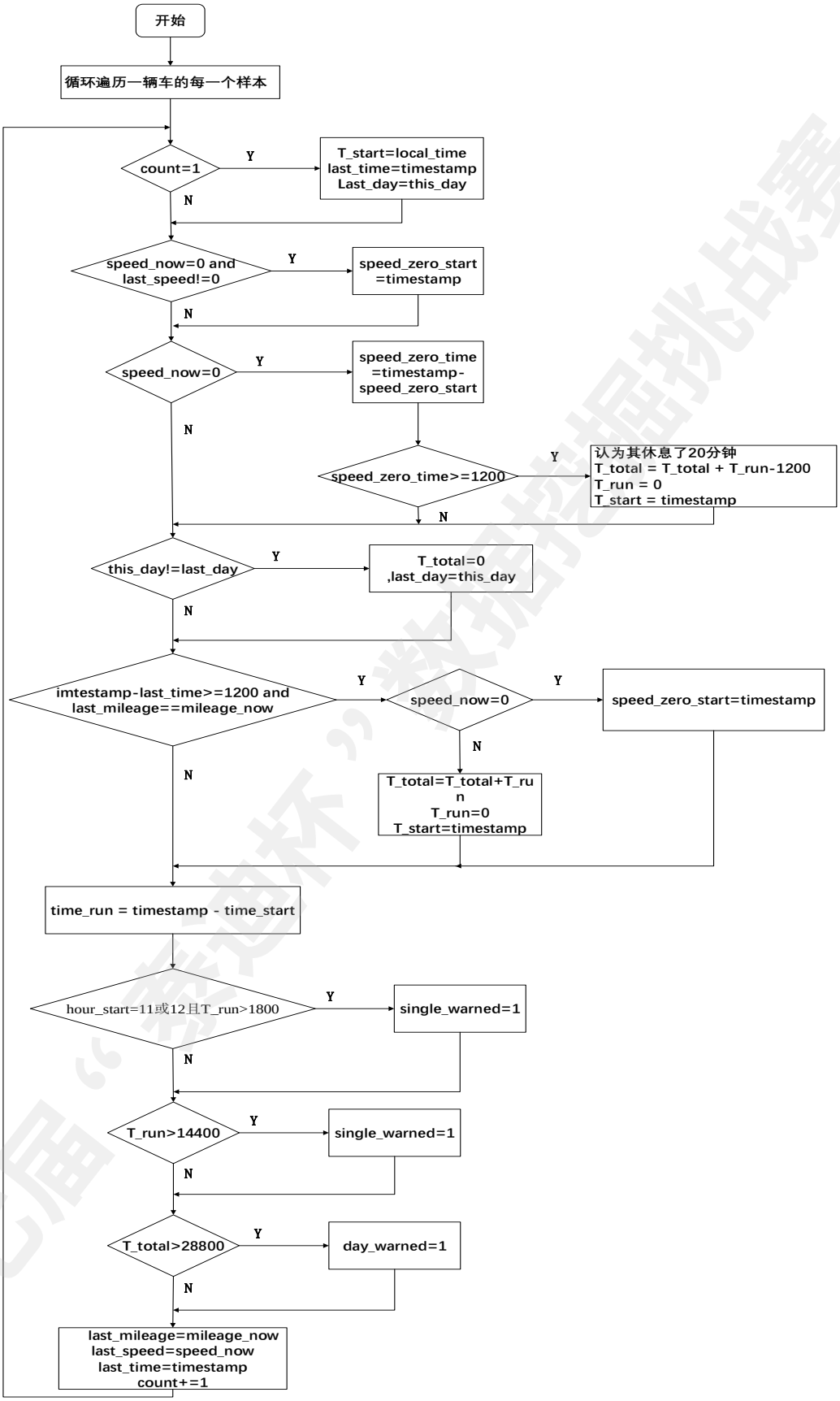

图 38　疲劳驾驶的判断流程图

（1）对原数据按行进行遍历，即遍历每一个样本，如果 count=1,即当前为第一个样本，T_start 最置为第一个样本的 local_time，last_time 也置为当前样本的时间戳，last_day 也置为当前样本的日期。

（2）如果 speed_now=0 且 last_speed!=0，说明此时速度刚开始为零，记录这个刻。令 speed_zero_start=timestamp

（3）如果 speed_now=0，则更新速度为零时的持续时间 speed_zero_time=timestamp-speed_zero_start，如果 speed_zero_time>=1200，即如果速度为零时间超过 20 分钟，认为其休息了 20 分钟。

更新当日累计驾驶时间，T_total = T_total + T_run-1200 ，将单次连续驾驶时间置为 0 ，T_run = 0，更新单次连续驾驶的起始时间为当前样本时间，T_start = timestamp 。

（4）如果 this_day!=last_day，当前样本的日期和上一个样本的日期不一样，则说明已经过了一天了，将当日累计驾驶时间置为 0，并将上一个样本的日期更新，以便于后面进行比较 T_total=0 ,last_day=this_day（注：我们的累计疲劳驾驶是按同一天内（比如 8 月 4 日）累计驾驶大于 8 小时算，而不是算一个完整的 24 小时）

（5）如果 timestamp-last_time>=1200 且 last_mileage==mileage_now，两个相邻样本的时间间隔大于 20 分钟且在这段间隔时间内里程数没有变化，则认为这段时间满足休息要求。

如果此时 speed_now=0，当前速度为零，则要将 0 速度的计时起点更新为当前时间（因为记录速度为零的时间也是为了判断是否休息了 20 分钟，这种情况下，已经休息了 20 分钟，如果速度为零的计时起点不更新，就会重复计算了），speed_zero_start=timestamp。

更新累计驾驶时间 ，T_total=T_total+T_run ，将单次连续驾驶时间置为 0，T_run=0 ，更新单次连续驾驶的起始时间为当前样本时间 T_start=timestamp 。

(6)经过上面几步的判断更新，最后判断 T_total 和 T_run 的值，如果 T_total>28800,则令 day_warned=1 当日累计疲劳驾驶；如果 T_run>14400,则令 single_warned=1 单次连续疲劳驾驶；如果 hour_start=11 或 12 且 T_run>1800,则令 single_warned=1 单次连续疲劳驾驶。

3.1.4 疲劳驾驶结果分析

对于日累计疲劳驾驶的情况，我们认为从整体上来看，与具体哪一天发生并没有什么关系，所以我们只给出样本中车辆的日累计疲劳驾驶情况，不做过多分析。

对于单次连续疲劳驾驶的情况，我们认为这种情况的发生与一天之内的具体时间段有关，所以主要关注的是在哪些时间段容易发生单次连续疲劳驾驶的情况。对此进行分析，将一天分成四个时间段 00:00-06:00，06:00-12:00，12:00-18:00，18:00-24:00（不过在对数据进行可视化时发现在 18:00-24:00 这个时间段没有发生单次连续疲劳驾驶）。

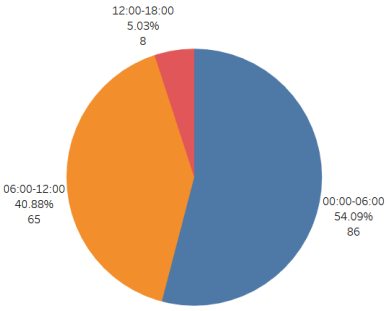

图 39　单次连续疲劳驾驶时间段分布图

由图 39 可知，单次连续疲劳驾驶主要发生在 06:00-12:00，00:00-06:00 这两个时段，其中 00:00-06:00 这个时间段发生的次数最多，而在 18:00-24:00 这个时间段发生的情况很少，几乎没有，可以不去分析。

在 00:00-06:00 夜间还在行驶，可能本身时间就比较急，不然也不会深夜还在行驶当中，可能有什么货物要第二天一早就要送到，或其他原因时间特别紧，所以行驶过程中就会减少休息时间，就容易发生单次连续疲劳驾驶，导致在夜间这个时间段疲劳驾驶比例很高。

在 06:00-12:00 这个时间段。有可能夜间凌晨一直赶时间在行驶，导致上午就出现疲劳驾驶的情况，也要可能是因为休息了一晚上，精神状态比较好，自己行驶过程中还没有感觉到特别劳累，休息的时间就比较少，造成了疲劳驾驶。这两种原因导致了这个时间段的单次疲劳驾驶情况比例比较高。

在 12:00-18:00 和 18:00-24:00 这个时间段，一般人下午的精神状态都没有上午好，所以大多数人中午都会选择午休。所以下午整个人容易犯困，驾驶员更容易感觉到自身的疲惫，也就会增加自身休息的时间，导致连续长时间驾驶情况减少，疲劳驾驶出现的情况就比较少。

3.2　怠速预热、超长怠速

3.2.1　定义简述

怠速预热：怠速预热并不都是不良驾驶行为，主要和预热的时间有关。车辆刚启动时。发动机类温度还没有达到正常工作时的温度（90-100 度左右），如果这时就行驶，会导致发动机类燃油燃烧不充分，不仅耗油，还会有很多为燃尽的杂质积累，损坏车辆。所以车辆刚启动有必要进行预热，但时间控制在 1 分钟以内比较合适，超过一分钟就会起到反作用，损毁车辆。所以我们定义怠速预热超过 1 分钟才为不良驾驶行为。,称其为不良怠速预热。

超长怠速：根据怠速的时间长短来定义超长怠速，我们认为 2 分钟以内怠速为正常情况，可能是在等红灯。2 到 10 分钟之间为长时间怠速，超过 10 分钟才为超长怠速。

为判断不良怠速预热，引入标记 abnormal_idling_prehea 其定义如下：

$$abnormal\_idling\_prehea = \begin{cases} 0, 未出现不良怠速预热 \\ 1, 不良怠速预热 \end{cases}$$

为了记录怠速预热时间，引入怠速预热时间 time_preheat，如果 time_preheat>60,则令 abnormal_idling_prehea =1。

为判断超长怠速，引入标记 idling_overtime 其定义如下：

$$idling\_overtime = \begin{cases} 0, 未出现超长怠速 \\ 1, 超长怠速 \end{cases}$$

为了记录怠速持续时间，引入怠速时间 time_overtime，如果 time_overtime>600,则令 idling_overtime =1。

3.2.2 问题初步分析

因为怠速判断与速度是否为真零关系很大，所以判断怠速的数据是进行过 0 速度填充后的数据，将那些可能大于 0，但小于 10，以致于设备无法显示的速度，修改成为其最可能的值。

怠速预热：

（1）考虑 acc_state 状态从 0 变到 1 的时刻，这时候车辆刚点火启动，怠速预热就发生在这个时刻往后 acc_state 为 1 的一段时间内。如果从这个时刻开始，速度为 0 状态持续时间超过 1 分钟，就认为其怠速预热。

（2）如果从上述状态开始之后，相邻两个速度为零的样本的 acc_state 状态都为 1，但如果这两个样本的的经纬度不相同，说明在这两个样本之间车辆至少有一段时间是在在行驶的，所以我们认为这不是是一个怠速过程。

超长怠速：

（1）判断 acc_state 为 1 时，速度为零的持续时间是否超过 10 分钟。

（2）同样的，如果相邻两个速度为零的样本的 acc_state 状态都为 1，但如果这两个样本的的经纬度不相同，说明在这两个样本之间车辆至少有一段时间是在在行驶的，所以我们认为这不是一个怠速过程。

3.2.3 相关变量的符号描述以及算法流程

怠速预热：本部分所用到的符号在表 4 中给出。

表 4 怠速预热所用到的的符号

符号	意义
count	标记当前为第几个样本
time_preheat	怠速预热的时间
time_preheat_start	怠速预热的起始时
time_now	当前样本的时间

timestamp	当前样本的时间戳
last_time	上一个样本的时间（时间戳）
speed_now	当前样本的速度
last_speed	上一个样本的速度
acc_now	当前样本的 acc 状态
last_acc	上一个样本的 acc 状态
lng_now	当前样本经度
last_lng	上一样本经度
lat_now	当前样本的纬度
last_lat	上一样本纬度
if_idling_prehea	标记车辆是否处于怠速预热状态
abnormal_idling_prehea	标记车辆为不良怠速预热

算法流程如图 40：

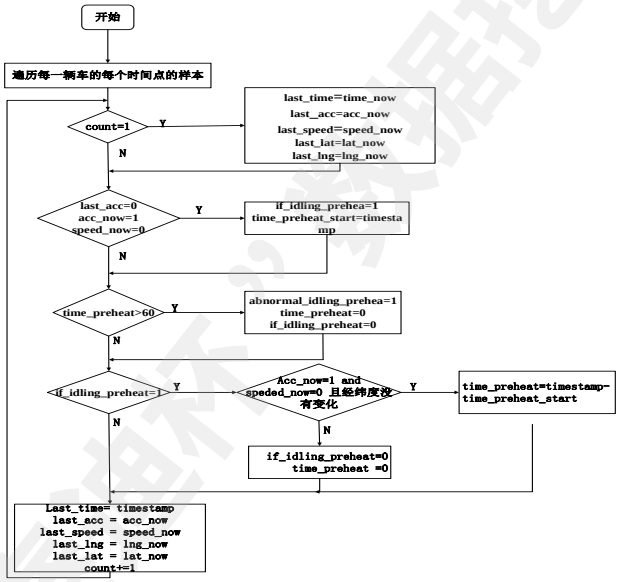

图 40　怠速预热判别算法流程图

（1）对原数据按行进行遍历，即遍历每一个样本，当前遍历的为第一个样本时，并将 last_time ，last_acc，last_speed,last_lat,last_lng 等值置为第一个样本对应的值。

（2）如果 last_acc=0 且 acc_now=1，还满足当前速度 speed_now=0，说明此时车辆刚启动，速度为 0，处于怠速状态。将标记值 if_idling_preheat 置为 1 ，开始怠速预热，并将怠速预热的起始时间 time_preheat_start 设为当前时间 timestamp。开始记录怠速持续的时间。

（3）如果怠速持续时间 time_preheat 大于 60 秒，则认为是不良驾驶行为。将 abnormal_idling_prehea 置为 1，输出怠速预热的这段时间，将怠速预热持续时间 time_preheat 置为 0，怠速预热标记值 if_idling_preheat 置为 0，重新开始判断下一个怠速预热时间段。

（4）如果 if_idling_preheat=1，说明此时已经开始了怠速预热状态，进而判断 acc_state 是否为 1 以及速度 gps_speed 是否为 0。如果不满足上述条件，则说明怠速预热状态结束，

将怠速预热持续时间 time_preheat 置为 0，怠速预热标记值 if_idling_preheat 置为 0，准备判断下一个怠速预热状态。如果满足上述条件，再判断相邻两个样本的经纬度是否一样，如果一样，则还在怠速预热状态，更新怠速预热持续时间 time_preheat，否则怠速预热状态结束，将怠速预热持续时间 time_preheat 置为 0，怠速预热标记值 if_idling_preheat 置为 0，准备判断下一个怠速预热状态。

超长怠速：本部分所用到的符号在表 5 中给出。

表 5　超长怠速所用到的的符号

符号	意义
count	标记当前为第几个样本
time_overtime	超长怠速的时间
time_overtime_start	超长怠速的起始时间
time_now	当前样本的时间
timestamp	当前样本的时间戳
last_time	上一个样本的时间（时间戳）
speed_now	当前样本的速度
last_speed	上一个样本的速度
acc_now	当前样本的 acc 状态
last_acc	上一个样本的 acc 状态
lng_now	当前样本经度
last_lng	上一样本经度
lat_now	当前样本的纬度
last_lat	上一样本纬度
if_idling	标记车辆是否处于怠速状态
idling_overtime	标记车辆超长怠速

算法流程如图 41：

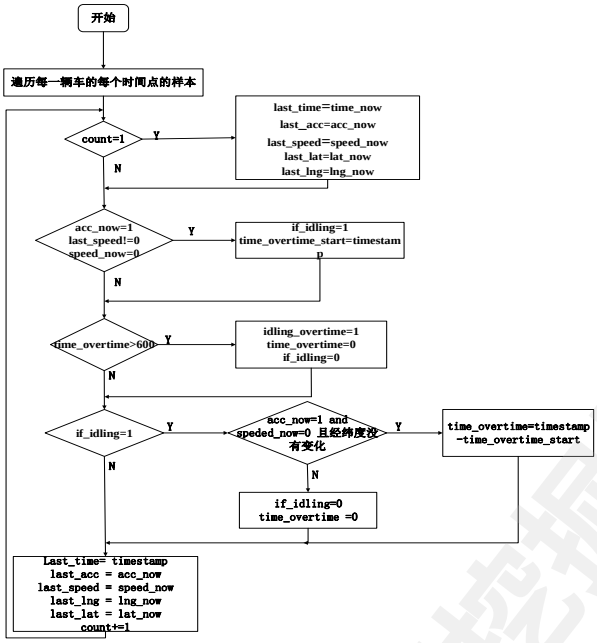

图 41　超长怠速判别算法流程图

（1）对原数据按行进行遍历，即遍历每一个样本，当前遍历的为第一个样本时，并将　last_time ，ast_acc，last_speed,last_lat,last_lng 等值置为第一个样本对应的值。

（2）如果 acc_now=1 且 last_speed != 0 且 speed_now = 0，说明开始怠速状态，将标记值 if_idling 置为 1 ，开始怠速预热，并将超长怠速的起始时间 time_overtime_start 设为当前时间 timestamp。开始记录怠速预热持续的时间。

（3）如果怠速预热持续时间 time_overtime 大于 600 秒，则认为是超长怠速。将 idling_overtime 置为 1，输出超长怠速的这段时间，将怠速持续时间 time_overtime 置为 0，怠速标记值 if_idling 置为 0，重新开始判断下一个怠速时间段。

（4）如果 if_idling=1，说明此时已经开始了怠速状态，进而判断 acc_state 是否为 1 以及速度 gps_speed 是否为 0。

如果不满足上述条件，则说明怠速状态结束，将怠速持续时间 time_overtime 置为 0，怠速标记值 if_idling 置为 0，准备判断下一个怠速状态。

如果满足上述条件，再判断相邻两个样本的经纬度是否一样，如果一样，则还在怠速状态，更新怠速持续时间 time_overtime，否则怠速状态结束，将怠速预热持续时间 time_overtime 置为 0，怠速标记值 if_idling 置为 0，准备判断下一个怠速状态。

3.2.4　结果分析

首先是部分判断的结果，如图 42，图 43：

车辆编号	车辆启动时间	达到不良怠速预热的时间
AA00001	2018-08-04 05:19:13	2018-08-04 05:20:15
AA00001	2018-08-04 11:23:23	2018-08-04 11:24:25
AA00001	2018-08-05 05:24:45	2018-08-05 05:25:47
AA00001	2018-08-06 00:57:15	2018-08-06 00:58:17
AA00001	2018-08-06 02:31:59	2018-08-06 02:33:01
AA00001	2018-08-06 05:07:07	2018-08-06 05:08:09
AA00001	2018-08-06 23:21:45	2018-08-06 23:22:47
AA00001	2018-09-09 00:25:19	2018-09-09 00:26:21
AA00001	2018-09-09 03:33:49	2018-09-09 03:34:51
AA00001	2018-09-10 05:04:30	2018-09-10 05:05:32
AA00001	2018-09-10 07:41:27	2018-09-10 07:42:29
AA00001	2018-09-10 07:48:16	2018-09-10 07:49:18
AA00001	2018-09-11 05:09:08	2018-09-11 05:10:10
AA00001	2018-09-11 05:30:33	2018-09-11 05:31:35
AA00001	2018-09-11 10:47:35	2018-09-11 10:48:37
AA00001	2018-09-12 23:31:01	2018-09-12 23:32:03
AA00001	2018-09-13 04:05:22	2018-09-13 04:06:24
AA00001	2018-09-13 09:29:04	2018-09-13 09:30:06
AA00002	2018-08-05 07:25:22	2018-08-05 07:26:24
AA00002	2018-08-05 08:18:52	2018-08-05 08:19:54
AA00002	2018-08-05 23:19:08	2018-08-05 23:20:10
AA00002	2018-08-06 07:09:16	2018-08-06 07:10:18
AA00002	2018-08-06 08:18:39	2018-08-06 08:19:41
AA00002	2018-08-07 00:00:22	2018-08-07 00:01:24
AA00002	2018-08-07 01:45:07	2018-08-07 01:46:09
AA00002	2018-08-07 03:34:43	2018-08-07 03:35:45
AA00004	2018-08-01 00:50:00	2018-08-01 00:51:02
AA00004	2018-08-02 09:06:06	2018-08-02 09:07:08
AA00004	2018-08-03 00:59:55	2018-08-03 01:00:57

图 42　不良怠速预热部分结果展示

车辆编号	车辆开始怠速时间	达到超长怠速的时间
AA00001	2018-09-13 06:04:46	2018-09-13 06:14:48
AA00036	2018-09-10 09:16:56	2018-09-10 09:26:59
AA00036	2018-09-12 02:03:36	2018-09-12 02:13:38
AA00036	2018-09-13 07:20:01	2018-09-13 07:40:02
AA00045	2018-09-08 01:49:14	2018-09-08 01:59:16
AA00052	2018-09-09 08:22:55	2018-09-09 08:32:58
AA00052	2018-09-09 23:40:01	2018-09-09 23:50:03
AA00055	2018-09-15 01:03:43	2018-09-15 01:13:45
AA00060	2018-09-12 07:41:27	2018-09-12 07:51:29
AA00061	2018-09-10 07:20:01	2018-09-10 07:30:03
AA00061	2018-09-11 07:51:47	2018-09-11 08:01:49
AA00061	2018-09-12 22:21:56	2018-09-12 22:31:58
AA00128	2018-09-13 11:49:22	2018-09-13 11:59:24
AA00143	2018-09-12 07:20:01	2018-09-12 07:30:03
AA00160	2018-09-09 02:18:33	2018-09-09 02:28:35
AA00173	2018-09-08 01:41:23	2018-09-08 01:51:25
AA00173	2018-09-11 01:17:38	2018-09-11 01:27:40

图 43　超长怠速部分结果展示

根据这两种行为判断的结果展示，我们很简单的就可以看出不良怠速预热发生的次数比超长怠速发生的次数要多很多。就 AA00001 这一辆车而言，在整个行驶过程中及发生了 20 次不良怠速预热，只发生了一次超长怠速。

这也与实际情况相符，因为大多驾驶员都认为行车前预热是很正常情况，所以很多人都会在行车前进行怠速预热，尤其在气温比较低时，但大多数人不知道怠速预热也需要一个合理时间，超过一分钟就会对车辆发动机造成轻微损害，减少寿命。对于超长怠速，基本上驾驶员也都是知道会对车辆造成损害，所以发生的情况就会少很多。

下面分析超长怠速和怠速预热主要发生在哪些时间段，分别对数据进行了统计分析，如图 44，图 45：

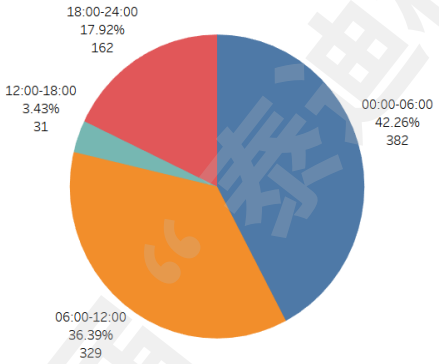

图 44　不良怠速预热时间段分布图

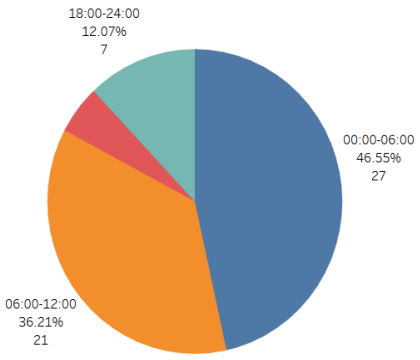

图 45　超长怠速时间段分布图

对于怠速预热情况，主要发生在上午和深夜凌晨：上午可能是一天之中第一次启动车辆，车辆停了一晚上，需要预热才能使发动机效率达到最大，所以热车情况比较多，导致怠速预热时间过长的情况也就比较多了；而在深夜的时间段，可能夜间气温就比较低，车辆启动前都需要进行预热，就导致怠速预热时间过长的情况发生也比较多。

对于超长怠速情况，也是主要发生在上午和深夜凌晨：夜间行车可能困了也不方便找地方休息，只能在车里短暂休息，同时车子也没熄火，发动机还保持运行状态，这就可能

导致超长怠速的情况发生比较多，除此之外也还可能是一些其他原因。

3.3 急加速急减速

在衡量车辆的行驶安全性时，车辆的急加速急减速行为是一项重要指标。尽管该行为本身并不违反交通规定，却可能蕴含着驾驶员在行车过程中出现的超车、变道、急停等不安全举动，不合理的急加速与急减速行为会产生较大的安全隐患，同时频繁的急加速预计减速也会造成能源的浪费及环境污染。

3.3.1 定义简述

在计算车辆加速度时，我们计算车辆在每个时刻的加速度的值，为判断急加速和急减速行为，按照行业经验，引入车辆的加速度阈值：

$$A = 3m/s^2 = 10.8km/h$$

对于车辆某一时刻的加速度 a，如果满足：

$$|a| \geq A$$

则判定处于急加速或急减速状态，再根据加速度的符号确定车辆处于急加速还是急减速：

$$\begin{cases} a > 0, & \text{急加速} \\ a < 0, & \text{急减速} \end{cases}$$

3.3.2 问题分析

在判断车辆的急加速/急减速状态时，我们程序的运行结果给出的是超出规定阈值 A 的车辆加速度、急加速/急减速的开始时间、持续时间。经过初步计算筛选，我们得到的持续时间均为一条记录对应的时间间隔。由于车辆的急加速/急减速行为可能连续发生，还需要解决急加速/急减速行为的合并问题。

1.加速度异常段和异常段的合并

我们的合并策略是，将急加速段和急减速段分别用列表 acc_list 和 dec_list 记录，判断各自的元素是否有时间相邻的情况，如果有，则将相邻时段合并，如图 46。

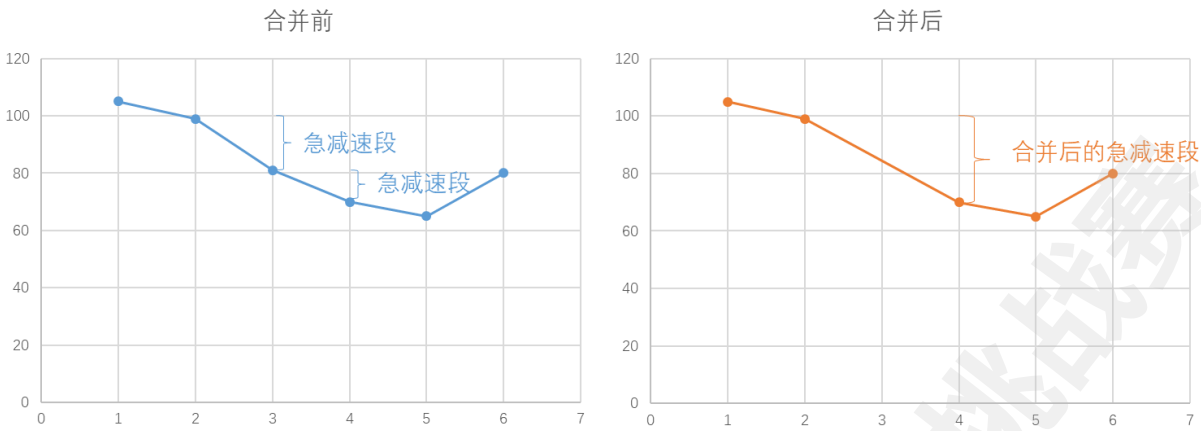

图 46　合并连续加速度异常段

由于在每个记录段的时间内都满足急加速/急减速情况，在合并后的记录区间时间内也一定满足急加速/急减速情况，证明过程如下：

对于两个相邻的急加速段，已知 $a_1 > 3m/s^2$，$a_2 > 3m/s^2$，持续时间分别为 t_1，t_2，有：

$$\begin{aligned}
&\frac{a_1 \cdot t_1 + a_2 \cdot t_2}{t_1 + t_2} \\
&= a_1 \cdot \frac{t_1}{t_1 + t_2} + a_2 \cdot \frac{t_1}{t_1 + t_2} \\
&> 3 \cdot \frac{t_1}{t_1 + t_2} + 3 \cdot \frac{t_1}{t_1 + t_2} \\
&= 3 \cdot \frac{t_1 + t_2}{t_1 + t_2} \\
&= 3m/s^2
\end{aligned}$$

2.加速度异常段和正常段的合并

在合并过相邻的急加速/急减速段后，我们发现，合并区间还可以进一步扩大。在一段连续的加速/减速过程中，可以有急加速/急减速，也可以有正常加速/减速，如果将正常和非正常的加速/减速状态合并，得到的平均加速度也有可能超出阈值 A，达到急加速/急减速的条件，如图 47。

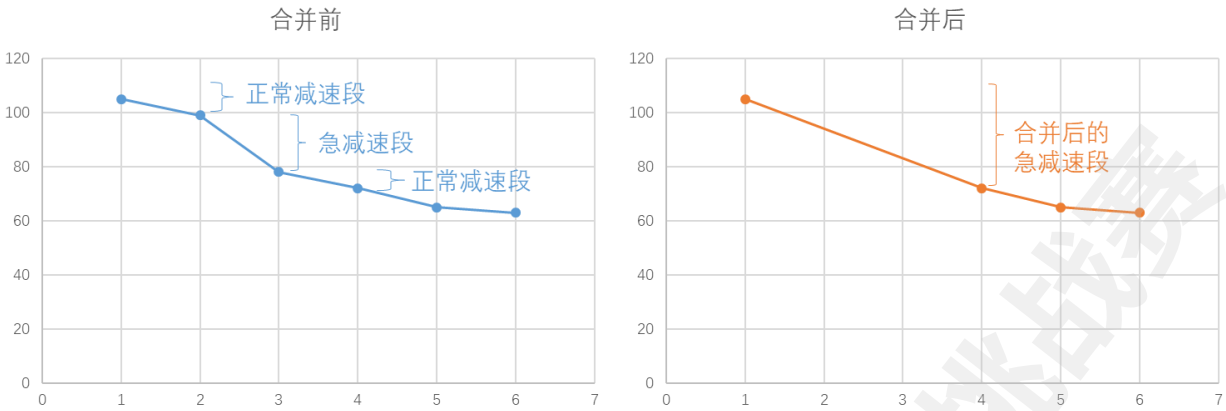

图 47　合并加速度异常段和正常段

在程序的具体实现过程中，我们发现，由于程序合并急加速/急减速段的方式为递归扩展，每个初始的急加速/急减速段最终都会扩展成一个急加速/急减速区间（时间可能不会延长）。

如果有临近但不直接相邻的急加速/急减速段，相互之间构成了合并条件，最终会合并成重复的急加速/急减速区间；有时合并后的区间可能不会重复，但会形成重叠，原理和程序模拟运行结果如图 48：

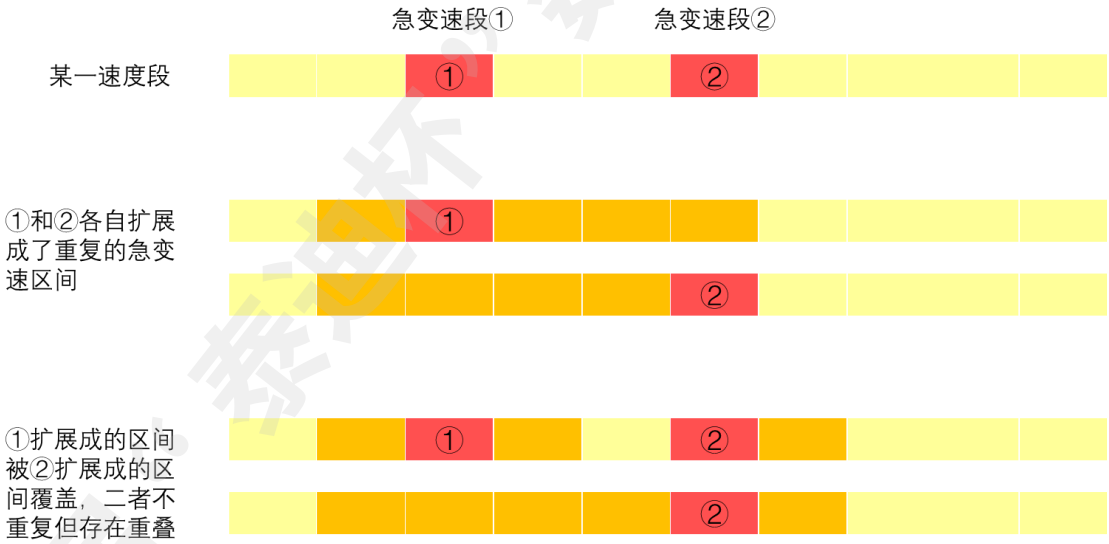

图 48　急加速急减速段的重叠扩展

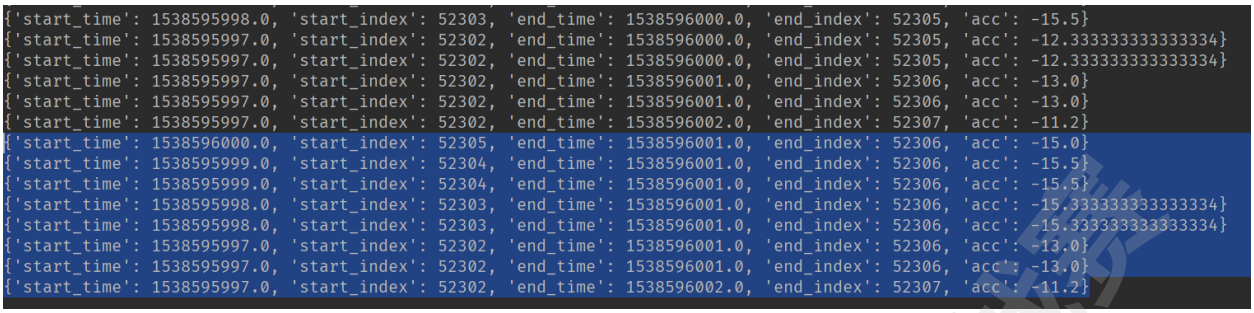

图 49　扩展程序运行结果

程序运行结果图图 49，选中部分（蓝色）和未选中部分的急变速区间分别从[52303, 52305]和[52305, 52306]各自扩展到了[52302, 52307]，扩展后的区间发生了完全重叠。

因此还需要对最终的合并结果判断是否出现重叠并去重。

3.3.3　相关变量的符号描述以及算法流程

本节给出急加速、急减速问题所涉及到的相关概念及形式化描述。文章所用到的符号在表 6 中给出。

表 6　判断急加速急减速所用到的符号

符号	意义
speed_list	记录车辆速度的列表
acc_list	记录急加速行为的列表
dec_list	记录急减速行为的列表
acc_record	一次连续急加速行为的记录
dec_record	一次连续急减速行为的记录
start_time	一个记录的起始时间
start_index	一个记录的起始索引
end_time	一个记录的终止时间
end_index	一个记录的终止索引
acc	急加速段的加速度
dec	急减速段的加速度

算法流程如图 50：

图 50　判断急加速急减速流程图

3.3.4　急加速急减速的时刻分布

在分析车辆急加速/急减速行为时，我们同时记录了车辆发生这类行为的时间（小时），并观察了发生急加速/急减速的时刻分布。图 51 是正式数据前三辆辆车发生急变速的时间统计：

```
file 1
[(11, 1), (7, 1), (22, 2), (6, 2), (8, 2), (5, 3), (4, 3), (1, 4), (10, 5), (9, 5), (23, 5), (3, 6), (12, 8)]
file 2
[(11, 3), (23, 11), (10, 18), (12, 25), (2, 30), (7, 41), (8, 41), (0, 44), (4, 45), (1, 55), (3, 65)]
file 3
[(22, 2), (10, 2), (4, 3), (1, 5), (11, 6), (23, 8), (2, 15), (0, 17)]
```

图 51　时刻分布统计

三辆车的急变速时刻分布如图 52：

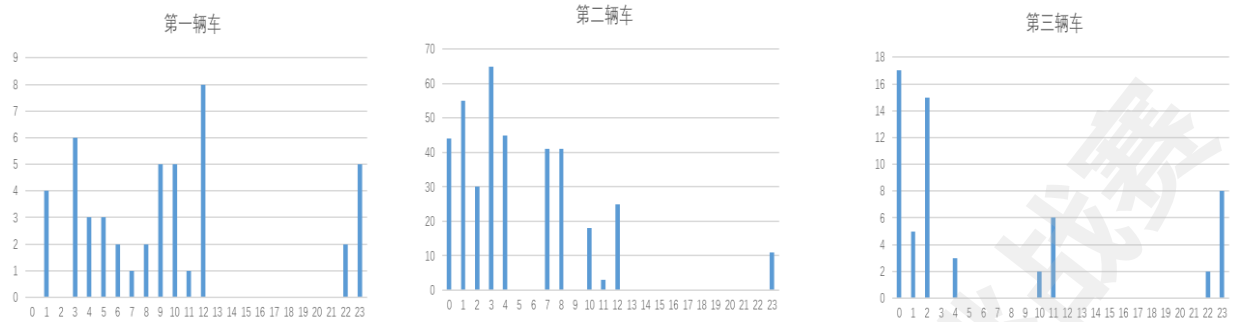

图 52　三辆车的急变速时间分布

从每一辆车发生急加速/急减速的时刻分布，可以更进一步分析车辆的行为模式，进而更好的帮助我们建立行车安全的评价模型。

3.4　熄火滑行

熄火滑行，是一类十分特殊的异常驾驶行为，根据字面含义，该行为是在引擎熄火、车辆关闭的情况下，车辆依然发生了一段滑行。对于大多数驾驶人员，等待车辆制动后再熄火才是正规操作，该异常行为的背后，蕴含着驾驶人员不正当操作带来的隐患，需要我们做出判断和记录。

3.4.1　相关定义

对于熄火滑行，我们定义 acc_state 为车辆引擎点火和熄火状态标志，acc_state = 0 表示车辆引擎处于熄火的状态，acc_state = 1 表示车辆引擎处于点火状态。而数据集中的属性 acc_state，恰好对应着这一标志。除此之外，我们定义 acc 为车辆当前加速度，speed 为当前速度。

我们定义，车辆发生了熄火滑行，当且仅当在 acc < 0 的过程中，acc_state = 0，且 speed ≠ 0。

3.4.2　问题分析

我们对没有进行零速度填充的 100 条原始示例数据集进行了判定，结果是，没有车辆发生过熄火滑行；接着，我们对填充过的示例数据集在此进行了判定，并提取出了 3 条发生熄火滑行的记录，均发生在车辆减速制动过程中。

我们对以上情况发生的原因进行了推理，分析结果如下：

对于原始数据，我们认为，采集车辆速度信息和点火信息的传感器工作状态，是和车辆本身状态有关的，对于一辆正常行驶的车辆，传感器工作状态如图 53：

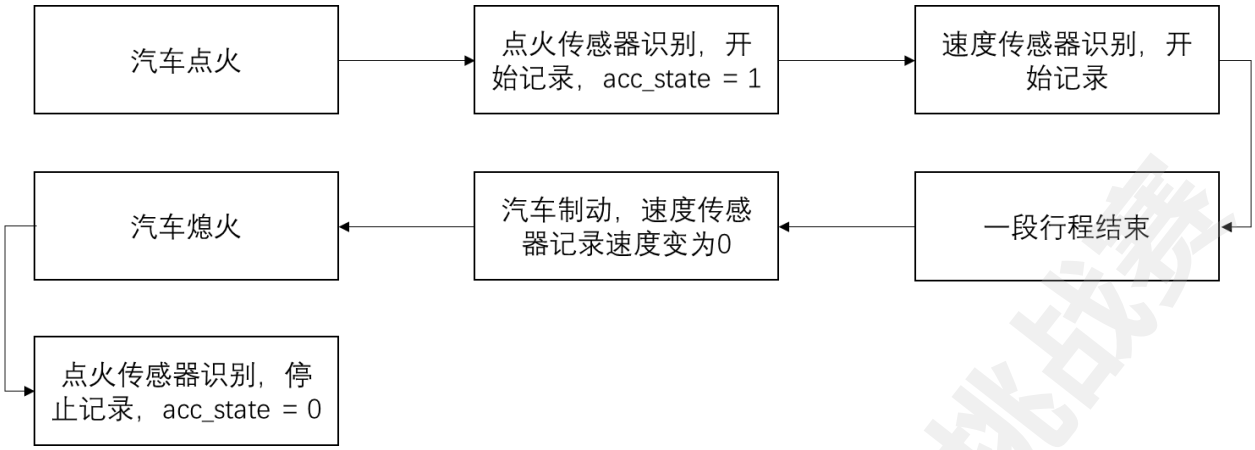

图 53　传感器正常工作状态

当车辆引擎处于关闭状态时，传感器不开启，当引擎启动，传感器才会开始采集车辆的速度信息（有可能会由于延迟而导致记录不全）。当车辆引擎关闭时，传感器也关闭（acc_state = 0），所以熄火前记录的最后一条数据速度一定为 0。

对于一辆发生熄火滑行的车辆，传感器工作状态如图 54：

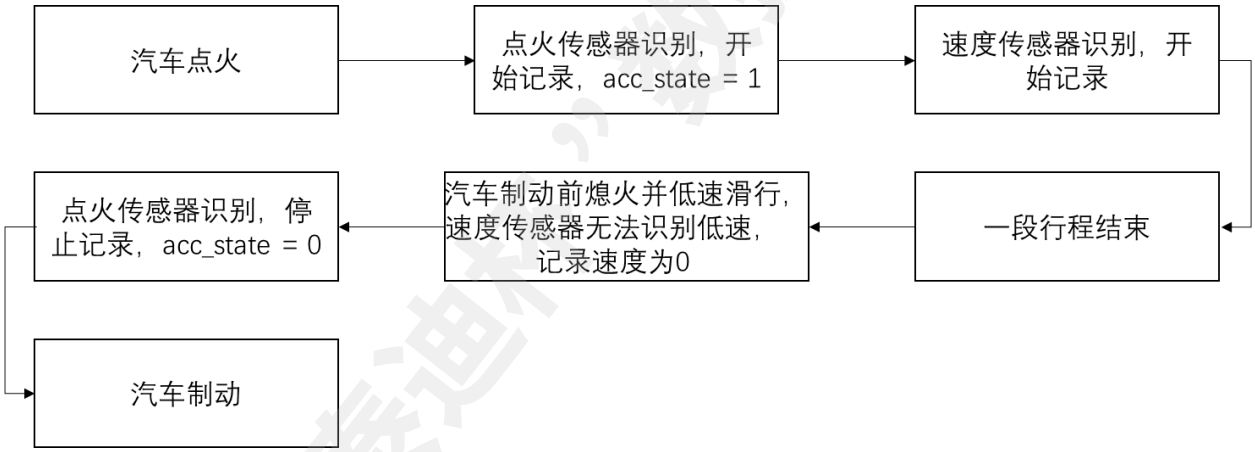

图 54　传感器熄火滑行工作状态

在车辆熄火滑行时，由于处于低速滑行状态，速度传感器无法识别低于 10km/h 的速度，记录仪记录值仍为 0。

因此，对于原始数据，不论车辆是否发生熄火滑行，速度传感器的记录值都在点火传感器变为前归零，acc_state = 0，而 gps_speed ≠ 0 的数据是不存在的。这就需要我们通过零速度填充挖掘出熄火滑行的异常。

图 55 的记录验证了我们推理的准确性（图中，第一列为 acc_state，第三列为速度）。

1	2018/10/8 22:56	13	21927	
1	2018/10/8 22:56	10	21927	
1	2018/10/8 22:56	10.5	21927	
1	2018/10/8 22:56	11	21927	
1	2018/10/8 22:56	7.333333	21927	
0	2018/10/8 22:56	3.666667	21927	
0	2018/10/8 22:56	0	21927	
0	2018/10/8 22:56	0	21927	
1	2018/10/9 1:36	0	21927	

图 55　熄火滑行结果展示

在对速度进行填充时，我们对于传感器无法识别的低速（速度小于 10km\h），模拟匀减速运动计算出速度，并进行填充。经过填充，减速阶段的速度被我们还原，这样，在 gps_speed $\neq$ 0 的时刻，出现了 acc_state = 0 的情况。熄火滑行的现象，就被成功地挖掘出来。

3.4.3　相关变量的符号描述以及算法流程

本节给出熄火滑行问题所涉及到的相关概念及形式化描述。文章所用到的符号在表 7 中给出。

表 7　判断熄火滑行所用到的符号

符号	意义
acc	车辆当前加速度
acc_state	车辆引擎点火和熄火的标志
gps_speed	车辆当前速度
speed_list	车辆的速度列表
state_list	车辆的启动关闭状态列表

算法流程如图 56：

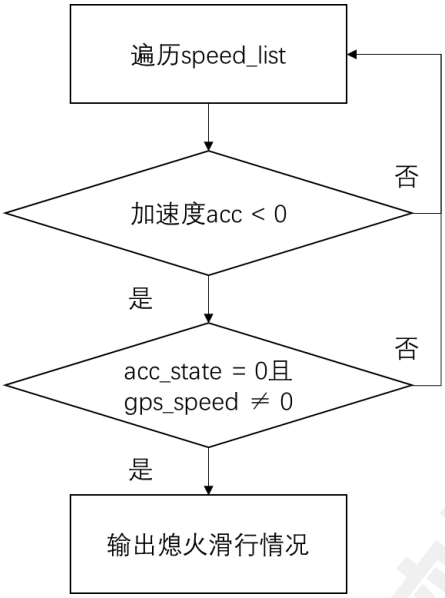

图 56　判断熄火滑行流程图

3.4.4　熄火滑行发生次数分布

我们对所有正式数据进行了分析，如图 57。统计频数后发现，对于大多数车辆，熄火滑行的行为并不会产生，或是发生次数很少，但是有三辆车的频数计数很高，分别是“AF00049”的 159 次，“AF00098”的 24 次，“AD00179”的 23 次。

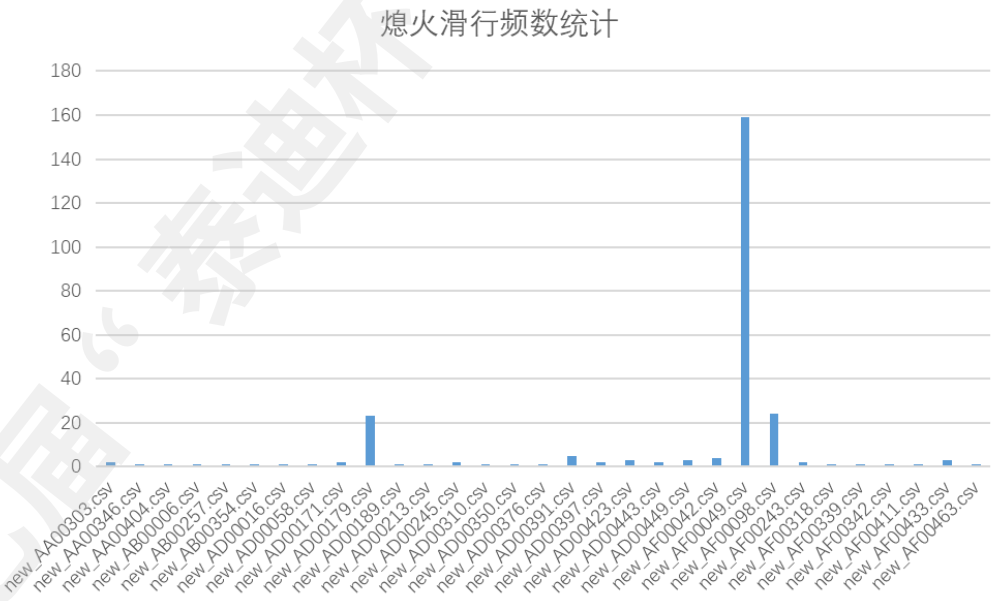

图 57　熄火滑行频数统计

由此可以看出，熄火滑行作为一种不良的驾驶习惯，拥有该习惯的驾驶员相比于没有此类习惯的驾驶员，驾驶车辆发生此类行为的概率要高很多。

3.5 急变道

急变道与急变速相似，属于行驶过程中的一种应激反应，往往发生于超车、避让、堵车时切换车道等情形，属于一种不安全的驾驶行为。

3.5.1 定义简述

急变道，是驾驶员在正常行驶过程中，由于某种原因，突然切换行车车道，变道后仍然沿原先道路方向行驶的一种行为。一次急变道行为如图 58：

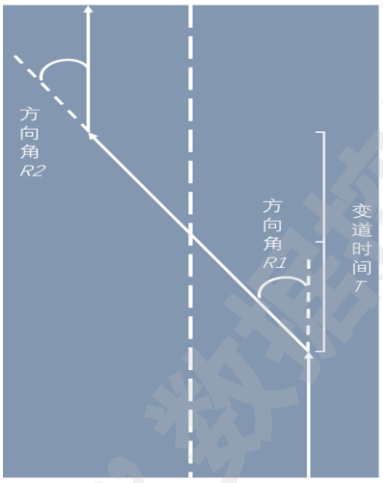

图 58　急变道行为示意图

根据这一定义，我们可以提取出急变道行为三个特征：

①行为发生时车辆速度 gps_speed 不等于 0；

②车头方向角 R 发生突变（变道角速度 angular_vlc 大于某一阈值 V ）；

③在一段时间内（汇入另一道路的时间 change_time 小于某一阈值 T ），车头方向角又恢复成原来的方向（变道后方向角与原方向角变化 delta_agl 小于某一阈值 D ）。

其中，

$$angular\_vlc = \frac{R}{t}$$

$$delta\_agl = \left|R_2 - R_1\right|$$

特征③是我们区别车辆急变道和急转弯的关键，在车辆转弯过程中，车头方向角也可能发生突变，导致角速度大于阈值但是不会在一段时间内再次突变回来。

3.5.2 问题分析

在判定急变道行为之前，我们对各阈值进行设定，规定：角速度阈值 $V = 20°/s$ ，时间阈值 $T = 10s$ ，方向角差异阈值 $D = 5°$ 。在设定好参数后，我们对数据集进行了分析，得到了车辆发生急变道的记录。

3.5.3 相关变量的符号描述以及算法流程

本节给出急变道问题所涉及到的相关概念及形式化描述。文章所用到的符号在表 8 中给出。

表 8 判断熄火滑行所用到的符号

符号	意义
direction_list	记录车辆方向角的列表
time_list	记录车辆行驶时刻的列表
speed_list	记录车辆速度的列表
angular_vlc	车头当前角速度
change_time	变道之后、方向角恢复之前的时间
delta_agl	一次变道后的角度变化
road_change	发生急变道的标志

算法流程如图 59：

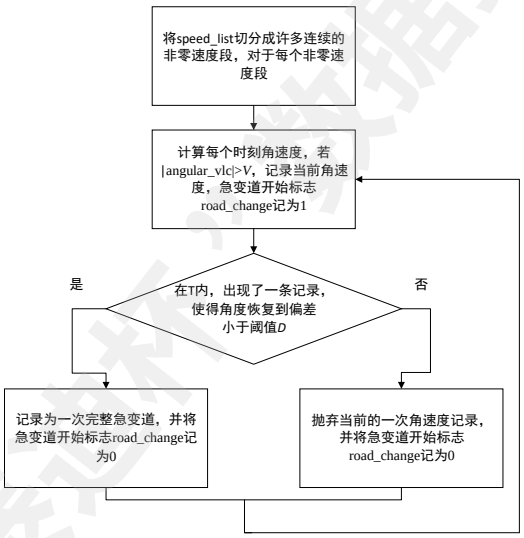

图 59 判断熄火滑行流程图

3.6 超速行驶

超速行驶是衡量车辆安全行驶的一项十分重要的指标，与其他许多指标不同的是，超速行驶是违反交通规则的异常驾驶行为，它直接威胁着驾驶人员和他人的生命财产安全，需要我们进行详细的分析和挖掘。

对于超速行驶，主要任务在于根据车辆当前经纬度坐标，判断出车辆当前行驶道路，并根据当前道路限速要求，与车辆当前速度结合判断其是否超速。

我们最初希望通过调用地图的 API 服务对车辆的超速行为进行判别，这样得到的当前限速阈值，是最符合当前路段限速要求的。但是，对于个人开发者，百度地图存在 API 调用次数的限制，其分析的效率远远达不到赛题数据规模的要求。因此，我们采取的评判策

略是，为超速行为定义速度和时间的阈值，从数据本身挖掘超速驾驶行为。

3.6.1 定义简述

在我国，对于不同道路，存在着不同的限速要求。

根据《城市道路工程设计规范》（CJJ37-2012），城市道路可以大致分为快速路、主干路、次干路和支路 4 种，并规定了不同的设计车速，大致范围为：快速路 60-100km/h，主干路 40-60km/h，次干路 30-50km/h，支路 20-40km/h；除了城市道路外，还有联通城市和城市的高速公路，车速范围在 60-120km/h[4]。

首先，我们定义车辆的 3 种超速模式，分别是城市道路超高速（Road_Over_Speed，简称 ROS），高速公路超低速（Highway_Low_Speed，简称 HLS），高速公路超高速（Highway_Over_Speed，简称 HOS）。

其次，我们定义 3 种道路等级，分别是城市低速路（Low_speed_road），城市快速路（Fast_speed_road）和高速公路（High_speed_road）。

最后，我们对车速定义 3 个阈值界限，分别是 60km/h，80km/h，120km/h，含义依次是：判别车辆低速行驶和高速行驶的界限，判断车辆在快速路和高速公路行驶的界限，高速公路超高速界限。

3.6.2 问题分析

我们遍历车辆的速度记录，得到车辆行驶过程中（gps_speed ≠ 0）的车速记录区间，每个非 0 速度区间基本可以视作车辆的一段行程。我们以 60km/h 为界限，将非 0 速度区间进行再划分，得到低速和高速行驶交替分布的区间，区间内部记录是连续的，如图 60。

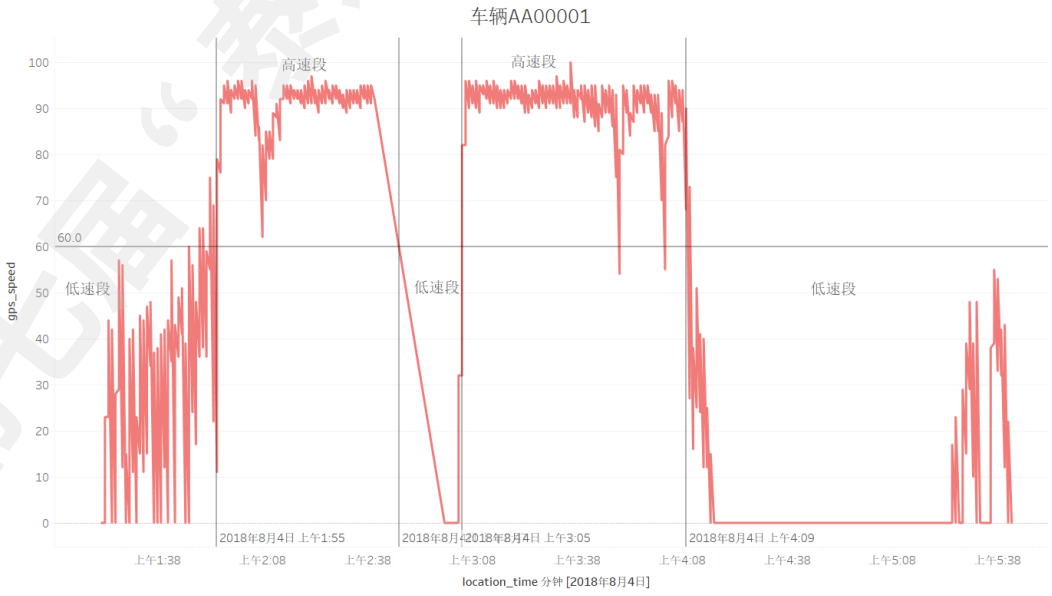

图 60 非 0 速度区间再划分

划分好区间后，进行如下步骤：

1.对于低速段：

如果区间对应行驶时间大于 60s，我们就认为该低速段发生在城市低速路上，属于正常行驶（因为在高速路段持续行驶超过 1 分钟是很危险的），否则转入步骤①；

①如果该低速段发生在一个非零速度区间的第一段，且时间小于 60s，则我们认为车辆初始在城市低速路上短暂行驶，属于正常行驶，否则转入步骤②；

②如果低速段发生在城市快速路段之后，且时间小于 60s，则我们认为车辆从城市快速路行驶到了低速路，属于正常行驶，否则转入步骤③；

③如果低速段发生在高速公路段之后，且时间小于 60s，则我们认为车辆在高速公路上发生了超低速行驶。

2.对于高速段：

如果区间对应行驶时间大于 600s，且最大速度不超过 120km/h，我们就认为该高速段发生在高速公路上，属于正常行驶（高速公路匝道收费站之间的距离一般是 15 公里，按照 90km/h 的平均速度，车辆行驶过一段高速公路的最短用时为 600s），否则转入步骤④；

④如果该高速段的最大速度不超过 80km/h（快速路的平均行驶速度应当比高速公路的平均行驶速度略小），且行驶时间小于 600s，我们认为车辆行驶在城市快速路上，属于正常行驶，否则转入步骤⑤；

⑤如果该高速段最大速度超过 80km/h，且行驶时间小于 600s，则我们认为车辆行驶在城市低速路上，属于超高速行驶，否则转入步骤⑥；

⑥如果该高速路段最大速度超过 120km/h，且行驶时间大于 600s，我们认为车辆行驶在高速公路上，且发生了超高速行驶。

3.6.3 相关变量的符号描述以及算法流程

本节给出超速驾驶问题所涉及到的相关概念及形式化描述。文章所用到的符号在表 9 中给出。

表 9 判断超速行驶所用到的符号

符号	意义
speed_list	记录车辆速度的列表
time_list	记录行车时间的列表
nzero_speed_list	记录非 0 速度区间的列表
low_speed_list	记录低速行驶的列表
high_speed_list	记录高速行驶的列表
nzero_interval	一段连续的非 0 速度区间
temp_record	记录当前行驶速度类型的起始时间和区间长度
start_index	记录的起始索引

time_continue	记录的区间长度

算法流程如图 61：

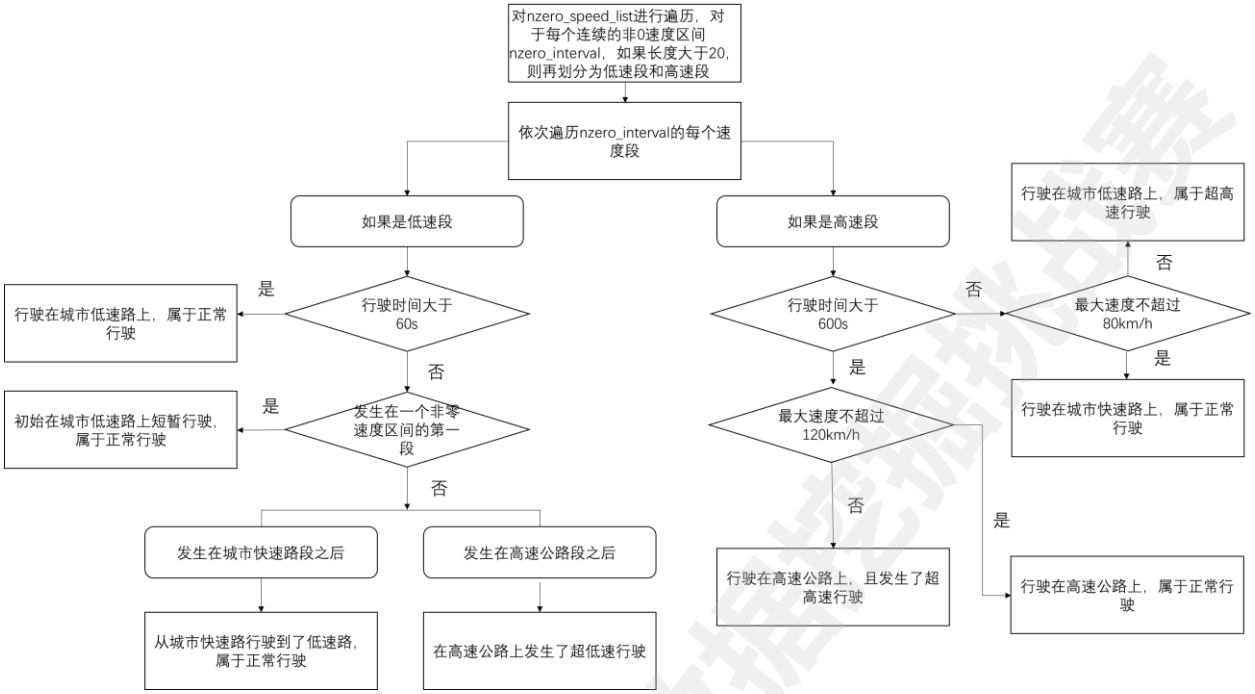

图 61　判断超速行驶流程图

3.6.4　超速驾驶判定结果验证

为了验证我们提取超速行驶特征方法的准确性，我们选择了一个数据量较小的车辆记录文件 AA00045.csv，调用百度地图 API，比较百度地图和以上方法的判定结果，进而验证我们模型判定的准确性。

我们编写的程序运行结果如图 62：

```
[[3823, 602, 'low'], [4439, 1, 'low'], [4702, 220, 'low'], [4937,
[[4425, 14, 'fast'], [4440, 26, 'fast'], [4922, 15, 'fast'], [494
['new_AA00045.csv', 'ROS', '2018-09-08 01:12:37', 392.0]
['new_AA00045.csv', 'ROS', '2018-09-08 09:31:01', 282.0]
['new_AA00045.csv', 'ROS', '2018-09-08 09:40:16', 204.0]
['new_AA00045.csv', 'HLS', '2018-09-08 01:39:11', 41.0]
['new_AA00045.csv', 'HLS', '2018-09-08 09:30:52', 9.0]
['new_AA00045.csv', 'HLS', '2018-09-08 09:35:43', 21.0]

Process finished with exit code 0
```

图 62　超速行驶判断程序运行结果

可以看出，我们的程序总共找出了 3 处城市道路超高速（ROS）、3 处高速公路超低速（HLS）行为。而百度地图总共找出了 15 处超速行为，其中，与程序运行结果基本符合的超速行为共有 3 处，依次是 9 月 8 日的 1:12:37、9:40:16 和 1:39:11 的数据，如图 63，图 64，图 65。

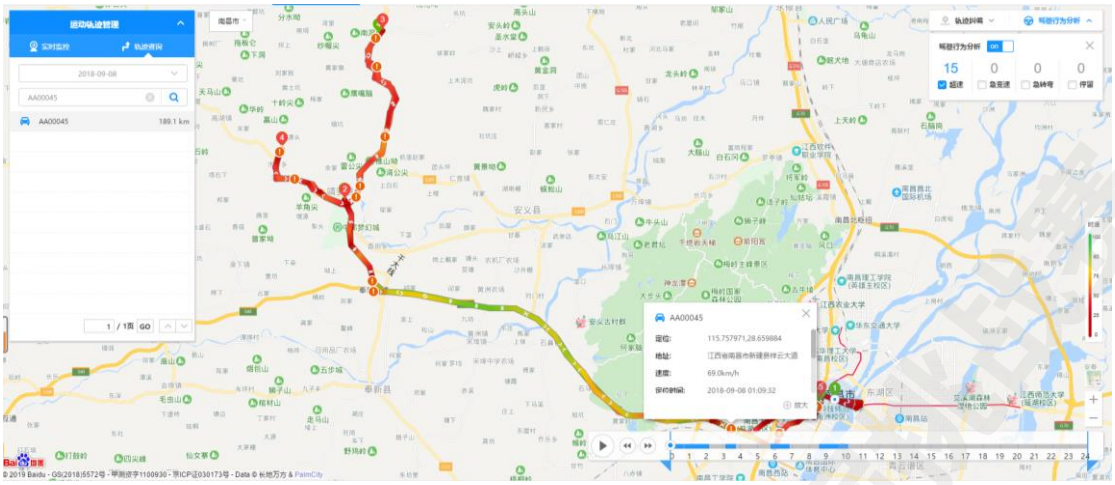

图 63　百度地图 API 判断超速结果一

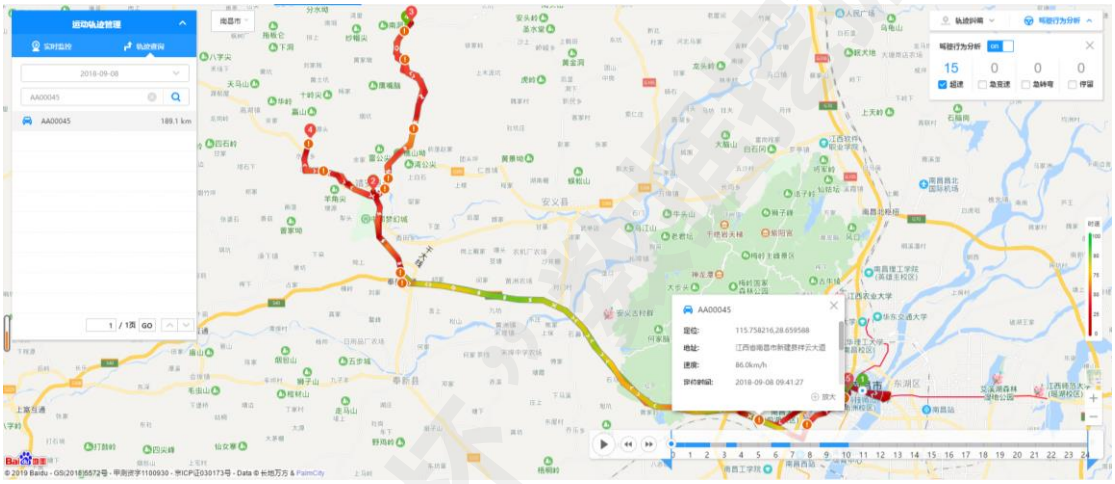

图 64　百度地图 API 判断超速结果二

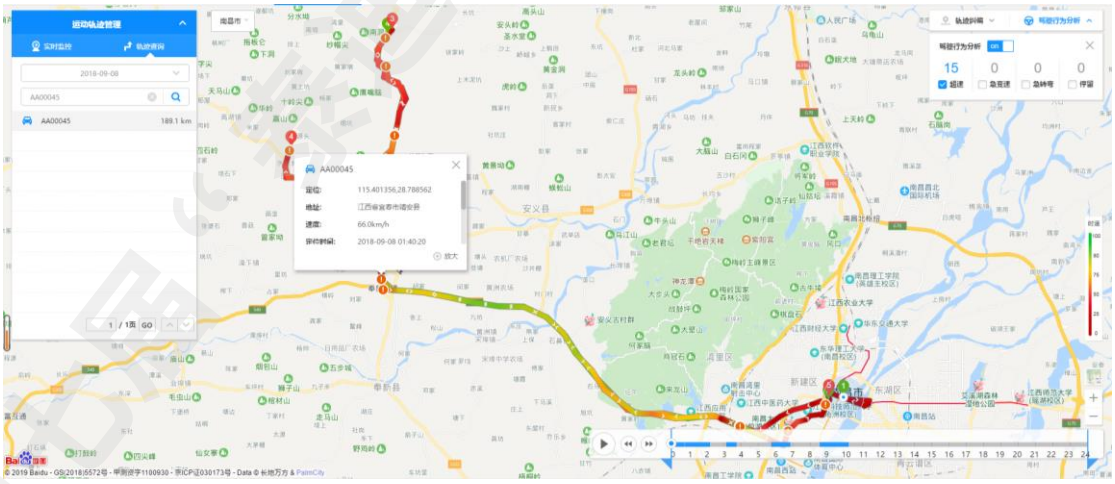

图 65　百度地图 API 判断超速结果三

对于文件 AA00045.csv 对应的车辆，相对于百度地图的分析结果，我们的超速特征提取正确率为 50%，考虑到判定条件的不同会导致结果不同，并不能认为程序中没有体现在百度地图上的超速信息就是错误的。而我们的超速特征提取覆盖率为 20%，主要原因是我

们的判别方法只适用于普遍情况，由于不能根据经纬度得到具体路段的限速信息，车辆在很多特殊路段的超速。

3.7 不良驾驶行为的特征提取结果展示

如图 66 所示，我们特征提取后的部分结果的展示。全部结果详见数据附件 2。

车辆编号	疲劳驾驶比例	不良怠速预热比例	超长怠速比例	急变道比例	急加急减速比例	熄火滑行比例	超速比例
AA00001	3	28	2	55	72	0	45
AA00002	27	107	0	80	5047	0	374
AA00004	0	51	0	787	972	0	503
AA00036	0	33	99	691	263	0	263
AA00045	0	93	47	90	721	0	270
AA00051	0	78	0	175	7358	0	205
AA00052	0	11	23	61	3364	0	476
AA00055	15	103	15	59	4846	0	103
AA00060	13	206	13	167	128	0	347
AA00061	0	84	42	320	306	0	125
AA00069	0	130	0	60	4309	0	405
AA00118	15	62	0	438	514	0	272
AA00125	0	285	0	300	9963	0	712
AA00128	8	45	8	60	2381	0	265
AA00143	11	45	11	221	376	0	66
AA00146	21	103	0	175	8828	0	339
AA00148	9	138	0	109	7664	0	264
AA00154	0	178	0	82	3492	0	238
AA00160	0	37	12	459	323	0	397
AA00164	10	142	0	814	382	0	291
AA00173	13	0	20	150	124	0	144
AA00188	32	0	0	142	379	0	379
AA00195	2	6	0	26	47	0	17
AA00202	0	14	4	11	21	0	0
AA00203	21	64	0	576	235	0	256
AA00211	0	15	4	118	251	0	49
AA00212	4	35	0	104	226	0	77
AA00219	8	17	0	151	205	0	415
AA00228	0	0	0	30	82	0	23
AA00232	26	26	26	607	660	0	264
AA00235	0	5	0	46	73	0	36
AA00238	3	0	3	6	56	0	34
AA00239	3	0	0	23	36	0	91

图 66 特征提取部分结果的展示

第四章　驾驶行为安全性多属性评价模型

在第三章，我们从车辆数据中分析了驾驶员的疲劳驾驶、急加速、急减速、怠速预热、超长怠速、熄火滑行、超速、急变道等不良驾驶行为，构建了 7 种属性的不良驾驶行为安全评价指标。本章节利用对前文所建立的不良驾驶行为表征指标对车辆的行车安全性进行分析：首先，分别采用主客观的方法为评价指标赋予权重，在此基础之上，通过最小二乘法对两种权重进行融合，形成兼顾专家意见与客观数据因素的评价指标权重；其次，根据理想逼近解法 TOPSIS 法对评价模型进行构建 ，并在真实数据上对研究对象进行安全等级排序；最后，我们通过 K-means 对数据进行粗略地二分类，将 TOPSIS 标记与粗略二分类的标记进行对比，选出标记相同的数据作为标注数据，训练二叉决策树评价模型，利用训练好的模型再对不确定数据进行评价，最终以投票的方式将不确定的数据进行安全性评价。具体流程如图 67：

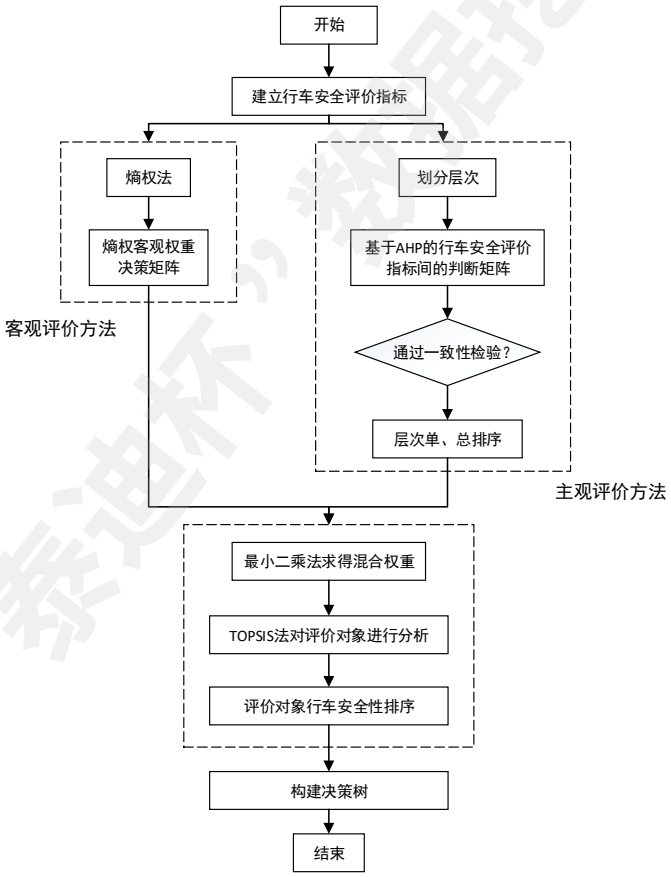

图 67　特征相关性分析结果

4.1　不良驾驶行为特征整合与分析

4.1.1　特征数据的整合

第三章不良驾驶行为特征提取时，主要得到的是每一辆车在数据集给出的整个行程中

发生不良驾驶行为的次数，但因为每一辆车的在整个行程中的总里程数也是不一样的，所以要以单位里程上发生不良驾驶行为的次数来标注每一个特征值。为此，引入变量：

count: 整个行程中发生不良驾驶行为的次数，

total_mileage: 整个行程中的总里程数，

count_permileage: 单位里程上发生不良驾驶行为的次数，则

$$count\_permileage = \frac{count}{total\_mileage}$$

以上述方式计算后发现有的为小数（比如 0.00002），有的为整数。需要进行规范化，使单位里程上的次数都为整数，于是在上述计算方式的基础上将结果扩大 10000 倍。即

$$count\_permileage = \frac{count}{total\_mileage} * 10000$$

数据整合后为 7 个特征值和 1 个目标值（待标记），为了叙述方便，给出对应的符号定义。本节所用到的符号在表 10 中给出。

表 10 特征提取所用到的符号

符号	意义
fatigue_drive	疲劳驾驶
rapid_speed_change	急变速
idling_preheat	不良怠速预热
idling_overtime	超长怠速
coast_off	熄火滑行
overspeed	超速
rapid_road_change	急变道
drive_safety	行车安全

以上述方式整合后的部分数据结果如图 68：

car_number	fatigue_drive	idling_preheat	idling_overtime	rapid_road_change	rapid_speed_change	coast_off	overspeed
AA00001	3	28	2	55	72	0	45
AA00002	27	107	0	80	5047	0	374
AA00004	0	51	0	787	972	0	503
AA00036	0	33	99	691	263	0	263
AA00045	0	93	47	90	721	0	270
AA00051	0	78	0	175	7358	0	205
AA00052	0	11	23	61	3364	0	476
AA00055	15	103	15	59	4846	0	103
AA00060	13	206	13	167	128	0	347
AA00061	0	84	42	320	306	0	125
AA00069	0	130	0	60	4309	0	405
AA00118	15	62	0	438	514	0	272
AA00125	0	285	0	300	9963	0	712
AA00128	8	45	8	60	2381	0	265

图 68 特征提取结果的部分展示

4.1.2 不同评价指标的相关性分析

在一个特征集合中，特征之间可能存在着某种联系，我们通过相关性分析，判断特征集中是否存在过强的关联性。如果指标之间有较强的关联性，则说明指标之间存在冗余。因此当建立一个特征矩阵时，检验其指标的相关性系数，保留最直接能体现评价结果的有效指标，减少工作量，利用 450 个待评价车辆的实验数据对不同表征的相关性进行分析，相关性的计算结果如表 11 所示：

表 11 特征相关性分析结果

序号	FD	IP	IO	RRC	RSC	CO	OS
FD	1						
IP	0.0535	1					
IO	-0.0035	0.0672	1				
RRC	0.054	0.1538	0.2169	1			
RSC	0.1306	0.1788	0.0031	0.2461	1		
CO	-0.0228	-0.0006	-0.0055	0.0296	0.0175	1	
OS	0.1526	0.3227	-0.0135	0.1572	0.2834	0.0435	1

表中缩写字母视图如下：

FD：疲劳驾驶　　IP：不良怠速
IO：超长怠速　　RRC：急变道
RSC：急加急减速　　CO：熄火滑行
OS：超速比例

由相关性分析结果显示个特征之间均不存在强相关性，相关性最大的超速和超长怠速两个特征之间的相关性也仅为 0.3。不存在冗余特征，所以应根据这所有的七个特征进行分类。

4.1.3 指标集合内在分布结构分析

我们假设对车辆的行车安全给出的评价结果是两类即安全和不安全，因此通过采用 k_means 聚类分析方法，令 k=2,对训练数据进行 2 分类，我们数据的内部分布情况，结果如图 69 所示：

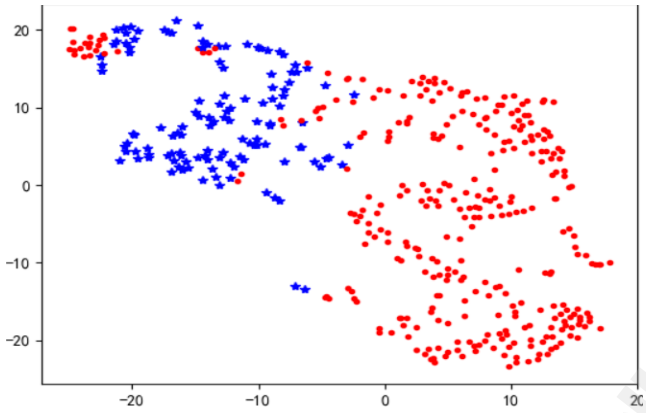

图 69　粗聚类结果展示

表 12 展示的是聚类的结果，两个簇的簇中心对应的七个特征的值，因为聚类前对数据进行了标准化，所以簇中心的特征值存在小于 0 的情况。从中明显的看出类别 0 的簇中心的特征值普遍比类别 1 的簇中心小，所以初步可以认为聚类的簇 0 为行车安全的类别，簇 1 为行车不安全的类别，这对我们后面的决策树分类也有一定参考价值。

表 12　聚类结果分析

	fatigue_drive	idling_preheat	idling_overtime	rapid_road_change	rapid_speed_chang	coast_off	overspeed	num
0	-0.2108	-0.2314	-0.0737	-0.2405	-0.3636	0.003	-0.4064	326
1	0.5541	0.6085	0.1938	0.6324	0.9558	-0.008	1.0683	124

根据上述图表的来看，训练数据的进行聚类行成 2 簇，两种类别比例大概为 1:2.5，数据的分布的相对偏差不大。

4.2　不良驾驶行为特征权重计算

车辆行为安全性的评价分析中，因多属性指标在评价中的贡献值不同，导致不同属性的评价指标对行车安全性的影响不同。因此，有必要对不良驾驶行为属性评价指标进行权重计算，建立最为合理化的评价方法。

熵信息法是确定指标权重的一种客观方法，此方法利用指标客观信息，规定同一指标间的波动越小，则该指标的权重越小，波动越大，则指标权重越大。但是这种方法容易受到数据的影响，数据值有变化（增加、删除或者改动）其相应会发生变化，层次分析法等主观方法对不良驾驶行为评价指标赋予权重，这类方法的优点是简单，所需定量数据信息较少，但是，定性成分多，不易令人信服，同时当指标过多时，数据统计量大，且权重难以确定。为了兼顾专家对指标的偏好和指标间的客观信息，采用层次分析法对不良驾驶行为指标进行主观赋值与评价，同时考虑各特征的熵信息赋予的权值，在此基础之上，通过最小二乘法确定最终各指标的权重[5]。

4.2.1　特征客观权重确定方法-熵值法

在信息论中，熵是系统无序程度的一种度量。熵越小，不确定性就越小；熵越大，不

确定性也越大。根据此性质，可以利用评价中各方案的固有信息，通过熵值法得到各个指标的熵，熵越小，其信息的效用值越大，指标的权重越大[6]。具体求解过程，如下所示：

1. 数据标准化

假设给定的数据矩阵包含 n 条数据，m 个属性，即其数据矩阵如公式 1 所示

$$X_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

采用最小-最大规范化规范数据，将各属性值映射到区间[0,1]上。假设规范化前 x_{ij} 表示第 i 条数据的第 j 个属性值，x_{ij} 表示其规范化之后的值。具体的规范公式如下：

$$x'_{ij}=\frac{x_{ij}-\min\{x_{1j},x_{2j},\cdots,x_{nj}\}}{\max\{x_{1j},x_{2j},\cdots,x_{nj}\}-\min\{x_{1j},x_{2j},\cdots,x_{nj}\}}$$

2. 求各属性的熵

对规范化之后的数据，计算各属性的信息熵。首先计算第 j 个属性下第 i 条数据所占该属性的比重，假设计算后的值为 p_{ij}。其计算公式为：

$$p_{ij}=\frac{x'_{ij}}{\sum_{i=1}^{n}x'_{ij}}$$

假设第 j 个属性的信息熵值用 e_j 表示，则其计算公式如下：

$$e_j=-k\sum_{i=1}^{n}p_{ij}\ln(p_{ij})$$

其中 $k=1/\ln(n)>0$，如果 $p_{ij}=0$，则定义 $\lim_{p_{ij}\to 0}p_{ij}\ln(p_{ij})=0$

3. 计算各指标权重

求出 m 个属性的信息熵依次为 $e_1,e_2,e_3\cdots e_m$，假设第 j 个属性计算出的权重为 w_j，其计算公式为：

$$w_j=\frac{1-e_j}{m-\sum_{j=1}^{m}e_j}$$

通过应用熵权法对 450 条数据，7 个属性进行处理与计算，计算出 7 个属性的权重如表 13 所示：

表 13　基于熵权值的指标权重

疲劳驾驶	不良怠速预热	超长怠速	急变道	急加急减速	熄火滑行	超速
0.073	0.079	0.258	0.084	0.086	0.352	0.068

4.2.2 特征主观权重确定方法-AHP

首先我们确定目标层为“行车不安全”，将“行车不安全”划分为“乘车辆危害”、“违规行为”、“驾驶员不良驾驶习惯”三个准则。根据准则代表的实际物理意义以及其可能包含的不良驾驶行为，构建出层次分析结构[7]，如图 70 所示。

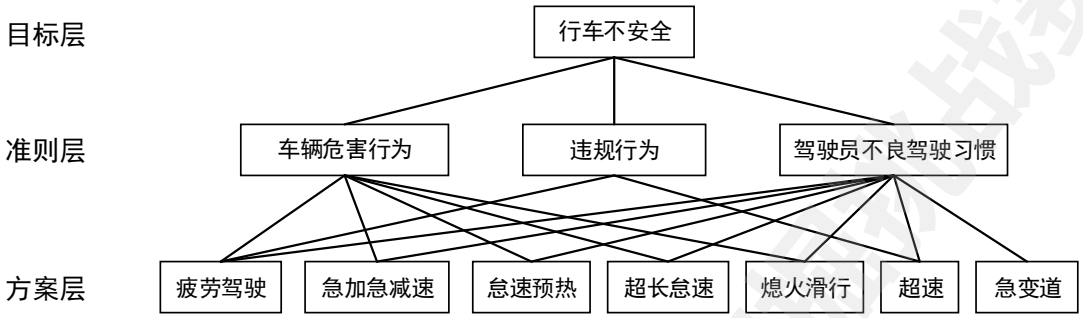

图 70 聚类结果分析

对于方案层各元素，给出其相对重要性，构建判断矩阵。

准则层元素对目标层的相对重要性判断矩阵：

$$A=\begin{bmatrix} 1 & 1/5 & 1/3 \\ 5 & 1 & 5/3 \\ 3 & 3/5 & 1 \end{bmatrix}$$

方案层相关元素对行车危害行为的相对重要性判断矩阵：

$$B=\begin{bmatrix} 1 & 7/3 & 7 & 7 & 7 \\ 3/7 & 1 & 3 & 3 & 3 \\ 1/7 & 1/3 & 1 & 1 & 1 \\ 1/7 & 1/3 & 1 & 1 & 1 \\ 1/7 & 1/3 & 1 & 1 & 1 \end{bmatrix}$$

方案层相关元素对“违规行为”的相对重要性判断矩阵：

$$B_2=\begin{bmatrix} 1 & 7/5 \\ 5/7 & 1 \end{bmatrix}$$

方案层相关元素对结束阶段不安全的相对重要性判断矩阵：

$$B_3=\begin{bmatrix} 1 & 7/3 & 7 & 7 & 7 & 7/5 & 7/5 \\ 3/7 & 1 & 3 & 3 & 3 & 3/5 & 3/5 \\ 1/7 & 1/3 & 1 & 1 & 1 & 1/5 & 1/5 \\ 1/7 & 1/3 & 1 & 1 & 1 & 1/5 & 1/5 \\ 1/7 & 1/3 & 1 & 1 & 1 & 1/5 & 1/5 \\ 5/7 & 5/3 & 5 & 5 & 5 & 1 & 1 \\ 5/7 & 5/3 & 5 & 5 & 5 & 1 & 1 \end{bmatrix}$$

由于 B_2 阶数为 2，不需要进行一致性检验。对于剩下的三个矩阵，进行单排序以及一

致性检验。通过检验，发现三个矩阵均具有较好的一致性，可进行后续计算。

将计算出的方案层所有元素对准则层各元素权重组成矩阵，进行层次总排序，如表 14。

表 14　基于层次分析的指标权重

	车辆危害行为	违规行为	驾驶员不良驾驶习惯	总排序
	0.111	0.556	0.333	
疲劳驾驶	0.538	0.583	0.304	0.458
急加急减速	0.231	0	0.130	0.069
怠速预热	0.077	0	0.043	0.023
超长怠速	0.077	0	0.043	0.023
熄火滑行	0.077	0	0.043	0.023
超速	0	0.417	0.217	0.304
急变道	0	0	0.217	0.072

通过总排序可以计算出方案层各元素对目标层影响程度的权值，各属性影响权值如表 15：

表 15　基于层次分析的指标权重总排序

疲劳驾驶	急加急减速	怠速预热	超长怠速	熄火滑行	超速	急变道
0.457	0.196	0.030	0.035	0.017	0.132	0.132

4.2.3　最小二乘法融合权重

假设各不良驾驶行为指标综合权重为$\eta=(w_1,w_2,\cdots w_n)^T$,为了达到主观和客观的统一，综合后的权重指标与主客观权重决策结果偏差越小越好。为此，建立如下最小二乘法决策模型：

$$Q=\sum_{i=1}^{n}[(y_{i1}-\hat{y}_i)^2+(y_{i2}-\hat{y}_i)^2]$$

$$=(y_1-X\cdot\eta)'(y_1-X\cdot\eta)+(y_2-X\cdot\eta)'(y_2-X\cdot\eta)$$

令$\frac{\partial Q}{\partial\eta}=0$，解得：

$$\eta=(\omega+\mu)/2$$

ω和μ分别为两种方法求出的权重列向量，η为混合权重向量，X是标准化之后的数据矩阵。

4.2.4　结果展示

利用梯度下降法计算混合权重的结果如表 16：

表 16　基于层次分析的指标权重总排序

	疲劳驾驶	不良怠速预热	超长怠速	急变道	急加急减速	熄火滑行	超速
主观权重	0.458	0.069	0.023	0.023	0.023	0.304	0.072
客观权重	0.073	0.079	0.258	0.084	0.086	0.352	0.068
混合权重	0.266	0.074	0.141	0.054	0.155	0.128	0.170

4.3　基于理想点逼近法（TOPSIS）构建评价模型

TOPSIS 是一种用于多目标决策的方法，通过检测评价对象与最优解、最差解的距离来排序。若评价对象最靠近最优解同时又最远离最差解为最好；否则为最差。考虑 TOPSIS 法具有原理简单并对实验样本要求不高，并且可生成一个明确的安全等级，因此选取 TOPSIS 来做为评价车辆行为安全性的方法[8]。

4.3.1　理想点逼近法（TOPSIS）

1. 构建初始评价指标矩阵

在对 450 辆车的 7 个不良驾驶行为进行统计分析后，我们给出了车辆不安全指数的评价矩阵 Z：

$$Z=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{17} \\ z_{21} & z_{22} & \cdots & z_{23} \\ \vdots & \vdots & \ddots & \vdots \\ z_{450,1} & z_{450,2} & \cdots & z_{450,7} \end{pmatrix}_{450\times 7}$$

由于各属性量纲相同，均为“次/千万米”，因此不需要对评价指标矩阵进行属性值的规范化。

2. 构建加权评价指标矩阵

在上一节中，我们使用了层次分析法、熵权法两种手段对各个异常行为特征的权重进行了计算，并使用最小二乘法将两种权重进行了融合，最终得到了综合权重向量 $w=[w_1,w_2,\cdots w_7]$，并根据该向量得到了综合权重矩阵 W：

$$W=\begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_7 \end{pmatrix}_{7\times 7}$$

在得到异常行为特征权重后，即可计算得到加权评价指标矩阵 F：

$$F = Z \cdot W$$
$$= \begin{pmatrix} w_1 z_{1,1} & w_2 z_{1,2} & \cdots & w_7 z_{1,7} \\ w_1 z_{2,1} & w_2 z_{2,2} & \cdots & w_7 z_{2,7} \\ \vdots & \vdots & \ddots & \vdots \\ w_1 z_{450,1} & w_2 z_{450,2} & \cdots & w_7 z_{450,7} \end{pmatrix}_{450\times 7}$$

其中，$f_{i,j} = w_j \cdot z_{i,j}, i = 1,2,\cdots,450; j = 1,2,\cdots 7$

3. 确定正负理想解

对于加权评价指标矩阵 F，找到第 j 列的最优记为 f_j^*，最劣为 $f_j^\wedge$，最终得到正理想解 $F^* = [f_1^*, f_2^*, \cdots f_7^*]$ 和负理想解 $F^\wedge = [f_1^\wedge, f_2^\wedge, \cdots f_7^\wedge]$。需要注意的是，对于加权评价指标矩阵 F，元素的值越大，说明该车辆样本在该异常特征下的表现越不安全。因此，每一列的最小值为最优，最大值为最劣。

4. 计算各样本与正负理想解的距离

对于每个车辆样本，依次计算它们与正负理想解的欧氏距离，如下：

$$S_i^* = \sqrt{\sum_{j=1}^{7} (f_{ij} - f_j^*)^2}, j = 1,2,\cdots 7$$

$$S_i^\wedge = \sqrt{\sum_{j=1}^{7} (f_{ij} - f_j^\wedge)^2}, j = 1,2,\cdots 7$$

5. 计算相对接近度并作出判断

对于每个样本，计算相对接近度，如下：

$$C_i = S_i^\wedge / (S_i^\wedge + S_i^*), i = 1,2,\cdots 450$$

我们需要明确，C_i 的取值范围为[0, 1]，取端点的情况为：

$$C_i = \begin{cases} 1, & S_i^* = 0 (\text{该车辆样本得分为最优}) \\ 0, & S_i^\wedge = 0 (\text{该车辆样本得分为最劣}) \end{cases}$$

相对接近度越大，说明车辆样本与正理想解距离约小，与负理想解距离越大，驾驶行为越安全。如果将样本按照相对接近度由大到小排序，那么设定阈值后取末尾若干位，得到的即是不安全行驶的车辆。

4.3.2 结果的展示与分析

我们通过 TOPSIS 算法，得到车辆安全性部分结果（全部结果见我们的数据附件 3）展示如图 71，图 72：

AD00083	8
AA00002	94
AD00053	125
AF00098	127
AD00013	138
AF00131	152
AD00003	242
AF00373	274
AB00006	275
AD00419	383

图 71　10 辆车的安全性排名

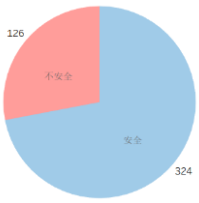

图 72　聚类结果展示

共有 126 辆车被标记为不安全，其余 324 辆车被标记为安全。对于评估中不安全车辆所占比例的合理性，解释如下：

在 TOPSIS 算法中，最后一步需要确定相对接近度阈值，这样才能对样本进行二分类，确定车辆的驾驶行为是否安全。但是，算法本身并没有给出确定阈值的具体方法，单凭主观判断给定阈值，无法判断划分得是否准确，这样就失去了算法根据 7 个异常行为特征的综合权重，对车辆进行分类的意义。

对此，我们希望寻找一种客观的方法，通过客观地分析来得到相对接近度下限 C_s。

在章节 4.1 中，我们使用聚类分析了样本二分类的均衡性，而聚类分析正是一种无监督学习，能够根据数据集中样本本身的相似度特性，分出不同的簇，且不需要人为干预。我们采取基于划分的 K-means 算法，得到了二簇聚类的结果：两个簇的样本数目分别为 124 和 326。

基于聚类分析，我们认为，如果分类的结果比例接近于聚类的结果，那么该分类是比较符合客观情况的，对此，我们不断调整相对接近度下限 C_s，最终发现，当 $C_s=0.96$ 时，分类的结果的比例最接近聚类结果，如表 17。

表 17　Cs 阈值选取

相对接近度阈值	0.94	0.95	0.96	0.97	0.98
分类比例	116:334	120:330	126:324	139:311	160:290

最终，我们确定了相对接近度阈值，得到了分类结果。图 73 是 450 辆车的相对接近系数分布及车辆的安全情况：

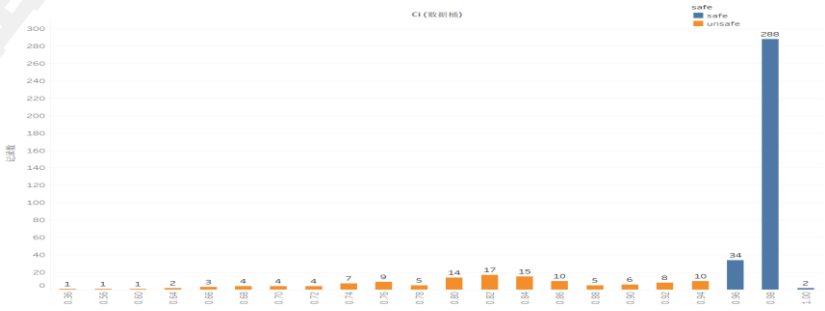

图 73　行车安全等级结果分析

表 18 为数据的部分结果展示：

表 18　TOPSIS 部分结果以及聚类结果的对比

车辆名称	Ci	聚类结果
AA00001	0.997125	0
AA00002	0.874787	1
AA00004	0.965082	1
AA00036	0.979013	1
AA00045	0.979362	0
AA00051	0.820137	1
AA00052	0.915047	1
AA00055	0.880128	1
AA00060	0.985877	1
AA00061	0.987642	0

通过该表格可知，我们用 k_means 聚类和理想点逼近法（Topsis）两种方法，对原数据进行标记，存在部分奇异。对于每一辆车的行车安全（drive_safety）得到两种标记。两种不同标记的比较如表 19 所示：

表 19　两种标记方式结果比较

Topsis \ k_means	0	1
0	295	29
1	48	78

其中，由上表可知，两种标记相同的样本数为 373 个，标记不同的样本数为 77 个。

4.4　评价决策树的构建

4.4.1　决策树评价模型

为了进一步确定 77 个不一致数据的行车安全性，我们的基本思想是在标记相同的数据集上即 373 个训练一个分类器，由于决策树分类器的可解释性以及经调参后的准确率优于其他分类器，通过下图通过调节 max_depth、min_impurity_split、max_leaf_nodes 以及 min_samples_leaf 如图 74 所示

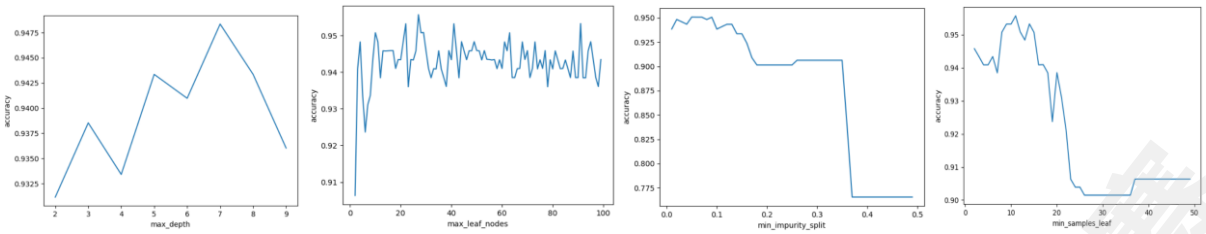

图 74　调参过程

综上，可知，使得模型最优的参数如表 20 所示：

表 20　参数的解释以及最优取值

参数	意义	取值
max_depth	树的最大高度	6
min_impurity_split	最小划分不纯度	0.1
max_leaf_nodes	最大叶子节点数量	28
min_samples_leaf	叶子节点最小样本数	11

4.4.2　决策树方法与结果分析

基于训练好的决策树，我们已标注的测试集进行预测，预测结果通过混淆矩阵的形式，如表 21 所示：

表 21　测试结果的混淆矩阵

	Predict_safe	Predict_not_safe
Actual_safe	62	7
Actual_not_safe	2	19

从上表可以计算出，模型的准确率为 0.969，召回率为 0.904，f 值为 0.809，因此该模型表现出了较好的评价能力。同时决策树分类器的另一优点在于可解释性比较强，同时评价规则也比较直观，如图 75 图 76，形象地展示出了我们的评价规则：

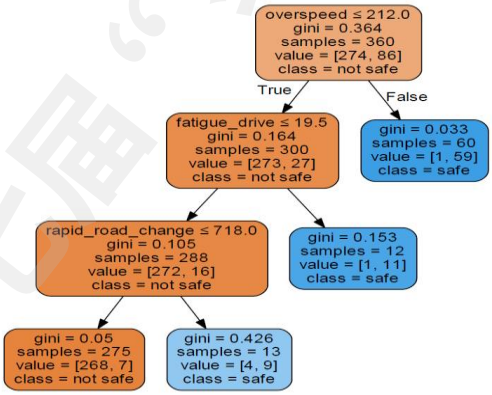

图 76　决策树评价模型

k_means	topsis	predict
0	1	1
0	1	1
0	1	1
0	1	1
0	1	1
0	1	1
0	1	1
0	1	1
1	0	1
0	1	0
0	1	1
0	1	1
0	1	1
1	0	0
1	0	0

图 75　部分结果展示

根据决策树输出的规则，可以看出以超速特征为属性进行划分的效果最好，即分类的基尼系数（Gini）最小，分类的不纯度最低，其次是疲劳驾驶特征，急加急减特征。这也

与我们的主观想法类似，超速行为对行车安全影响最大，然后是疲劳驾驶和急加急减速。而其他一些特征，怠速、熄火滑行等可能是对车辆本身的伤害比较大，对行车过程的安全的影响相比于上面几个特征要小得多。

基于 4.3 的 TOPSIS 评价模型和聚类结果分析，最终发现有 77 个车辆行车安全的评价是不一致的，因此，可以通过 4.4 训练的决策树评价模型进一步修正，最后得到的部分结果如图所示：

基于投票的方式，将之前不一致的评价结果进行修正。由最终预测结果可知，对于两种方式标记结果不同的样本，最终预测结果大部分与 Topsis 方法标记的结果相同。也可以认为最初的两种标记方法，Topsis 的效果比较好，同时 Topsis 能够给出相应的行车安全性等级。

4.5 本章小结

本章节首先分别采用主客观想结合的方法对权重进行融合，形成兼顾专家意见与客观数据因素的评价指标权重；其次，根据理想逼近解法 TOPSIS 法对评价模型进行构建 ，并在真实数据上对研究对象进行安全等级排序；最后，我们通过 K-means 对数据进行粗略地二分类，将 TOPSIS 标记与粗略二分类的标记进行对比，选出标记相同的数据作为标注数据，训练二叉决策树评价模型，利用训练好的模型再对不确定数据进行评价，最终以投票的方式将不确定的数据进行安全性评价。具体全部评价结果请参看数据附件 3。

第五章　综合评价指标体系与综合评价模型

在第三章和第四章的基础上，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，首先，为运输车辆管理部门建立行车安全的综合评价指标体系；其次，结合第四章的评价模型，构建综合评价模型。

5.1　综合评价指标体系的构建

在第三章，我们已经构建好了描述车辆不良驾驶的六种行为特征。在此基础之上，我们综合分析附件 2 给出的自然气象数据以及自己额外获取的关于车辆路线的道路状况数据，同时构建了行车天气环境指标以及道路状况指标。

5.1.1 天气环境指标

（1）附件 1 的天气环境数据分析

分析给出的气象数据，我们发现“relative_humidity”和“precipitation”这两个属性存在很多缺失值，经过统计发现其缺失率分别为 89.8%和 90.1%。由于这两个属性缺失率高于 50%，所以我们抛弃掉这两个属性。：

通过分析，我们发现“wind_direction”和“wind_power”这两个属性存在异常或无实际意义的值。在整个数据集中，存在部分数据项的“wind_direction”和“wind_power”值为“nan 转 nan”，而这些值是没有意义的错误值，如图 77。对于这部分数据项，由于数量不多，小于总数据量的 1%，我们选择删除掉这部分数据项。

1	province	prefecture_city	county	wind_direction	wind_power	high_temp	low_temp	conditions	record_date
15384	浙江	湖州	湖州	nan转nan	nan转nan	34℃	26℃	多云	18/8/2018
15385	浙江	湖州	长兴	nan转nan	nan转nan	34℃	25℃	多云	18/8/2018
15386	浙江	湖州	安吉	nan转nan	nan转nan	35℃	24℃	多云	18/8/2018
15387	浙江	湖州	德清	nan转nan	nan转nan	34℃	26℃	多云	18/8/2018
15403	浙江	绍兴	绍兴	nan转nan	nan转nan	34℃	26℃	多云	18/8/2018
15404	浙江	绍兴	诸暨	nan转nan	nan转nan	34℃	26℃	多云	18/8/2018
15405	浙江	绍兴	上虞	nan转nan	nan转nan	34℃	26℃	多云	18/8/2018
15461	浙江	绍兴	新昌	nan转nan	nan转nan	34℃	25℃	多云	18/8/2018
15462	浙江	绍兴	嵊州	nan转nan	nan转nan	34℃	25℃	多云	18/8/2018
15478	浙江	丽水	丽水	nan转nan	nan转nan	36℃	25℃	雷阵雨转阴	18/8/2018
15479	浙江	丽水	遂昌	nan转nan	nan转nan	34℃	25℃	雷阵雨转阴	18/8/2018
15480	浙江	丽水	龙泉	nan转nan	nan转nan	35℃	25℃	雷阵雨转阴	18/8/2018
15481	浙江	丽水	缙云	nan转nan	nan转nan	35℃	24℃	雷阵雨转阴	18/8/2018
15482	浙江	丽水	青田	nan转nan	nan转nan	34℃	25℃	雷阵雨转阴	18/8/2018
15483	浙江	丽水	云和	nan转nan	nan转nan	35℃	25℃	雷阵雨转阴	18/8/2018
15484	浙江	丽水	庆元	nan转nan	nan转nan	32℃	23℃	雷阵雨转阴	18/8/2018
15757	江苏	镇江	镇江	nan转nan	nan转nan	33℃	27℃	小雨转多云	18/8/2018
15758	江苏	镇江	丹阳	nan转nan	nan转nan	33℃	27℃	小雨转多云	18/8/2018

图 77　天气数据的缺失值

（2）天气特征的提取

由于行车过程中车辆的方向角一直在变化，且气象数据中的“wind_direction”属性值情况太多，不好给出其优劣等级，所以我们做天气特征提取时选择抛弃掉这一属性，只提取“wind_power”、“conditions”、“high_temp”和“low_temp”这四个属性。

根据每辆车记录的经纬度数据进行逆地址解析协议，查找出其对应的省、市、县（区），再加上对应的时间信息（只需要年、月、日部分信息），在给出的气象数据中进行匹配。将每辆车匹配到的所有数据分别保存到不同的表格中。由于一辆车匹配到的气象数据存在大量重复值，所以再对匹配出的数据进行去重。最终得到每辆车不同天对应的气象数据。

将每辆车匹配出的气象数据连接为一个表，分析该表可以发现其对应的“wind_power”以及“conditions”属性所包含的种类比初始的天气数据表中少了许多，这也减少了我们后续的工作。

（3）特征的分析

1.“wind_power”属性分析

通过分析，我们发现”wind_power”属性只包含三类值，对此我们根据其代表的风级强度给出其对应的数值，方便后续数据的处理。其对应关系如表 22 所示：

表 22　风级属性对应关系

风级属性值	对应数值
<3 级	1
1-2 级	1.5
3-4 级	3.5
4-5 级	4.5

2.“conditions”属性分析

由于每辆车对应的“conditions”属性包含值的类型较多，依次给出每类值较为复杂，所以我们采用合并部分属性值得方法来简化“conditions”属性与数值的对应关系。属性值对应的数值也考虑到其气候的恶劣情况给出的不同的值，数值越大，表面气候越恶劣。具体的合并与对应关系见表 23：

表 23　天气状况属性对应关系

合并后类	晴	多云	阴	小雨
具体包含的属性值	晴	多云 多云转晴 晴转多云	阴 阴转多云 多云转阴 晴转阴 阴转晴	小雨 小雨转阴 小雨转晴 小雨转多云 多云转小雨 晴转小雨 阴转小雨
对应数值	1	2	3	4
合并后类	中雨	阵雨	雷阵雨	暴雨
具体包含属性值	中雨 中雨转小雨 中雨转阴	阵雨 中雨转阵雨 阵雨转中雨	雷阵雨 雷阵雨转多云 雷阵雨转小雨	暴雨转小雨

	中雨转多云 小雨-中雨转多云 小雨转中雨	阵雨转阴 多云转阵雨 阵雨转多云 阵雨转晴	雷阵雨转阴 雷阵雨转阵雨 雷阵雨转中雨 晴转雷阵雨 多云转雷阵雨 中雨转雷阵雨	
对应数值	5	6	7	8

3.“high_temp”和“low_temp”属性分析

由于这两个属性属性值格为数值加单位，所以直接去掉单位即可得到数值部分值，不需要做其他处理。

4.获得单个属性综合值

现在我们已经将每辆车匹配到的气象数据全部转化为对应数值表示且均没有被删除。为了方便后期构建属性矩阵，我们将每辆车的 4 个气象属性统一用其平均值代替。

5.1.2 道路状况数据分析

在分析车辆行驶过程中的道路状况时，我们引进特征“道路曲折程度”，用来刻画道路包含转弯个数的多少，进而表现道路状况对行驶安全的影响。

1.车辆转弯行为的定义

分析道路转弯的个数实际是分析车辆发生转弯的次数。我们在分析车辆异常驾驶行为一部分定义了“急变道”的行为，而“转弯”实际上是“急变道”的过程的一半，它的行为定义如下：

当车辆处于行驶状态时(gps_speed $\neq$ 0)，如果方向角发生改变，且当前角速度 A_0 大于阈值 A_m，则认为车辆开始发生转弯，并记录接下来一段时间内的角速度。对于下一时刻的角速度 A_1，如果 A_1 大于阈值 A_m 且满足：

$$A_0 \cdot A_1 > 0$$

即角速度方向相同，则计入同一次转弯中，直到下一时刻角速度不在阈值范围内，或者方向与上一时刻保持一致或相反时，一次转弯记录结束，计入一次车辆转弯次数。

2.道路曲折程度的计算

我们定义道路曲折程度，为一段里程内的平均转弯次数，计算方法为，一段行程的转弯次数除以该行程的总里程数：

$$road\_twist = \frac{turn\_time}{mile}$$

通过计算每辆车每段行程的平均转弯次数，就得到了该行程的“道路曲折程度”。

5.1.3 综合指标的相关性分析（的下面的表格）

根据以上分析建立基于行车安全性的 7 种不良驾驶行为指标（超速、熄火滑行、急加速急减速、急变道、超长怠速、怠速预热以及疲劳驾驶）；基于天气环境状况的 4 种指标（天气状态、风级、最高温度、最低温度）；基于道路状况的 1 种指标（道路曲折度）共 12 种驾驶行为表征指标，建立如图 78 所示的驾驶行为安全性指标集。

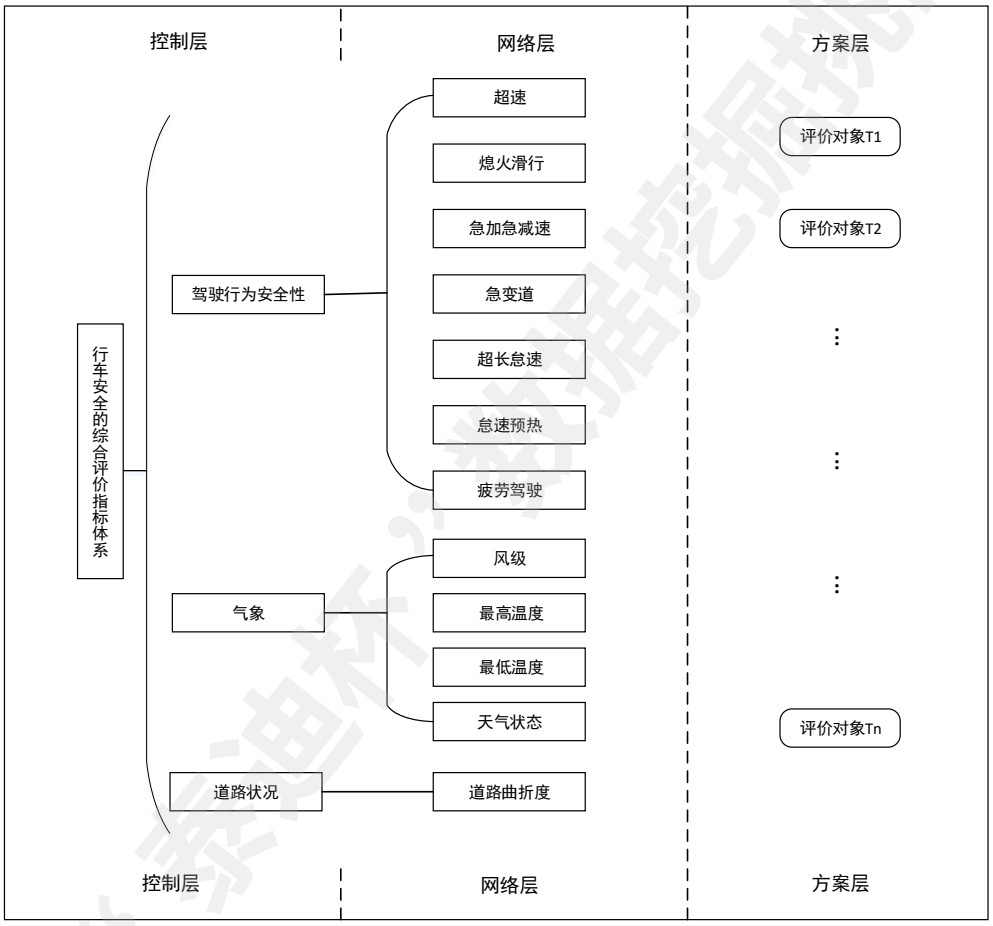

图 78　驾驶行为安全性指标集

5.2　综合评价模型构建

为了更真实准确地反应出驾驶行为评价指标的相互反馈关系以及本研究中评价指标所体现的分布特点，采用网络层次分析法对驾驶行为安全指标权重赋值进行主观性评价，其中客观权重赋值方法与第四章的客观赋权重的方法相同，其他的思路与第四章基本完全相同。

5.2.1 网络层次分析

网络层次分析（ANP）考虑到一般层次分析各层次结构内部存在相互关系，将元素划

分为两大部分，控制层和网络层。控制层包括研究的问题目标和决策准则，认为各准则之间是相互独立的；网络层由控制层准则之下的各特征元素组成，考虑这些元素之间的相互关系，所以每个控制层准则之下都是一个具体的特征元素网络[9]。

我们这里通过网络层次分析法，根据车辆的驾驶行为、天气情况和道路状况，综合考虑节能和效率，构建行车安全综合评价体系。主要分为两个步骤：

①根据本题数据，构造 ANP 典型结构；

②构造 ANP 超级矩阵计算特征权重。

先给出本节出现的符号定义，如表 24：

表 24　本节所用到的符号

符号	意义
safty(P1)	评价体系中安全方面
efficiency(P2)	评价体系中效率方面
energy(P3)	评价体系中节能方面
drive(D)	驾驶行为
road(R)	道路状况
weather(W)	天气情况
fatigue(D1)	疲劳驾驶
overspeed(D2)	超速
accelerate(D3)	急变速
roadchange(D4)	急变道
idling(D5)	怠速预热
coast(D6)	熄火滑行
overtime(D7)	超长怠速
twist(R1)	道路曲折度
weatherType(W1)	天气类型
wind(W2)	风速等级
highTemperature(W3)	高温
lowTemperature(W4)	低温

（1）构造 ANP 结构

a)给出评价行车安全的指标体系，如表 25：

表 25　行车安全的指标体系

驾驶行为（D）	疲劳驾驶（D1）
	超速（D2）
	急变速（D3）
	急变道（D4）
	怠速预热（D5）
	熄火滑行（D6）
	超长怠速（D7）
道路状况（R）	道路曲折度（R1）

天气情况（W）	天气类型（W1）
	风速等级（W2）
	高温（W3）
	低温（W4）

b)构建元素间相互关系

控制层各准则的权重可以根据 AHP 方法决定，本文上一章已经介绍过。网络层次分析还需要确定各特征元素之间的相互关系，这部分由专家以一个二联表的形式给出，本题我们只能根据主观判断给出各指标间是否相关，结果如表 26（说明：顶部元素为被影响的特征因素，左列元素为引起顶部特征因素的因素，打 √ 的表明左列的因素会影响顶部对应的特征）：

表 26　各特征元素之间相互关系二联表

被影响因素 / 影响因素		Project (P)			Drive (D)							Road (R)	Weather (W)			
		P1	P2	P3	D1	D2	D3	D4	D5	D6	D7	R1	W1	W2	W3	W4
Project (P)	P1				√	√	√	√	√	√	√	√	√	√	√	√
	P2				√	√	√	√	√	√	√	√	√	√	√	√
	P3				√	√	√	√	√	√	√	√	√	√	√	√
Drive (D)	D1	√	√	√			√	√	√							
	D2	√	√	√	√			√								
	D3	√	√	√	√	√										
	D4	√	√	√	√	√	√									
	D5	√	√	√												
	D6	√	√	√							√					
	D7	√	√	√												
Road (R)	R1	√	√	√	√		√	√								
Weather (W)	W1	√	√	√	√		√	√		√	√				√	√
	W2	√	√	√									√		√	√
	W3	√	√	√	√								√			
	W4	√	√	√	√					√	√		√			

由二联表得到的行车安全综合评价体系 ANP 结构图如图 79：

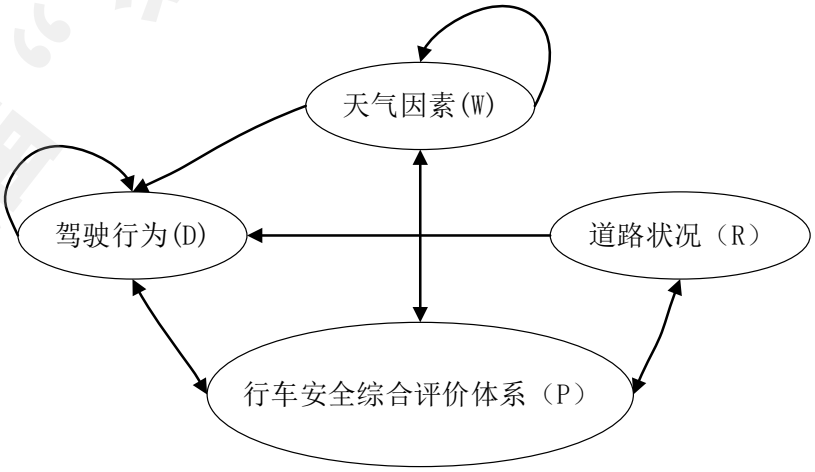

图 79　行车安全综合评价体系 ANP 结构图

（2）构造 ANP 超级矩阵计算权重

a)基于网络层中各因素之间的相互关系，进行两两比较，得到判断矩阵，如图 80

被影响因素 / 影响因素	Project (P)	Drive (D)	Road (R)	Weather (W)
Project (P)		21	3	12
Drive (D)	21	11		
Road (R)	3			
Weather (W)	12	9		7

图 80　判断矩阵

该表展示的是控制层准则之间的关联情况，网络层中各元素之间的关联情况也以这样的二联表格式表示。根据这些二联表就可以得出元素之间两两判断矩阵。

b)确定未加权超矩阵，如图 81（基于两两判断矩阵，使用特征向量法获得归一化特征向量值，填入超矩阵列向量，这部分以及下面几步都是用 ANP 计算工具——SuperDesicion 实现，这里都直接给出结果）：

	D1	D2	D3	D4	D5	D6	D7	P1	P2	P3	R1	W1	W2	W3	W4
D1	0	0.8	0.66667	0.53961	0	0	0	0.35428	0.35428	0.35428	0.62501	0.49633	0	1	0
D2	0	0	0.33333	0.29696	0	0	0	0.23993	0.23993	0.23993	0	0	0	0	0
D3	0.53961	0	0	0.16342	0	0	0	0.15865	0.15865	0.15865	0.23849	0.23595	0	0	0
D4	0.29696	0.2	0	0	0	0	0	0.10362	0.10362	0.10362	0.1365	0.15285	0	0	0
D5	0.16342	0	0	0	0	0	0	0.06757	0.06757	0.06757	0	0	0	0	0
D6	0	0	0	0	0	0	0	0.04477	0.04477	0.04477	0	0.0688	0	0	0.66667
D7	0	0	0	0	0	1	0	0.03118	0.03118	0.03118	0	0.04607	0	0	0.33333
P1	0.63698	0.63698	0.63698	0.63698	0.63698	0.63698	0.63698	0	0	0	0.63698	0.63698	0.63698	0.63698	0.63698
P2	0.10473	0.10473	0.10473	0.10473	0.10473	0.10473	0.10473	0	0	0	0.10473	0.10473	0.10473	0.10473	0.10473
P3	0.25829	0.25829	0.25829	0.25829	0.25829	0.25829	0.25829	0	0	0	0.25829	0.25829	0.25829	0.25829	0.25829
R1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
W1	0	0	0	0	0	0	0	0.54244	0.54244	0.54244	0	0	0.68334	0.8	1
W2	0	0	0	0	0	0	0	0.08542	0.08542	0.08542	0	0	0	0	0
W3	0	0	0	0	0	0	0	0.15885	0.15885	0.15885	0	0.33333	0.11685	0.2	0
W4	0	0	0	0	0	0	0	0.21329	0.21329	0.21329	0	0.66667	0.19981	0	0

图 81　未加权超矩阵

c)计算加权超矩阵，如图 82：

	D1	D2	D3	D4	D5	D6	D7	P1	P2	P3	R1	W1	W2	W3	W4
D1	0	0.19999	0.16665	0.13489	0	0	0	0.25886	0.25886	0.25886	0.15624	0.12819	0	0.25827	0
D2	0	0	0.08333	0.07423	0	0	0	0.1753	0.1753	0.1753	0	0	0	0	0
D3	0.13489	0	0	0.04085	0	0	0	0.11592	0.11592	0.11592	0.05962	0.06094	0	0	0
D4	0.07423	0.05	0	0	0	0	0	0.07571	0.07571	0.07571	0.03412	0.03948	0	0	0
D5	0.04085	0	0	0	0	0	0	0.04937	0.04937	0.04937	0	0	0	0	0
D6	0	0	0	0	0	0	0	0.03271	0.03271	0.03271	0	0.01777	0	0	0.17218
D7	0	0	0	0	0	0.24998	0	0.02278	0.02278	0.02278	0	0.0119	0	0	0.08609
P1	0.47775	0.47775	0.47775	0.47775	0.63698	0.47775	0.63698	0	0	0	0.47775	0.40576	0.54704	0.40576	0.40576
P2	0.07855	0.07855	0.07855	0.07855	0.10473	0.07855	0.10473	0	0	0	0.07855	0.06671	0.08994	0.06671	0.06671
P3	0.19372	0.19372	0.19372	0.19372	0.25829	0.19372	0.25829	0	0	0	0.19372	0.16453	0.22182	0.16453	0.16453
R1	0	0	0	0	0	0	0	0.08096	0.08096	0.08096	0	0	0	0	0
W1	0	0	0	0	0	0	0	0.1022	0.1022	0.1022	0	0	0.09648	0.08378	0.10473
W2	0	0	0	0	0	0	0	0.01609	0.01609	0.01609	0	0	0	0	0
W3	0	0	0	0	0	0	0	0.02993	0.02993	0.02993	0	0.03491	0.0165	0.02095	0
W4	0	0	0	0	0	0	0	0.04018	0.04018	0.04018	0	0.06982	0.02821	0	0

图 82　加权超矩阵

d)计算指标的全局权重和综合权重，如图 83，图 84：

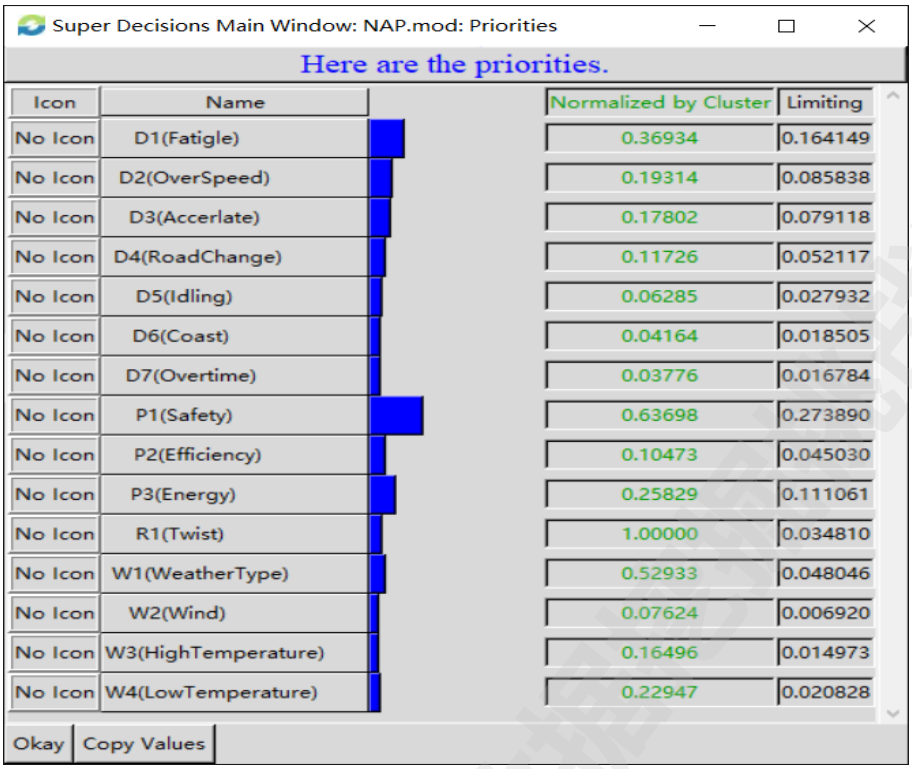

图 83　指标的全局权重

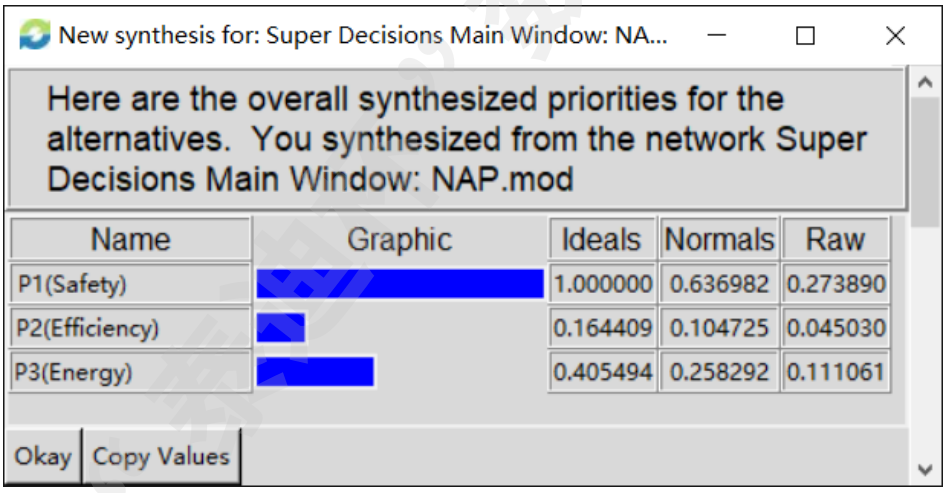

图 84　指标的综合权重

5.3　结果分析

本章我们根据气象条件和道路状况的数据，提取出温度高低、风速、道路曲折度等几个特征，再结合第三章提取的驾驶行为的七个特征，一共 12 个特征。综合考虑行车安全、效率和节能，给出了一个行车安全的综合评价指标体系。

根据网络层次分析的结果，由指标的全局权重可以看出，疲劳驾驶、超速、急变速、急变道、天气类别这五个指标是评价指标体系中占权重最大的五个指标，这也与我们的主观感觉相符。而对于安全、节能、效率这三个方面，根据指标的综合权重结果可以看出，

我们给出的这个综合评价指标体系，关注的程度最大的还是行车安全，然后是节能，最后才是效率。

最后建立这个综合评价模型的方法与第四章建立行车安全评级模型的相同，只是数据特征由之前的七个驾驶行为的特征，变为综合考虑到天气和道路状况后的 12 个特征，这样就可以得到我们的本章的行车安全综合评价指标体系。

总结

本文主要运用机器学习评价权重和决策树的算法构建行车安全性多属性综合评价模型。

首先，我们对对原始数据集的每个属性的数据特点进行了分析，为每个属性制定了不同的数据预处理策略，尽可能全面地填补数据的缺失值，筛选数据的异常值。

其次，我们对车辆行驶过程中的驾驶行为信息进行了详细的分析和提取，包括车辆的行驶路线、行驶里程、平均行驶速度、发生急加速急减速的情况等。并通过可视化手段对分析的结果进行了全面的展示。

之后，我们对车辆行驶过程中的不良驾驶行为，包括疲劳驾驶、超速、急加速、急减速、怠速预热、超长怠速、熄火滑行、急变道等，建立了具体的定义和判别策略，并通过多种思路验证判别结果的准确性。

在提取不良驾驶行为的基础上，我们通过主观和客观两种方法为 7 种特征赋权值，构建了 450 辆车的评价矩阵，并使用 TOPSIS 算法，对车辆的行驶安全进行了标记，判断数据集中所有车辆的行驶安全性。同时使用 K-means 聚类算法为决策树分类提供了借鉴和指导，最终构建了安全评价模型。

最后，我们综合考虑安全、效率、环保等因素，通过 ANP 算法得到了不同指标的权重，最终构建了车辆行驶安全综合评价模型。

参考文献

[1] pyx61198.GIS 算法--已知一点经纬度，方位角，距离求另一点[EB/OL].https://blog.csdn.net/pyx6119822/article/details/52298037,2016-08-24.

[2] 丁琛.基于车辆动态监控数据的异常驾驶行为识别技术研究[D].北京:北京交通大学,2015.

[3] Ann Williamson. Review of on-road driver fatigue monitoning devices[C].NSW Injury Risk Management Research Centre University of New South Wales April 07nd 2002.

[4] CJJ37-2012,城市道路工程设计规范[S].北京:中华人民共和国住房和城乡建设部,2016.

[5] 飘云飘云飘云.计算权重的方法[EB/OL].https://wenku.baidu.com/view/de9d0b31580216fc700afd2d.html,218-06-26.

[6] 飞龟道人.Python熵权法确定权重[EB/OL].https://blog.csdn.net/weixin_40450867/article/details/81226705,2018-07-26.

[7] 晓窗读易 101.AHP层次分析法及应用基础教程[EB/OL].https://wenku.baidu.com/view/dc752f4ceffdc8d376eeaeaad1f34693daef1098.html,218-08-13.

[8] zhangyue_lala.理想解法TOPSIS评价[EB/OL].https://blog.csdn.net/zhangyue_lala/article/details/70149388,2017-04-13.

[9] Weilai0131.用SuperDecision进行网络层次分析的应用实例[EB/OL].https://wenku.baidu.com/view/84a98622360cba1aa911dad5.html,2018-07-01