2019 年（第六届）全国大学生统计建模大赛

基于 FAHP 与 GA-BP 神经网络的行车安全评价

参 赛 单位: 东北大学秦皇岛分校

参赛者姓名:岳淼聪、吴自强、张雪

目　录

表格和插图清单

摘　要

本文以第七届“泰迪杯”数据挖掘挑战赛C题所提供的数据集为基础，针对如何建立行车安全评价模型进行研究，对驾驶员的驾驶行为做出安全评估。经过初步探索数据与一系列数据预处理后，运用百度API绘制行车路线图并结合数据分析出驾驶员的行车习惯，通过R型聚类分析得到安全评价指标体系，同时基于模糊层次分析法（FAHP）采集到的样本分别用偏最小二乘回归模型、PSO-BP 神经网络模型与GA-BP神经网络模型得到评价结果，选择评价效果最好的模型对行车安全做出评价。

首先对数据集进行初步探索，分析数据集的质量。之后进行数据规约与数据清洗，同时求出相邻经纬度之间实际距离与对应速度的一阶差商，采用K-Means算法与箱线图识别异常点，初步用平均插值法填补异常点，用百度API绘制出行车路线图后，发现行车路线存在大量漂移过程，接着用db3小波分析对数据进行奇异点检测和剔除，但去噪的同时也丢失了原本的有效信息，得到的行车路线图仍不理想。将原始数据与经纬度时序图对比分析后发现异常数据是成段出现的，采用三次样方插值拟合模型对数据进行整段拟合修复，最终得到了理想的行车路线图。

对于行车安全评价模型，首先设计算法从处理后的数据集中识别出超速行驶、疲劳驾驶、急加速、急减速等 8 个安全评价指标，通过 R 型聚类分析对 8 个指标中存在较强相关性的指标进行取舍，最终得到 6 个评价指标。构建 FAHP 模型对指标进行赋权，求得6个指标的权值向量，由此来计算全部驾驶员的评价结果，并对其基于留出法进行划分。根据所采集的样本，建立偏最小二乘回归模型、PSO-BP 神经网络模型与 GA-BP 神经网络模型对行车安全进行评价。利用测试集进行多次试验，并由最大似然估计得到均方根差、平均绝对值误差、平均绝对偏差百分比等指标的无偏估计值对三个模型的评价结果进行比较，得出GA-BP神经网络模型的评价精度优于其他两个模型，之后通过K-Means算法划分安全等级，使安全评价更有依据，最后给出了模型的不足之处。

关键词：行车安全评价模型　FAHP　偏最小二乘回归模型　PSO-BP神经网络模型　GA-BP神经网络模型

一、引言

（一）研究背景

交通运输行业的迅猛发展在给社会带来巨大便利的同时，严峻的交通安全问题也越来越引起人们的关注。车辆是否被安全驾驶，不仅关系到司机和乘客的安全问题，危险驾驶行为也可能干扰道路上其余车辆的正常行驶，甚至危及路人的生命安全。国内外道路交通事故统计结果显示，由驾驶人直接原因造成的事故占到70%以上。而对事故成因的分析表明，驾驶行为与交通事故发生有着很强的相关性。因此，对运输车辆安全驾驶行为进行分析建模，以便提前为运输车辆不安全行为做出干预，对于日常交通安全显得尤为重要。

数据中蕴含有很多信息，丰富的数据能够大大提高评价的客观性。车联网[1]系统中被自动采集的数据，对安全驾驶行为的分析提供了很大帮助。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利用车联网等系统数据，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。

（二）研究现状

目前，世界上有很多汽车安全评价方法，部分以技术标准和碰撞试验数据来规定和评价汽车安全性，部分通过新车安全评价程序(New Car Assessment Program,NCAP)、相关零部件的安全试验及技术法规等来评价，经查阅资料可知，在目前的研究中，主要有以下几种研究方法：

李平凡[2]基于模糊ANP理论的驾驶行为安全性评估方法基于驾驶行为表征指标，提出了基于模糊AP理论的驾驶行为安全性评估方法。针对各个驾驶环节进行评估，利用模糊理论建立驾驶行为表征指标与行为安全性的模糊隶属关系，确定驾驶感知环节、决策环节、操控环节间的相对重要度权重。

刘茜[3]采用模糊综合评价法的方法建立汽车安全状态模型，并选取具体汽车采取预设故障的方法进行道路试验，分析评价方法的准确性。

许潇潇[4]建立了汽车安全性综合评价模型并计算出各环境类型对于汽车安全性影响的权重并利用Euro NCAPI的试验数据作为汽车被动安全性的参考数据，建立了一个能够将汽车安全性与其行驶环境特点相结合的综合评价模型及指标

体系。

但目前看来，还没有一个比较成熟的方法针对车联网中的大数据进行分析，进而构建对驾驶员驾驶行为安全性评价的合理模型。

二、建模准备

（一）数据来源

数据来自第七届“泰迪杯”数据挖掘挑战赛 C 题所提供的数据集（数据集说明见表 1）。数据为行车过程中车辆的电子标签自动采集的当前驾驶行为下的行车状态信息。数据集包含 450 辆运输车辆的行车轨迹采集数据。由于采集设备精度，存在异常数据。

表 1　数据集说明

序号	指标名称	指标说明	具体说明
1	vehicleplatenumber	车牌号码	
2	device_num	设备号	
3	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）
4	lng	经度	东经
5	lat	纬度	北纬
6	acc_state	ACC 状态	点火 1/熄火 0
7	right_turn_signals	右转向灯	灭 0/开 1
8	left_turn_signals	左转向灯	灭 0/开 1
9	hand_brake	手刹	灭 0/开 1
10	foot_brake	脚刹	无 0/有 1
11	location_time	采集时间	
12	gps_speed	GPS 速度	单位：Km/h
13	mileage	GPS 里程	单位：Km

（二）数据探索

抽取 10 辆运输车辆的行车数据进行探索性分析，包括统计量分析、速度分布分析和经纬度时序图分析。

1.统计量分析

①速度四分位间距：将所有 GPS 速度数值由小到大排序并分成四等份，处于第一个分割点位置的数值是下四分位数，处于第三个分割点的数值是上四分位数。四分位间距是上四分位数和下四分位数之差，其值越大，说明 GPS 速度的变异程度越大，即行车速度越不稳定；反之，则越小。

②速度众数：每辆车所测 GPS 速度中出现次数最多的数值，反映汽车总体的行车状态。

经过 MATLAB 数值计算，得到 10 辆汽车的统计量（见表 2）。

表 2　10 辆汽车的统计量

车牌号	速度四分位间距	速度众数
AA00002	84	0
AB00003	58	0
AD00006	31	0
AD00013	45	0
AD00053	49	0
AD00083	54	0
AD00419	62	0
AF00098	49	0
AF00131	46	0
AF00373	54	0

2.速度分布分析

根据 10 辆车全部的 GPS 速度数值分布情况，取组距为 20 将速度分组，并绘出速度频率分布直方图（见图 1）。

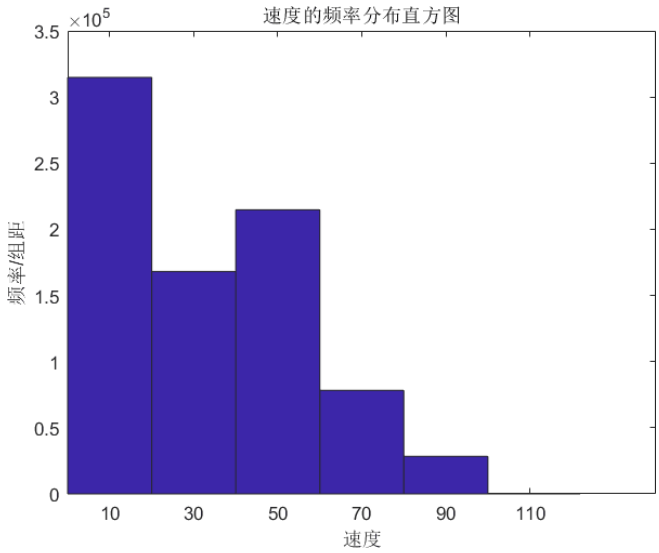

图 1　汽车速度频率分布直方图

速度频率分布直方图能够更好地体现 10 辆车的速度分布情况，可见速度大多集中在[0,20]区间内，结合速度众数分析，发现这 10 辆车大部分时间速度为 0，即没有行驶。

3.经纬度时序图

运用 MATLAB 画出车辆的经纬度时序图，以车牌号 AA00002 为例（见图 2、图 3）。

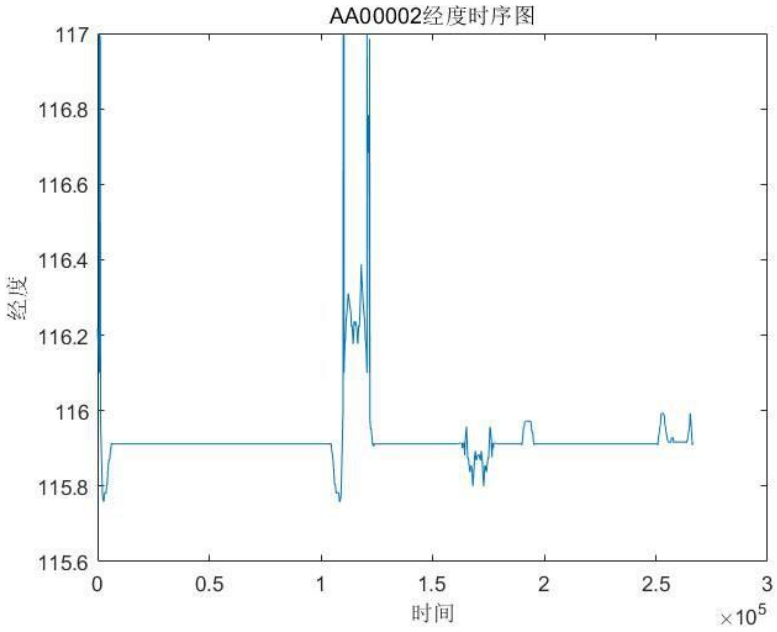

图 2　AA00002 经度时序图

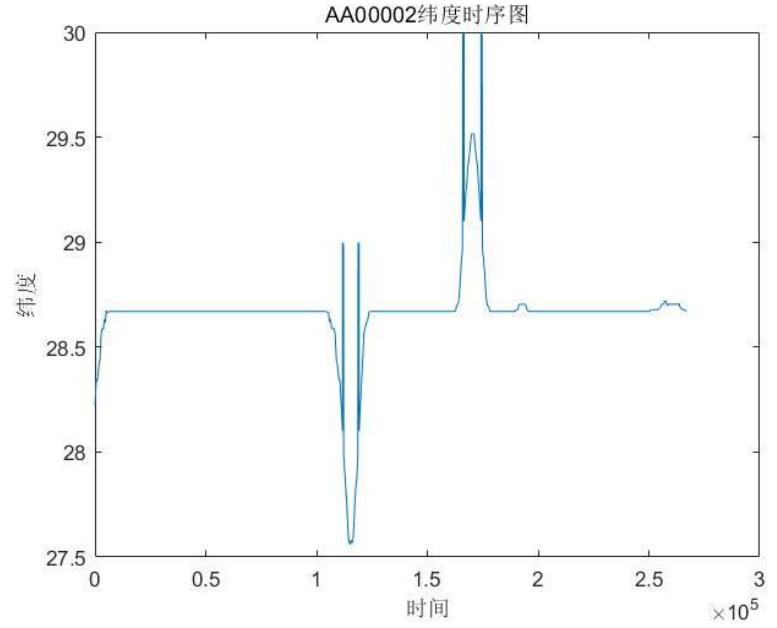

图 3　AA00002 纬度时序图

从图 2、图 3 可以发现 AA00002 的经纬度时序图像大部分时间呈现一条直线，即经度和纬度大部分处于一个数值附近，个别时间段会出现经纬度数值骤变的情况，则在数据清洗中需要对具体一段时间内的经纬度坐标数值进行处理。

（三）数据变换及预处理

1. 数据变换

运用坐标转换[5]接口将 84 经纬度坐标转换成百度经纬度，从而用百度 API 绘制 10 辆车辆的行车路线图，验证数据预处理的效果。

2.数据规约

通过初步的数据探索，发现有些数据指标对于行车安全评价模型的建立没有帮助，可以删掉无用的数据指标，即对于反映相同状况的数据指标可以进行数据规约处理。处理后保留属性结果见表 3：

表 3　数据集保留属性

属性名称	属性说明
车牌号码	车辆的标识号码
方向角	汽车当前时间下方向盘的角度，用于判断急变道情况
经度	汽车当前时间下的坐标经度，用于判断行车距离
纬度	汽车当前时间下的坐标纬度，用于判断行车距离
ACC 状态	汽车当前时间下发动机状态，1 表示点火，汽车有正常供电行驶；0 表示熄火，汽车没有正常供电
采集时间	该数据采集的时间
GPS 速度	汽车当前时间下的 GPS 速度
GPS 里程	汽车到目前时间为止的行驶里程

3.数据清洗

①重复值剔除

由初步探索发现大部分运输车辆的速度为 0 且经纬度没有发生变化，说明运输车辆有很长时间属于静止状态，这对于速度和路线的研究没有指导作用。为减少大量无用数据造成的工作量，清洗掉冗余数据，具体流程见图 4：

图 4　重复值剔除流程图

流程说明：输入车辆的 GPS 速度、经度及纬度。当汽车速度为零时，判断该辆车相邻两点的经纬度是否发生变化，若不发生变化则当作重复值剔除；若经纬度发生变化则保留数值。

②异常值剔除

第一步：指标构造。

构造差分。由于所给数据是离散的，构造速度的一阶差分，更能反映离散量之间的一种关系。

$$\Delta v = v_n - v_{n-1} \tag{1}$$

进行经纬度变换得到两点之间的距离为：

$$D = cos\left(\frac{latM}{57.2958}\right) \times cos\left(\frac{lngM-lngN}{57.2958}\right) + sin\left(\frac{latM}{57.2958}\right) \times sin\left(\frac{latN}{57.2958}\right) \quad (2)$$

$$D_{mn=} = R_E \times arccos(D) = 6370.856\, arccos(D) \quad (3)$$

式（2）中$(lngM, latM)$和$(lngN, latN)$是两点的经纬度坐标，式（3）中R_E是地球的平均半径。

第二步：异常值识别。

速度异常值识别。绘制出速度差分Δv的箱线图，在箱线图中定义异常值的标准小于$Q_L - 1.5IQR$或大于$Q_U + 1.5IQR$（其中Q_U是上四分位点，Q_L是下四分位点，IQR是四分位间距），自动标识出数据异常值，以车牌号 AA00002 为例（见图 5）。

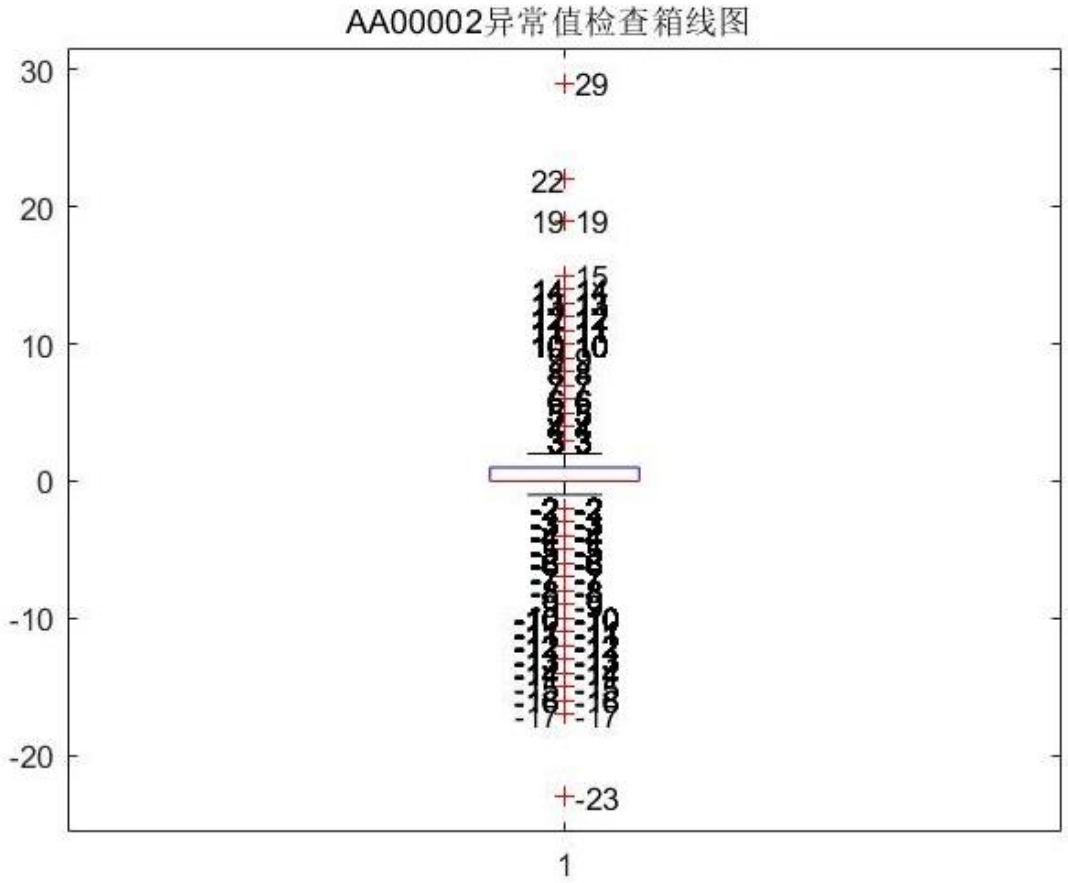

图 5　AA00002 速度异常值检验箱线图

坐标异常值识别。运用 K-Means 算法进行聚类，具体识别步骤如下：

- 运用 K-Means 算法将样本聚为 1 簇；
- 计算各坐标到它质心的相对位置；
- 给定阈值为 70m。若某坐标的距离大于该阈值，则判定该坐标为离群

点即异常点。以车牌号 AA00002 为例（见图 6）。

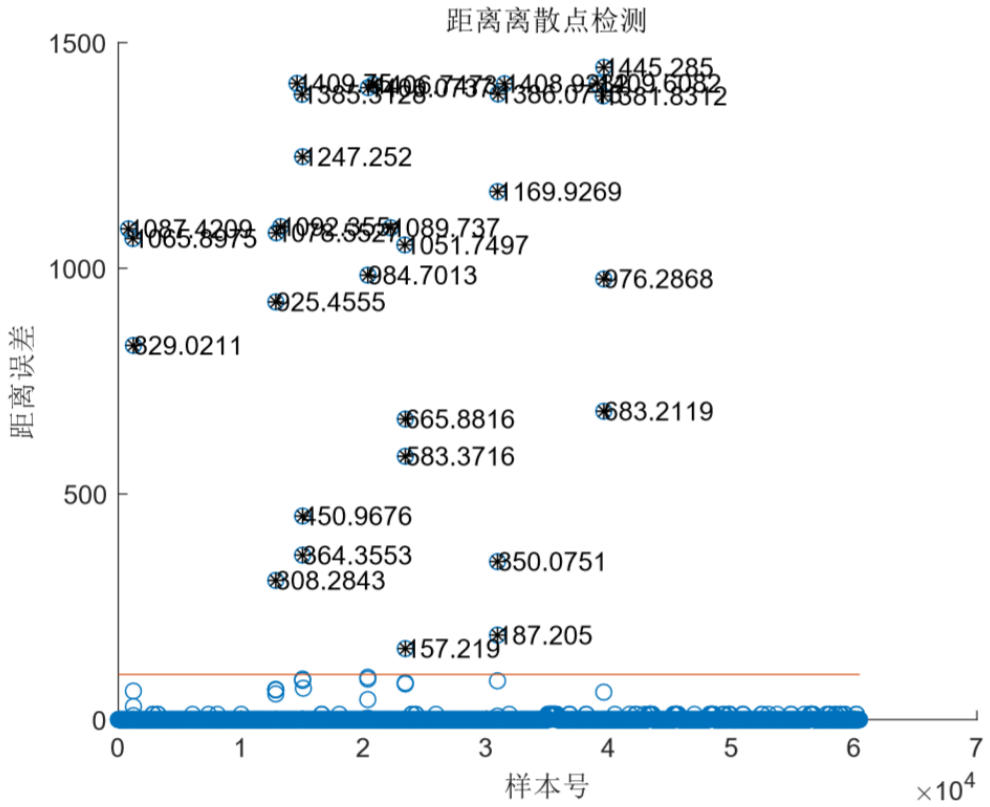

图 6　AA00002 坐标离散点检测

同时在处理过程中发现了一辆车的数据存在 60%以上的异常，于是将其从数据集中剔除。

③异常值补充。

方法一：平均插值法。

对于速度和经纬度异常点，运用平均插值法，补充被剔除的速度与经纬度异常值。

$$v_n{}' = \frac{v_{n+1}+v_{n-1}}{2} \tag{4}$$

$$\left(lng_n{}', lat_n{}'\right) = \frac{(lng_{n+1}, lat_{n+1})+(lng_{n-1}, lat_{n-1})}{2} \tag{5}$$

经过初步修正后，观察车牌号为 AA00002 的车辆的经纬度时序图（见图 7），可见经初步修正后数据的异常点相对减少了，但异常点仍存在。用百度 API 画出路线图（见图 8），发现 AA00002 的行车路线图与实际道路相比仍存在较多偏移，则说明平均插值法对剔除异常值的补充效果并不理想。

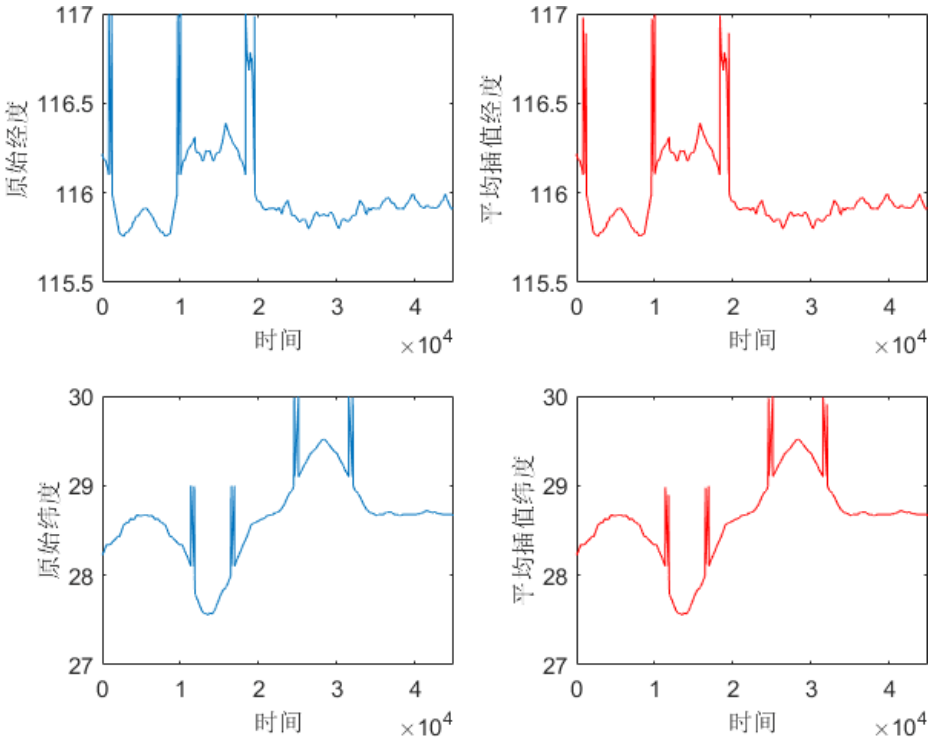

图 7　初步修正后 AA00002 车的经纬度时序对比图

图 8　AA00002 车的整体与局部路线图

方法二：小波分析检测奇异点并剔除。

运用小波变换，通过对时间频率的局部化分析，对经纬度的时间序列信号逐步进行多尺度细化，最终达到高频处时间细分，低频处低频率细分，能够自动适应时频信号分析的要求，从而可聚焦到信号的任意细节，利用 db3 小波对信号进行 6 层分解，发现奇异值点包含在细节信号的 d1、d2 和 d3，且与原信号中的奇

异点是同步的。为了消除奇异点，重构信号时，令细节信号 d1、d2、d3 和 d4 都等于零。画出小波变换后 AA00002 的经纬度时序图（见图 9）。

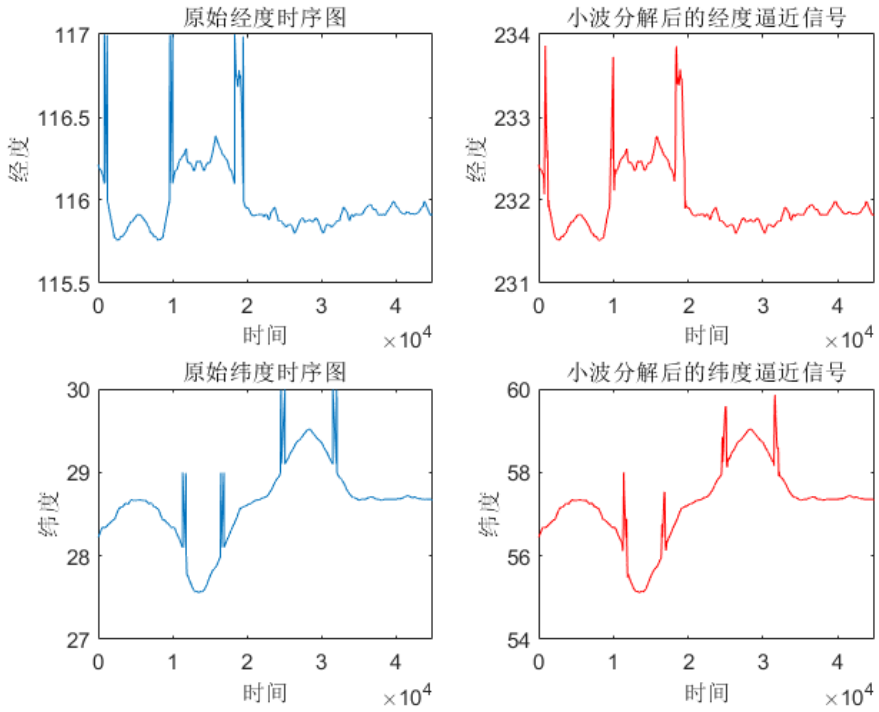

图 9　小波变换后 AA00002 的经纬度时序对比图

经观察，AA00002 的经纬度骤变部分已经相对平滑，奇异点数目减少，此时利用百度 API 画出路线图（见图 10）。发现虽然该车的行驶路线平滑了许多，但行车路线图仍与实际道路有偏离，对剔除异常值的处理仍不理想。

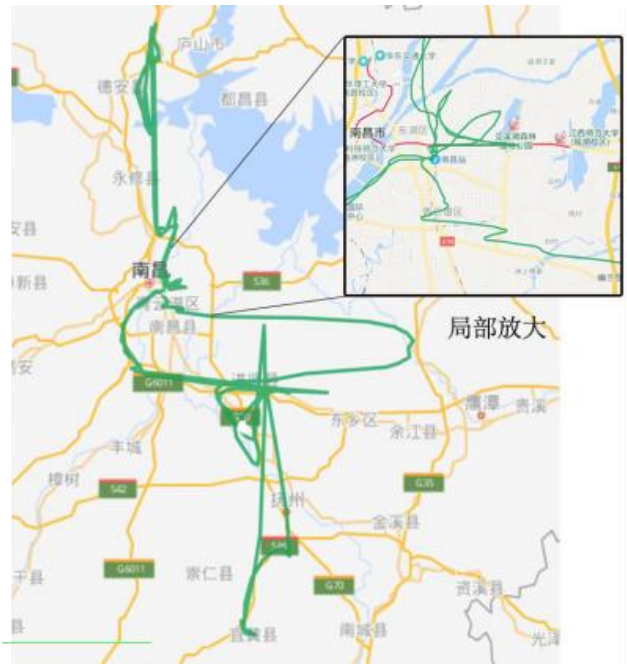

图 10　AA00002 车小波变换后的整体和局部路线图

经分析得知因为小波变换在剔除奇异值的同时也删除了原始数据中的有用信息，这样修正后的数据对于绘制行车路线图是没有帮助的。平均插值法和小波分析都是从一个异常点出发，清洗异常值，但是从百度 API 绘制出来的路线图来看，这两种方法的剔除效果并不好，对单个异常值点的处理对路线图绘制是没有帮助的，再次对距离的原始数据进行观察，发现异常值出现的那一段时间内距离数值都表现为异常，如表 4 中 AA00002 两点间行车距离异常段原始数据。

表 4　AA00002 车原始异常数据段

经度	纬度	行车距离（m）
116.10	28.36	86965
116.99	28.36	2397.3
116.91	28.36	2367.9
116.90	28.36	2319
116.87	28.36	2309.2
116.85	28.36	2299.4

由表 4 分析，AA00002 车在两个较短时间内移动了 2000 多米甚至 8000 多米显然是不正常的。则可以推断采集设备在汽车行驶过程中记录数据有延误或者出现设备故障等问题。

方法三：样方插值拟合模型

为解决一段时间内的数据异常值，采用样方插值拟合模型（见图 11）来对整段的异常数据进行填补。

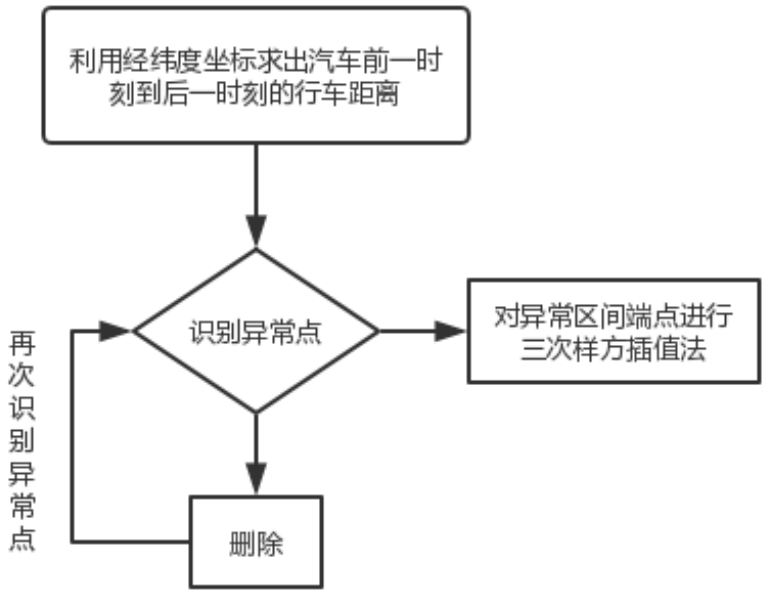

图 11　样方插值拟合模型流程图

步骤一：识别异常值，对经纬度异常区间进行局部放大（见图 12）。

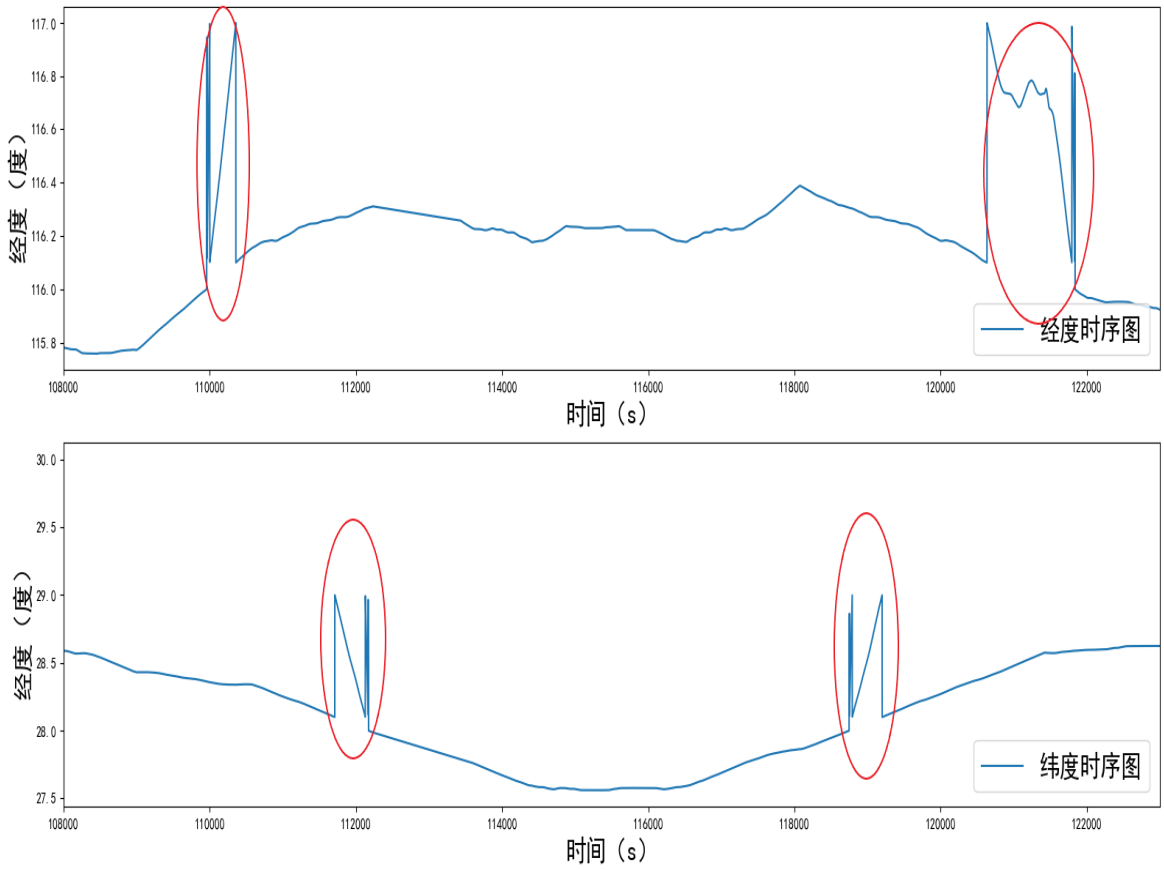

图 12　AA00002 车经纬度局部时序图

步骤二：剔除异常值。通过异常值识别算法，可以得到汽车经纬度数值异常段，从而对数据进行剔除。

步骤三：三次样方插值。再次识别异常值，找出异常段的端点，运用三次样方插值法拟合出异常段的曲线，用拟合曲线代表原本的行车曲线，用百度 API 绘制行车路线图（见图 13）。

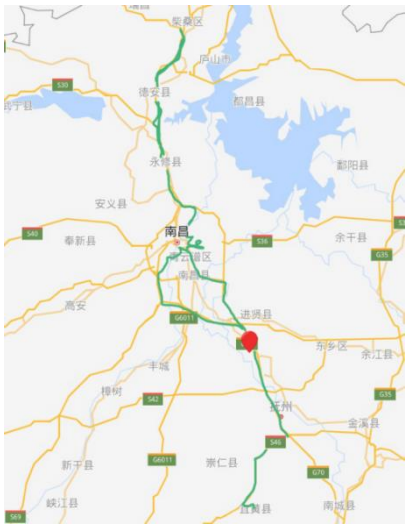

图 13　AA00002 的行车路线图

观察图 13 可以看出行车路线已经非常贴合实际路线，即可认为图 13 就是 AA00002 的实际行车路线图。

通过对数据的充分处理，最终可得到车辆的行车路线图（部分见图 14），整个过程提高了数据集的质量，进而提高行车安全评价模型的精度。

图 14　AB00006、AD00419、AF00373 的行车路线图

三、行车安全评价模型的构建

（一）构造行车安全评价指标

1.超速行驶

超速行驶[6]是指驾驶员在驾车行驶中，以超过法律、法规规定的速度进行行驶的行为。超速行驶评价模型见图 15。

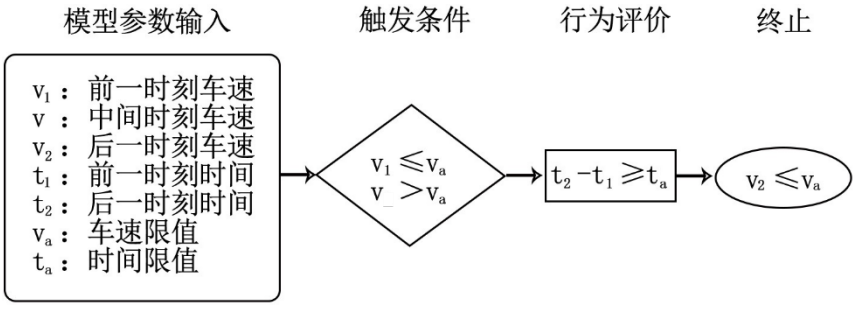

图 15　超速行驶评价模型

在超速算法识别流程中，根据行业标准设定最高车速限值v_a为 100km/h，超速时间最低限值t_a为 3s，此处设置时间限值是为了避免卫星定位数据漂移所造成的不必要的误判。超速行驶识别算法[7]流程步骤说明如下:

步骤 1：获取采样数据后，对数据进行预处理（后续指标不再叙述此步骤）；

步骤 2：判断是否连续的两条数据中，前者小于等于速度限值v_a，后者大于速度限值v_a，若是，则记录后者为v_1，并记录此条数据定位时间为超速发生时间t_1，并进行步骤 3，若否，则继续循环；

步骤 3：判断在t_1时间后速度是否大于v_a，若是，则累加两条数据的间隔时间，若否，则判断已累计时间t是否大于等于 3s，若是，则记录下时间t，并将t重置为 0，若否，则退出循环，并将t重置为 0；

步骤 4：判断是否为最后一条数据，若否，则继续循环，若是，则输出已记录所有时间t的总和(超速累计时长)和个数(超速次数)，算法结束。

2. 急减速

急减速是指描述车辆刚起步或行驶过程中猛踩制动的动作，其对驾驶安全有直接影响。急减速行驶评价模型见图 16。

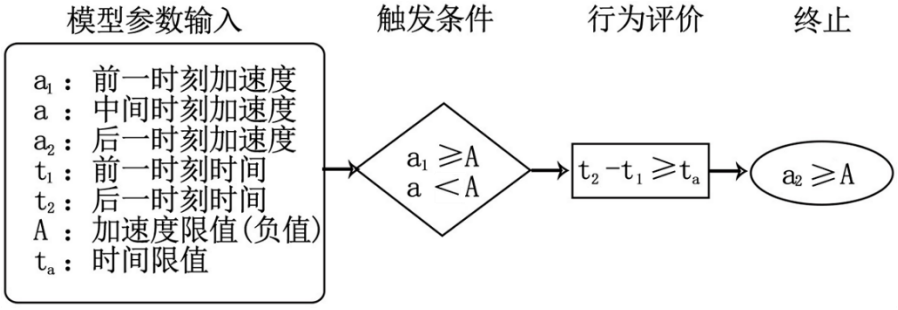

图 16　急减速行车评价模型

引入急减速行为界定限值A，按照行业经验预取A为$-1.6m/s^2$，时间限值t_n为最低限值 3s。急减速行为识别算法流程步骤说明如下：

步骤 1：判断是否连续的两条数据中，前者大于等于加速度限值A，后者小于加速度限值A，若是，则记录后者为a_1，并记录此条数据定位时间为急减速发生时间t_1，并进行步骤 2，若否，则继续循环；

步骤 2：判断在t_1时间后加速度是否小于A，若是，则累加两条数据的间隔时间，若否，则判断已累计时间t是否大于等于 3s，若是，则记录下时间t，并将t重置为 0，若否，则退出循环，并将t重置为 0；

步骤 3：判断是否为最后一条数据，若否，则继续循环，若是，则输出已记录所有时间t的总和(急减速累计时长)和个数(急减速次数)，算法结束。

3. 急加速

急加速是指描述车辆刚起步或行驶过程中猛踩加速踏板的动作，其对驾驶安

全有直接影响。急加速行驶评价模型见图 17:

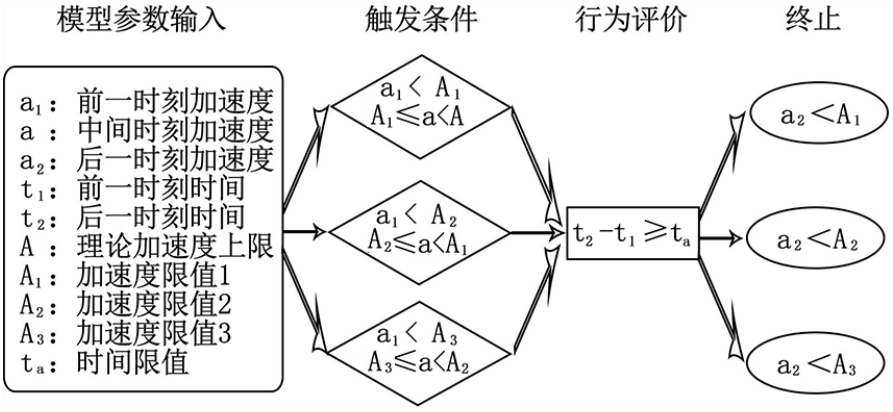

图 17　急加速行车评价模型

按照车辆危险行为考核标准[8]，引入急加速行为界定限值$A_1 = 2.78m/s^2$，$A_2 = 2.22m/s^2$，$A_3 = 1.67m/s^2$，时间限值$t_a = 2s$。此外，按照行业经验预取理论加速度上限$A = 5.88m/s^2$(即 0.6 倍重力加速度)。急加速行为识别算法流程步骤说明如下:

步骤 1：判断是否连续的两条数据中，前者小于A_1，后者大于等于A_1并小于A，若是，则记录后者为a_1，并记录此条数据定位时间为急加速一级行为发生时间t_1，并进行步骤 7，若否，则进行步骤 2；

步骤 2：判断是否连续的两条数据中，前者小于A_2，后者大于等于A_2并小于A_1，若是，则记录后者为a_1，并记录此条数据定位时间为急加速二级行为发生时间t_1，并进行步骤 6，若否，则进行步骤 3；

步骤 3：判断是否连续的两条数据中，前者小于A_3，后者大于等于A_3并小于A_2，若是，则记录后者为a_1，并记录此条数据定位时间为急加速三级行为发生时间t_1，并进行步骤 4，若否，则继续循环；

步骤 4：判断在t_1时间后加速度是否大于等于A_3并小于A_2，若是，则累加两条数据的间隔时间，若否，则判断已累计时间t是否大于等于 2s，若是，则记录下时间，并将t重置为 0，若否，则退出循环，并将t重置为 0；

步骤 5：判断在t_1时间后加速度是否大于等于A_2并小于A_1，若是，则累加两条数据的间隔时间，若否，则判断已累计时间是否大于等于 2s，若是，则记录下时间，并将t重置为 0，若否，则退出循环，并将t重置为 0；

步骤 6：判断在t_1时间后加速度是否大于等于A_1并小于A，若是，则累加两条数据的间隔时间，若否，则判断已累计时间是否大于等于 2s，若是，则记录下时间t，并将t重置为 0，若否，则退出循环，并将t重置为 0；

步骤 7：判断是否为最后一条数据，若否，则继续循环，若是，则分别输出一、二、三级行为已记录所有时间t的总和(急加速一、二、三级累计时长)和个数(急加速一、二、三级次数)，算法结束。

4.疲劳驾驶

疲劳驾驶评价模型见图 18。

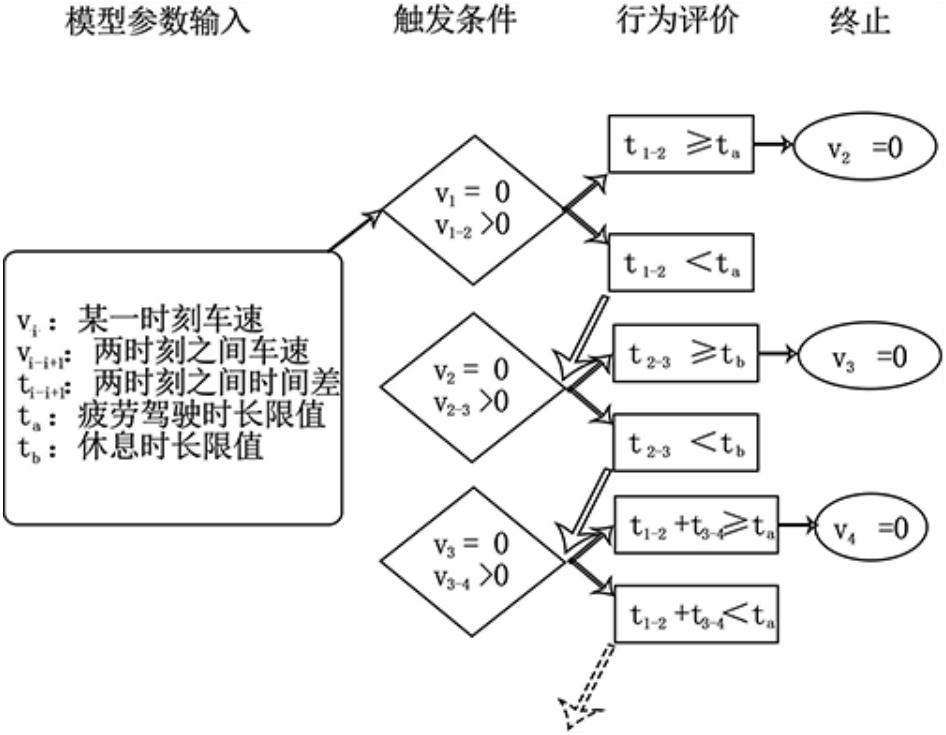

图 18　疲劳驾驶评价模型

根据道路运输行业相关法规和规范，驾驶员行车连续驾驶时间不得超过 4 小时，每次休息时间不得少于 20 分钟[9]。疲劳驾驶行为识别算法流程步骤说明如下：

步骤 1：判断是否连续的两条数据中，速度均大于 0，若是，则累积两条数据的间隔时间t，若否，则进行步骤 2；

步骤 2：判断已累计时间t是否大于 4 小时，若是，则记录下此时的累积时间t，并将t重置为 0，若否，则进行步骤 3；

步骤 3：判断下一个速度大于 0 的数据与当前数据的时间差T是否小于 20 分钟，若是，则继续累积速度大于 0 的两条数据的间隔时间t，并继续循环，若否，则将T与t重置为 0；

步骤 4：判断是否为最后一条数据，若否，则继续循环，若是，则输出已记录所有时间t的总和(疲劳驾驶累计时长)和个数(疲劳驾驶次数)，算法结束。

5.熄火滑行

熄火滑行是指将发动机熄火，变速箱置于空挡，利用汽车前进的惯性滑行。熄火滑行评价模型见图19。

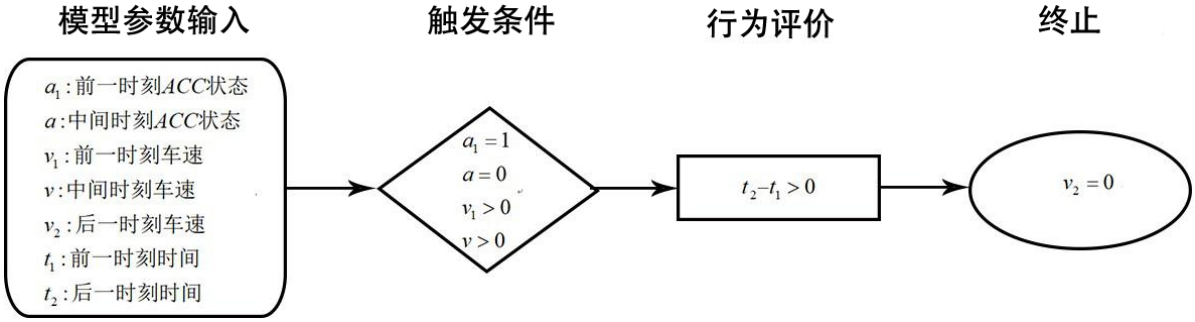

图 19　熄火滑行评价模型

熄火滑行行为识别算法流程步骤说明如下：

步骤1：判断是否连续的两条数据中，前者ACC状态为1，后者ACC状态为0，并且速度始终大于0，若是，则记录后者时间为熄火滑行发生时间t_1，并进行步骤2，若否，则继续循环；

步骤 2：判断在t_1时间后速度是否大于A，若是，则累加两条数据的间隔时间，若否，则记录下时间t，并将t重置为0；

步骤3：判断是否为最后一条数据，若否，则继续循环，若是，则输出已记录所有时间的总和(熄火滑行累计时长)和个数(熄火滑行次数)，算法结束。

6.车速稳定性

车速稳定性[10]是指车辆行驶过程中的稳定程度，本文提取车速标准差来表示车速的离散程度，标准差越大，离散程度越大，车速变化频率就越大。车速标准差计算公式为：

$$\bar{v}=\frac{1}{N}\sum_{i=1}^{N}v_i \tag{6}$$

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(v_i-\bar{v})^2} \tag{7}$$

7.方向稳定性

方向稳定性[11]是指车辆行驶过程中方向角的稳定程度，本文提取方向角的相对变化量的绝对值的标准差[12]来表示方向的离散程度。方向稳定性评价模型见图20。

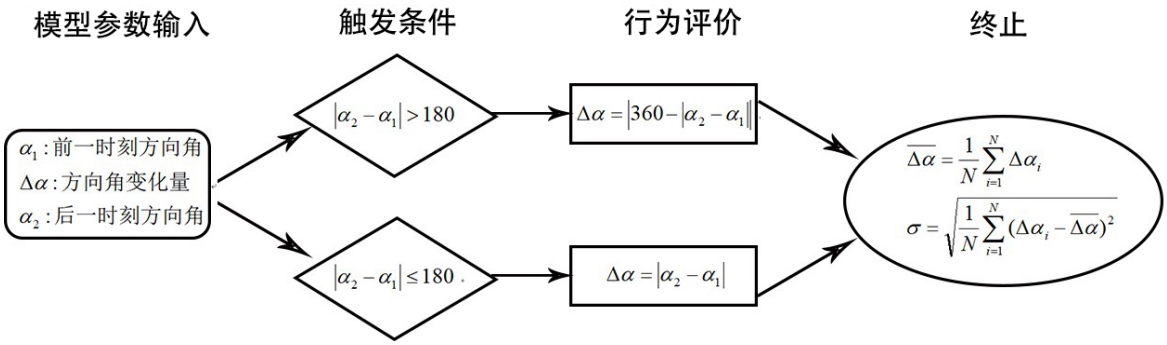

图 20　方向稳定性评价模型

8. 方向操控紧急程度

方向操控紧急程度是指驾驶员在行车过程中操纵方向盘的快慢程度，本文提取方向角的相对一阶均差的绝对值的标准差来表示。标准差越大，紧急程度越大，方向变化频率就越大。方向操控紧急程度评价模型见图 21。

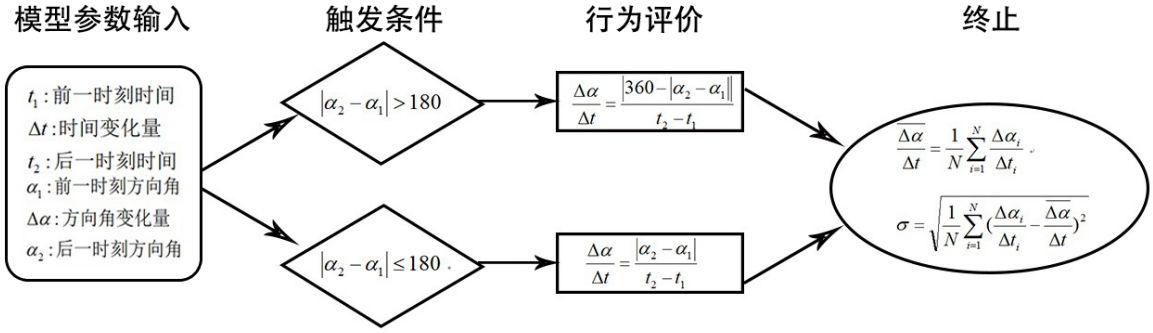

图 21　方向操控紧急程度评价模型

结合以上 8 种评价指标[13]，各安全指标的得分标准见表 5。

表 5　各评价指标得分标准

评价指标	得分标准		
车速稳定性	标准差	$y=\begin{cases}100; & 0\le\sigma\le20\\80; & 20<\sigma\le40\\60; & \sigma>40\end{cases}$	y
超速行驶	超速累积时长	$y_1=\begin{cases}100; & 0\le t\le20\\90; & 20<t\le50\\70; & 50<t\le100\\50; & 100<t\le200\\30; & t>200\end{cases}$	$y=\frac{1}{2}y_1+\frac{1}{2}y_2$
	超速次数	$y_2=\begin{cases}100-5n; & 100-5n\ge0\\0; & 100-5n<0\end{cases}$	
急减速	急减速累积时长	$y_1=\begin{cases}100-t; & 100-t\ge0\\0; & 100-t<0\end{cases}$	$y=\frac{1}{2}y_1+\frac{1}{2}y_2$
	急减速次数	$y_2=\begin{cases}100-3n; & 100-3n\ge0\\0; & 100-3n<0\end{cases}$	

续表

评价指标	得分标准		
急加速	危险行为一级时长	$y_1=\begin{cases}100-t_1;100-10t_1\geq 0\\0;\qquad 100-10t_1<0\end{cases}$	$y=\frac{1}{4}y_1+\frac{1}{4}y_2+\frac{1}{6}y_3+\frac{1}{6}y_4+\frac{1}{12}y_5+\frac{1}{12}y_6$
	危险行为一级次数	$y_2=\begin{cases}100-n_1;100-10n_1\geq 0\\0;\qquad 100-10n_1<0\end{cases}$	
	危险行为二级时长	$y_3=\begin{cases}100-t_2;100-10t_2\geq 0\\0;\qquad 100-10t_2<0\end{cases}$	
	危险行为二级次数	$y_4=\begin{cases}100-n_2;100-10n_2\geq 0\\0;\qquad 100-10n_2<0\end{cases}$	
	危险行为三级时长	$y_5=\begin{cases}100-t_3;100-10t_3\geq 0\\0;\qquad 100-10t_3<0\end{cases}$	
	危险行为三级次数	$y_6=\begin{cases}100-n_3;100-10n_3\geq 0\\0;\qquad 100-10n_3<0\end{cases}$	
疲劳驾驶	疲劳驾驶时长比例	$y_1=100(1-k)$	$y=\frac{1}{2}y_1+\frac{1}{2}y_2$
	疲劳驾驶次数	$y_2=\begin{cases}100-20n;100-20n\geq 0\\0;\qquad 100-20n<0\end{cases}$	
熄火滑行	熄火滑行时长比例	$y_1=100(1-k)$	$y=\frac{1}{2}y_1+\frac{1}{2}y_2$
	熄火滑行次数	$y_2=\begin{cases}100-50n;100-50n\geq 0\\0;\qquad 100-50n<0\end{cases}$	
方向稳定性	标准差排序序号	$y=\begin{cases}100; & 0<rank\leq\frac{1}{10}\times 450\\90; & \frac{1}{10}\times 450<rank\leq\frac{2}{10}\times 450\\80; & \frac{2}{10}\times 450<rank\leq\frac{3}{10}\times 450\\70; & \frac{3}{10}\times 450<rank\leq\frac{4}{10}\times 450\\60; & rank>\frac{4}{10}\times 450\end{cases}$	y
方向盘操控紧急程度	标准差排序序号	$y=\begin{cases}100; & 0<rank\leq\frac{1}{10}\times 450\\90; & \frac{1}{10}\times 450<rank\leq\frac{2}{10}\times 450\\80; & \frac{2}{10}\times 450<rank\leq\frac{3}{10}\times 450\\70; & \frac{3}{10}\times 450<rank\leq\frac{4}{10}\times 450\\60; & rank>\frac{4}{10}\times 450\end{cases}$	y

（二）样本采集及划分

对于从数据集中挖掘出的大量的评价指标，若对其一一进行评价，将消耗大量的时间和成本。经过数据探索与指标分析发现部分指标有较强的相关性。为使评价指标可以简明地反映行车安全性，对具有较大相关性的指标加以筛选，重新确定评价指标。

1.R 型聚类分析

通过定性考察以上 8 项指标，可以看出一些指标存在较强相关性，考虑从中选出几个具有代表性的指标，于是对 8 项指标根据相关性进行 R 型聚类，步骤如下：

步骤 1：对每个指标进行标准化处理；

步骤 2：运用类平均法进行类间相似性度量。

记指标$G_i(i=1,2,\cdots,8)$，$x_{ik}\in G_i(k=1,2,\cdots,449)$ 为该指标下车辆得分。

$$D(G_i,G_j)=\frac{1}{n_i\cdot n_j}\sum_{x_{ik}\in G_i}\sum_{x_{jk}\in G_j}d(x_{ik},x_{jk}) \qquad (8)$$

其中$D(G_i,G_j)$是G_i,G_j中两样本点距离的平均，n_i,n_j分别为G_i,G_j中样本点个数。经过处理，得到聚类树形图（见图 22）。

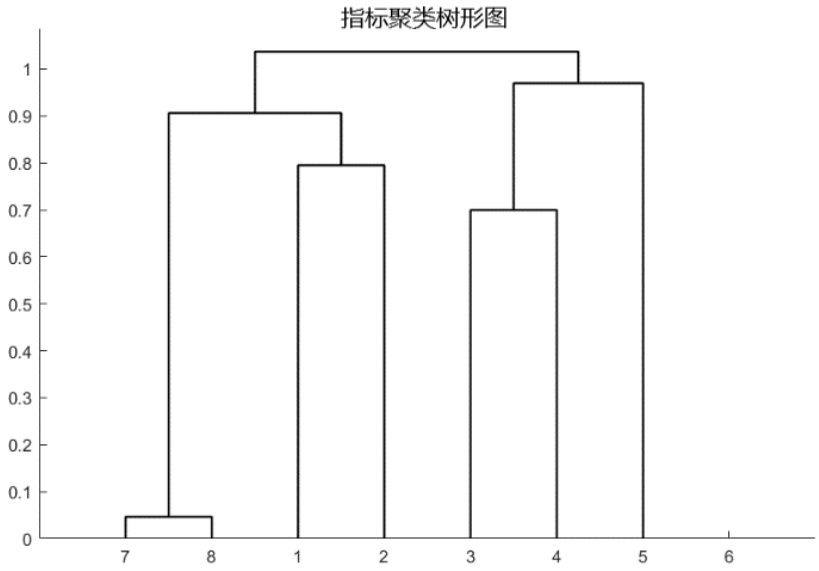

图 22　评价指标聚类树形图

从聚类图可以看出指标之间的相关性，最后得出车速稳定性、超速行驶、急减速、急加速、疲劳驾驶、方向操控紧急程度这 6 个精简的指标。

2.样本采集——FAHP

由于传统层次分析法在判断矩阵的建立过程中没有考虑人思想的模糊性，即只考虑了判断的两种极端情况。于是，本文采用传统的层次分析法结合模糊思想，引进 FAHP，即模糊层次分析法[14]。

在模糊层次分析法中，构造矩阵时使用模糊数来代替传统方法中的标度值，由于采用模糊数来构造矩阵，所以给出的判断值不再是具体的数值，而是一个区间数或者模糊数的形式，本文选用三角模糊数互补判断矩阵。三角模糊数定义如下：

记$G(R)$为R上全体模糊数，设$M \in G(R)$，若M的隶属函数$u_m: R \to [0,1]$可表示为：

$$u_m(x) = \begin{cases} \frac{x}{m-c} - \frac{l}{m-l}, x \in [l,m) \\ \frac{x}{m-r} - \frac{r}{m-r}, x \in [m,r) \\ 0, x \in (-\infty, l) \cup [r, +\infty) \end{cases} \quad (9)$$

式中l和m表示M的上界和下界值，m为M的隶属度为l的中值，三角模型数M表示为(l,m,r)。

通过三角模糊数来构造判断矩阵，即三角模糊矩阵互补判断矩阵。在判断矩阵建立的过程中，中值由相关事故的统计结合专家经验给出，而上下界的值l，r 则由专家根据判断自行给出。模糊标度见表 6。

表 6　模糊标度

标度	含义
0.1	两个元素相比，后者比前者极端重要
0.2	两个元素相比，后者比前者强烈重要
0.3	两个元素相比，后者比前者明显重要
0.4	两个元素相比，后者比前者稍微重要
0.5	两个元素相比，两者同等重要
0.6	两个元素相比，前者比后者稍微重要
0.7	两个元素相比，前者比后者明显重要
0.8	两个元素相比，前者比后者强烈重要
0.9	两个元素相比，前者比后者极端重要

模糊层次分析法赋权过程如下：

①根据 4 位专家的评判，构造三角模糊数互补判断矩阵$A = (a_{ij})_{n \times n}$（n 为行车安全指标数），其中$a_{ij} = (l_{ij}, m_{ij}, r_{ij})$；

②计算概率矩阵$B=(b_{ij})_{n\times n}$，其中$b_{ij}=\frac{l_{ij}+4m_{ij}+r_{ij}}{b}$；

③构造模糊判断矩阵$S=(s_{ij})_{n\times n}$，其中$s_{ij}=1-\frac{c_{ij}}{2(l_{ij}+m_{ij}+r_{ij})},c_{ij}=r_{ij}-l_{ij}$；

④通过计算$T=B\times S=(b_{ij}\times s_{ij})_{n\times n}$，通过$e_{ij}=\frac{1}{2}(1+t_{ij}-t_{ji})$把矩阵$T$变换为模糊互补判断矩阵$E$；

⑤模糊互补判断矩阵一致性检验

引入特征矩阵和相容性的概念，设$W=(w_1,w_2,\cdots,w_3)^T$为E的排序向量，则矩阵$W=(W_{ij})_n\times n=[a(w_i-w_j)+0.5]_n\times n$为$E$的特征矩阵，且$w_i=\frac{1}{n}-\frac{1}{2a}+\frac{1}{na}\sum_{k=i}^{n}c_ik$，其中$a\geq\frac{n-1}{2}$，而相参性指标$FC(E,S)=\sum_{i=1}^{n}\sum_{j=1}^{n}|c_{ij}-w_{ij}|$。

检验模糊互补判断矩阵的一致性，可通过矩阵E与其特征矩阵W的相容性来判断E的一致性。

作偏差矩阵$D=(d_{ij})_n\times n,d_{ij}=c_{ij}-w_{ij},i\in N,j\in N$，由于$E$与$W$对角线上元素为0，所以$d_{ij}=0$，即对$FC(W,S)$有影响的偏差仅有$n^2-n$个。则可用指标$I(E,W)=\frac{FC(E,W)}{n(n-1)}$来刻画$E$的一致性质，也即$I(E,W)$为$E$的一致性指标。取一个临界值为0.1，即当$I(E,W)<0.1$时，认为$E$具有满意的一致性。

⑥模糊互补矩阵的一致性改进

经过上述步骤的检验，当$I(E,W)>0.1$时，需要对E进行一致性改进。首先，计算偏差矩阵$D=E-W$，若$d_{ij}>0$，则适当减少c_{ij}；反之，若$d_{ij}<0$，则适当增大c_{ij}。本文调整幅度为0.2。重复过程⑤和⑥，最终得到四位专家的一致性I均小于0.1，即具有满意一致性。

⑦求指标权重

由式$H_i=\frac{1}{n}-\frac{1}{2a}+\frac{1}{na}\sum_{j=1}^{n}e_{ij}$,其中$a\geq\frac{n-1}{2},i=1,2,\cdots n$，计算各指标权重。

由于专家的经验各不相同，评价能力也参差不齐。于是，本文采用信噪比方法来检测专家的主观评价能力。同时，这里认为大部分专家对评价指标的理解是相同的，所以各专家的权重之间应该出现较高的一致性。为了对每位专家做出评

价，本文对 4 个专家的权重进行 4 次信噪比分析，即分别剔除一位专家对其余的三位专家进行信噪比分析。通过信噪比大小的比较，就能对每位专家的能力做出评价。最终得到的信噪比分析结果如表 7 所示：

表 7　信噪比分析结果

剔除的专家	专家 1	专家 2	专家 3	专家 4
信噪比	0.7045	0.8023	0.7547	0.9265

由上表可以看出，剔除专家 4 后余下的专家信噪比最高，由此可以判断专家 4 的评价能力最差。于是，为了保证权重的可信性，本文仅采用前三位专家的评价结果。

根据前三位专家的指标权重，可以画出行车安全主观评价的雷达图，见图 23。

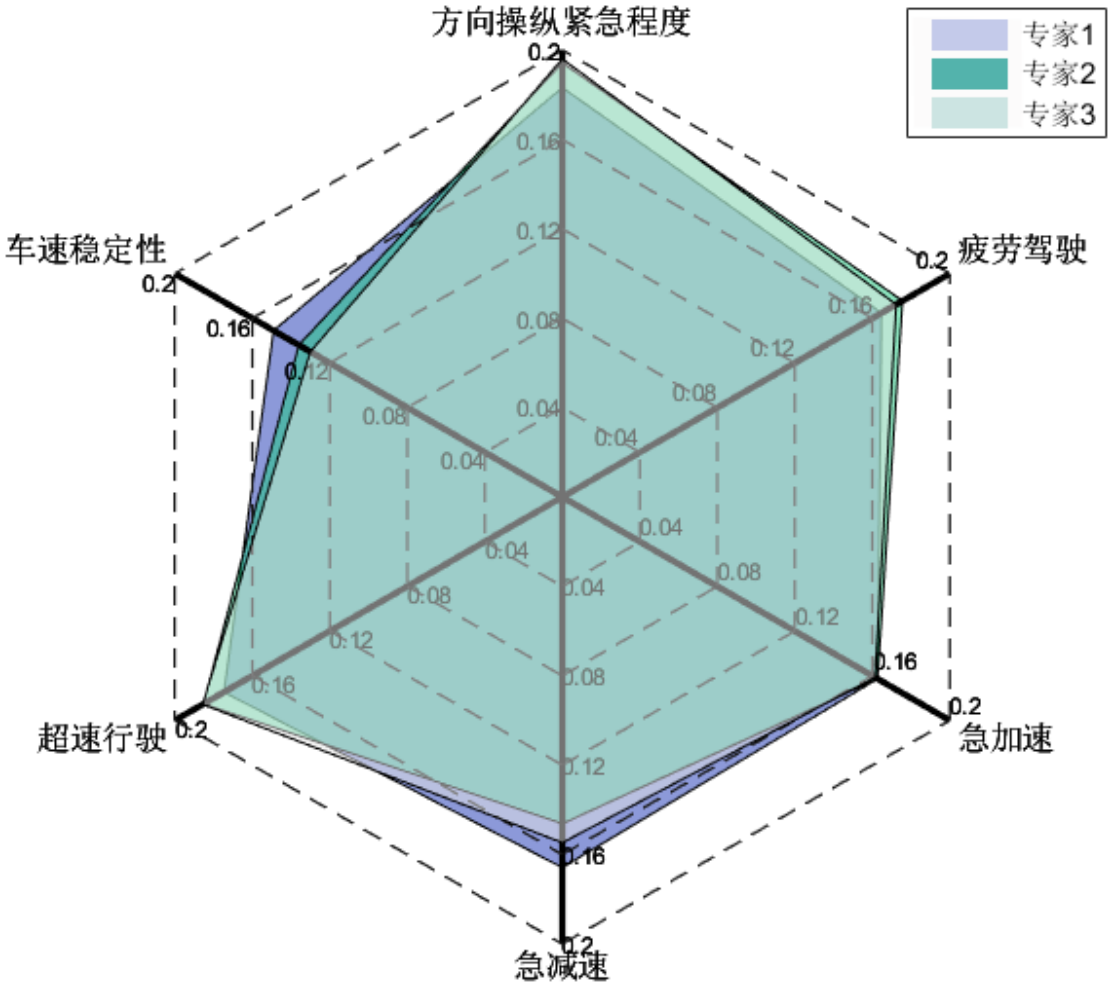

图 23　行车安全主观评价雷达图

由雷达图可以进一步验证三个专家对于关键指标的评价并无明显差异，其中方向操控紧急程度和超速行驶两个指标的权重较高，平均值分别达到了 1.911

和 0.1819。

最后，由模糊层次分析结合行车安全指标体系，得到最终的行车安全评价指标表（见表 8）。

表 8　行车安全评价表

目标层	准测层	相对目标层权重	子准则层
行车安全性	车速稳定性	(0.149,0.136,0.13)	标准差
	超速行驶	(0.1746,0.1852,0.1856)	超速累计时长/超速次数
	急减速	(0.1658,0.1462,0.1547)	急减速累计时长/急减速次数
	急加速	(0.1627,0.1627,0.1614)	危险行为一、二、三级时长/次数
	疲劳驾驶	(0.1647,0.1756,0.1723)	疲劳驾驶时长比例/次数
	方向操纵紧急程度	(0.1832,0.1943,0.196)	标准差排序序号

通过每个指标所对应的三个权重，并结合每个指标的得分算法，分别计算出每辆车在不同权重下的三个综合得分，最终，我们得到了 1347 个样本数据。

3.样本划分——留出法

在本研究中，需要一个训练样本集用来建立模型以及一个测试样本集来检验模型的评价效果。为了使样本集的划分尽可能保持数据分布的一致性，避免样本划分过程中引入额外的偏差，于是采用保留类别比例的采样方式——分层采样。首先对样本数据中的综合得分以 60 为分割点进行分层，之后通过对样本集进行分层采样而获得含 70%样本的训练集（包含 943 个样本）和含 30%样本的测试集（包含 404 个样本）。

（三）偏最小二乘回归模型

偏最小二乘回归模型在建模过程中结合主成分分析、典型相关分析和线性回归分析等方法的特点，在结果的分析中，不仅可以提供一个更为合理的回归模型，还可以完成类似主成分分析等方法的研究结果，从而提供更丰富的信息。在这里，主要想通过偏最小二乘回归模型与之后的神经网络模型得到的结果作对比。

首先以x_1,x_2,x_3,x_4,x_5,x_6分别表示自变量指标车速稳定性、超速行驶、急减速、急加速、疲劳驾驶、方向操纵紧急程度，y表示综合得分。这里以留出法得到的 943 个训练样本进行拟合，则自变量的实际数字矩阵为$A=(a_{ij})_{943\times6}$，y的数字矩阵为$B=(b_{ij})_{943\times1}$。偏最小二乘回归分析步骤如下：

①数据标准化。即将指标值a_{ij}与得分值b_{ij}转换成标准化指标值$\bar{a}_{ij}$与$\bar{b}_{ij}$；

②求出 6 个变量的简单相关系数矩阵；

③提取出自变量组的成分。这里通过 MATLAB 得到前四个成分解释自变量的比率已经超出了 90%，所以只要前四个变量即可[15]；

④求四个成分对标准化指标变量和成分变量之间的回归方程[15]；

⑤求出y与自变量之间的回归方程；

⑥检验模型。

（四）PSO-BP 神经网络模型

在与偏最小二乘回归模型相同的训练集中，把经过 R 型聚类筛选出来的 6 个指标作为 BP 神经网络的输入层节点，将最终得到的综合评分作为唯一输出。通过 MATLAB 中的 mapminmax()函数对数据进行归一化处理，同时对训练后的输出数据进行反归一化处理。之后，根据大量的实验对模型的参数进行了设定。具体的算法流程见图 24：

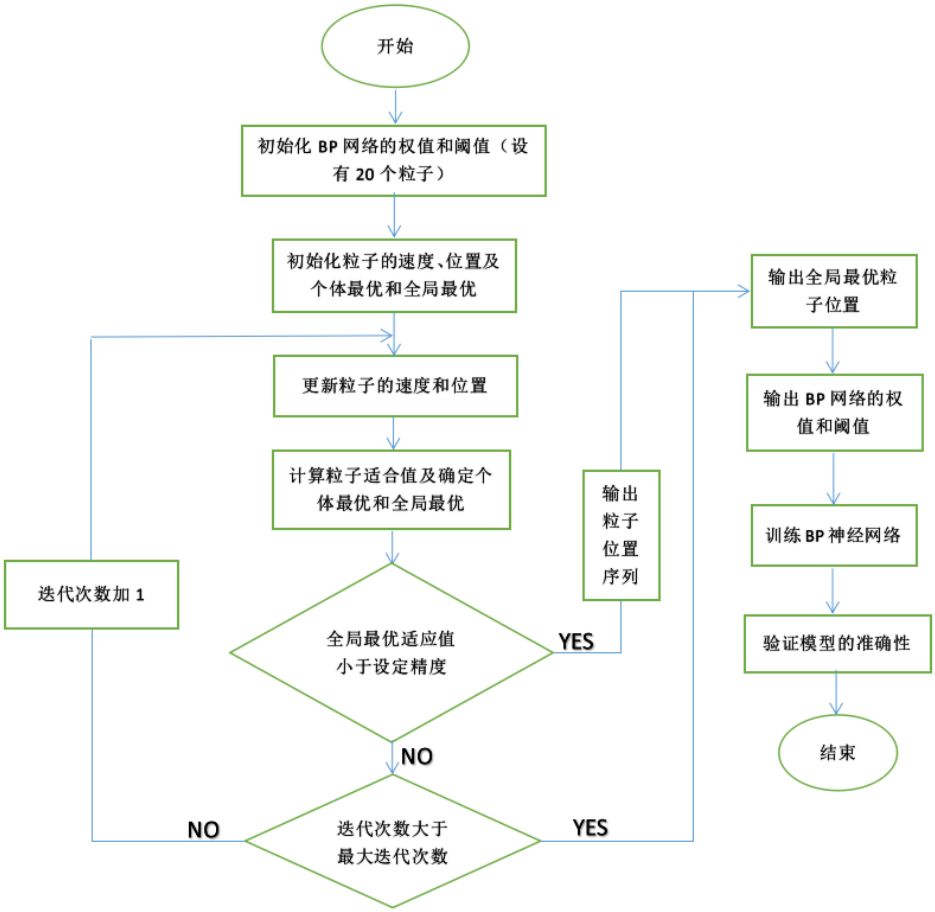

图 24　PSO-BP 神经网络算法流程图

1.BP 神经网络具体参数设定

①隐含层传递函数选择双曲正切 S 型函数（tansig）；

②输出层传递函数选择线性函数（purelin）；

③训练的终止采用早终止（Early Stopping）方法，以防止过拟合现象；

④在 BP 神经网络的构建过程中，如果隐含层含节点数太少，BP 网络不能建立复杂的映射关系，网络的预测误差会比较大。但是，如果节点数过多，可能会出现过拟合的现象，同时导致网络的学习时间增加[16]。

为求得最佳的隐含层节点数，下面考虑算法的期望泛化误差，由于问题的特殊性，将样本数据划分出三个训练集（分别包含 393 个样本）和一个测试集（包含 168 个样本）。

对于测试样本$\boldsymbol{x}$，令y_D为$\boldsymbol{x}$在数据集中的标记，y为的真实标记，$f(\boldsymbol{x};D)$为训练集D上学得模型f在$\boldsymbol{x}$上的预测输出，所以学习算法的期望预测为

$$\overline{f}(\boldsymbol{x}) = E_D[f(\boldsymbol{x};D)] \tag{10}$$

使用样本数相同的不同训练集产生的方差为

$$var(\boldsymbol{x}) = E_D[(f(\boldsymbol{x};D) - \overline{f}(\boldsymbol{x}))^2] \tag{11}$$

噪声为

$$\varepsilon^2 = E_D[(y_D - y)^2] \tag{12}$$

期望输出与真实标记的差别称为偏差，即

$$bias^2(\boldsymbol{x}) = (\overline{f}(\boldsymbol{x}) - y)^2 \tag{13}$$

其中噪声为随机误差，服从高斯分布$\varepsilon \sim N(0,\sigma^2)$，于是算法的期望泛化误差可进行如下分解（推导过程较复杂，这里仅给出结果）：

$$\begin{aligned} E(f;D) &= \mathrm{E}_D[(f(\mathbf{x};D) - y_D)^2] \\ &= \mathrm{E}_D[(f(\mathbf{x};D) - \overline{f}(\mathbf{x}))^2] + (\overline{f}(\mathbf{x}) - y)^2 + \mathrm{E}_D[(y_D - y)^2] \\ &= var(\mathbf{x}) + bias^2(\mathbf{x}) + \varepsilon^2 \end{aligned} \tag{14}$$

也就是说，泛化误差可分解为偏差、方差与噪声之和。噪声表达了当前任何学习算法所能达到的期望泛化误差的下界，跟学习算法无关，所以忽略其影响，对测试集所有样本求其偏差与方差之和，于是得到图 25：

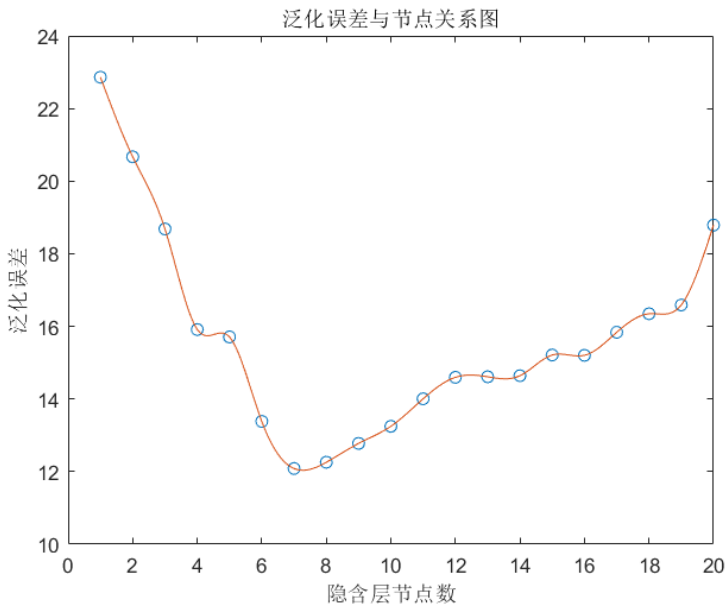

图 25　泛化误差与隐含层节点数关系图

通过上图可以看出，BP 神经网络的泛化误差随节点数的变化有一个明显的最小值（节点数=7）。于是，最终确定隐含层节点数为 7 个。得到 BP 网络的结构（见图 26），即输入节点 6 个，隐含层节点 7 个，输出层节点 1 个的 BP 网络模型。

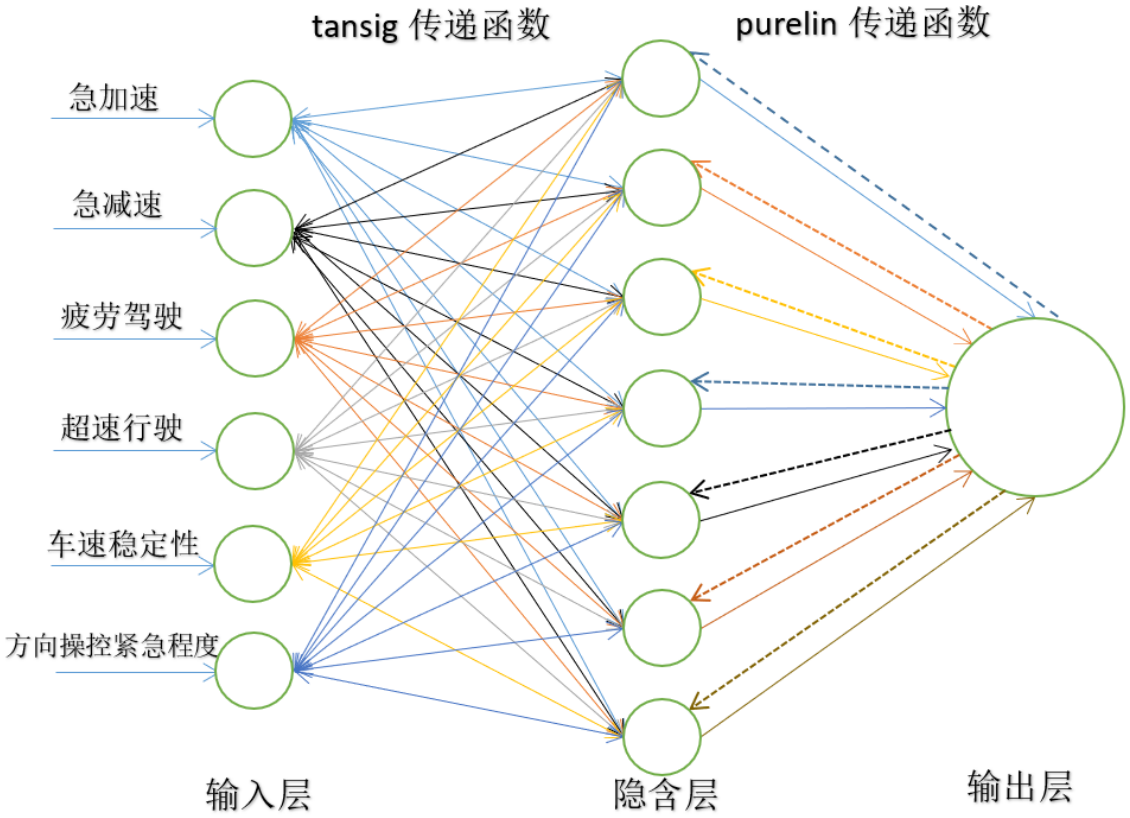

图 26　BP 神经网络模型

2. PSO 具体参数设定

①种群规模$N = 20$；

②进化次数为 120；

③学习因子$C_1 = C_2 = 1.49445$；

④粒子最大速度V_{max}，它决定了粒子在一个循环中的最大移动距离[17]，根据经验来选取V_{max}；

⑤终止条件，达到最大迭代次数或全局最优位置满足适应阈值[17]。

（五）GA-BP 神经网络模型

由于 BP 网络容易陷入局部极小，造成评价结果的偏差较大。希望通过遗传算法来得到更好的网络初始权值和阈值，从而进一步提高 BP 网络的评价精度。同时，由于需要和上述的 PSO-BP 神经网络的评价结果进行比较，所以采用和 PSO-BP 网络模型相同的训练集和 BP 参数设定。具体的算法流程（见图 27）：

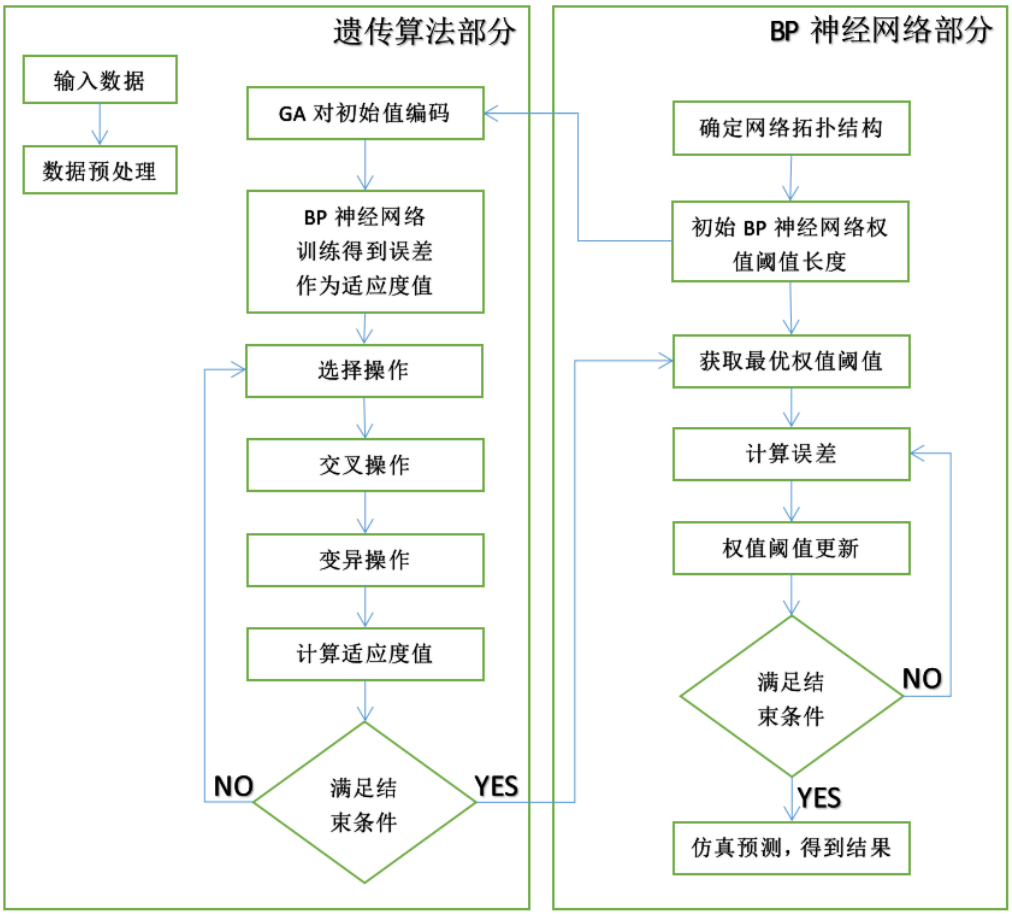

图 27　GA-BP 神经网络算法流程图

GA 具体参数设定：

①群体规模M。群体规模越大，越容易找到最优解，但是受到计算机运算能力的限制，群体规模越大，计算所需要的时间也会增加，经过多次试验，我们选取M=20。

②进化次数为 120。

③交叉率$p_c = 1$。交叉率为 1 可以保证种群的充分进化[15]。

④变异率p_m。一般而言，变异发生的可能性较小，本文取 0.2。

四、结果

（一）模型的结果

1.偏最小二乘回归模型

通过偏最小二乘回归分析最终得到综合得分y与自变量指标之间的回归方程如下：$y = 0.0912x_1 + 0.2398x_2 + 0.6674x_3 + 0.3373x_4 + 0.0592x_5 + 0.1644x_6$。

为了更加直观地表示各个自变量在解释综合得分时的边际作用，绘制了回归系数直方图（见图 28）。从回归系数图中可以看出急减速变量在解释综合得分时起到了重要的作用，同时，急加速对于得分的解释能力也比较高。

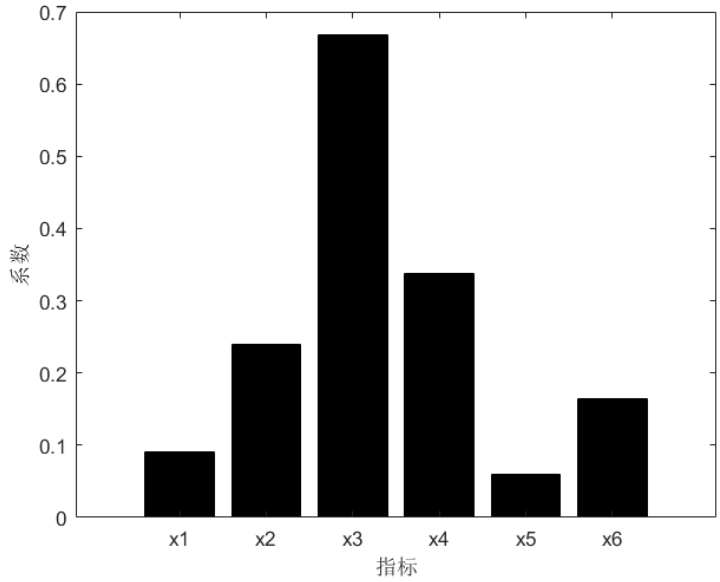

图 28　回归系数直方图

为了考察这个回归模型的精度，以$(\hat{y}_i, y_i)$为坐标值，对测试集中的样本点绘制评估图（见图 29）。其中$\hat{y}_i$是因变量在第i个样本点的预测值[15]。在这个评估

图上，如果所有点都能在该图的对角线上均匀分布，则方程的拟合值与实际值差异较小，即拟合结果是满意的。

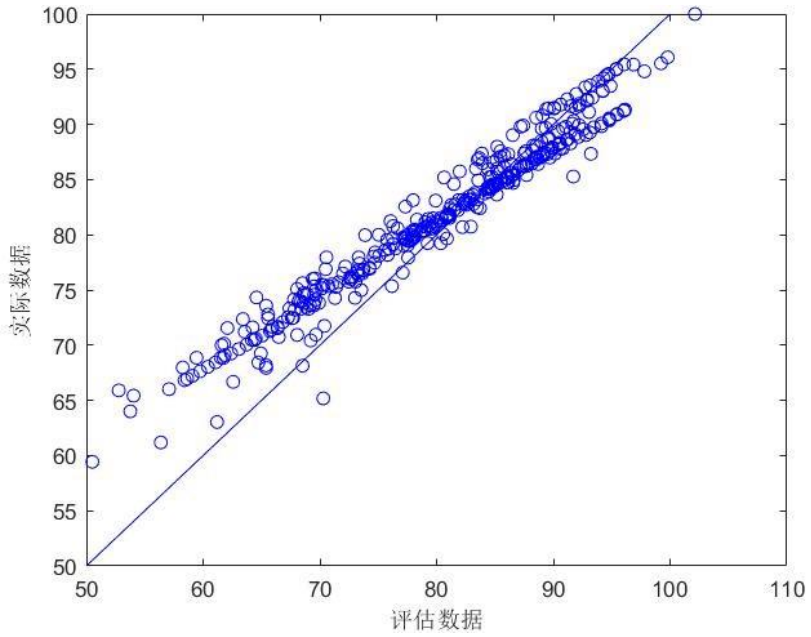

图 29　回归分析评估图

从上图可以看出样本集中的所有点近似分布在对角线周围，因此回归方程的拟合结果是可以接受的。

2. PSO-BP 神经网络模型

粒子群算法优化过程中最优个体适应度值变化过程（见图 30）：

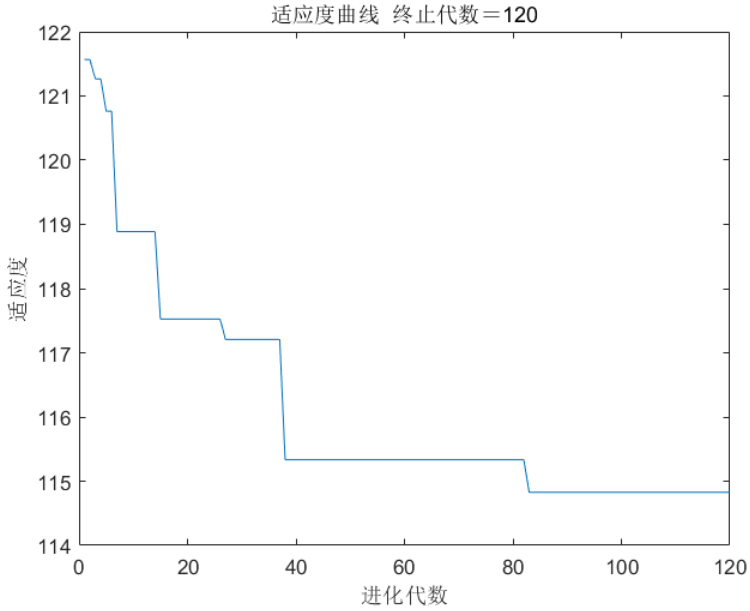

图 30　PSO 最优个体适应度值

通过粒子寻优找到了网络最佳的初始权值和阈值，把最优权值和阈值赋给神经网络。运用训练数据反复训练，最终得到了评价结果。由于测试集数据过多，

这里仅展示部分评价结果（见图 31）。从图 31 可以看出，PSO-BP 网络模型的评价结果与实际得分的误差可以控制在 2 分左右，已经符合对于评分精度的要求，因此 PSO-BP 网络模型能够较好地对汽车得分进行综合评价。

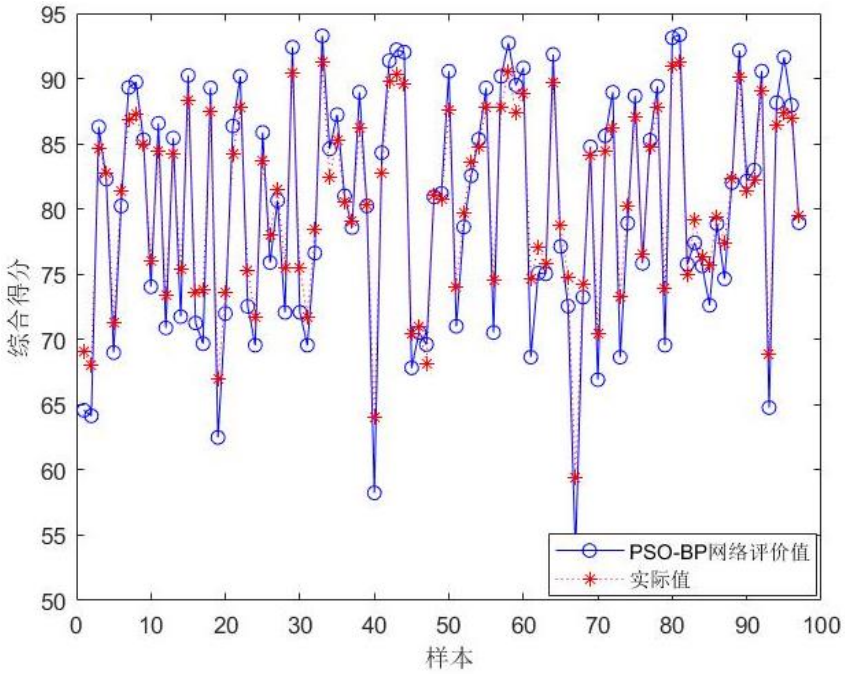

图 31　PSO-BP 神经网络模型评分值与实际值

3.GA-BP 神经网络模型

遗传算法优化过程中最优个体适应度值变化（见图 32）：

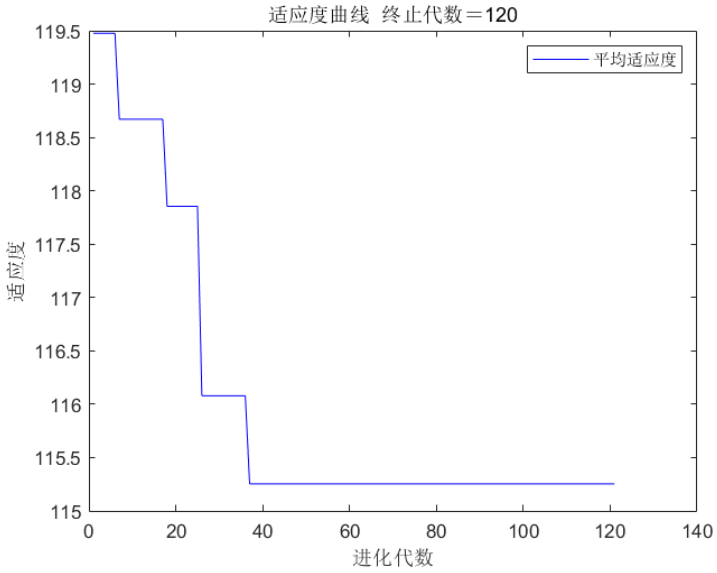

图 32　GA 最优个体适应度值

图 32 为遗传算法的优化过程，随着迭代次数的增加，平均适应度值变小并趋于稳定，最终于 120 次截止。之后把遗传算法得到的最优权值和阈值赋给 BP

神经网络，进一步训练得到评价结果，在这里也仅展示测试集的部分评价结果（见图 33）。

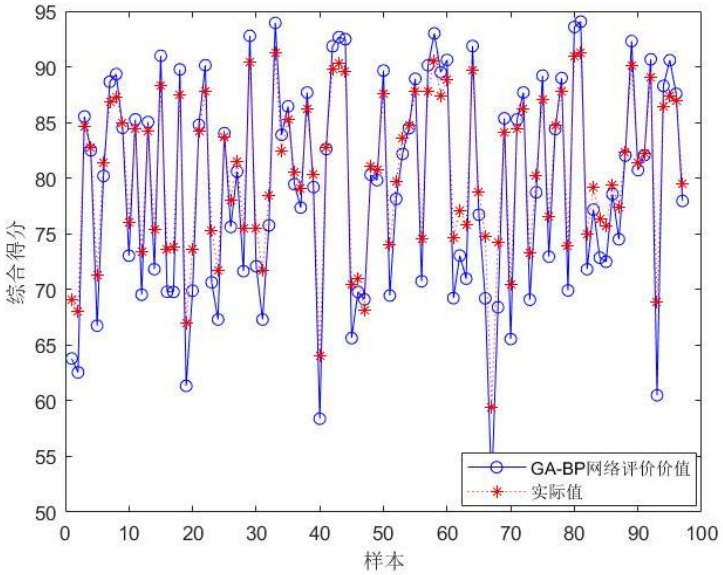

图 33　GA-BP 神经网络模型评分值与实际值

图 33 直观地反映了 GA-BP 网络的评分结果，可以看出评分结果也是满意的，但是不能直接判断出 GA-BP 网络、PSO-BP 网络以及回归模型的优劣。因此，需要对三者的评价效果做出进一步的讨论。

（二）评分效果的比较

利用上述三个模型对测试集中的数据进行评价，并对这三个模型的评价值与实际值的误差进行比较分析，为了便于观察，这里仅展示部分评价误差（见图 34）。

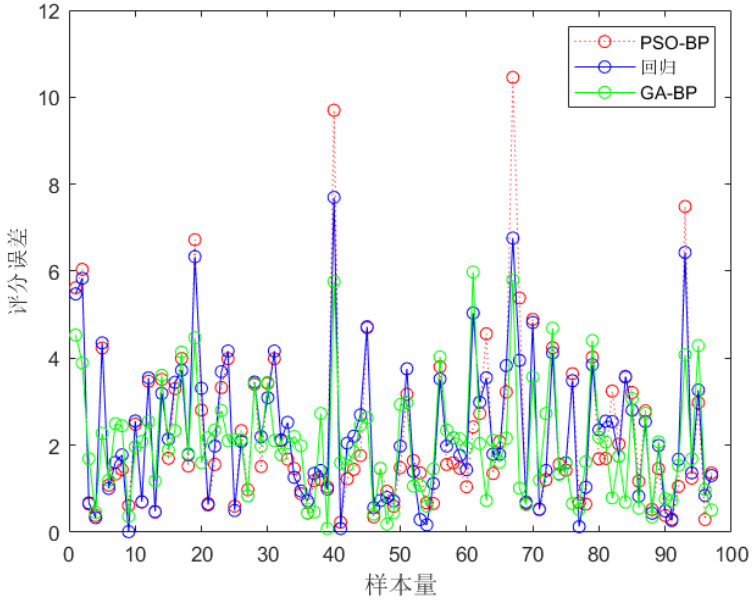

图 34　模型评价误差

由图 34 可得，GA-BP 神经网络模型和偏最小二乘回归模型的评分误差相对于 PSO-BP 神经网络的评分误差显得更加稳定，因此可以认为 GA-BP 网络和偏最小二乘回归模型的拟合效果是优于 PSO-BP 网络的。同时，可以直观地发现 GA 算法优化的 BP 神经网络的评价精度是三者中最优的。

为了更具体地说明 GA-BP 神经网络模型的评价效果，本文分别研究了三个模型的测试样本的均方根差（R），平均绝对值误差（M），平均绝对偏差百分比（P）三个评价指标。

$$R = \sqrt{\frac{\sum_i^N (y_i - x_i)^2}{N}} \tag{15}$$

$$M = \frac{\sum_i^N |y_i - x_i|}{N} \tag{16}$$

$$P = \frac{\sum_i^N |y_i - x_i|}{\sum_i^N |x_i|} \tag{17}$$

其中y_i为模型评价值，x_i为实际值。

由于测试集样本的选择具有随机性，同时机器学习算法的结果也带有一定的随机性，所以如果通过一次实验的结果就直接断定一个模型的好坏，会存在较大的偶然性。为了消除这种随机因素的影响，分别对三个模型进行 50 次实验，得到了一个 9 维随机向量$X = (Y_R, Y_M, Y_P, U_R, U_M, U_P, Z_R, Z_M, Z_P)'$，其中$Y, U, Z$分别代表回归模型，PSO-BP 模型，GA-BP 模型；R, M, P分别代表三个评价指标（例如Y_R代表回归模型的均方根差）。通过对随机向量和实际问题的分析，X近似服从多元正态分布，即$X \sim N_9(u, \sum)$，设$X_{(i)} = (y_{i1}, y_{i2}, y_{i3}, u_{i1}, u_{i2}, u_{i3}, z_{i1}, z_{i2}, z_{i3})'(i = 1 \cdots 50)$为$X$的简单随机样本，令$n = 50$，则观测数据阵为：

$$X = \begin{bmatrix} y_{11} & \cdots & z_{13} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & z_{n3} \end{bmatrix} \tag{18}$$

则u的极大似然估计为$\hat{u} = \bar{X} = \frac{1}{50}\sum_{i=1}^{50} X_{(i)}$。

最终，通过计算得到模型评分比较结果（见表 9）：

表 9　模型评分效果比较

	R	M	P
回归模型	3.9958	3.0984	0.0392
PSO-BP	5.9782	7.2978	0.0393
GA-BP	2.2384	2.0178	0.0257

由上表可以看出，GA-BP 神经网络模型的均方根差，平均绝对值误差，平均绝对偏差百分比均比 PSO-BP 模型和回归模型小。因此，GA-BP 神经网络模型的评价精度是优于 PSO-BP 模型和回归模型的，该模型具有较好的可靠性，而且可以通过以往专家的评价不断学习，不断优化模型，为行车安全提供准确的评价。

（三）安全等级的划分

通过上述的神经网络模型可以得到每一位驾驶员的综合得分，但是由于没有比较，仅仅根据这个得分无法直接判断驾驶员的行车是否规范，也无法通过综合得分直接看出一个整体的驾驶水平，通过 GA-BP 网络的评分也存在两分左右的误差，所以需要通过聚类的方法对驾驶行为进行分级。在众多聚类算法中，K-Means 算法的适用性比较强，不妨采用 K-Means 算法对综合得分进行分级。在聚类之前，先通过轮廓值对 K-Means 算法得到的聚类结果进行评价，并由此来确定最佳的类别数。

K-Means 算法的步骤如下：

①从i个数据中选择j个对象作为初始聚类中心；

②重复步骤③和④直到每个聚类不再变化为止；

③根据中心对象，计算对象到均值的距离，根据最小距离划分对象；

④重新计算中心对象，直到聚类中心不再变化为止。

通过上述的方法，可得到综合得分的平均轮廓值和各类的轮廓值分布（见图 35、图 36）。

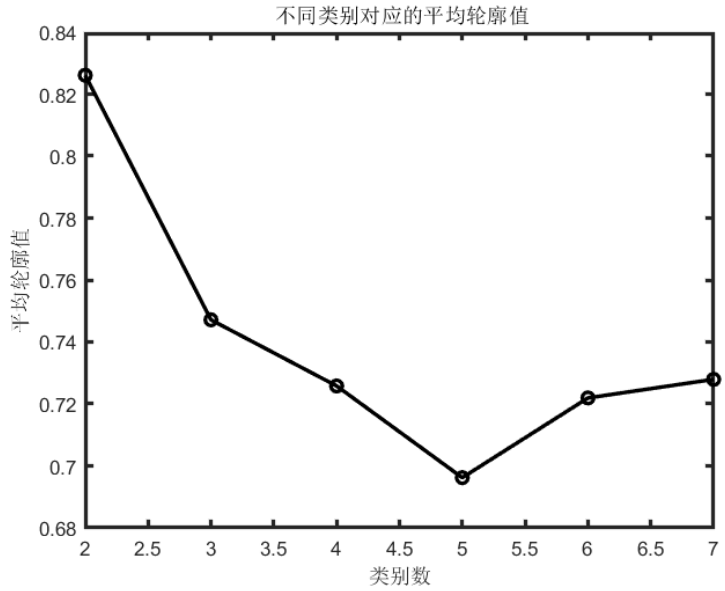

图 35　轮廓值与类别数的关系

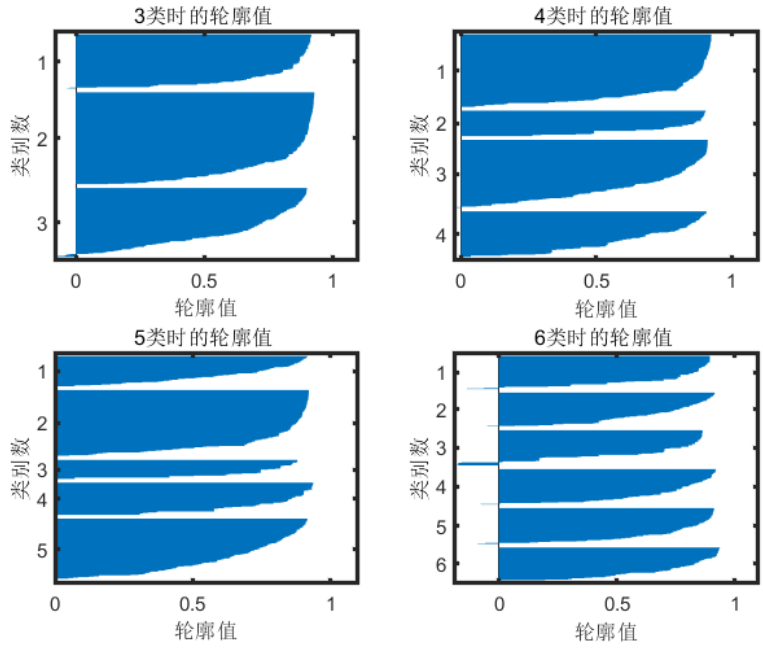

图 36　不同类别下的轮廓值分布图

对于该问题，既希望聚类的数目比较适中，又希望每个样品的轮廓值尽量高。通过图 35 我们发现类别为 2 的平均轮廓值最高，但是如果只分两类，无法很好地对驾驶行为进行区分。同时，由图 36 可以看出分 3 类时轮廓值的分布并不均匀。于是我们选择了类别数和轮廓值都较为合适的 6 类。

当类别数为 6 时，可得到 K-Means 算法的聚类结果（见图 37）。

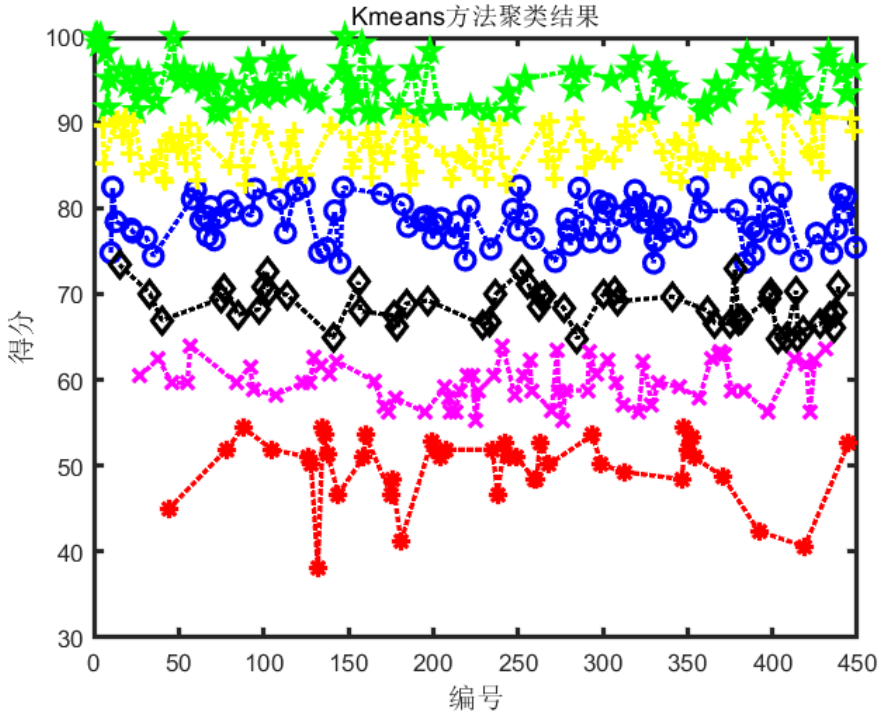

图 37　综合得分的 K-Means 算法聚类结果

通过 K-Means 算法，最终得到了六类中心点和每一类对应的样本，由于样本总数较多，不进行罗列，采用样本数量代替（见表 10）。

表 10　K-Means 算法聚类结果

	第一类	第二类	第三类	第四类	第五类	第六类
中心点	49.902	78.4274	94.6009	68.6073	86.8637	59.7483
样本数量	41	91	96	50	107	64

通过表 10 可以将驾驶行为的安全性分为 6 个等级，其中最安全的为第三类，仅占总人数的大约 21%；二、五类可以认为是比较安全的，大约占 44%；而一类、四类和六类则代表较为危险的驾驶行为，大约占 35%。通过划分的安全等级，可以对每一个驾驶员的驾驶行为进行具体的划分，并由此给出一些合理的建议。同时，该安全等级也为驾驶员的评估提供了依据。

五、主要结论及不足

（一）结论

本文通过数据挖掘技术和 R 型聚类算法，从 450 辆车的数据集中得到了这样一些结论。首先从数据集中提取出 6 个评价指标，分别为：车速稳定性，超速行驶，急减速，急加速，疲劳驾驶，方向操控紧急程度。

通过模糊层次分析法得到各个指标的权重分布，通过 GA-BP 网络、PSO-BP 网络以及偏最小二乘回归分析建立评价模型，并经过反复试验与比较，得到了一个准确评价驾驶行为的神经网络模型。通过 K-Means 算法，将最终的综合评价得分分成了 6 大类，以便于准确地进行安全等级划分。

（二）模型的不足

本文还存在很多不足之处，数据集的数据量大，而且数据来源于接收器的实际接受信号，受到了天气、地点、路况等多种因素的干扰。虽然对数据进行了深入的探索和多角度的预处理，但是仍然难以避免规约掉一些有价值的数据，可能会导致最终的评价效果不是很理想。同时，本文通过层次分析法对数据样本进行采集，作为神经网络的训练集，在采集过程中可能难以全面地表达专家的意见，所以在之后的研究中可能需要对专家的意见进一步进行划分。本文的目的是做出一个能够通过汽车行驶信息来对驾驶员进行安全评估的模型，从而能够让驾驶员准确了解自己的驾驶行为安全性，以便于及时给出恰当的建议。但是由于时间紧迫，以及数据集的复杂性，难以从数据集中完全提取出所有的有效可靠的评价指

标，而且在短时间内难以找到更多的专家和更多的数据样本，所以神经网络的学习效果受到了一定的限制，可能导致结果不是很准确。之后可以通过得到更多的指标和数据样本，由神经网络进行学习和积累评分经验值，不断完善神经网络的精度。

参考文献

[1]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].长安大学,2015.

[2]李平凡.驾驶行为表征指标及分析方法研究[D]吉林大学,2010.

[3]刘茜.关于汽车安全综合评价模型的研究[J].自动化与仪器仪表,2017(08):12-13+16.

[4]许潇潇.汽车安全性指数系统研究[D].沈阳航空工业学院,2009.

[5]舒进.四轮转向车辆运动仿真分析[J].汽车科技,200(06):6-8.

[6]余卓平,高晓杰.车辆行驶过程中的状态估计问题综述[J].机械工程学报,2009,45(05):20-33.

[7]许昶.车联网系统中云端的算法研究与车载终端的软件实现[D].电子科技大学,2017.

[8]姜杰.城市道路交通安全评价研究[D].山东科技大学,2008.

[9]赵前军.道路交通安全综合评价方法研究[D].天津大学,2011.

[10]邢如飞.乘用车操纵稳定性主观评价方法研究[D]吉林大学,2010.

[11]汪伟,赵又群,许健雄,等.汽车转向操纵动态特性评价[J].华中科技大学学报(自然科学版),2014,42(04):16-21.

[12]白海英,王秀英,张云辉.汽车侧向加速度软测量方法研究[J].汽车技术,2011(10):42-45.

[13]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].北京交通大学,2016.

[14]公彦德,曾雪兰.模糊互补判断矩阵的一致性检验及其调整方法[J].

[15]司守奎.数学建模算法与应用[M].北京:国防工业出版社,2016-1-2.

[16]MATLAB神经网络43个案例分析[M].北京:北京航空航天大学出版社,2013-8.

[17]王爱平,江丽.基于PSO的BP神经网络学习算法[J].计算机工程,2012,38(21):193-196.

附　录

1.由于数量巨大，只列出综合评分之后的数据集中的20辆车（见表10）。

表 11　汽车的综合评分

车辆	车速稳定分	超速行驶分	急减速分	急加速分	疲劳驾驶分	方向紧急程度打分	综合得分
AD00112	100	100	100	100	100	100	100
AF00105	100	100	100	100	100	100	100
AM00176	100	100	100	97.5	100	99.6	99.1
AF00094	100	100	100	100	90	100	100
AD00099	100	100	76.5	92.5	100	100	89.6
AD00291	80	100	77.5	87.5	100	89.5	85.2
AD00443	100	95	97	97.5	98	100	98.1
AA00036	80	100	86.5	97.5	100	100	91.6
AD00247	80	67.5	94	100	100	100	95.0
AD00112	100	82.5	100	100	100	100	100
…	…	…	…	…	…	…	…
AF00199	80	100	87.5	67.5	100	60	81.6
AF00348	100	100	97	90	100	60	95.5
AF00042	80	86.7	57.5	90	100	60	79.5
AF00442	80	100	71	82.5	100	60	81.3
AD00331	100	100	85	95	100	60	93.3
AD00117	80	56	70.5	0	100	60	52.6
AD00450	80	100	93	87.5	100	66.1	90.3
AD00248	100	100	94	95	82.7	60	96.2
AD00297	100	100	77	90	100	60	88.9
AF00342	80	100	53	82.5	100	60	75.4

2.第七届“泰迪杯”数据挖掘挑战赛——C题：运输车辆安全驾驶行为的分析

问题背景：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利用车联网等系统数据，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。

某运输企业所辖各车辆均存在常规运输路线与驾驶人员。在驾驶员每次运输过程中，车辆均可自动采集当前驾驶行为下的行车状态信息并上传至车联网系统。驾驶行为可能随气象、路况等因素的变化而变化，进一步影响行车安全、运输效率与节能水平请根据该运输企业所采集的数据，分析车辆行驶过程中的驾驶行为对行车安全、运输效率与节能情况的影响，运用数据挖掘的方法，建立有效的数学模型进行评价。

致　谢

本文从选题、立论到撰写的整个过程，得到了指导教师组以及相关专业教授的悉心指导，并为该文参考资料的查阅提供了诸多方便。在此，对各位老师及提供过帮助的专家表示最真挚的感谢和最崇高的敬意！

最后，对参加本论文评阅、答辩和对本论文提出宝贵意见的所有老师和同学表示诚挚的谢意！