杭州电子科技大学

硕 士 学 位 论 文

题目：基于驾驶风格的重型货车油耗预测模型研究

研 究 生 林若奇

专　　业 物流工程与管理

指导教师 曾鸣 副教授

完成日期 2022 年 10 月

杭州电子科技大学硕士学位论文

基于驾驶风格的重型货车油耗预测模型研究

研 究 生：林若奇

指导教师：曾鸣 副教授

2022 年 10 月

Dissertation Submitted To Hangzhou Dianzi University for

The Degree of Master

Research on Fuel Consumption Prediction Model of Heavy Truck Based on Driving Style

Candidate：Ruoqi Lin

Supervisor：A.P. Zeng Ming

October, 2022

摘要

重型货车的驱动离不开燃油。随着国内物流行业迅速发展，重型货车的保有量不断增加，因此对于燃油的进口需求也不断提高。但国际燃油价格不断上涨，造成物流企业的运输成本持续上升，因此如何提升重型货车的燃油经济性，成为了众物流企业亟需解决的一个问题。燃油经济性与道路条件、车辆设备和驾驶行为等因素有关。相较于道路条件、车辆设备等客观因素，驾驶员可以对驾驶行为进行更好地调控。通过对行驶数据挖掘可提取不同的驾驶行为，驾驶行为可以侧面体现驾驶员的驾驶风格，基于不同驾驶风格的油耗分析将为驾驶员的决策提供参考。鉴于此，本文的工作内容如下：

首先，对原始数据进行预处理操作。通过数据合并、去除重复值、GPS 异常值筛选以及数据标准化的方法对原始数据进行预处理，并基于预处理的数据提取四大类指标。随后，通过主成分分析法将复杂的多维数据降为数据量较少的低维数据，从 28 个指标中提取出 6 个主成分，有效降低了后续实验所需要的时间和计算量。

其次，在预处理的数据的基础上，提出一种基于聚类和支持向量机的驾驶风格分类识别模型。通过提取的 6 个主成分聚类获得三个簇，分析三个簇有关的驾驶行为数据，打上三类驾驶风格标签，分别是猛踩油门型、高速行驶型和频繁变速型；在此基础上构建了一个三分类的支持向量机分类模型，该分类模型可以为后续新加入的数据进行驾驶风格识别，经过反复验证，该模型的准确率为 96.5%。

最后，在驾驶风格分类的基础上提出一种油耗预测模型。以不同驾驶风格为基础，分别基于随机森林 RF 算法和极端梯度增强 XGBoost 算法，构建了两个油耗预测模型，将不同驾驶风格的数据集分别输入到两个模型，拟合获得油耗的预测值，通过回归评价指标计算每个算法的预测准确率。通过比较，XGBoost 算法准确率比 RF 算法高 5%，因此选择了性能更好的 XGBoost 算法用于实验仿真。考虑到 XGBoost 算法符合黑箱理论特征，因此选用了 Shapley value 算法来分析特征变量对于油耗预测值的影响机制，进而提出节能驾驶策略建议和管理启示。

关键词：数据挖掘，驾驶风格，集成学习，油耗预测，Shapley value

ABSTRACT

The drive of heavy trucks is inseparable from fuel. With the rapid development of the domestic logistics industry, the number of heavy trucks is increasing, so the demand for fuel imports is also increasing. However, the rising international fuel prices have caused the transportation costs of logistics enterprises to continue to rise. Therefore, how to improve the fuel economy of heavy trucks has become an urgent problem for logistics enterprises. Fuel economy is related to road conditions, vehicle equipment and driving behavior. Compared with objective factors such as road conditions and vehicle equipment, drivers can better regulate driving behavior. Different driving behaviors can be extracted by driving data mining. Driving behavior can reflect the driver 's driving style. Fuel consumption analysis based on different driving styles will provide reference for driver 's decision-making. In view of this, the work of this article is as follows :

Firstly, the original data is preprocessed. The original data is preprocessed by data merging, removing duplicate values, GPS outlier screening and data standardization, and four categories of indicators are extracted based on the preprocessed data. Subsequently, the complex multi-dimensional data was reduced to low-dimensional data with less data volume by principal component analysis, and 6 principal components were extracted from 28 indicators, which effectively reduced the time and calculation required for subsequent experiments.

Secondly, based on the preprocessed data, a driving style classification and recognition model based on clustering and support vector machine is proposed. Three clusters are obtained by extracting six principal component clusters, and the driving behavior data related to the three clusters are analyzed. Three types of driving style labels are marked, namely, slam on the accelerator, high-speed driving and frequent variable speed ; on this basis, a three-class support vector machine classification model is constructed. The classification model can identify the driving style for the newly added data. After repeated verification, the accuracy of the model is 96.5%.

Finally, a fuel consumption prediction model is proposed based on driving style classification. Based on different driving styles, two fuel consumption prediction models are constructed based on random forest RF algorithm and extreme gradient

enhancement XGBoost algorithm respectively. The data sets of different driving styles are input into the two models respectively, and the predicted values of fuel consumption are obtained by fitting. The prediction accuracy of each algorithm is calculated by regression evaluation index. By comparison, the accuracy of XGBoost algorithm is 5 % higher than that of RF algorithm, so XGBoost algorithm with better performance is selected for experimental simulation. Considering that the XGBoost algorithm conforms to the characteristics of the black box theory, the Shapley value algorithm is selected to analyze the influence mechanism of the characteristic variables on the fuel consumption prediction value, and then the energy-saving driving strategy suggestions and management enlightenment are put forward.

Keywords: Data Mining, Driving Styles, Ensemble Learning, Fuel Consumption Prediction, Shapley Value

目　录

1 引言

1.1 研究背景及意义

近几年，交通领域的能源消耗问题逐步成为影响国家综合实力和居民生活水品的重要因素。据统计，我国 2020 年燃油进口量为 5.4 亿吨，较 2019 年的 5.06 亿吨同比增长了 7.3%，创造历史新高。同年我国机动车燃油消耗量为 6.5 亿吨，其中，柴油重型货车因其大排量的发动机结构、较长的运输距离以及高载重的运输量，其油耗普遍高于其他类型机动车，导致许多物流企业的运输成本居高不下。

由于国际燃油价格不断上涨，政府和企业管理日益聚焦于提高重型货车燃油经济性。政府和企业管理人员提出了各项行政法规和管理措施来约束驾驶员的驾驶行为，如积极开发新能源汽车、限制路段车速等方法。因此，针对燃油消耗的影响因素以及油耗预测模型的研究，将有助于优化管理者的管理方法。

现有研究表明燃油消耗量是一个受多维度影响的概念，最早被研究且最易于理解的就是与车辆特性相关的机械参数，如发动机功率、扭矩、后桥速比、发动机转速、进气压力等，这些参数的初始规格通常由汽车出厂时设置而成，后续不易受外界环境影响，且和油耗有着密切联系，随着后续研究，越来越多的专家学者开始探索性的考虑不同影响因子对于油耗的影响，比如环境因素中的天气、温度、湿度等[1][2]，还有部分专家提出不同道路类型或者交通状况对于燃油消耗量的影响[2][3]，同时，有些文献认为驾驶员是车辆行驶的主要操作者，因此有必要探究驾驶员行驶的驾驶行为和驾驶风格对于油耗的影响[4][5]。

考虑到车辆机械参数与油耗关系的研究需要一定的汽车动力学和物理知识，构建的模型结构往往比较复杂，输入参数较多，不易于复杂环境的推广，且此类研究出现较早，现有结论比较丰富，因此可以考虑探究其他影响因子对于油耗的作用，现有文献中，不少专家学者在考虑油耗问题时，会引入驾驶行为的数据，比如：行驶速度、行驶加速度等[1]，但在实际的驾驶过程中，一些驾驶员会受自身驾驶风格的影响，做出一些激进的驾驶行为，如猛踩油门超车，或者因为当天心情等主观因素的影响，做出一些不同往常的驾驶行为，因此在考虑驾驶行为对于油耗的影响前，需要识别不同的驾驶风格类别，因此本文将基于驾驶行为和驾驶风格对油耗的影响进行研究，为车队管理者在做出优化策略时提供一些理论依据。

近几年，随着数据的爆炸增长，越来越多的专家学者开始思考传统油耗模型是否适合复杂多变的环境，因此，基于机器学习的油耗模型开始进入大家的视野，

但同时便产生了新问题，即模型的可解释性，机器学习中大部分算法模型都属于黑箱模型，其本质的运作机制无法被观测，只有少部分较为简单的机器学习模型如逻辑斯蒂模型、决策树模型等可以通过输入和输出的结果推测其运作机制。为了解决此类问题，一些专家学者进行了探索性的研究，如通过修改输入变量的大小值，从而分析出对于燃油消耗量影响最大的因素[6]，这样做可以一定程度的提高模型的透明度，但实际上却改变了原始数据的大小，缺乏真实性。

综上所述，本文在识别驾驶风格的前提下，探讨驾驶行为对油耗的影响。具体做法是通过数据挖掘手段构建基于 PCA 和聚类分析的驾驶风格分类识别模型，从而识别不同行程段的驾驶风格，再通过集成学习算法构建油耗预测模型，并根据 Shapley value 讨论驾驶行为对于油耗的影响，为重型货车驾驶员提供驾驶行为意见。

1.2 国内外研究综述

对于提高燃油经济性的问题，国外学者相较于国内学者的起步较早，相关研究内容比较充实，尤其是在传统汽车动力学方面的油耗模型，而国内学者则倾向于大数据背景下的油耗分析，利用多维度的影响因子构建油耗预测模型。本文基于现有研究，将文献综述分为两部分，分别是车辆燃油影响因素相关文献和车辆燃油消耗模型相关文献。

1.2.1 车辆燃油影响因素

许多变量会影响车辆燃油的消耗量，这些变量大致可以分为五大类：天气因素、车辆因素、道路因素、交通因素和驾驶员相关因素：

（1）天气因素

天气相关因素包括环境温度、湿度和风速的影响。这些因素通常会间接影响车辆的燃油消耗量，冯闪等[7]发现冬天温度低时，燃油消耗量显著提高，其次车辆在逆风行驶时，风阻变大，此时的油耗也高于正常行驶时的油耗；Shang R 等[2]发现与晴朗天气条件相比，下雨天气条件下的燃油消耗和排放往往较低。

（2）车辆因素

车辆相关因素通常是指发动机和整车结构。发动机是影响燃油经济性的关键因素，发动机的尺寸、功率、使用的燃料类型直接决定了发动机的燃油消耗[8]。Benajes J 等[9]比较了两种燃油模式下的燃油消耗效率，发现柴油-汽油模式的效率会高于柴油模式，具有一定的节油潜力。张登等[10]从车型、发动机、变速箱和轮胎等方面出发构建了指标，通过这些指标建立油耗预测模型，结果表明这些指标有利于提高模型的准确率。

（3）道路因素

道路相关因素主要指道路的物理特性，如道路坡度、道路类型等。研究表明平坦路线的燃油经济型比丘陵路线高 5-20%。张贤彪[11]认为不同道路其路况情况不同，一些道路是新修的道路，车道宽敞且平坦，在这些路段上行驶时少了一些加速或者减速的行为，油耗自然较低，一些道路比较老旧，坑坑洼洼多，开起来自然费油。郑天雷等[12]在探究重型货车油耗影响影响因素的实验时发现，相同的参数变化在市区、公路和高速公路上的燃油消耗存在显著差异。Kamal 等[13]开发了一种基于模型控制的生态驾驶系统，以使车辆在上坡和下坡道路上实现生态驾驶，他们在虚拟道路上的实验表明，与对照组车辆相比，生态驾驶车辆的燃油节省率在 5%至 7.04%之间。Jianqiang W 等[14]使用全局和局部方法优化了上坡和下坡的燃油消耗，仿真获得了生态速度曲线，使得燃油减少了 5.5%。Faria M V 等[15]利用真实世界驾驶数据库的数据，探究道路坡度和道路类型对于燃油消耗的影响，实验结果表明，燃油消耗率随着道路坡度的增加而增加，最多为 3 倍。

（4）交通因素

交通相关因素通常指的是道路拥挤程度。Xu J 等[3]探究了道路交通状态对于燃油效率的影响，结果表明，在纳入交通状况后，生态驾驶的评分系统得到了改善，因为出行在拥堵条件下时需要经常性的走走停停，获得较高生态评分的可能性较小，驾驶员此时不一定会因为在拥堵条件下加速或者减速多而获得较低的生态评分。Wu Y 等[16]在探究行驶速度和道路拥堵情况对于燃油消耗的影响时发现，在低速拥挤的交通情况下，燃油消耗变得比较高，比正常路况高 18%-28%。

（5）驾驶员因素

驾驶员相关因素主要指驾驶员的驾驶行为，通常通过速度和加速度曲线来识别。与没有经验的驾驶员相比，有经验的驾驶员可以通过巧妙地调节车速来避免在交通信号灯处等车以及在其他区域的急加速和减速，从而节省燃料。Barth M 等[17]构建了一个动态驾驶辅助系统，实时地向驾驶员提供驾驶意见，在不显著增加行驶时间的情况下，可节省约 10%-20%的燃油消耗。Carrese S 等[18]通过一系列试验后得到了以下结论：合理的驾驶行为最高可以节省 27%的燃油消耗量。Zheng F 等[19]研究发现不同驾驶员类型之间的燃油消耗存在显著差异，谨慎的新手司机油耗往往较低，经验丰富的驾驶员更勇于做出一些激进的驾驶行为，如急加速，因此具有较高的燃油消耗。Díaz-Ramirez J 等[20]通过比较驾驶员生态驾驶训练前后的燃油消耗量，发现其燃油消耗降低了 6.8%，出行的加速、制动和超速等驾驶行为次数减少了 96%。Berry I M 等[21]通过实验发现，不同驾驶风格的驾驶员的节油策略具有差异性，激进的驾驶员应转专注于降低加速度，而较为平稳的驾驶员则应专注于以较低的速度行驶，处于两者之间的驾驶员以较低的加速度行驶，则可实现最大的燃油节省。

综上，尽管现有文献认为驾驶风格对于燃油的影响十分显著，但其影响往往会被忽略[22]。大部分文献仅基于外部因素（天气和道路交通情况等）或仅基于车辆机械特性构建燃油消耗模型，没有考虑驾驶员的个人特性，因此在后续的研究中有必要探究驾驶风格对于油耗的影响。

1.2.2 车辆燃油消耗模型

现有燃油消耗模型有许多分类依据，根据其输入数据类别可以分为基于物理数据驱动的油耗模型和基于行驶数据驱动的油耗模型[23]，根据输入数据维度可将油耗模型分为微观、中观和宏观模型[24]，Zhou 等[25]提出了另一种燃油消耗模型分类依据，该分类基于模型对大量数据的依赖程度以及对估算燃油消耗时机械细节的理解，根据计算的透明度，有三种类型的模型：白箱、灰箱和黑箱模型。白箱模型主要来源于相关理论知识，要求深入了解发动机和子系统的运作原理。与白箱模型相比，黑箱模型在其结构中缺乏数学或者物理的理论背景，在黑箱模型中，发动机本身被认为是一个黑箱，其内部发生的过程除了可测量的数据外，其他都不知道。灰箱模型介于两种方法之间，即基于发动机的基础理论以及实验收集的数据进行分析。

（1）白箱模型

燃油消耗白箱模型是基于车辆发动机中发生的物流和化学过程，使用数学方程描述燃料吸入、压缩、燃烧和蒸发过程。Cachón L 等[26]通过碳平衡法与车辆纵向动力学模型相结合，根据实际测量结果预测压缩天然气车辆的燃油消耗量。Heywood J B[27]开发了一个基于内燃知识的平均值现象学模型，以预测消耗的燃油量、产生的废气和扭矩。该模型由四个主要子系统组成：进气歧管系统、燃料输送系统、扭矩产生系统和排气系统。

总的来说，开发白箱模型需要了解整个发动机系统及其子系统，在燃油消耗白箱模型中需要确定的参数数量通常很多，在某些情况下，由于燃料系统的复杂性，此类模型将过于复杂，建模效率低。

（2）黑箱模型

燃油消耗黑箱模型通常基于统计和机器学习方法，使用真实收集的实验数据进行训练[28]。它们的主要缺点是开发过程需要大量数据，且在结果的可解性方面存在不足。基于此问题，Walnum H J 等[4]将一组驾驶指标作为解释变量，构建多元回归模型进行油耗预测，并采用平均弹性分析获取对油耗影响较大的变量。Chen 等[29]提出了一种油耗模型，采取多元自适应回归样条方法，通过改进传统多元线性回归模型，使得准确率高且可解释性更强。这些模型一定程度上增加了燃油消耗黑箱模型的透明度，但其本质仍与白箱模型不同，不能描述具体的物理或者化学含义。

在真实的行驶过程中，油耗受多个因素的影响，例如天气因素、车辆因素、驾驶员因素、道路因素以及交通因素等，具有非线性，结构简单的回归模型无法满足实际需要，随着大数据时代的到来以及计算机算力的提升，越来越多的专家学者开始研究深度学习算法。深度学习有很强的拟合能力，可以从复杂的数据集中学习有用的信息[30]。其中运用最多的深度学习模型就是各类神经网络模型。Wysocki O 等[31]通过发动机转速和扭矩数据，分别利用多项式回归、K 近邻和人工神经网络构建了油耗预测模型，结果表明，人工神经网络的模型准确率远高于另外两种模型。Xu Z 等[32]利用广义回归神经网络研究司机驾驶行为和燃油消耗的关系，所提出来的模型在预测燃油消耗方面具有更强的性能。Kanarachos S 等[33]提出了一种基于递归神经网络的瞬时油耗预测模型，结果证实了其优越的性能，可以用于车辆燃料消耗的大规模精确检测。

除开深度学习算法以外，集成学习算法也属于黑箱模型。集成算法是指将多个弱学习器集合成一个强学习器，这些弱学习器可以是决策树、支持向量机等，在油耗预测中就是对回归问题进行集成。集成学习又分为 Bagging 与 Boosting。Bagging 是一种并行的集成学习方法，各弱学习器之间同时运行，以随机森林为代表。Yao Y[34]等利用多个机器学习模型构建了油耗预测模型，其中包括支持向量机、BP 神经网络和随机森林，结果表明随机森林具有最高的预测准确率和最快的运行速度，同时可根据损失函数评估影响油耗的重要指标。Boosting 则是一种串行的集成学习方法，以梯度提升和极端提升为代表。Bousonville T 等[35]采用总重量、平均速度、地形、天气条件等指标，利用三种机器学习算法进行油耗预测，结果表明梯度提升算法是研究货车油耗的最佳算法，且除了总重量和地形等特征显著影响油耗外，天气条件数据也有助于改进预测结果。Wang Q 等[36]基于多维特征和 XGBoost 算法，提出了矿用货车燃料消耗的预测模型，并确定了 R^2 和平均绝对百分比误差分别是 0.93 和 8.78%。上述算法，只有少部分可分析输入变量对于油耗的影响，大部分算法只是通过输入变量拟合出油耗预测值，本质上属于黑箱模型，缺乏算法可解释性。

（3）灰箱模型

燃油消耗灰箱模型可以视为燃油消耗白箱和黑箱模型的组合，具体而言，与白箱模型相比，灰箱模型不需要完全了解发动机的工作原理，这使得模型更易于开发，灰箱模型具有一定的物理意义，而不是由其输入和输出决定的纯数学模型，其模型结构比白箱模型简单，但比黑箱模型复杂。Pelkmans L 等[37]开发了一种称为 VeTESS 的模拟工具，用于模拟真实交通瞬态车辆运行的燃油消耗和排放，它在给定速度曲线上，通过发动机转速、发动机扭矩和扭矩变化，逐秒计算特定车辆的燃油消耗量。

综合上述分析，白箱模型和灰箱模型都需要一定物理知识，针对驾驶行为数据的研究不易于展开，而基于统计和机器学习的黑箱模型不要求了解发动机内部的运作机制，建模成本低，模型准确率高，但对于数据的质量和数量有一定的要求，且模型解释性弱，因此，可在黑箱模型的基础上探究提高模型透明度的方法。

1.3 研究内容

目前国内外对于车辆油耗分析的研究大部分从天气因素、道路因素、交通因素、车辆因素等角度入手，但并未探究不同驾驶风格对于油耗的影响，以及驾驶行为与油耗之间的相关性。基于此本文提出以下研究内容：

（1）因为不同车载设备具有不同的采集频率，且本文收集的数据存在缺失或者异常的情况，因此在实验开始前需要对原始数据进行预处理操作，消除误差数据。

（2）考虑到构建的驾驶行为指标数量较多，直接进行信息提取时间复杂度较高，因此需要利用降维的方式，将高维数据替换为低维数据，用较少的数据量反映大致的数据信息，这里具体的做法就是采用 PCA 算法，接着采用 PSO-Kmeans 的算法对降维后的数据进行聚类，获取数据的驾驶风格类别，最后为后续新加入的数据构建了一个驾驶风格识别模型，模型基础采用 SVM 支持向量机算法。

（3）在驾驶风格分类后，通过集成学习算法拟合获取预测油耗值，并通过 Shapley value 对影响油耗的驾驶行为进行分析。

1.4 论文结构

论文在总结国内外现有文献的基础上，考虑现有研究的不足之处进行了相关的实验，具体结构如下所示：

第一章：引言。介绍本文的研究背景和意义，并通过两个方面总结国内外文献，敲定了本文的研究方向，即考虑驾驶风格对于重型货车油耗的影响，同时分析驾驶行为对于油耗预测值的影响。

第二章：相关理论概述。对驾驶风格的概念进行界定，明确了它和驾驶模式、驾驶行为、驾驶条件、驾驶事件和驾驶技能之间的联系和区别；接着介绍本文后续构建指标需要的数据；最后对本文所用到的所有理论进行了简单概述。

第三章：数据预处理与指标构建。本章首先介绍了数据来源和数据类型，后续对数据进行了预处理；接着，考虑到原始数据记录了驾驶员一天所有的行车数据，因此还需对原始数据进行切分；最后，介绍了所有驾驶行为指标的简介和具体计算方式。

第四章：基于车联网数据的驾驶风格分类识别。本章的主要工作就是界定不

同数据集的驾驶风格，这里需要结合驾驶行为指标进行讨论，首先需要通过 PCA 主成分分析对数据进行降维；再通过 PSO-Kmeans 对降维后的数据进行聚类，获得三类不同的驾驶风格；最后基于支持向量机算法构建驾驶风格识别模型。

第五章：基于驾驶风格的油耗预测。本章运用随机森林 RF 和极端梯度提升算法 XGBoost 算法进行油耗预测，期间用回归模型评价指标对两个模型性能进行比较，最后通过 Shapley value 算法进行驾驶行为分析，讨论不同驾驶行为对于油耗的影响，从而提出节能驾驶策略。

第六章：总结和展望。总结了本文主要的研究内容和重要的实验结果，并提出了研究过程中存在的不足，针对不足之处罗列了几点可在未来改进的地方。

1.5 论文研究技术路线

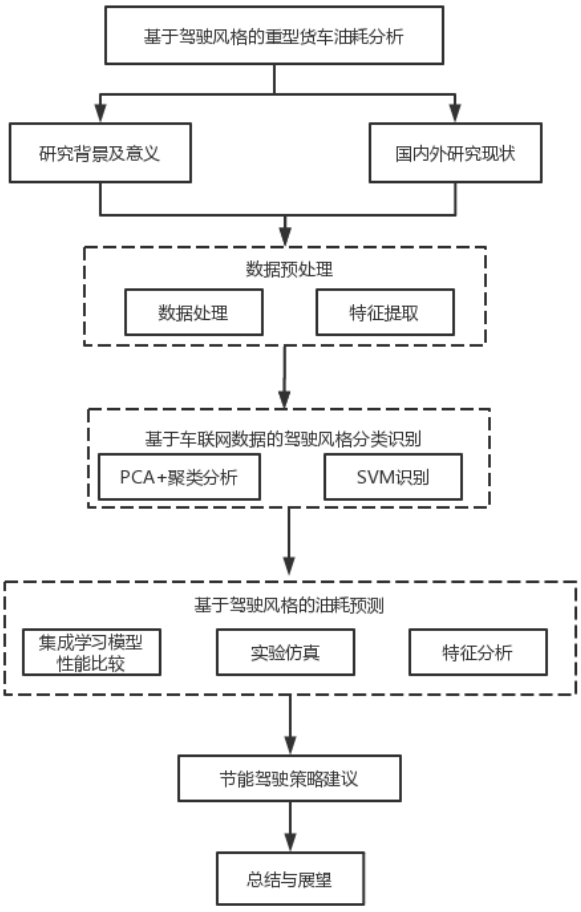

图 1.1 论文技术路线图

1.6 本章小结

本章介绍了本文的研究背景和现实意义，分别从影响因素和油耗模型出发总结现有文献研究，并针对现有文献存在的不足提出了本文重点研究的内容，即基

于不同驾驶风格，探究不同驾驶行为对于车辆油耗的影响，接着简单总结了下本文具体的研究内容，最后通过论文结构安排和技术路线图大致介绍了本文的整体框架安排。

2 相关理论概述

本章主要从三个角度出发，对论文涉及的理论进行概述，分别是驾驶风格概念界定和与其相关概念介绍、探究驾驶风格的相关参数以及相关技术概述。驾驶风格界定主要阐述了其跟驾驶技能、驾驶行为、驾驶条件、驾驶事件以及驾驶模式之间的联系和区别，主要需要知道不同驾驶风格会产生不同驾驶行为，同时驾驶模式中的速度加速度分析图可反映具体的驾驶行为，因此对分析图的研究可获取具体的驾驶行为和驾驶风格；探究驾驶风格的相关参数主要阐明了本文所选取的参数，分别是油门踏板开度、速度、加速度，而后续构建的指标则都是这三个参数的衍生；相关技术概述则简单介绍了本文所选取的算法。

2.1 驾驶风格界定

驾驶风格是一个复杂的概念，受到许多因素的影响，使其描述变得复杂。这导致了许多术语的出现，这些术语通常缺乏一致的定义[38]。因此需要简明扼要的描述以避免混淆。本文所理解的术语之间的关系如图 2.1 所示：

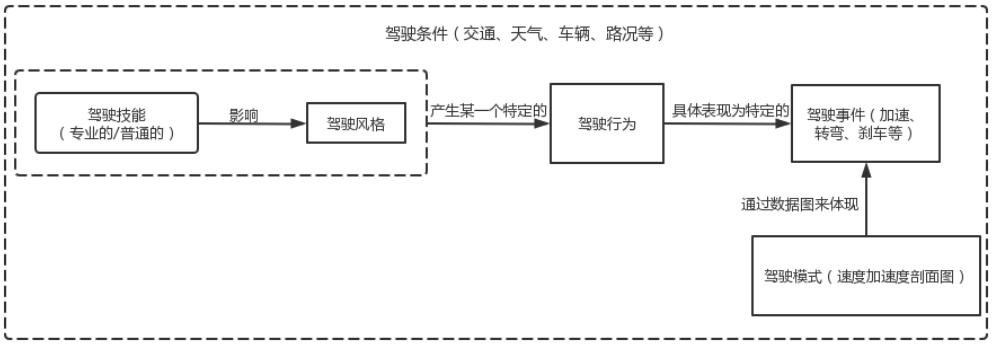

图 2.1 驾驶风格界定图

驾驶事件通常被理解为驾驶任务期间发生的操纵，例如：加速、减速、转弯和变道，可用于识别驾驶风格[39]。驾驶模式定义为速度曲线，这个描述包括可从速度分析获得的所有附加信息，部分文献将其定义为：速度曲线和计算出的功率需求[40]、加速次数和匀速时间[41]。驾驶模式与道路因素、天气因素和驾驶风格密切相关[42]，但不体现这些因素的信息。驾驶行为则往往只关注于驾驶员的操作，而忽略了外部因素[41][43]。驾驶技能指得是驾驶员对于车辆的控制能力，不同驾驶员可能会表现出专业的和普通的驾驶技能，同时驾驶技能则会影响驾驶员的驾驶风格[44]。

之前的文献中没有关于驾驶员驾驶风格的一致定义。这主要是因为驾驶风格

是一个多维度概念，受主观因素和客观因素的影响，如：驾驶员当天的情绪、疲劳度、道路拥堵程度等。Dörr 等人提出了一种简单却贴合实际的描述，认为驾驶风格就是驾驶员完成相应驾驶任务的驾驶方式。这个概念可以理解为驾驶员操作方向盘、油门和制动踏板等的方式，这是一个区别于驾驶模式的概念界定[45]。同时驾驶员的驾驶风格并不只受驾驶员本身影响，它同时也会受外界环境影响[45][46]。因此，驾驶员驾驶风格被理解为驾驶员在自身主观因素以及外部环境影响下操作车辆的方式，也就是说，驾驶风格在不同的场景会呈现出不同的结果，哪怕是同一个司机也可能因为路况的变化而发生驾驶风格的改变。该定义与前面的描述一致，并考虑了图 2.1 驾驶风格相关术语。

2.2 探究驾驶风格的相关参数

通过上一节对驾驶风格定义的界定，我们了解到驾驶风格是一个受多维度影响的概念，比如天气因素、驾驶员因素、交通因素等。通过驾驶模式中的速度及从速度分析出来的附加信息可反映驾驶员的驾驶行为，从而分析驾驶员此时的驾驶风格。考虑到现有文献和已有数据，本文最后敲定从三个角度分析驾驶行为并讨论驾驶风格，分别是油门踏板开度、速度、加速度。

（1）油门踏板开度

踩油门是驾驶行为最直接的反映[47]，速度和加速度的增加通常需要踩油门踏板，因此本文将对油门踏板开度数据进行分析。

（2）速度

速度是观察行驶车辆状态不可或缺的一个参数，通过对平均速度、速度标准差和匀速情况的分析，可初步了解到驾驶员整体行程的速度分布，是属于高速驱动驾驶还是低速驱动驾驶，其值的大小有助于后续对于驾驶风格的分类讨论。

（3）加速度

加速度数据可通过速度数据计算获得，其反映了车速的变化情况，过急的加速度或者减速度可能会导致发动机燃油消耗激增，同时具有一定的危险性，通过对加速度的分析也可以很好的识别驾驶员的驾驶风格，如加减速次数多的驾驶员可定义为频繁变速型驾驶风格。

2.3 相关技术概述

本文用于驾驶风格分类识别模型和油耗预测模型的理论分别是主成分分析、聚类分析、支持向量机、集成学习以及 Shapley value，为了方便后续的展开讨论，这里先对这些概念进行一次简单的介绍，在后面使用这些方法时会具体介绍其计算过程。

2.3.1 主成分分析概述

主成分分析(Principal Component Analysis,PCA)是一种常用的数据分析方法。PCA 通过线性变化将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维，解决高维数据计算复杂度过高的问题。

2.3.2 聚类分析概述

聚类分析（Cluster analysis）就是针对大量数据或者样本，根据数据本身的特性研究分类方法，并遵循这个分类方法对数据进行合理的分类，最终将相似数据分为一组，也就是“同类相同，异类相异”。聚类分析方法是定量地研究事物分类问题和分区问题的重要方法。

目前常用的聚类方法有两种，分别为系统聚类和 K-means 聚类。

（1）系统聚类

系统聚类也称为层次聚类，分类的单位由高到低呈树形结构，且所处的位置越低，其所包含的对象就越少，但这些对象间的共同特征就越多。该聚类方法只适合在小数据量的时候使用，数据量大的时候速度会非常慢。

（2）K-means 聚类

K-means 聚类是一种无监督机器学习方法。目标是针对给定的样本，根据它们的特征相似度或距离，将其归类到若干“簇”中。根据距离计算方式与聚类中心的不同，存在很多种类的聚类算法。其中，最常用的是 K 均值聚类算法。该算法实现较为简单、收敛速度快，弥补了系统聚类处理大批量数据的缺点，因此得到广泛运用。

2.3.3 支持向量机概述

支持向量机(Support Vector Machine,SVM)是一种二分类模型，在机器学习、计算机视觉、数据挖掘等领域中广泛应用，主要解决数据的分类问题，目的是寻找一个超平面对样本进行分割，分割的原则是间隔最大化。通常 SVM 用于二元分类问题，对于多元分类可将其分解为多个二元分类问题，再进行分类，支持向量机就是对应着将数据正确划分并且间隔最大的直线。

2.3.4 集成学习概述

集成学习（Ensemble learning），并不是一个单独的机器学习算法，而是通过构建并结合多个机器学习器（基学习器，Base learner）来完成学习任务。集成学习根据其损失函数的不同，可分别应用于分类问题和回归问题，且决策树特有的信息增益使得集成学习模型可以输出特征分析图，得益于此，许多医学、金融学等领域的问题都会选择集成学习算法构建模型，用于实际的变量分析。对于训练集数据，我们通过训练若干个弱学习器（Weak learner），通过一定的结合策

略，就可以最终形成一个强学习器（Strong learner），以达到博采众长的目的。其中普遍运用的就是随机森林（Random Forest，简称 RF）算法和极端梯度增强（Extreme Gradient Boosting，简称 XGBoost）算法。

（1）决策树

不管是随机森林还是极端梯度增强算法，其所用的弱学习器都是决策树模型。决策树是一种监督学习算法，是一种基本的分类和回归方法，因其决策图类似树状而得名。

决策树最重要的一步就是特征选择，通常用来筛选特征的算法是ID3算法、C4.5 算法和 CART 算法。ID3 算法通过信息增益划分特征；C4.5 算法在ID3的基础上提出了信息增益率作为划分依据，避免取数较多的属性信息增益对选取特征的影响；前两种算法都只能处理分类问题，而 CART 算法则能同时处理分类和回归问题，在分类问题中 CART 算法通过基尼指数划分特征，在回归问题中则会根据不纯度函数划分特征。

（2）集成学习

集成学习的基本思路是将多个弱学习器的结果通过投票或求平均值的方法获得一个综合结果。集成学习主要分为两类，分别是基于并行的 Bagging 算法和基于串行的 Boosting 算法。

Bagging 的算法原理如图 2.2 所示：

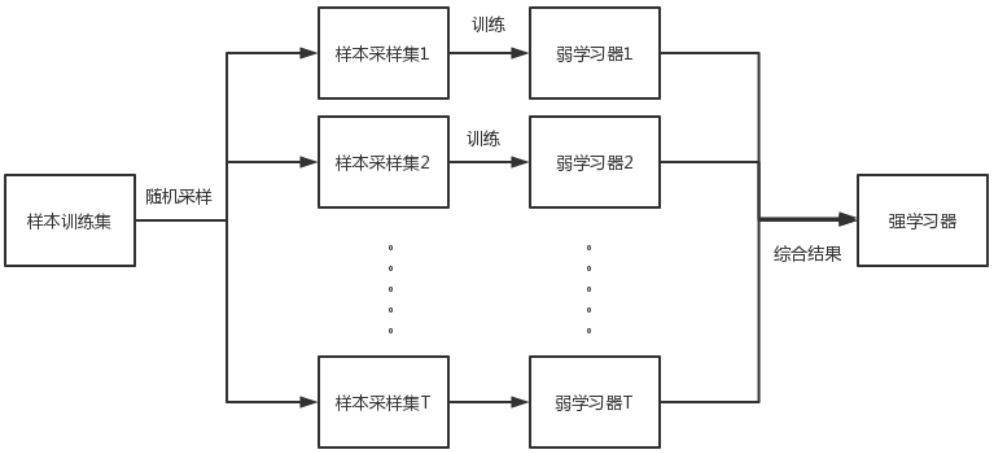

图 2.2 Bagging 原理图

可以看出 Bagging 中的弱学习器在训练时互不关联，运行效率高。同时 Bagging 算法还提到了一个随机采样，一般情况下，都会选择 Bootsrap 抽样法，具体做法就是先随机抽取样本放入采样集，接着再把样本放回，也就是说被采集过的样本可能还会再被采集，因此，采集获取的数据集与原始样本不同，和其他采样集也不同，这样就可以得到多种弱学习器。

Boosting 的算法原理如图 2.3 所示：

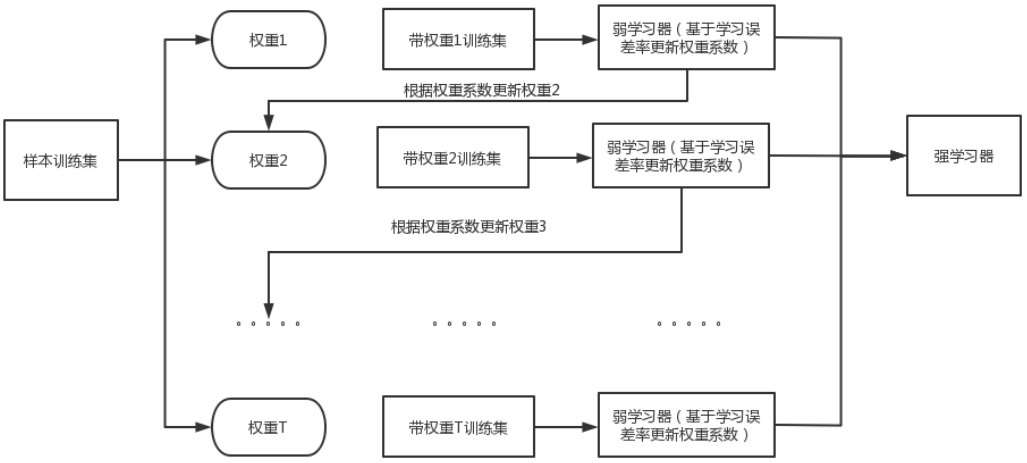

图 2.3 Boosting 原理图

从图中可以看出 Boosting 算法的工作机制是首先选取样本训练集，通过初始化权重 1 训练模型，并根据学习误差率表现来更新权重，获取权重 2，使得之前在训练集 1 中的误差率高的训练样本权重变高，将权重 2 代入训练集 2，依次重复执行，直至弱学习器数量达到预先设定值，最终将 T 个弱学习器的结果结合获取最终结果。

2.3.5 Shapley value 概述

Shapley value，通常被翻译为夏普利值或沙普利值，来源于合作博弈理论，被用于贡献分配方式的计算，其基本思路是考虑某个成员加入工作而带来的边际贡献。随着机器学习算法的不断发展，人们在追求模型准确率的同时，也开始考虑模型的透明度，比如在医学领域、保险领域、金融领域等，我们需要知道模型输入变量对于模型输出结果的影响，从而做出决策，而比较遗憾的是大部分机器学习模型都属于黑箱模型，即我们并不清楚其内部的运作规律，只是简单的通过输入 x 计算出输出 y，基于此，越来越多的专家学者开始将目光放在模型可解释性的研究上。Lundberg S M 首次将 Shapley value 的方法运用在模型可解释性研究上，其核心思想是探究某个特征加入模型时做出的边际贡献，计算方式为求该特征在所有的特征序列中的平均边际贡献[48]。

2.4 本章小结

本章首先界定了驾驶风格的概念，并讨论了和它相关概念的联系和区别;接着选取了三类行车数据，用于后续驾驶行为的分析，进而讨论不同的驾驶风格；最后简单介绍了本文所用的相关技术。其中包括主成分分析（PCA）、聚类分析、支持向量机、集成学习以及 Shapley value。

3 数据预处理与驾驶行为指标构建

真实可靠的实验数据能使得研究的结论具有一定的科学性，如果实验数据质量得不到保证，那么实验所运用的各种方法就无法得到验证[49]。因此，我们需要真实且有质量的实验数据，确保实验数据能用来研究驾驶行为、驾驶风格以及油耗之间的关系，本文将采集器获取的数据进行了合并汇总，进行了一些预处理操作，并提取出一些与驾驶行为相关的指标，然后基于这些指标构建模型。因此，本章的主要工作是数据的预处理以及指标构建。

3.1 数据来源与介绍

本文的数据来源于 T 车联网物流公司，该公司的所有重型货车均已实现了由车载设备向联网数据库远程传送数据的功能，各类车辆传感器可定期向数据库终端提供行车数据，本文选用了 8 辆发动机规格以及车型一致的重型货车进行研究，通过访谈企业负责人了解到，一般情况下，该公司下的所有重型货车在运输货物时为了实现运输的最大效益，会选择略低于满载的规格运输，因此本文认为所选车辆之间的重量差距不大，并未就车重对于油耗的影响进行后续讨论。共收集 2020 年 9 月这 30 天里所有的车联网数据，数据量达到 400 万条。收集的数据包括采集时间点、行驶总里程，仪表盘车速、发动机转速、行驶里程、累计油耗、油门踏板开度、GPS 数据等，这些数据可基本反映车辆的行驶状况。如表 3.1 所示是速度传感器收集的部分数据实例：

表 3.1 速度传感器部分数据集表

数据编号	采集时间	累计里程	仪表盘车速	发动机转速
1	20200901120039	15818.5	1	600
2	20200902073355	16476.8	40	1340
3	20200913020034	20337.8	58	1180
4	20200915044642	21233	49	1300
5	20200917123048	22154.1	50	1300
6	20200919043830	22727.4	63	1300
7	20200923131120	24811.2	54	1100
8	20200925173544	25969.1	55	1120
9	20200927170304	26521.3	69	1100
10	20200928183341	27119.8	20	1460

为了方便后续的实验讨论，这里简单介绍一下部分采集信息的含义。

（1）以速度传感器数据集为例：采集时间表示该数据点当前的时间节点，比如 20200901120039 就表示 2020 年 09 月 01 日 12 时 00 分 39 秒；累计里程表

示当前仪表盘记录的车辆累计行程，单位是 km；仪表盘车速表示当前仪表盘记录的车辆行驶速度，单位是 km/h；发动机转速表示发动机曲轴每分钟的回转数，单位是 rpm。

（2）以累计油耗传感器数据集为例：累计油耗表示当前仪表盘记录的车辆累计油耗，其单位为升（L）。

（3）以油门踏板开度传感器数据集为例：油门踏板通常用于控制节气门开度大小，因此油门踏板开度越大，节气门开度越大，进入燃油发动机的空气越多，其值的单位为%。

3.2 数据预处理

由于车辆传感器采集数据的频率各不相同，且存在部分传感器受信号干扰的情况，为了消除这些异常数据对于实验结果的影响，在实验开始前需要进行数据预处理操作。

（1）数据同步

本文中采集的数据来自不同的车载设备，其采集频率从 2s 到 60s 不等，因此需要根据时间序列对所有数据进行同步，这里主要需要合并三个数据集，分别是速度数据集、行驶油耗数据集和油门踏板开度数据集，其具体情况如表 3.2、表 3.3 和表 3.4 所示：

表 3.2 速度数据表

数据编号	采集时间	累计里程	仪表盘车速	发动机转速
1	20200901000000	31181.5	51	1360
2	20200901000002	31181.5	51	1340
3	20200901000004	31181.5	52	1360
4	20200901000006	31181.6	53	1400
5	20200901000008	31181.6	54	1420
6	20200901000010	31181.6	55	1460
7	20200901000012	31181.7	56	1480
8	20200901000014	31181.7	57	1100
9	20200901000016	31181.7	57	1180
10	20200901000018	31181.8	57	1160

表 3.3 行驶油耗数据表

数据编号	采集时间	累计油耗
1	20200901000034	11280.5
2	20200901000135	11280.5
3	20200901000236	11281
4	20200901000337	11281
5	20200901000437	11281.5
6	20200901000538	11281.5
7	20200901000638	11282
8	20200901000738	11282.5
9	20200901000839	11283
10	20200901000939	11283

表 3.4 油门踏板开度数据表

数据编号	采集时间	油门踏板开度
1	20200901000001	24.8
2	20200901000004	43.6
3	20200901000006	39.6
4	20200901000008	35.6
5	20200901000010	51.2
6	20200901000012	56.4
7	20200901000014	0.8
8	20200901000016	14.4
9	20200901000018	11.2
10	20200901000020	23.2

累计油耗的采集频率最长，为 60s 采集一次，是最容易产生误差的一项数据集，但考虑到本文研究的行程片段多为长途行驶，一分钟内的油耗误差并不会过多影响整体的实验结果，因此选择通过同一分钟进行时间序列合并，即只观察累计油耗数据集中采集时间的分钟位，将同一分钟的数据写入速度数据集中。

至于油门踏板开度的数据，其采集频率也是 2s 一次，同速度数据集一样，因此合并这两类数据集则选择同一采集时间的数据直接写入速度数据集，如出现无法相等匹配的情况则采取移动平均法填补缺失值，具体做法是设置一个时间窗口，记作 k，选取某个数据点前 k 个数据，将其取平均值，所求值即为该点的数据，可用式（3.1）表示：

$$x_{n+1} = AVERAGE(x_n + x_{n-1} + \cdots + x_{n-k+1}) \tag{3.1}$$

考虑到数据集具有较强的时序性，且采集频率为 2s，综合考虑后设置 k=3 的窗口进行实验。

（2）重复数据处理

本文采集的数据是按秒依次产生的，出现同样时间的数据显然是有问题的，因此通过观察采集时间确定数据集中是否存在重复的数据，对此，本文对同一时

间的数据进行合并处理，去除重复数据只保留一条。如下表 3.5 所示，采集时间为 20200920145421 的数据有多达 5 个相同的数据，因此需要删除重复内容并只保留一条数据。

表 3.5 部分重复数据表

数据编号	采集时间	累计里程	仪表盘车速	发动机转速	累计油耗
1	20200920145415	23131.2	65	1380	7976.5
2	20200920145417	23131.3	67	1380	7976.5
3	20200920145419	23131.3	68	1380	7976.5
4	20200920145421	23131.3	68	1380	7976.5
5	20200920145421	23131.3	68	1380	7976.5
6	20200920145421	23131.3	68	1380	7976.5
7	20200920145421	23131.3	68	1380	7976.5
8	20200920145421	23131.3	68	1380	7976.5
9	20200920145423	23131.4	68	1380	7976.5
10	20200920145425	23131.4	68	1380	7976.5

（3）gps 数据筛选

根据开源数据库 OSM（open street map）获取的城市路网数据，利用 Arcmap 软件剔除部分脱离路网的数据点以及不符合实际情况的数据点，如果某段行程超过 5%的 gps 坐标点脱离路网数据，则在后续的讨论中不再考虑这段行程。如图 3.1 所示是一段 gps 数据异常的行程图。

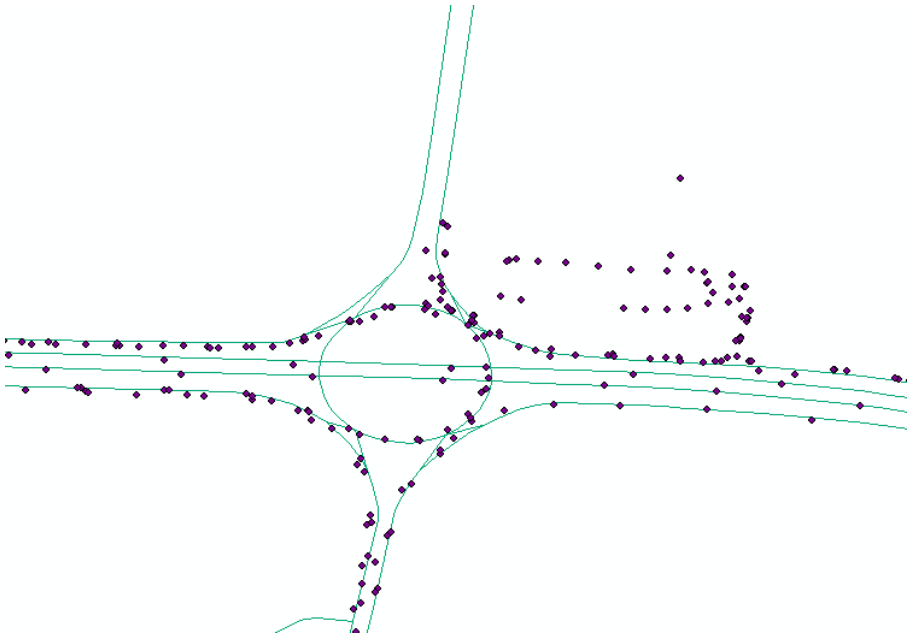

图 3.1 gps 数据异常图

（4）数据标准化

不同的数据量纲会导致后续的实验产生不同的结果，比如支持向量机就对不

同量纲的数据十分敏感，如果采用原始数据分析，模型的准确率将会有所下降，因此我们需要对数据进行标准化处理。常用的数据标准化处理方式有两种，分为离差标准化和标准差标准化，下面将分别进行讲解。

（a）离差标准化

离差标准化又称 0-1 标准化或者归一化，计算过程需要用到原始数据和数据的最大最小值，使所有数据映射到[0，1]区间，可用如式（3.2）表示：

$$y = \frac{x - min}{max - min} \tag{3.2}$$

其中 x 为该段行程中某一特征的参数值，max 表示该特征参数的最大值，min 表示该特征参数的最小值。离差标准化结构易于理解，操作起来不复杂，因此运用较为广泛，但考虑到离差标准化在有新数据加入后，标准化结果会受参数最大值和最小值的改变而改变，运用起来不够稳定，且标准化之后只能比较数据的分布情况，使用场景单一。

（b）标准差标准化

标准差标准化又称 Z-score 标准化，是对数据列中每个数据点作减去均值除以方差的操作，使得处理后的数据近似符合标准正态分布，可用式（3.3）表示：

$$y = \frac{x - \mu}{\sigma} \tag{3.3}$$

其中 x 是该段行程中某一特征的参数值，μ是该特征所有参数值的平均值，σ是该特征参数值的标准差。

这里需要注意的是，有些参考资料将标准差标准化误叫做归一化，这是不对的，标准差标准化获取的标准化值不一定处于 0 到 1 之间，当样本小于或者远大于参数平均值时，会出现小于 0 以及大于 1 的标准差值。考虑到离差标准化容易受最大值和最小值的影响，当数据发生变动时标准化结果会不稳定，因此本文选择标准差标准化对后续数据进行处理。

3.3 行程片段切分

行程片段的切分没有统一的标准，应具体问题具体分析，通过对现有文献进行总结，主要分为以下两种方法：行程分析法和等长或等时截取法。行程分析法通过定义完整运动学片段来划分行程，这里的运动学片段可分为两种，分别是以发动机熄火为依据的运动学片段和车速降为零的运动学片段，前者认为的运动学片段是车辆从起步点火到熄火停车的整个过程，后者则以车辆速度从零开始到下次降为零的整个过程，其相较于前者片段较短，且无法探究车辆怠速这一重要的驾驶模式，这两类运动学片段的车辆都会经历多个加速匀速再减速等一系列过程，司机的驾驶行为也会直观的体现在车速、加速度、油门踏板、油耗等参数上。

等长或等时截取法通过相等的行驶时间或者相等的行驶距离分割某一整段

行程，再对分割后的行程提取行驶参数，这一方法的前提过于理想，不适用现实中复杂多变的驾驶环境，多用于一些模拟驾驶的实验。考虑到本文选取的数据源自现实世界，且在后续的讨论中需要探讨怠速驾驶这一驾驶模式，因此选用行程分析法中基于车辆熄火的运动学片段来截取行程段，这里熄火的判定也很简单，只需观察发动机转速的值是否为零即可，具体切分方法见流程图 3.2。

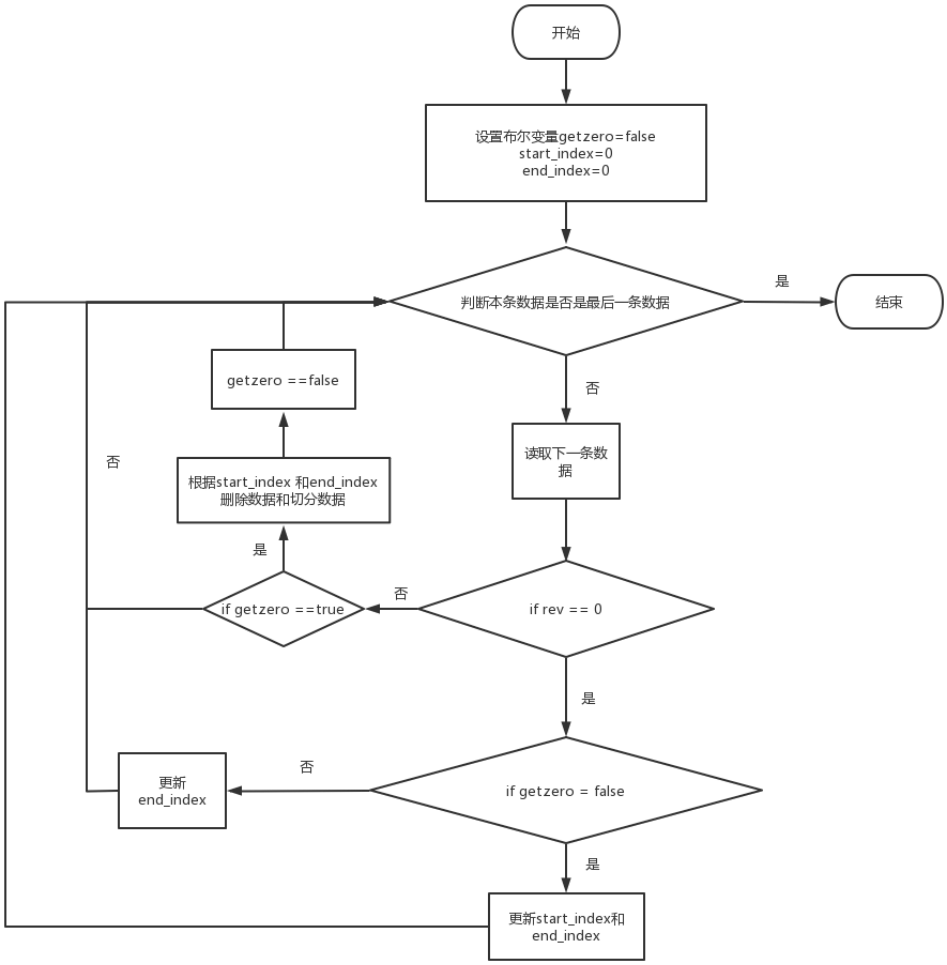

图 3.2 行程片段切分流程图

具体步骤如下：

Step1：设定初始布尔变量 getzero=fasle，start_index=0，end_index=0，其中 getzero=fasle 表示读取数据的前一个数据转速不等于 0，start_index 和 end_index 分别表示读取数据的序列值，其初始值均为 0。

Step2：判断是否已经遍历所有数据，如果是，则循环结束，如果否，判断当前数据的发动机转速是否为零，如果是，再判断其上一条数据的发动机转速是否是 0，如果是，更新 start_index 和 end_index，如果否，只更新 end_index，再回到发动机转速判断，如果发动机转速不是零，再判断上一条数据是否是零，如果是，则给根据 start_index 和 end_index 删除数据和切分数据，同时另

getzero==false，如果否，则继续下一条数据的循环。

Step3：循环步骤 2 直至再无可用数据。

如图 3.3 所示就是某段运动学片段的速度-时间图，可以看见车辆在这运动过程中会经历多个加速、减速过程，同时还夹杂着若干匀速的过程。

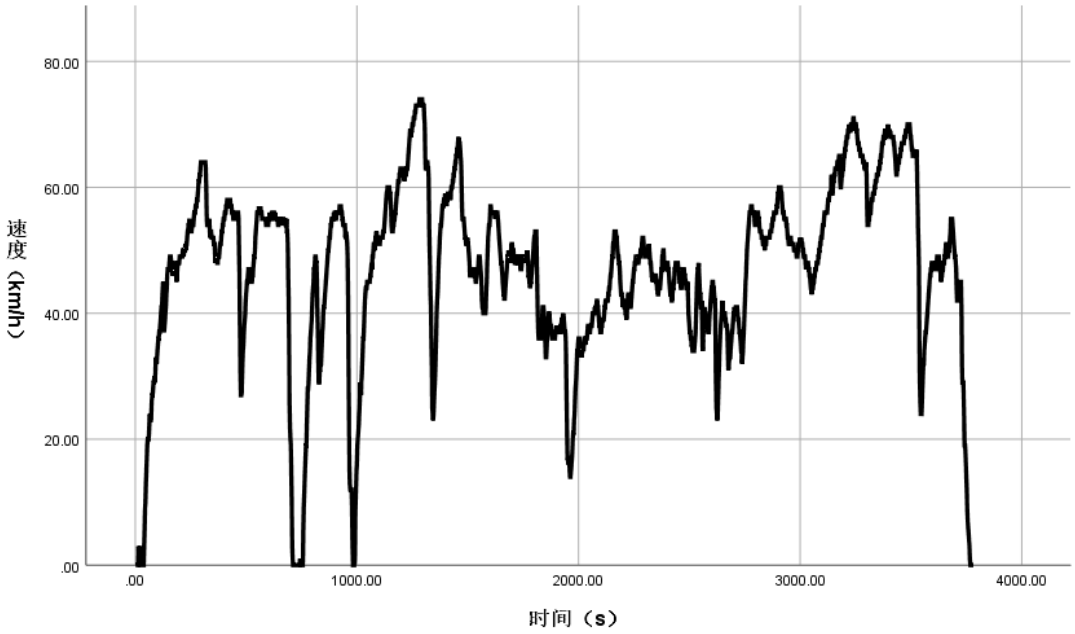

图 3.3 速度-时间图

同时考虑到重型货车可能会在中转中心卸货或者装货，使得有些行程片段基本全程处于怠速或低速状态，这类片段没有实际的研究意义，因此挑选了平均速度大于 20km/h 的行程[50]。

3.4 驾驶行为指标构建

影响车辆燃油经济型的因素有很多，包括道路因素、天气因素、交通因素、车辆因素和驾驶员因素等，其中道路交通因素以及天气因素是人为难以调控的，而车辆因素则根据车辆的实际情况而定，也是客观存在的，因此我们能做的就是在车辆运行的过程中，通过改进驾驶员的驾驶行为来提高燃油经济性，所以我们有必要分析影响燃油经济性的驾驶行为指标。国外文献常定义 15 到 30 个指标进行研究[51]，本文以此为基础并考虑油门踏板开度构建了 28 种指标，每个指标均由一段行程统计获得，根据不同指标所属类别具体可以分为以下四类：与油门踏板开度相关、与速度相关、与加速度相关以及与驾驶模式相关的指标。下面将详细介绍每个指标的具体含义和计算方式。

3.4.1 油门踏板开度相关指标

驾驶员驾驶行为最直接的表现就是对于油门踏板和刹车踏板的控制[52],从而才能间接反映在速度和加速度数值上，同时考虑到刹车踏板对于油耗的影响会间

接体现在油门踏板对于油耗的影响上，比如踩完刹车后通常会伴随踩油门的操作，综合考虑，本小节选取了几个有关油门踏板开度的指标。表 3.6 是指标的概念描述和计算方法，其中 m 表示运动学片段的数据量,i 表示第 i 个数据，T 表示行程的行驶总时长。

表 3.6 与油门踏板开度相关的指标表

序号	驾驶行为指标	描述	计算方法
1	油门踏板开度平均值 p	重型货车行驶时油门踏板开度的平均值	$\bar{p}=\frac{1}{m}\sum_{i=1}^{m}p_i$
2	油门踏板开度整体标准差$\sigma(p)$	重型货车行驶时油门踏板开度的整体标准差	$\sigma(p)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(p_i-\bar{p})^2}$
3	油门踏板角速度平均值$\overline{pa}$	重型货车行驶时油门踏板角速度的平均值	$\overline{pa}=\frac{1}{m}\sum_{i=1}^{m}pa_i$
4	油门踏板角速度整体标准差$\sigma(pa)$	重型货车行驶时油门踏板角速度的整体标准差	$\sigma(pa)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(pa_i-\overline{pa})^2}$
5	低油门踏板开度 p_{0-25}	重型货车行驶时油门踏板开度处于 0-25%所占比例	$p_{0-25}=\frac{t_{0-25}}{T}\times 100\%$
6	中油门踏板开度 p_{25-50}	重型货车行驶时油门踏板开度处于 25-50%所占比例	$p_{25-50}=\frac{t_{25-50}}{T}\times 100\%$
7	高油门踏板开度 p_{50-75}	重型货车行驶时油门踏板开度处于 50-75%所占比例	$p_{50-75}=\frac{t_{50-75}}{T}\times 100\%$
8	超高油门踏板开度 p_{75-100}	重型货车行驶时油门踏板开度处于 75-100%所占比例	$p_{75-100}=\frac{t_{75-100}}{T}\times 100\%$

$$pa_i=\frac{p_{i+1}-p_i}{\Delta t} \tag{3.4}$$

其中p_i表示第 i 个数据点的油门踏板开度，Δt表示两个数据点之间的时间差。

3.4.2 速度相关指标

目前，大部分国内外专家学者在研究驾驶行为的时候都会考虑到行驶车速。Wickramanayake S 等[53]在构建油耗预测模型时，选择了平均速度作为其指标之一；赵晓华等[54]在研究驾驶行为变化时，选择速度标准差用来衡量驾驶员速度稳定性； Marafie Z 等[55]在研究驾驶行为时将车速细分为低速、中速、高速和超

高速，用不同区间的速度判断驾驶员属于什么驾驶风格。因此平均速度可以用来反映行程的整体速度;速度整体标准差体现了行驶片段整体的速度变化情况，其值越大，反映整段行程中车辆的速度落差越大，波动性较大；不同的速度区间反映了车速的分布情况。综上，本文在现有研究的基础上，选取了 8 个有关速度的指标，需要一提的是这里的行驶速度表示车辆行驶的车速，因此行驶速度不考虑速度为 0km/h 的数据点。通过仪表盘车速，根据不同运动学片段统计平均速度、速度标准差以及速度分布情况，具体如表 3.7 所示：

表 3.7 与速度相关的指标表

序号	驾驶行为指标	描述	计算方法
1	平均速度$\bar{v}$	重型货车速度的平均值	$\bar{v}=\frac{1}{m}\sum_{i=1}^{m}v_i$
2	速度整体标准差σ（v）	重型货车速度的整体标准差	$\sigma(v)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(v_i-\bar{v})^2}$
.3	平均行驶速度$\bar{V}$	重型货车行驶速度的平均值	$\bar{V}=\frac{1}{m}\sum_{i=1}^{m}V_i$
4	行驶速度整体标准差σ(V)	重型货车行驶速度的整体标准差	$\sigma(V)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(V_i-\bar{V})^2}$
5	低速占比$v_{低}$	重型货车行驶速度处于 0-30km/h	$V_{低}=\frac{t_{低}}{T}\times 100\%$
6	中速占比$v_{中}$	重型货车行驶速度处于 30-60km/h	$V_{中}=\frac{t_{中}}{T}\times 100\%$
7	高速占比$v_{高}$	重型货车行驶速度处于 60-90km/h	$V_{高}=\frac{t_{高}}{T}\times 100\%$
8	超高速占比$v_{超高}$	重型货车行驶速度处于 90-120km/h	$V_{超高}=\frac{t_{超高}}{T}\times 100\%$

其中低速，中速，高速以及超高速反映了驾驶员不同速度区间的驾驶情况，这里具体展示下低速的识别步骤，其他三类的识别过程基本一致，具体步骤如下：

Step1：获取当前第 i 条数据的速度v_i，低速总时长 t，处理数据总量 m。

Step2：判断是否已经遍历所有数据，若否，继续判断当前数据是否满足对于速度的要求：$0km/h<v_i\leq 30km/h$。若是，$t=t+2$，$i=i+1$。若否，则$i=i+1$，继续判断下一条数据。

Step3：循环执行步骤 2，直到所有数据遍历完成，输出低速总时间t。

3.4.3 加速度相关指标

车辆加速度的值可以有效反映驾驶员车速变化情况，较高或较低的加速度反映了车辆的急加速或急减速驾驶行为，平稳的加减速可以有效降低油耗，同时急加速会加速车辆发动机的损耗，且有一定的安全隐患，急减速会加速车辆硬件的磨损，不利于车辆的使用寿命[56]。现有数据中并没有加速度数据，但我们可以通过速度速度及其采集频率计算出每个数据点的加速度值，加速度的计算公式如式（3.5）所示：

$$a_i = \frac{v_{i+1} - v_i}{\Delta t} \tag{3.5}$$

其中v表示第 i 个数据点的速度值，Δt表示两个数据点之间的时间差。

同时根据设定的加速度阈值，可识别正常加速、中等加速、急加速、正常减速、中等减速、急减速等驾驶行为，查阅相关资料发现，加速度或者减速度区间阈值的设定并无一个统一的标准，大部分文献会将车辆急加速或者急减速的加速度阈值设置在 1-2 m/s^2[51][57][58]，再向下均分获取正常加速和中等加速的阈值，表 3.8 为本文数据集获取的所有正加速度值，图 3.4 为正加速度柱状图。

表 3.8 正加速度频数表

加速度	频数	百分比	累计百分比
0.14	845889	60	60
0.28	329996	23.4	83.3
0.42	130411	9.2	92.6
0.56	56042	4	96.6
0.69	26160	1.9	98.4
0.83	12240	0.9	99.3
0.97	5690	0.4	99.7
1.11	2707	0.2	99.9
1.25	1160	0.1	100
1.39	422	0	100
1.53	154	0	100
1.67	50	0	100
1.81	19	0	100
1.94	3	0	100
2.22	1	0	100
3.19	1	0	100
3.61	1	0	100
3.89	1	0	100
5.97	1	0	100
总计	1410948	100	100

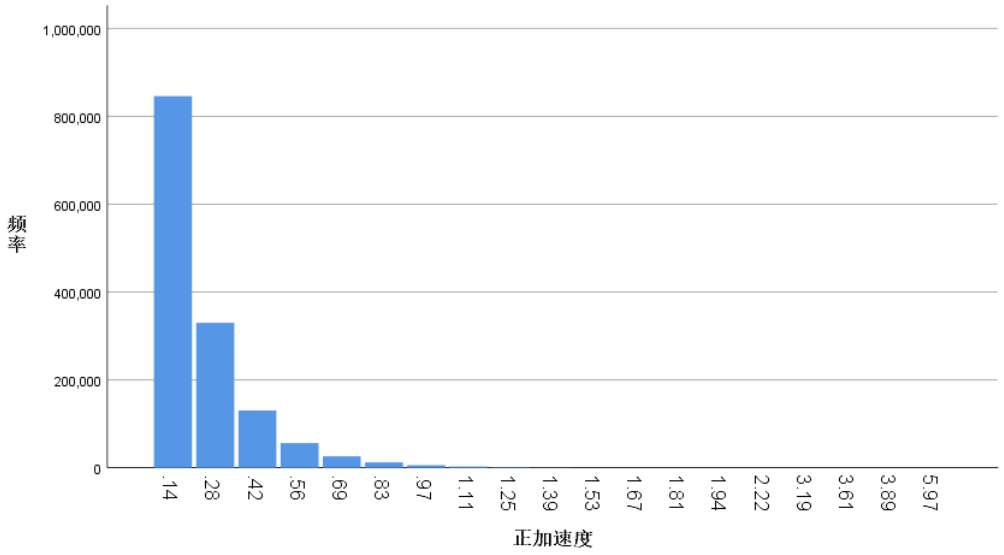

图 3.4 正加速度柱状图

结合表 3.8 和图 3.4 可以看出本文的正加速度值大于 1 m/s² 的数据集较少，如果盲目选择之前文献的阈值数可能无法正确判断驾驶员的驾驶行为，因此本文选择第 95 分位数据值作为急加速和急减速的识别阈值[1]，观察发现急加速的阈值为 0.56 m/s²，急减速的阈值为-0.83 m/s²，再向下选取正常和中等加减速的阈值。

除此以外，通过加速度值还可间接统计获得平均加速度、平均减速度、加速度标准差、减速标准差这几个驾驶行为指标，这些指标的计算方法如表 3.9 所示：

表 3.9 与加速度相关的指标表

序号	驾驶行为指标	描述	计算方法
1	平均加速度$\bar{a}$	重型货车行驶时加速度的平均值	$\bar{a}=\frac{1}{m_a}\sum_{i=1}^{m_a}a_i$
2	平均减速度$\bar{d}$	重型货车行驶时减速度的平均值	$\bar{d}=\frac{1}{m_d}\sum_{i=1}^{m_d}d_i$
3	加速度整体标准差 $\sigma(a)$	重型货车行驶时所有加速度的整体标准差	$\sigma(a)=\sqrt{\frac{1}{m_a}\sum_{i=1}^{m_a}(a_i-\bar{a})^2}$
4	减速度整体标准差 $\sigma(d)$	重型货车行驶时所有减速度的整体标准差	$\sigma(d)=\sqrt{\frac{1}{m_d}\sum_{i=1}^{m_d}(d_i-\bar{d})^2}$
5	正常加速占比 $a_{正常加速}$	重型货车行驶时加速度大于 0.13m/s² 小于 0.28m/s² 所占比例	$a_{正常加速}=\frac{t_{正常加速}}{T}\times 100\%$
6	中等加速占比	重型货车行驶时加速度大于 028m/s² 小于	$a_{中等加速}=\frac{t_{中等加速}}{T}\times 100\%$

序号	驾驶行为指标	描述	计算方法
	$a_{中等加速}$	等于 0.56m/s² 所占比例	
7	急加速占比$a_{急加速}$	重型货车行驶时加速度大于 0.56m/s² 所占比例	$a_{急加速}=\frac{t_{急加速}}{T}\times 100\%$
8	正常减速占比 $a_{正常减速}$	重型货车行驶时加速度小于-0.13m/s² 大于等于-0.56m/s² 所占比例	$a_{正常减速}=\frac{t_{正常减速}}{T}\times 100\%$
9	中等减速占比 $a_{中等减速}$	重型货车行驶时加速度小于-0.56m/s² 大于等于-0.83m/s² 所占比例	$a_{中等减速}=\frac{t_{中等减速}}{T}\times 100\%$
10	急减速占比$a_{急减速}$	重型货车行驶时加速度小于-0.83m/s² 所占比例	$a_{急减速}=\frac{t_{急减速}}{T}\times 100\%$

根据不同加速度阈值设置的指标可以细化对驾驶员加速度的研究，这里我们以急加速识别为例，其他指标的识别过程基本一致，其具体流程步骤如下：

Step1：获取当前第 i 条数据的加速度a_i，急加速总时长 t，处理数据总量 m。

Step2：判断是否已经遍历所有数据，若否，继续判断当前数据是否满足对加速度值的要求：$a_i > 0.56m/s^2$。若是，$t=t+2, i=i+1$。若否则，$i=i+1$，继续判断下一条数据。

Step3：循环执行步骤 2，直到所有数据遍历完成，输出急加速总时间t。

3.4.4 驾驶模式相关指标

在一些文献中还会对行程的怠速比例、匀速比例、加速比例和减速比例进行提取，考虑到加速度指标中的各类加减速时间占比已反映了行程的加速比例和减速比例，因此这里只对怠速比例和匀速比例进行提取，具体如表 3.10 所示：

表 3.10 与驾驶模式相关的指标表

序号	驾驶行为指标	描述	计算方法
1	匀速比例 Cruise	实际驾驶中加速度为 0m/s² 的数据点	$Cruise=\frac{t_c}{T}\times 100\%$
2	怠速比例 Ldle	重型货车行驶时，速度为 0km/h 但发动机仍在运行的状态	$Ldle=\frac{t_l}{T}\times 100\%$

（1）匀速行为特征提取

驾驶员在实际驾驶的过程中，若能将加速度控制在一个稳定的区间，保证车辆长期处于较为匀速的行驶状态，可有效降级车辆各种硬间的损耗，提高燃油经济性。本文将满足式（3.6）的数据点记为匀速驾驶：

$$a_i = 0 \tag{3.6}$$

具体步骤如下：

Step1：获取当前数据对应的加速度a_i、处理数据量m、匀速总时间t。

Step2：判断是否已经遍历所有数据，若否，继续判断当前数据是否满足对加速度值的要求：$a_i = 0m/s^2$。若是，$t = t + 2$，$i = i + 1$。若否则，$i = i + 1$，继续判断下一条数据。

Step3：循环执行步骤 2，直到所有数据遍历完成，输出匀速总时间t。

（2）怠速行为特征提取

重型货车在行驶过程中，会因为交通拥堵或者红路灯而选择停车等待，驾驶员通常在此时会出于方便而选择脚踩刹车或者挂空挡，此时发动机仍处在运转状态，但车辆实际没有运动，虽然转速不高，但实际也会造成不必要的燃料消耗。怠速行为常发生在道路拥挤的路段，驾驶员需要经常性的停车等红灯或者排队通行，有些驾驶员行车时会考虑这一问题，在主观认为需要较长时间停车等待时会选择熄火停车。综上所述，本文认为的怠速就是车辆速度为 0，但转速不为 0 的情况，基于此本文将满足式（3.7）的行驶过程记为一次怠速

$$R_i \neq 0, v_i = 0 \tag{3.7}$$

具体流程步骤如下：

Step1：获取第 i 条数据的速度值v_i、转速值R_i、处理数据量m、怠速总时间t。

Step2：判断是否已经遍历所有数据，若否，继续判断当前数据是否同时满足对速度值和转速值的要求：$R_i \neq 0\, r/min$，$v_i = 0km/h$。若是，$t = t + 2$，$i = i + 1$。若否则，$i = i + 1$，继续判断下一条数据。

Step3：循环执行步骤 2，直到所有数据遍历完成，输出怠速总时间t。

3.4.5 指标总结

综上，我们总共获得了 28 个指标，为了后续实验方便讨论，这里给出每个指标的中英文及其缩写，具体如表 3.11 所示：

表 3.11 指标中英文及缩写表

中文	英文	缩写
油门踏板开度平均值	Average opening of accelerator pedal	AP
油门踏板开度整体标准差	Overall standard deviation of accelerator pedal	STDP
油门踏板角速度平均值	Average accelerator pedal angular speed	APA

中文	英文	缩写
油门踏板角速度整体标准差	Overall standard deviation of accelerator pedal angular speed	STDPA
低油门踏板开度占比	Low pedal	LP
中油门踏板开度占比	Medium pedal	MP
高油门踏板开度占比	High pedal	HP
超高油门踏板开度占比	Super high pedal	SHP
平均速度	Average speed	AS
速度标准差	Overall standard deviation of speed	STDS
平均行驶速度	Average driving speed	ADS
行驶速度标准差	Overall standard deviation of driving speed	STDDS
低速占比	Low speed	LS
中速占比	Medium speed	MS
高速占比	High speed	HS
超高速占比	Super high speed	SHS
平均加速度	Average acceleration	AA
平均减速度	Average deceleration	AD
加速度整体标准差	Overall standard deviation of acceleration	STDA
减速度整体标准差	Overall standard deviation of deceleration	STDD
正常加速占比	Nomal acceleration	NA
中等加速占比	Medium acceleration	MA
急加速占比	Shap acceleration	SA
正常减速占比	Nomal deceleration	ND
中等减速占比	Medium deceleration	MD
急减速占比	Shap deceleration	SD
匀速比例	Cruise	C
怠速比例	Ldle	L

3.5 本章小结

本章首先介绍了数据来源，接着考虑到原始数据存在缺失和误差值，引入了

一系列方法对数据进行预处理操作；然后对不同指标进行了详细的介绍和统计，其中包括油门踏板开度指标、速度指标、加速度指标以及驾驶模式指标。

4 基于车联网数据的驾驶风格分类识别

在考虑驾驶行为对于燃油经济性的影响时，我们应综合考虑驾驶风格，就比如驾驶风格较为平稳的驾驶员，其在行驶的过程中车辆速度波动不大，很少有急加速或者急踩油门的驾驶行为，其燃油消耗普遍低于频繁变速或者猛踩油门的驾驶风格，如果不考虑驾驶风格对于油耗预测模型的影响，是十分不合理的。同时根据第二章的驾驶风格界定我们可以知道，不同驾驶风格会产生相应驾驶行为，具体驾驶行为则由各项指标来体现。因此我们有必要在节能分析前对驾驶员的驾驶风格进行确认，且通过第三章构建的指标分析驾驶员的驾驶风格。

4.1 驾驶风格分类识别模型整体框架

通过对驾驶行为的研究可识别不同的驾驶风格，如胡浈等[59]就将油门踏板开度的相关指标用于驾驶风格的识别，吴丽宁[60]则选取了速度和加速度的相关指标用于驾驶风格分类。

如图 4.1 给出了驾驶风格分类识别模型的整体框架。首先，通过车载设备和终端数据库获得车辆行驶数据，通过第三章的数据预处理和指标构建获取驾驶行为相关指标，用于后续驾驶风格分类，这里我们总共提取了 28 个指标，这些指标从速度、加速度、油门踏板开度反映了驾驶员行程段的驾驶行为。理论上来说，指标构建的越多，涉及维度越广，我们后续在做聚类的时候获得的结果越准确，但考虑到实际运用时，高维数据的运算成本太高，因此这里选择 PCA 主成分分析将多维的数据集降为低维的数据集，然后通过 PSO-Kmeans 聚类算法进行聚类，获取不同重型货车行程片段的驾驶风格类别。通过 PCA 和聚类分析，我们已经初步获取了每段行程的驾驶风格类别，最后，构建基于支持向量机算法的驾驶风格识别模型用于后续新数据集加入时的驾驶风格识别。

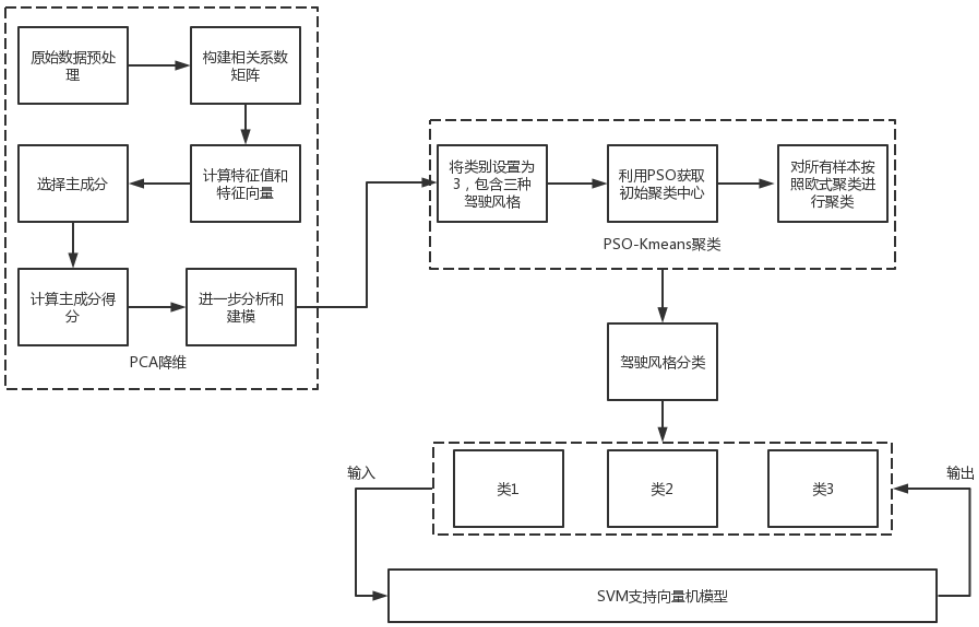

图 4.1 模型整体框架图

4.2 车联网数据分析

4.2.1 驾驶行为指标降维

在科学研究中，往往需要通过多个角度来思考问题，基于此，我们可能需要对多个数据变量进行观察，多变量毫无疑问可以带来更高的准确率，但当变量过多时可能会带来样本点之间的界限难以区分以及高计算量的问题。PCA 就是一种可以有效降低数据维度的方法，其基本原理是把多维指标转化为几个综合性的指标，这些综合性指标就称为主成分，这些主成分统计学上是原始变量的线性组合，且主成分之间不具有相关性，这就使得主成分相比较原始变量特征有着更好的优越性。

一个主成分对于原始变量的解释力度往往是不够的，因此我们就需要计算获得更多主成分，同时我们需要让主成分之间互不相关，具体做法就是使得各主成分协方差等于 0。

设Z_i表示第i个主成分，i=1,2,3,…,p，可设：

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p \end{cases} \tag{4.1}$$

其中对每一个i，均有$c_{i1}^2 + c_{i2}^2 + \cdots + c_{ip}^2 = 1$

PCA 的基本步骤如下：

（1）为了防止不同量纲的数据集影响后续实验，这里需要对数据进行标准

化处理，这一步在 SPSS 中会自动进行，计算方式就是标准差标准化，为了后续介绍，这里再具体讲解一下其计算过程。

假设进行主成分分析的指标变量有m个：$x_1, x_2, \dots, x_m$,共有n个评价对象，第i个评价对象的第 j 个指标的取值为x_{ij}。将各指标值x_{ij}转换成标准指标$\tilde{x}_{ij}$：

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{4.2}$$

其中$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$，$s_j = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$，$j = 1,2,\dots,m$，即$\bar{x}_j$和$s_j$为第$j$个指标的样本均值和样本标准差。对应地，称式（4.3）为标准化指标变量。

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{s_i}, i = 1,2,\dots,m \tag{4.3}$$

（2）建立变量之间的相关系数矩阵 R

相关系数矩阵$R = (r_{ij})_{m\times m}$，其中的r_{ij}如式（4.4）所示：

$$r_{ij} = \frac{\sum_{k=1}^{n} \tilde{x}_{ki} \cdot \tilde{x}_{kj}}{n-1}, (i, j = 1,2,\dots,m) \tag{4.4}$$

式中$r_{ii} = 1, r_{ij} = r_{ji}, r_{ij}$是第$i$个指标与第$j$个指标的相关系数。通常我们认为变量之间的相关系数越高越适合做 PCA，即出现数据冗余的情况，若相关系数过低则数据不适合做主成分分析

（3）计算相关系数矩阵 R 的特征值和特征向量

计算相关系数矩阵 R 的特征值$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_m \geq 0$，通常我们只选取特征值大于 1 的主成分，如果特征值小于 1，表明该主成分的解释力度不大，不可代表原始数据信息。对应的特征向量为$u_1, u_2, \dots, u_m$,其中$u_j = (u_{1j}, u_{1j}, \dots u_{nj})^T$,由特征向量组成$m$个新的指标变量如下式所示：

$$\begin{cases} y_1 = u_{11}\tilde{x}_1 + u_{21}\tilde{x}_1 + \cdots + u_{n1}\tilde{x}_n \\ y_2 = u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \cdots + u_{n2}\tilde{x}_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = u_{1m}\tilde{x}_1 + u_{2m}\tilde{x}_2 + \cdots + u_{nm}\tilde{x}_n \end{cases} \tag{4.5}$$

式中$y_{1,2,\dots,m}$表示第 1,2,…,m 个主成分，u_{nm}代表协方差矩阵的特征向量，$\tilde{x}_n$表示 n 个指标的标准化值。

（4）写出主成分并计算综合得分

计算特征值$\lambda_j (j = 1,2,\dots,m)$的信息贡献率和累积贡献率。主成分$y_j$的信息贡献率如式（4.6）所示：

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} (j = 1,2,\dots,m) \tag{4.6}$$

主成分$y_1, y_2, \dots, y_p$的累计贡献率如式（4.7）所示：

$$a_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k} \tag{4.7}$$

当a_p接近于 1 时，则选择前p个指标变量$y_1, y_2, \ldots, y_p$作为p个主成分，代替原来m个指标，从而可对p个主成分进行综合分析。

通过第三章的数据预处理和指标构建后，我们获取了 28 个有关驾驶行为的指标，如果全部用来实验理论会获取更好的实验结果，但考虑到时间复杂度这里有必要运用 PCA 算法进行降维处理，尽可能地用较少地数据体现较多的信息量。

前文说过，数据之间的相关系数如果过低，数据不适合做主成分分析，本文选取两个方法判断数据是否适合做 PCA，分别是观察 KMO 值和进行 Bartlett 球形检验，本文数据集的 KMO 值为 0.566>0.5，球形检验 P 值<0.05，因此我们认为本文选取的数据的指标之间存在相关性，比较适合做主成分分析[54]。

表 4.1 是通过式（4.6）和式（4.7）计算得到的主成分贡献率和累计贡献率。观察累计贡献率可以发现前六个主成分的累计贡献率达到了 83.567%，大于 80%，可大致反映原始数据信息，且这六个主成分的特征值均大于 1，因此在后续的聚类分析实验中将选取这 6 个主成分的得分作为输入变量。

表 4.1 主成分特征值及贡献率表

主成分	特征值	贡献率/%	累计贡献率/%
F1	8.198	29.279	29.279
F2	5.13	18.322	47.601
F3	4.251	15.181	62.783
F4	2.893	10.331	73.114
F5	1.684	6.014	79.128
F6	1.243	4.439	83.567
F7	0.876	3.127	86.694
F8	0.777	2.777	89.47
F9	0.559	1.998	91.469
F10	0.525	1.877	93.345
F11	0.424	1.514	94.859
F12	0.396	1.416	96.275
F13	0.275	0.982	97.258
F14	0.204	0.729	97.987
F15	0.163	0.581	98.568
F16	0.123	0.438	99.005
F17	0.065	0.234	99.239
F18	0.058	0.208	99.448
F19	0.043	0.155	99.603
F20	0.036	0.127	99.73
21	0.026	0.092	99.822
22	0.021	0.076	99.897
23	0.015	0.054	99.951

主成分	特征值	贡献率/%	累计贡献率/%
24	0.01	0.036	99.987
25	0.002	0.006	99.993
26	0.001	0.005	99.998
27	0.001	0.002	100
28	0.000	0.000	100

碎石图是根据各主成分对数据变异的解释程度绘制的图。图 4.2 是本文获取的碎石图，图中的每个点表示一个主成分极其特征，我们认为当碎石图趋向平缓时的主成分点即为选取的主成分数。在本研究中，我们可以发现当主成分的数达到 6 时，图形趋于平缓，因此我们认为可以提取前 6 个主成分。

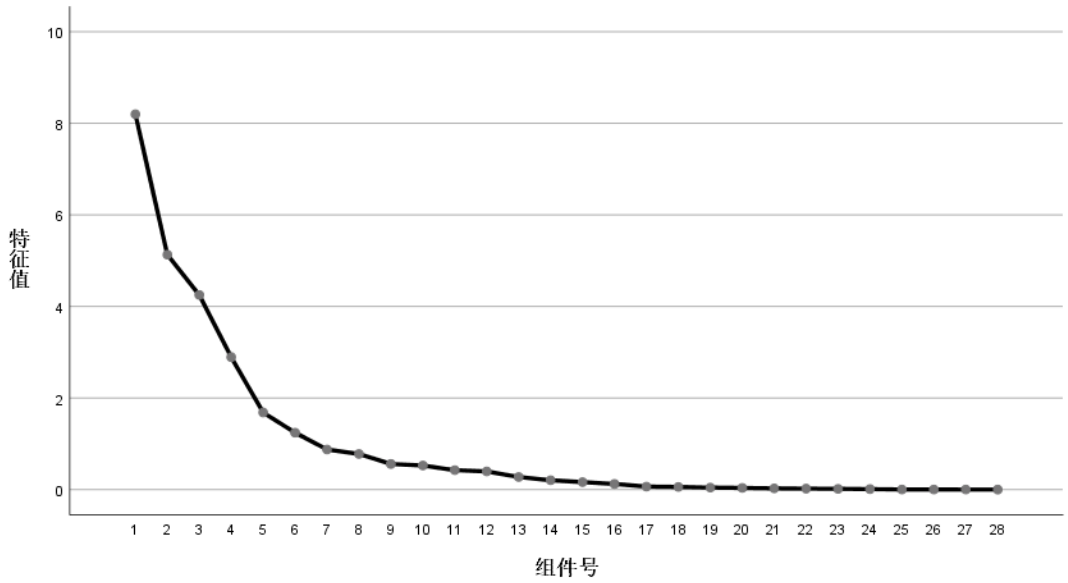

图 4.2 碎石图

经过以上处理，原有的 28 个变量被替换成了 6 个主成分，表 4.2 为计算得出的载荷表，主成分的载荷表示各变量和主成分的关系，载荷的绝对值越大，其与该主成分的相关性就越强。本文认为相关系数绝对值大于 0.5 就说明两者之间存在一定相关性，可以得到以下结论：主成分 F1 和 F5 中相关系数绝对值大于 0.5 的指标有匀速比例、正加速度标准差、中等加速度、急减速、平均减速度、中游门踏板开度、平均加速度、负加速度标准差、中等减速占比、速度平均值和正常减速占比，这些指标除去速度平均值以外都直接或间接反映了正负加速度值及变化情况，它们可以代表高加速度或高减速度驾驶行为；主成分 F2 中相关系数绝对值大于 0.5 的指标有油门踏板开度平均值、油门踏板开度标准差、高油门踏板开度占比、低油门踏板开度占比、超高油门踏板开度占比、油门踏板角速度平均值和油门踏板角速度标准差，值得注意的是低油门踏板开度占比的载荷值是-0.765，表明低油门踏板开度占比越小，该主成分的最终得分越高，因此整体来

看主成分 F2 反映了油门踏板开度和角速度的值，代表了行程的高油门驾驶行为；主成分 F3、F4 和 F6 中相关系数绝对值大于 0.5 的指标有高速占比、中速占比、速度标准差、行驶速度平均值、行驶速度标准差、速度平均值、低速、怠速比例和正常减速占比，同样的中速和低速的载荷值为负数，因此除去怠速比例和正常减速占比这两个指标，其他指标均反映了高速度值和高速度变化情况，可以代表行程的高速度驾驶行为。

表 4.2 载荷表

参数	成分					
	F1	F2	F3	F4	F5	F6
速度平均值	-0.514	0.345	0.588	0.504	0.043	-0.01
速度标准差	0.155	0.362	0.654	-0.506	0.052	-0.095
行驶速度平均值	-0.42	0.419	0.73	0.224	0.076	-0.143
行驶速度标准差	0.157	0.358	0.608	-0.447	0.012	0.088
低速	0.317	-0.294	-0.515	-0.194	-0.136	0.636
中速	-0.102	-0.27	-0.52	0.517	0.075	-0.5
高速	-0.322	0.422	0.744	0.152	-0.009	0.054
超高速	-0.302	0.22	0.427	0.004	0.061	0.415
平均加速度	0.927	-0.063	0.236	0.161	0.111	0.087
平均减速度	-0.81	-0.247	-0.036	-0.052	0.492	0.074
正加速度标准差	0.878	-0.077	0.291	0.147	0.157	0.064
负加速度标准差	0.703	0.251	0.19	0.14	-0.447	-0.151
正常减速占比	0.101	-0.239	0.153	0.613	0.649	0.173
中等减速占比	0.628	0.087	-0.244	0.218	-0.17	0.127
急减速	0.813	0.17	0.126	0.313	-0.339	0.007
正常加速占比	-0.442	0.301	-0.362	0.454	-0.313	-0.089
中等加速占比	0.875	0.012	0.135	0.32	-0.028	0.108
急加速	0.823	-0.078	0.296	0.253	0.039	0.148
油门踏板开度平均值	-0.449	0.682	-0.228	0.179	-0.198	0.237
油门踏板开度标准差	0	0.865	-0.41	-0.061	0.122	0.056
油门踏板角速度平均值	0.486	0.689	-0.175	0.137	0.276	-0.172
油门踏板角速度标准差	0.378	0.764	-0.246	0.058	0.311	-0.099
低油门踏板开度	0.065	-0.765	0.2	0.257	0.028	0.054
中油门踏板开度	-0.57	-0.183	0.401	0.198	-0.451	0.039
高油门踏板开度	-0.216	0.736	-0.209	0.116	-0.147	0.04
超高油门踏板开度	-0.002	0.655	-0.482	-0.064	0.181	0.232
怠速比例	0.415	-0.006	0.037	-0.812	0.042	-0.253
匀速比例	-0.89	-0.078	0.003	0.048	-0.119	0.16

通过式（4.8）可计算不同主成分参数的系数值，其中PC表示各参数的系数矩阵，l_k表示第k个主成分下所有参数的载荷值，λ_k表示第 k 个主成分的特征值，

$$PC = \frac{l_k}{\sqrt{\lambda_k}} \tag{4.8}$$

每个主成分还可以计算其相应的主成分得分。主成分的得分等于主成分系数矩阵乘以数据标准化矩阵，计算方式为式（4.9），其中$Score$为主成分得分，Y为数据标准化矩阵，PC为主成分系数矩阵，计算获得的主成分得分将用于后续的聚类分析。

$$Score = Y \times PC \tag{4.9}$$

4.2.2 驾驶行程片段聚类

在现实的行驶过程中，驾驶员往往会根据当时的驾驶条件呈现不同的驾驶风格，通过驾驶风格的判定，可分类讨论不同情况下的驾驶行为对于油耗的影响。本文通过聚类分析，总结有类似特征的行程，将其归为一类，具体做法就是通过K-means 聚类算法与 PSO 算法结合实现。

K-means 均值聚类算法的核心思想是把用于聚类的数据集中的数据代表的点划分到目标数量的聚类中，计算出每个聚类的均值（此即聚类中心），使得每个数据点都属于离它距离最近的聚类中心对应的聚类。在本研究中选择欧式距离作为距离的度量。设 n 为数据空间维数，聚类中心$K = (k_1, k_2, \dots, k_n)$，任意一数据点$X = (x_1, x_2, \dots, x_n)$到 K 的欧式距离如公式（4.10）所示：

$$d(X, K) = \sqrt{\sum_{i=1}^{n} (x_i - k_i)^2} \tag{4.10}$$

K-means 算法虽然操作起来方便且原理易于理解，但其初始聚类中心是随机选择，后续的聚类结果与初始聚类中心的选择有很大的关系，因此本文选用 PSO 粒子群算法对初始聚类中心进行迭代选择。粒子群算法始于鸟群捕食模拟研究。假定一个场景，一个区域内有一块食物，鸟儿不知道食物的具体位置，但大致知道自己和食物之间的距离，寻找该食物最简单的方式就是寻找距离食物最近的鸟的周围区域。粒子群算法受此启发，将每个鸟看作是一个粒子，所有粒子都有一个用于优化的适应度函数，每个粒子还有特定的速度决定它们的飞翔方向，然后就不断向着自己的最优目标进行搜寻[61]。

PSO 首先会初始化一些初始解，然后通过迭代寻找最优解，在每一次迭代中，粒子会通过两个极值来更新自己的速度和距离，第一个就是粒子依靠自身记忆找到的最优解，记为个体极值$pBest$，另一个是通过集群共享信息找到的最优解，记为全局极值$gBest$。

具体步骤如下：

Step1：设置粒子的个数，随机产生一些初始粒子，这些粒子的初始速度v_i和初始位置X_i也随机产生，作为当前解集，并设置最大迭代次数。

Step2：计算当前种群里每个解的适应度，即目标函数值。

$$f(x)=\sum_{j=1}^{k}\sum_{S_i\in C_j}\left\|S_i-Z_j\right\| \tag{4.11}$$

其中C_j为 k 个聚类中心对应的 k 个类别；S_i为类C_j中的其他所有点；Z_j为聚类中心。通过判断适应度方差和阈值θ的大小来判断是否结束循环，方差σ^2计算公式如下：

$$\sigma^2=\frac{1}{m}\sum_{i=1}^{m}\left[f\left(x_i-f_{avg}\right)\right]^2 \tag{4.12}$$

其中$f(x_i)$为粒子i的适应度值，f_{avg}为所有粒子的适应度均值。

Step3：利用每个粒子的个体极值$pBest$和全局极值$gBest$更新粒子的速度和距离，具体如式（4.13）和式（4.14）所示：

$$v_i=\omega\times v_i+c_1\times rand()\times(pBest_i-x_i)+c_2\times rand()(gBest_i-x_i) \tag{4.13}$$

$$X_i=X_i+v_i \tag{4.14}$$

其中，c_1为惯性因子，c_2为约束因子，ω为惯性权重，通常其取值是一个常数项，主要目的是用于约束函数，$rand()$为 0 到 1 之间的随机数。

Step4：计算当前种群离每个解的适应度值。

Step5：循环执行步骤 3-4，直至达到最大迭代次数或满足适应度函数的约束条件。

综上 PSO-Kmeans 具体流程图 4.3 如下：

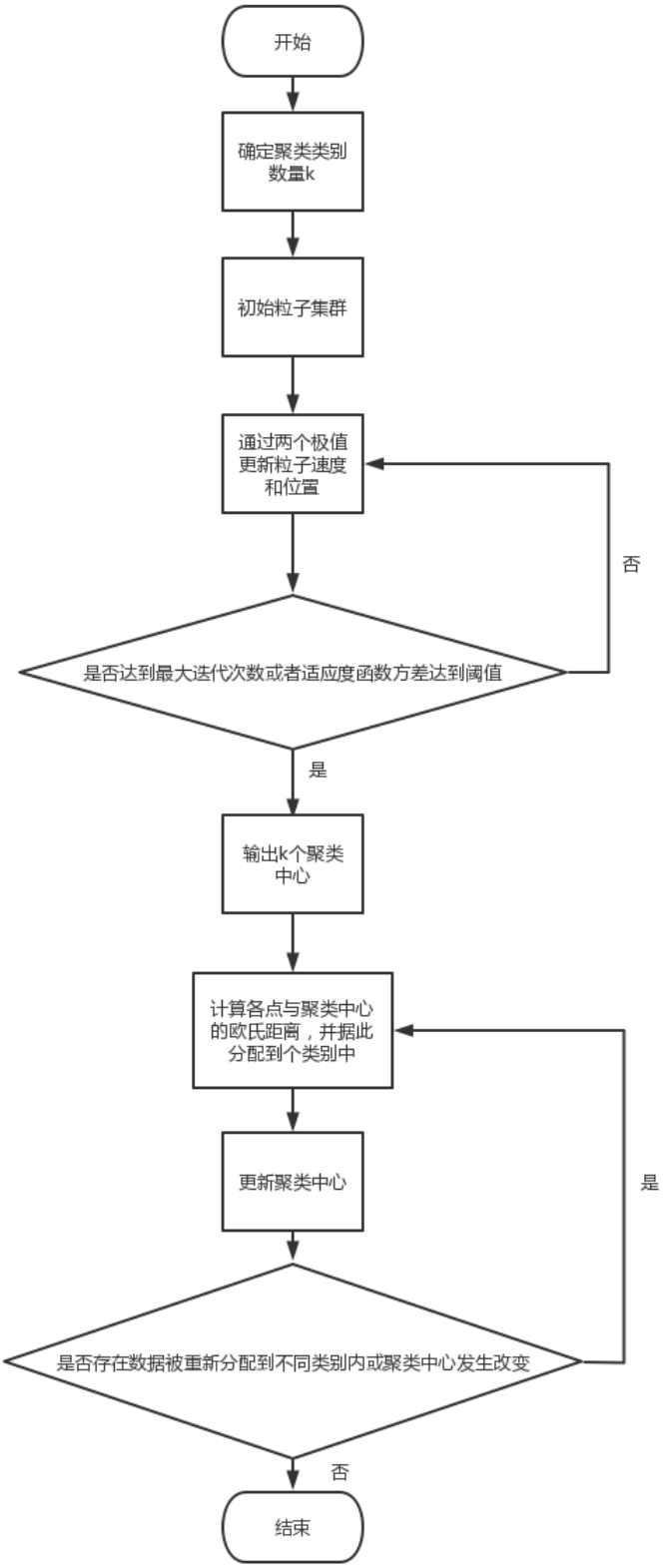

图 4.3 PSO-Kmeans 算法流程图

PSO-Kmeans 算法中有一些参数需要人工赋值，比如c_1和c_2的大小，一般取值都为 2.0，本文最终选取的c_1和c_2大小均为 1.49；ω的值本文采用模糊控制的方

式设定，即将其确定为 0.72；种群规模大小为 30；最大迭代次数为 1000 次；聚类类别 k=3。

对获取的数据进行预处理，提取有关驾驶员驾驶行为的各类指标，以此为基础通过 PCA 主成分分析获取新的主成分以代替原有数据，再通过 PSO-Kmeans 进行聚类分析讨论获得数据集驾驶风格标签。基于 PCA 分析后的主成分再进行聚类分析，将所有样本聚类为三类，考虑到聚类可视化结果在多维指标的情况下无法展示，因此挑选了贡献率最大的 F1、F2 和 F3 分别展示聚类的二维以及三维结果图，具体结果如图 4.4 和图 4.5 所示：

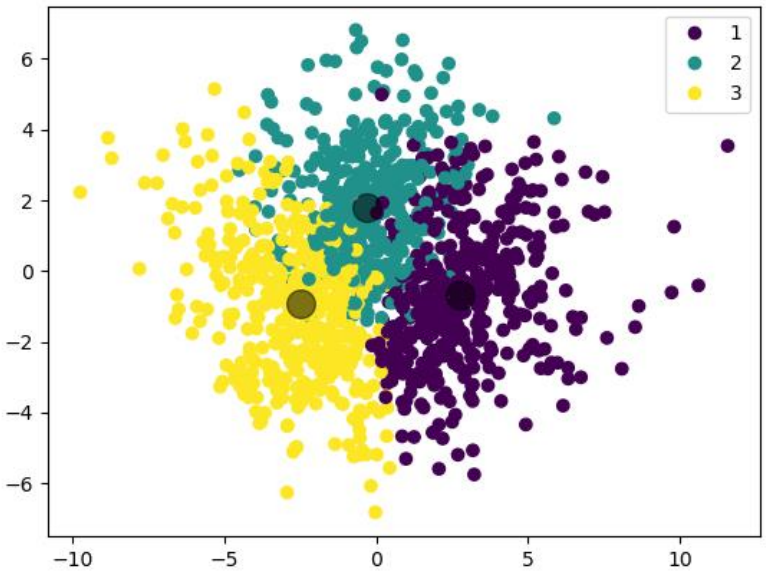

图 4.4 二维聚类结果图

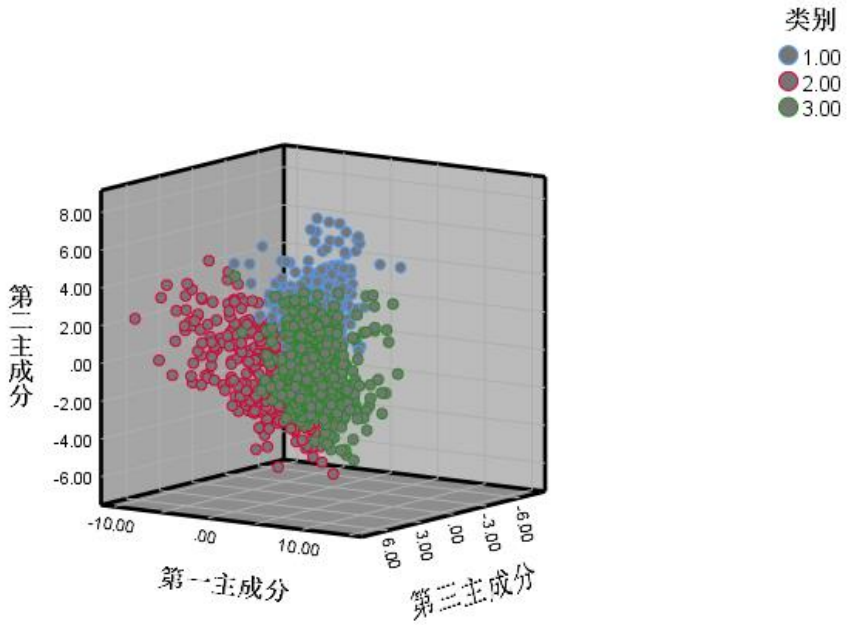

图 4.5 三维聚类结果图

表 4.3 为基于不同类别下的各指标平均统计值，后续将参考这些数据分析不同类别的驾驶风格。

表 4.3 各指标平均值表

参数（平均值）	1	2	3
速度平均值	36.28	42.77	35.16
速度整体标准差	21.63	21.74	23.27
行驶速度平均值	41.77	47.19	41.94
行驶速度整体标准差	18.47	18.53	19.79
低速占比	0.24	0.2	0.25
中速占比	0.46	0.45	0.4
高速占比	0.16	0.24	0.18
超高速占比	0	0.02	0
平均加速度	0.22	0.2	0.31
平均减速度	-0.3	-0.25	-0.34
加速度标准差	0.13	0.1	0.21
减速度标准差	0.26	0.22	0.31
正常减速占比	0.19	0.2	0.23
中等减速占比	0.04	0.03	0.04
急减速占比	0.03	0.02	0.04
正常加速占比	0.3	0.28	0.23
中等减速占比	0.04	0.02	0.08
急减速	0.01	0	0.03
油门踏板开度平均值	24.57	19.76	15.08
油门踏板开度整体标准差	28.93	18.8	19.4
油门踏板角速度平均值	7.97	4.88	6.98
油门踏板角速度整体标准差	8.49	5.13	6.87
低油门踏板占比	0.15	0.26	0.27
中油门踏板占比	0.16	0.3	0.18
高油门踏板占比	0.14	0.07	0.06
超高油门踏板占比	0.09	0.01	0.02
怠速比例	0.14	0.1	0.16
匀速比例	0.27	0.35	0.19

可以看出类 1 和类 3 有着差不多速度分布，且相对速度较低，但类 3 却有着最高的加速和减速度，匀速行驶占比较少，趋向于变速驾驶，同时类 1 有着较高的油门踏板开度，因此可以认为类 1 为猛踩油门型驾驶风格，其占比是 30.4%；类 2 关于油门踏板和加速度的数值较小，但有着最高的车速，因此可以认为是高速行驶型驾驶风格，占比为 34.7%；类 3 的加减速度变化最明显，因此为频繁变速型驾驶风格，占比为 34.9%。

4.3 驾驶风格识别模型构建与仿真实验

4.3.1 驾驶风格识别模型构建

建立驾驶风格识别模型主要是为了在油耗分析时，排除不同驾驶风格对油耗分析结果的影响，因此需要建立驾驶风格识别模型，以识别未知行程片段的驾驶风格类型。本文通过支持向量机算法来实现驾驶风格的识别。

（1）基本原理

支持向量机以统计学习理论为基础，在处理小样本、非线性及高位空间模式中的分类问题时，表现出良好的分类效果。SVM 是从线性可分情况下的最优分类平面发展而来的，其基本思想是将数据映射到某个高位空间中，增强数据的可分性，使得原本线性不可分的样本在高维空间中线性可分，在高位空间构造最优超平面完成分类。

其基本形式如下：

$$\min_{\omega,\ b}\frac{1}{2}\|\omega\|^2 \tag{4,15}$$

$$s.t.\ y_i(\omega\cdot x_i+b)-1\geq 0, i=1,2,\dots,N$$

其中ω为法向量，b 为固定常数，其本质是一个凸二次规划问题。

在实际运用中，大部分研究问题都很难是线性的，当遇到非线性的问题时就需要引入误差的概念，也就是通过加入松弛变量ξ_i，$i=1,2,\dots,N$和惩罚参数 C>0 来解决这个问题。C 值大时对误分类的惩罚增大，C 值小时对误分类的惩罚减小。

在本文中，我们为了研究驾驶风格影响因子而构建了 28 个维度的指标，同时这些指标之间也存在着相互关联，因此，我们可将行程片段分类问题看作是一个非线性支持向量机分类问题。

对于非线性问题，通常我们会引入核函数，转化为在其他高位特征空间中的线性的问题，而核函数的选择则会影响模型最后的分类结果 ，常用的核函数有以下两种：

（a）多项式核函数$K(x,z)=(x\cdot z+1)^p$

（b）高斯径向基核函数$K(x,z)=exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$

对于上述两类核函数，多项式核函数虽然能够将原始空间映射到高位空间，但是当样本的维数较高时，分类算法的性能都不到改善，因此选用高斯径向基核函数对原始样本进行空间映射。

（2）支持向量机多分类问题

本文的驾驶风格分为了三类，分别是猛踩油门型、高速行驶型和频繁变速型驾驶风格，这是一个三分类问题，无法用传统的支持向量机二分类思路来解决，当处理多分类问题时，就需要构造合适的多类分类器。构造多分类器可以采用直接法或者间接法。若采用直接法对目标函数进行修改，涉及的计算量较大，不适合数据量大的问题。因此本文将选取间接法进行多分类。间接法目前分为两类，

分别是一对多和一对一。

一对多（one versus rest，简称 OVR SVMs），此方法训练时把某个类别的样本归为一类，其他剩余的样本归为另一类，这样 k 个类别的样本就构造出了 k 个 SVM。分类时将位置样本分类为具有最大分类函数值的那类。举个例子，假如我有三类要划分，它们是 1，2，3，于是抽取训练集的时候，分别选取三个训练集如下：

（a）1 所对应的向量为正极，2，3 所对应的向量为负极。

（b）2 所对应的向量为正极，1，3 所对应的向量为负极。

（c）3 所对应的向量为正极，1，2 所对应的向量为负极。

使用这三个训练集分别进行训练，然后得到三个训练结果，在测试的时候，把对应的测试向量分别利用这三个训练结果进行测试，最后模型都会有一个决策函数值$f_1(x)$，$f_2(x)$，$f_3(x)$，最终结果便从这三个中选取最大的值所对应的类别作为分类结果。

一对一（one versus one，简称 OVO SVMs），其做法是在任意两类样本之间设计一个 SVM，因此 k 个类别的样本就要设计$k(k-1)$个 SVM，当对一个位置样本进行分类时，最后得票最多的类别即为该未知样本的类别。举个例子，假如有三类，别分是 1，2，3，这三类可组成的训练集为 12，13，23 三种，因此可以得到三个训练结果，此时再将测试集的向量代入训练结果进行测试，采取投票形式，最后得到一组结果。具体情况可以概述成这样：12 分类器中，如果测试集被分为 1 则 1 的投票加一，否则 2 的投票加一；13 分类器中，如果测试集被分为 1 则 1 的投票加一，否则 3 的投票加一；23 分类器中，如果测试集被分为 2 则 2 的投票加一，否则 3 的投票加一，这样的分类结果与一对多相比不会产生有样本不属于任何一类的问题，且本文中研究的问题是三分类问题，符合一对一方法对于类别数目不能过多的要求，因此选择一对一方法进行分类。

4.3.2 驾驶风格识别仿真实验

SVM 常用的核函数有多项式核函数和 RBF 核函数，通过上述比较，本文选择了 RBF 核函数。在实验之前，考虑到 SVM 容易受特征变量量纲不同的影响，因此这里对 28 个驾驶行为指标数据进行了标准化处理。经过参数优化，惩罚系数为 100，gamma 为 0.001，对数据集进行随机划分，训练集占 70%，测试集占 30%，测试集的准确率为 96.5%，具有良好的预测准确率，其混淆矩阵如表 4.4 所示：

表 4.4 混淆矩阵表

驾驶风格	高速行驶型	频繁变速型	猛踩油门型	行总计
高速行驶型	175	1	7	183
频繁变速型	1	114	2	117
猛踩油门型	2	0	74	76
列总计	178	115	83	376

4.4 本章小结

本章通过主成分分析 PCA，实现数据降维，达到降低时间复杂度的目的，并且用 PSO-Kmeans 聚类算法对驾驶风格进行分类，将驾驶风格分为猛踩油门型、高速行驶型以及频繁变速型，然后选取 SVM 支持向量机算法进行驾驶风格识别实验，利用混淆矩阵评估模型的准确性，实验结果表明模型的准确率高达 96.5%。

5 基于驾驶风格的油耗预测

通过第四章的驾驶风格分类，我们提取了三种极具特性的驾驶风格，分别是高速行驶型、频繁变速型以及猛踩油门型，基于这三类驾驶风格的油耗模型可帮助物流企业分析不同驾驶员的燃油经济性，同时考虑到驾驶行为会对车辆的燃油消耗产生一定影响[1]，因此结合驾驶行为和驾驶风格进行油耗分析后，可根据实际情况提出节能驾驶策略。

5.1 基于驾驶风格的油耗预测模型整体框架

图 5.1 给出了油耗预测模型的整体框架，从整个框架可以看出，该模型主要分成三部分：数据预处理、建模以及特征重要性分析。其中数据预处理环节主要是在第三章构建的 28 个指标的基础上，加入百公里油耗值。至于建模部分采用了集成学习算法构建油耗预测模型。特征重要性分析主要是运用Shapley value算法研究油耗影响因子。

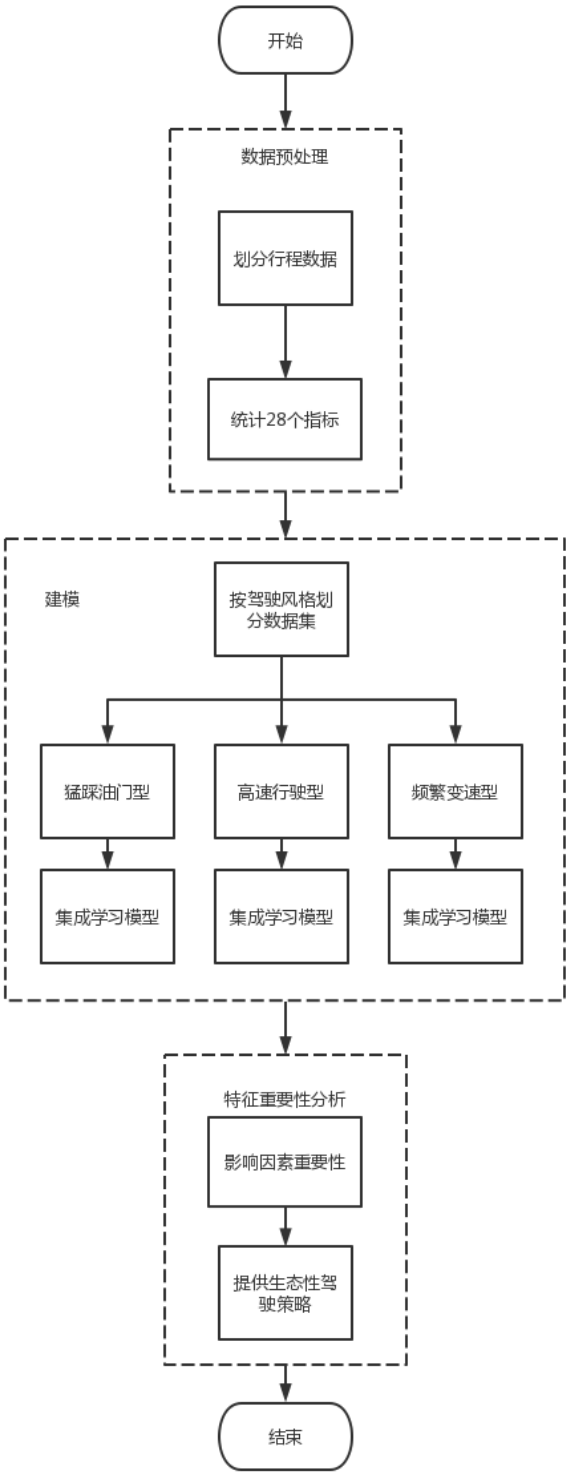

图 5.1 模型整体框架图

5.2 不同驾驶风格的油耗预测模型构建

5.2.1 随机森林模型

随机森林模型因其并行的集成结构，具有运算效率快和运算精度高的特点，

将其运用在油耗预测的问题上，能快速获取精确的预测值，同时考虑到模型本身的基分类器是决策树模型，因此还可以通过计算特征重要性来分析驾驶行为对于油耗的影响。

随机森林是一个基于决策树的集成学习模型。基础随机森林的基分类器是决策树，使用 Bagging 抽样的方法建立多颗随机划分特征的决策树，能有效抵抗过拟合，因此分类效果要好于其他分类算法，随机森林的预测结果通常由多个决策树同时决定，即取每个弱学习器预测结果的平均值，式（5.1）为其具体表达式，其中K表示K个决策树，$\mathcal{F}$代表一组决策树,$f_k(x_i)$表示第k个决策树的预测结果：

$$\hat{y}_i = \frac{1}{K}\sum_{K=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{5.1}$$

考虑到本文研究的是回归问题，因此在特征选取和内部结点切分时应选取 CART 算法。针对于切分变量和切分点的选择，本文采用穷举法，即遍历每个特征和每个特征的所有取值，最后从中找出最好的切分变量和切分点；针对于切分变量和切分点的好坏，一般以切分后节点的不纯度来衡量，即各个子节点不纯度的加权和$G（x_i, v_{ij}）$，其计算公式如下式（5.2）所示：

$$G(x_i, v_{ij}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H(X_{right}) \tag{5.2}$$

其中，x_i为某一个切分变量，v_{ij}为切分变量的一个切分值，n_{left}，n_{right}，N_s 分别为切分后左子节点的训练样本个数、右子节点的训练样本个数以及当前节点所有训练样本个数，X_{left}，X_{right}分为左右子节点的训练样本集合，$H(X)$为衡量节点不纯度的函数(impurity function/criterion)，本文中我们使用平方误差 MSE 做为不纯度函数，其计算公式如下式（5.3）所示：

$$H(X) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{5.3}$$

代入上式可得某切分点的不纯度函数为式（5.4）：

$$G(x, v) = \frac{1}{N_s}\left[\sum_{\hat{y}_i \epsilon X_{left}} (\hat{y}_i - y_{left})^2 + \sum_{\hat{y}_i \epsilon X_{right}} (\hat{y}_i - y_{right})^2\right] \tag{5.4}$$

5.2.2 XGBoost 模型

极端梯度增强模型是一种串行的集成学习模型，相比较随机森林模型，其计算效率较低，但考虑到其模型内部是一个不断迭代升级的过程，因此其模型准确率普遍优于随机森林模型，得益于此，该模型在金融、交通、天气等领域运用广泛，同样的，其也可以通过计算特征重要性来分析驾驶行为对于油耗的影响。

XGBoost 是基于 CART 树的一种 Boosting 算法，该算法的目标函数采用的

是加法训练，多个决策树串联在一起，利用梯度提升算法减少前一棵决策树的损失，生成一棵新的决策树，从而形成新的预测模型，循环持续到损失函数小于期望值为止，与传统机器学习方法相比，XGBoost 算法有着更高的准确率，因此在能源行业运用广泛。

XGBoost 首先是一种基于决策树的集成模型，假设有K棵 CART 树，则集成的预测结果为式（5.5）：

$$\hat{y}_i = \sum_{K=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{5.5}$$

K代表了决策树的数量，x_i代表第i次的输入变量，$\mathcal{F}$代表一组决策树，f_k代表每棵树的预测结果。

本文的目的是预测重型货车的油耗，XGBoost 算法在训练模型的过程中需要使用回归型的目标损失函数，目标损失函数越小，预测准确率越高。目标损失函数不仅会考虑训练误差，还增加了一个正则化式（5.6），其中$\Omega(f_k)$的表达式为式（5.7）：

$$f_{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{n} \Omega(f_k) \tag{5.6}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{5.7}$$

f_{obj}代表目标损失函数，$l(y_i, \hat{y}_i)$代表训练误差，$\hat{y}_i$和y_i分别是样本i的预测值和实际值，$\Omega(f_k)$是第k个回归树的正则项，T代表 CART 树的叶节点数，ω_j是第j个叶节点的权重，γ 表示节点切分的难度，λ 表示 L2 正则化系数。

XGBoost 目标函数的优化过程是逐级进行的，即通过前一棵树的学习结果优化后一颗树的训练样本权重，不存在两棵树同时运行的情况，具体计算过程如下所示：

$$\hat{y}_i^{(0)} = 0 \tag{5.8}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{5.9}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{5.10}$$

…..

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5.11}$$

当 XGBoost 使用训练数据集进行学习时，原始模型每次都保持不变，并以迭代方式添加新函数。如果新添加的函数可以降低损失函数的分数，则会将其添加到模型中，直到生成最终的集成模型。将该式代入目标损失函数可得：

$$f_{obj}{}^{t} = \sum_{i=1}^{n} l[(y_i, \hat{y}_u^{(t-1)}) + f_t(x_i)] + \sum_{i=1}^{n} \Omega(f_k) \tag{5.12}$$

$$= \sum_{i=1}^{n} l[\left(y_i, \hat{y}_u^{(t-1)}\right) + f_t(x_i)] + \sum_{i=1}^{n} \Omega(f_k) + Constant \tag{5.13}$$

其中$f_t\ (x_i)$代表添加了t次新函数，$Constant$是一个独立于f的常数项，一般用字母C代替。

引入二阶泰勒公式对目标损失函数进行计算。其定义如下式（5.14）所示：

$$f(x + \Delta x) \cong f(x) + \dot{f}(x)\Delta x + \frac{1}{2}\ddot{f}(x)\Delta x^2 \tag{5.14}$$

其中$\dot{f}(x)$表示一阶导数，$\ddot{f}(x)$表示二阶导数，展开后获得损失函数的近似表达式（5.15）：

$$f_{obj}{}^{t} \cong \left[l\left(y_i, \hat{y}_u^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \sum_{i=1}^{n} \Omega(f_k) + C \tag{5.15}$$

$$g_i = \partial_{\hat{y}^{(t-1)}}\left(y_i, \hat{y}_i^{(t-1)}\right), h_2 = \partial^2_{\hat{y}^{(t-1)}}(y_i, \hat{y}_i^{t-1}) \tag{5.16}$$

$$G_j = \sum_{i \epsilon I_j} g_i\,, H_j = \sum_{i \epsilon I_j} h_i \tag{5.17}$$

简化后得到目标损失函数的最终形式：

$$f_{obj}{}^{t} = \sum_{j=1}^{T}\left[G_j \omega_j + \frac{1}{2}\sum_{i \epsilon I_j}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma \mathrm{T} \tag{5.18}$$

对目标损失函数关于ω_j求偏导等于 0，此时损失函数达到最小值，可得：

$$\omega_j^* = \frac{-G_j}{H_j + \lambda} \tag{5.19}$$

最终得到目标损失函数的最优形式：

$$f_{obj}{}^{t} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma \mathrm{T} \tag{5.20}$$

同样的，对于 XGBoost 我们仍然使用了平方误差 MSE 作为其不纯度函数。

5.2.3 模型比较

（1）数据预处理

通过数据预处理获取 28 个驾驶行为特征指标，并评价燃油经济性。燃油经济性是指以最小的燃油消耗量完成单位运输工作的能力。如表 5.1 所示，燃油经济性有三个评价指标：单位行驶里程的燃料消耗量、单位运输工作量的燃料消耗

量、消耗单位燃油所行驶的里程，我国通常选取单位行驶里程的燃油消耗量作为燃油经济性的评价指标。

表 5.1 燃油经济性评价指标表

名称	单位	说明
单位行驶里程的燃油消耗量	L/100km	车辆行驶百公里的单位油耗
单位运输工作量的燃油消耗量	L/100t·km	车辆以单位工作量和行驶里程的油耗
消耗单位燃油所行驶的里程	km/L	车辆单位油耗所行驶的里程

本文以单位行驶里程的燃料消耗量作为车辆燃油经济型的评估指标，具体计算方法如下：

式（5.21）为行驶总油耗计算公式：

$$F = Fuel_{end} - Fuel_{start} \tag{5.21}$$

式中，$Fuel_{end}$是行程结束时仪表盘的油耗值，$Fuel_{start}$是行程开始时仪表盘的油耗值，这三个值的单位都是L。行驶距离计算过程也类似，其单位为km。式（5.22）为总距离计算公式：

$$D = D_{end} - D_{start} \tag{5.22}$$

则某行程片段的百公里油耗计算公式如式（5.23）所示：

$$f_d = \frac{F}{D} \times 100 \tag{5.23}$$

同时在第四章中我们根据400万条车辆行驶数据截取了1252条有效行程段，并对其进行了驾驶风格分类，其中猛踩油门型驾驶风格的行程段有381个，高速行驶型驾驶风格的行程段有434个，频繁变速型驾驶风格的行程段有437个。分别对三个数据集进行训练集和测试集的切分，切分比为 7：3，获得三类训练数据集和测试集。

（2）模型最优参数搜索

均方根误差（RMSE）和 R 平方值（R^2）通常用作模型预测准确率的评估指标。RMSE 则是预测值和实际值的均方根误差，RMSE 使得模型预测准确率结果看起来更加直观，R^2 用于评估模型与数据的拟合程度，R^2 值的范围为 0 到 1，R^2 值越大，模型的预测性能越好。这两种评价指标的计算方式如下所示：

式（5.24）为均方根误差的计算公式：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5.24}$$

决定系数（R-Squared score R2 score）。决定系数反映了因变量 y 的波动，有多少百分比能被自变量 x 的波动所描述。该参数可以用来判断统计模型对数据的拟合能力。设样本数据的实际值为$y_1, y_2, \ldots, y_n$，通过模型拟合出来的预测值为$\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n$。在计算决定系数时，需要求得实际值的平均值，计算公式如下：

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.25}$$

计算决定系数 R2 score 首先需要求得回归平方和与总离差平方和，两者的比值即为决定系数，式（5.26）为其计算公式：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2} \tag{5.26}$$

通过网格搜索的方式穷举遍历参数区间内的参数，利用十折交叉验证计算的 R^2 判断参数是否为最优。本文共构建了 3 个数据集，每个数据集都进行了 7：3 的训练集数据集切分，需要注意的是调参用到的训练集和测试集与后续构建模型用的是同一组，保证实验前后的逻辑一致性。采用十倍交叉验证选取随机森林模型和极端梯度增强模型最优参数，具体做法如下：

（a）随机森林 RF 参数调优

首先对随机森林进行参数调优，得益于 Python 的 SKlearn 模块的 Random Forest Regressor 模型，我们可以很方便地进行参数调优。网格搜索使用 GridSearchCV 函数对弱学习器个数 n_estimators、决策树最大深度 max_depth、内部节点再划分最小样本数 min_samples_split 进行搜索。n_estimators 的大小决定了弱学习器的数量，一般其设置的值越大，模型的准确率越高，但当 n_estimators 大于特定的值时，其带来的提升效果会变弱；max_depth 表示每棵树的深度，默认值为 None，如果选择默认值，其会一直分裂直至每个叶子都是“纯”的或者叶子中包含于 min_samples_split 个样本；min_samples_split 表示每个内部节点包含的最少的样本数。

依次对上述参数进行调参，获取最优参数，需要的注意的是每次获取一个最优参数后需及时对模型的参数进行更新。先对 n_estimators 从 20 到 200 以 20 为间隔进行参数搜索。然后对 max_ depth 从 5 到 30 依次进行调参。最后对 min_samples_split 从 2 到 10 依次进行参数搜索。最终结果如表 5.2 所示。

表 5.2 RF 最优参数表

RF	n_estimators	max_depth	min_samples_split	R^2score
猛踩油门型	60	22	3	72.14%
高速行驶型	140	7	2	64.84%
频繁变速型	160	21	5	74.56%

（b）极端梯度增强 XGBoost 参数调优

接着对 XGBoost 算法进行参数调优，同样的，在 SKlearn 模块中已经有独立的 XGBoost 包，可直接调包实现运算过程，本文的问题是回归问题，因此选用 XGBRegressor 模块包。与随机森林类似，极端梯度增强也需要选取弱学习器个数 n_estimators、最大深度 max_depth 以及最小子节点权重阈值 min_child_weight。

其中 n_estimators 在这里也表示弱学习器的个数，但考虑到 XGBoost 是串行的结构，因此也可以理解为迭代次数；同时与随机森林不同的是，XGBoost 判断子树停止分裂的依据是看最小子节点权重 min_child_weight 是否达到阈值。

与随机森林调参类似，依次对上述参数进行调参，获取最优参数。先对 n_estimators 从 20 到 200 以 20 为间隔进行参数搜索。然后对 max_ depth 从 5 到 30 依次进行调参。最后对 min_samples_weight 从 1 到 10 依次进行参数搜索。

表 5.3 XGBoost 最优参数表

XGBoost	n_estimators	max_depth	min_samples_weight	R^2score
猛踩油门型	100	8	2	77.71%
高速行驶型	80	5	9	69.17%
频繁变速型	100	7	8	75.29%

在网格调参得到最优模型参数之后，通过十折交叉验证获取了每个模型的决定系数 R2 score。可以看出，极端梯度提升算法 XGBoost 对于三个数据集的模型准确率均大于随机森林算法 RF，尤其是基于猛踩油门型的驾驶风格，XGBoost 的表现远优于 RF，因此在后续的仿真实验中，本文选取 XGBoost 算法构建最终的油耗预测模型，并基于此进行后续的驾驶行为节能分析。

5.3 不同驾驶风格的油耗预测仿真实验

5.3.1 模型可解释性实验设计

在研究中可以发现，对于传统的机器学习算法，如线性回归、逻辑回归模型等，针对简单的问题可以得到较为准确的计算结果，在应用中也可直观地看到哪个指标影响了结果，但其解决非线性问题时的能力较差。XGBoost 模型或其他复杂机器学习模型，虽有着较强的预测能力，但算法本质是黑箱模型，可解释性比较弱，在实际运用中，从驾驶员角度来看，模型只可以给出油耗预测值，无法向驾驶员解释什么样的驾驶操作会使得其油耗变高，这就要求模型具有很好的可解释性。因此如何透明化复杂的黑箱模型，也是本文研究的问题之一。

（1）特征重要性

不管是 XGboost 的打分函数还是 random forest 的不纯度函数，其本质上都可用来查看特征对于森林中所有树的影响，从而来衡量特征的重要性，它在训练后自动计算每个特征的得分，并对结果进行标准化，以使所有特征的重要性总和等于 1。但两种算法计算特征重要性的方法也有所不同，本文中 XGboost 使用了梯度提升算法计算了特征用于分割的平均信息增益（gain），从而得出特征重要性为式（5.27）：

$$importance = \frac{total_{gain}}{wight} \tag{5.27}$$

其中total_gain是在所有树中该特征用作分裂节点带来的 gain 值之和，weight

是在所有树中，该特征用作分裂节点的总次数。

随机森林则根据不纯度进行特征重要性计算，首先计算某一节点特征k的重要性为式（5.28）：

$$n_k = \omega_k \times G_k - \omega_{left} \times G_{left} - \omega_{right} \times G_{right} \tag{5.28}$$

其中，$\omega_k, \omega_{left}, \omega_{right}$分别为节点 k 以及其左右子节点中训练样本个数与总训练样本数目的比例，G_k, G_{left}, G_{right}分为为节点 k 以及其左右子节点的不纯度。知道每一个节点的重要性之后，即可通过以下式（5.29）得出某一特征的重要性：

$$f_i = \frac{\sum_{j \in nodes\ spilt\ on\ feature\ i} n_j}{\sum_{k \in all\ nodes} n_k} \tag{5.29}$$

基于决策树的集成模型确实能较为直观的展示出模型中的特征重要性，让我们知道特征变量对于模型预测值的影响程度，这虽然从一定程度上解决了部分机器学习模型不可解释的诟病，但这个影响程度结论只限于全局且无正负相关性，这启发了我们开展了后续工作，即探究局部个体中特征变量的重要性以及这些变量对于预测值的正负影响。

（2）Shapley value

Shapley value 起源与合作博弈论，描述了一群拥有不同技能的参与者为了集体奖励而相互合作最后实现奖励公平分配的现实场景，2017 年 Lundberg S M 首次将概念引入机器学习领域，其中参与者是模型的输入变量，集体奖励则是模型的预测值，分配奖励则是模型中的特征重要值，即 Shapley 值为式（5.30）：

$$\emptyset_i(v) = \sum_{S \subseteq n \backslash \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \big(v(S \cup \{i\}) - v(S)\big) \tag{5.30}$$

该式子经过拆解重写后，可得到式（5.31）：

$$\emptyset_i(v) = \frac{1}{|N|} \sum_{S \subseteq n \backslash \{i\}} \binom{|N| - 1}{|S|}^{-1} \big(v(S \cup \{i\}) - v(S)\big) \tag{5.31}$$

其中i代表第i个特征，N表示所有特征，$S \subseteq N \backslash \{i\}$表示除$i$以外的所有特征子集，v(S)给出了该子集的大小，$(v(S \cup \{i\}) - v(S))$表示特征 i 添加到该子集 S 的边际贡献，$\binom{|N|-1}{|S|}^{-1}$表示除特征 i 以外的所有剩余子集的排列可以有多少个的倒数，$\frac{1}{|N|}$则是为了平均特征规模的影响，表示特征i贡献了多少与特征规模无关。

举个例子，假设某个模型可以根据一个人的年龄 Age、性别 Gender 和工作 job 来预测该人的收入，具体可简化为如图 5.2 所示的结构，其种 f 代表选取的特征数量：

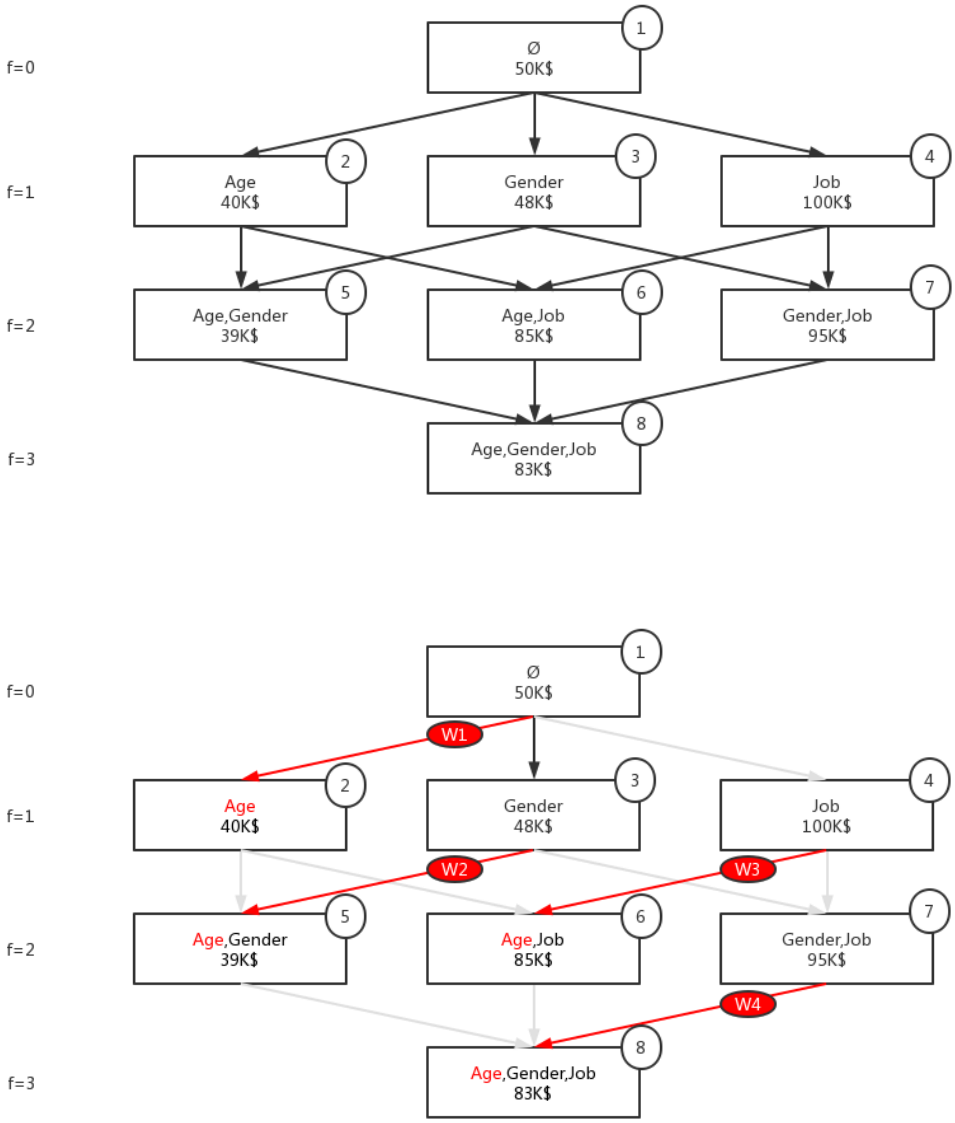

图 5.2 例子图

根据上式我们可以知道此时的N=3，S的排列组合可以有 4 种，分别是{∅}，{$Gender$}，{Job}，{$Gender$，Job}，首先计算边际贡献的权重，即式（5.32）：

$$\omega = \frac{1}{|N|}\sum_{S\subseteq n\backslash\{i\}}\binom{|N|-1}{|S|}^{-1} \tag{5.32}$$

再计算每种情况下的 Age 的边际贡献，最终将权重和其累成并累加获得了 Age 对模型的SHAP值：

$$SHAP_{Age} = \frac{1}{3}\times（-10）+\frac{1}{6}\times（-9）+\frac{1}{6}\times（-15）+\frac{1}{3}\times（-13）$$
$$= -14.68$$

5.3.2 油耗预测仿真实验

选取了 8 辆重型货车 2020 年 9 月三十天的车联网数据，对原数据切分后获取了 1253 段有效行程，统计每个行程有关驾驶行为的变量和百公里油耗值，根据驾驶风格和预测变量可获取 3 类数据集，它们分别是（1）猛踩油门型驾驶风格油耗数据集。（2）高速行驶型驾驶风格油耗数据集。（3）频繁变速型驾驶风格油耗数据集。选取 XGBoost 算法进行实验，并根据 Shapley value 分析不同驾驶风格下的影响因子。

（1）基于猛踩油门型驾驶风格的仿真实验

模型参数选取 5.3.5 中调好的数值，同时加入 Shapley value 算法分析不同特征对于预测结果的影响，测试集的回归预测评价指标如表 5.4 所示。

表 5.4 猛踩油门型驾驶风格的回归预测评价指标表

评价指标	RMSE	R^2
准确度（油耗）	7.96	0.81

测试集真实值和预测值散点图和折线图如图 5.3 和图 5.4 所示，其中靠近斜率为$k = 1$的直线的点越多，模型整体准确度越高。

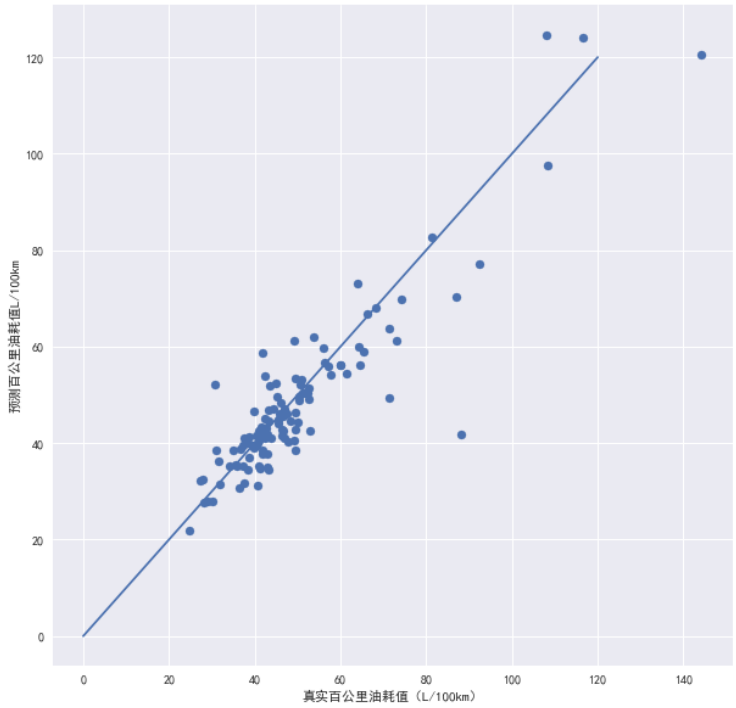

图 5.3 油耗预测值和真实值散点图

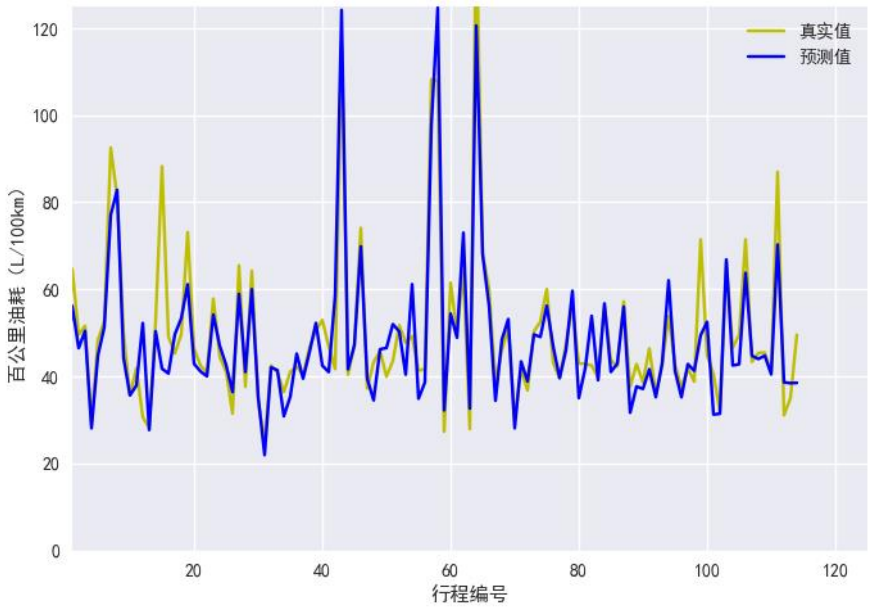

图 5.4 油耗预测值和真实值折线图

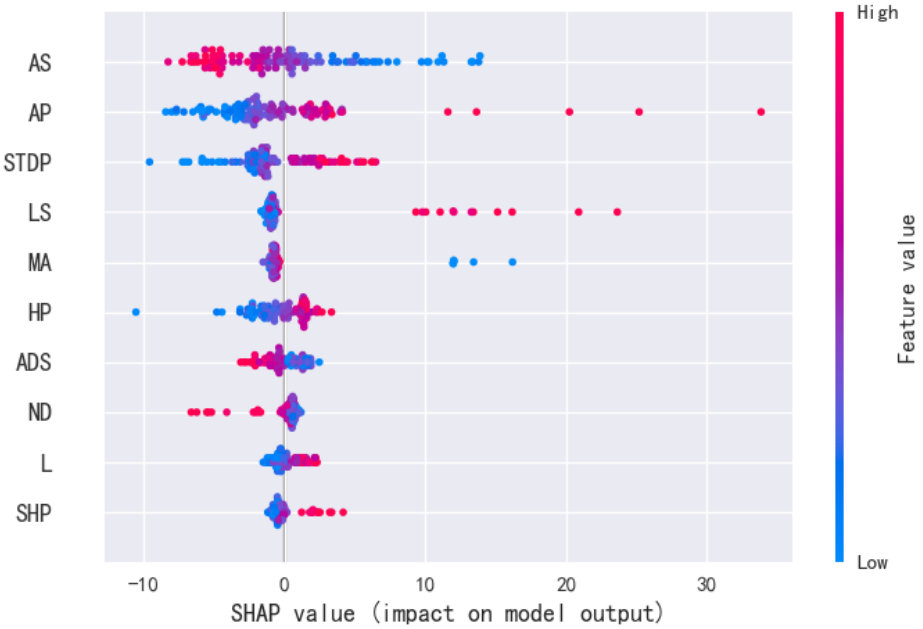

图 5.5 Shapley value 油耗特征分析图

（2）基于高速行驶型驾驶风格的仿真实验

测试集的回归预测评价指标如表 5.5 所示。

表 5.5 高速行驶型驾驶风格的回归预测评价指标表

评价指标	RMSE	R^2
准确度（油耗）	3.59	0.77

测试集真实值和预测值散点图和折线图如图 5.6 和图 5.7 所示：

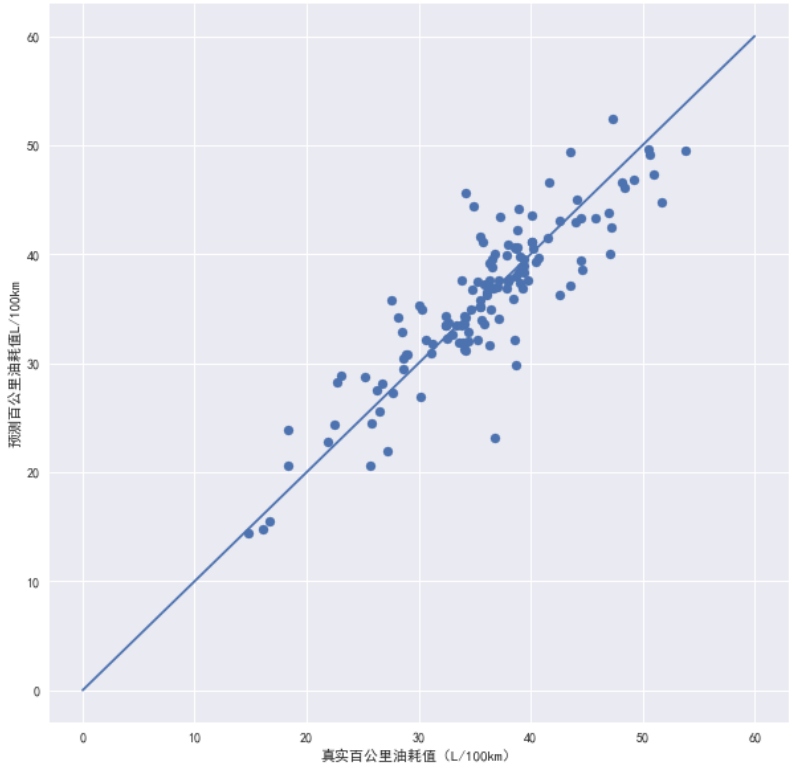

图 5.6 油耗预测值和真实值散点图

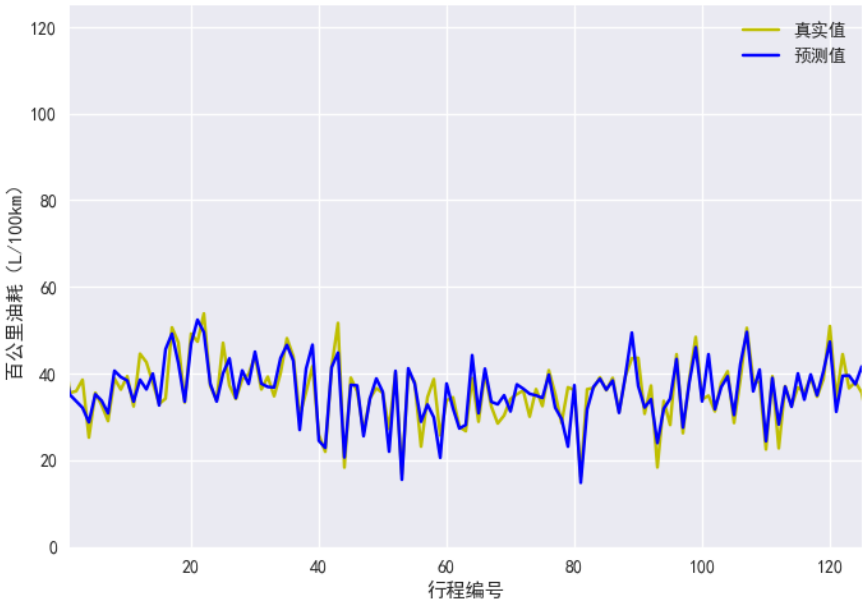

图 5.7 油耗预测值和真实值折线图

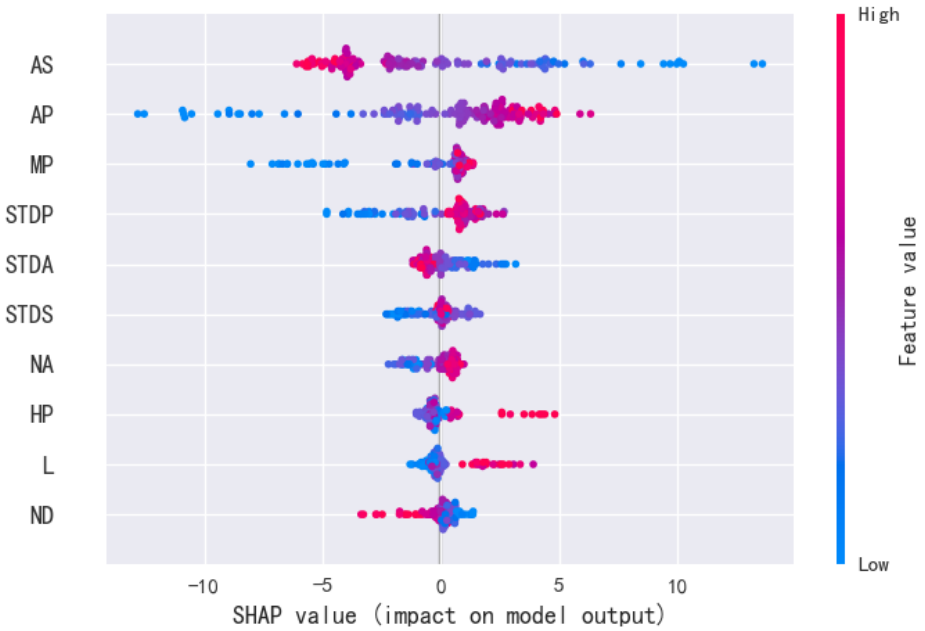

图 5.8 Shapley value 油耗特征分析图

（3）基于频繁变速型驾驶风格的仿真实验

测试集的回归预测评价指标如表 5.6 所示。

表 5.6 频繁变速驾驶风格的回归预测评价指标表

评价指标	RMSE	R^2
准确度（油耗）	4.17	0.77

测试集真实值和预测值散点图和折线图如图 5.9 和图 5.10 所示：

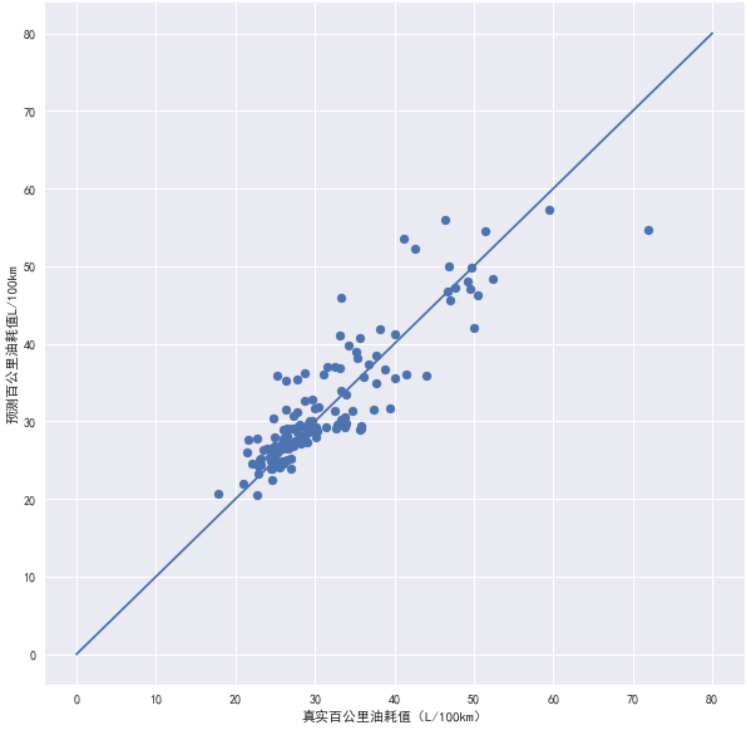

图 5.9 油耗预测值和真实值散点图

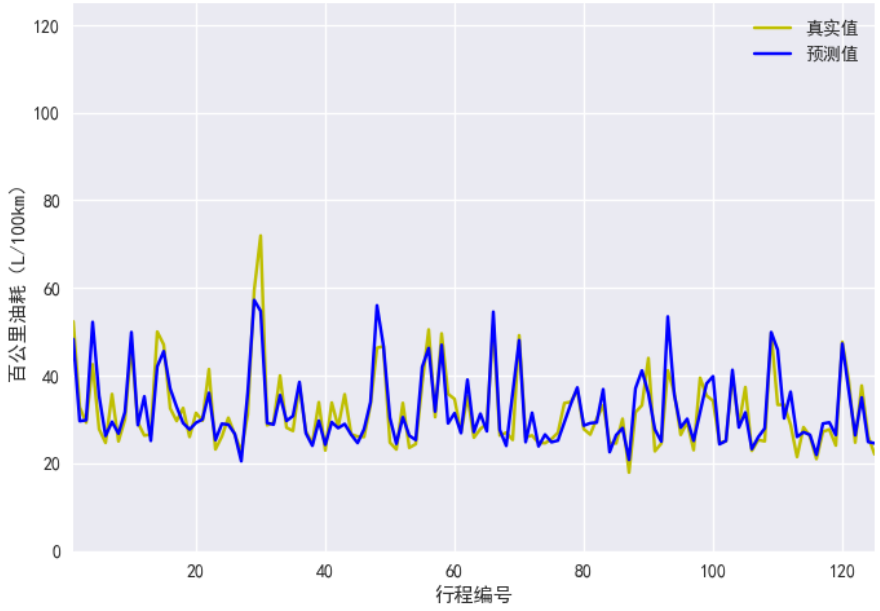

图 5.10 油耗预测值和真实值折线图

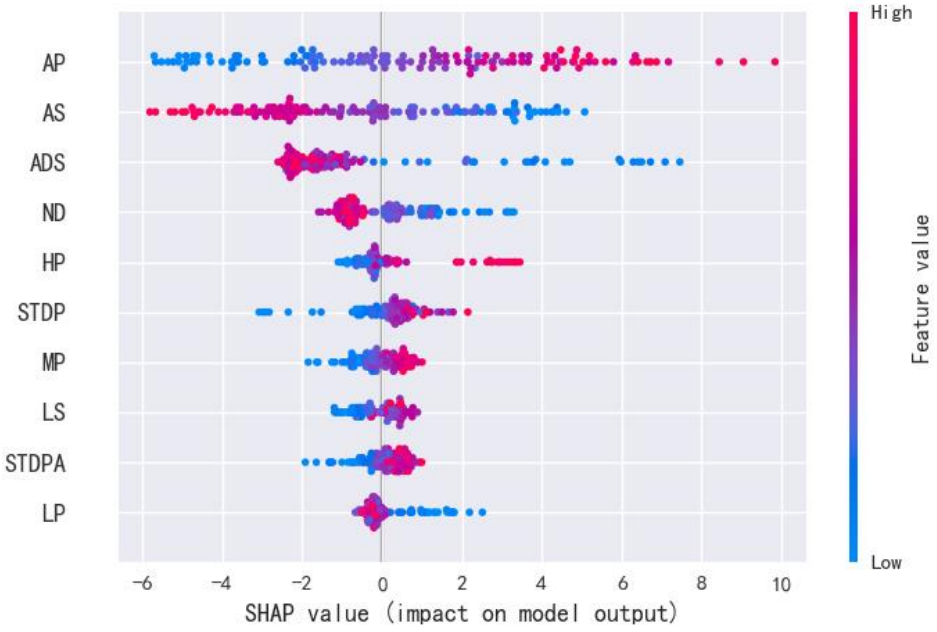

图 5.11 Shapley value 油耗特征分析图

从上述表 5.4-5.6 的算法性能对比解结果中可以看出，猛踩油门型驾驶风格的油耗预测普遍偏高，其预测值和实际值的拟合程度较强。整体来说模型的准确率达到预期，可用于实际运用。

图 5.3、图 5.6 和图 5.9 反映了测试集预测值和真实值散点图，其中靠近斜率为 k 的直线的点越多，模型整体准确度越高。

图 5.4、图 5.7 和图 5.10 反映了测试集真实值和预测值的折线图，折线的拟

合效果越好，说明模型的整体预测准确率越高。

至于图 5.5、图 5.8 和图 5.11，这些图是通过 Python 中的 Shap 包得出的特征鸟瞰图，这里只展示了排名前十的特征图，图中的每个点都具有三个含义：竖直坐标是说明它属于哪个特征、颜色代表了这个特征对应某个样本的数值是高还是低、水平位置代表了这个特征对应某个样本是提高该样本的预测值还是降低预测值。此图可以至少从三方面解释：特征按所有样本的 SHAP 值大小绝对值之和从上到下进行排序，反映了特征重要性；单独观察某一特征，可以得出预测结果随特征值变动的变化趋势；特征在图中的横坐标区间越大，表明该特征对预测的影响更强。可以发现，在三种驾驶风格中，平均速度和油门踏板开度对于最后的油耗影响最强烈，且平均速度的影响是负向的，油门踏板的影响是正向的，这启示驾驶员若想要减少油耗需要避免大踩油门的操作，同时平均速度的提高可以有效降低油耗，但重型货车司机在行驶过程中应结合实际路况将车速控制在安全车速内。除此以外，不同数据集的整体特征图除去前两个最重要的指标，后面指标的排序均有所不同，这反映了不同驾驶风格下驾驶行为对于油耗影响的差异性。

5.4 节能驾驶策略

（1）微观策略

在车队管理中，通常需要对某个行程片段进行分析，从而提出相应的管理策略。在本文，我们需要关注每趟行程中驾驶员的驾驶行为和驾驶风格对于油耗的影响，因此我们还需计算出单个行程片段下各指标的 shapley value 值。得益于 Shap 包的丰富功能，我们可以直接输出模型的局部特征重要性图，如图 5.12 是某段行程的 Shapley value 局部特征图，图中所有特征的 Shapley value 值加上基准线 Base value 的值就是模型的预测值，即式（5.33）：

$$\hat{y}_i = y_{base} + f(x_i, 1) + f(x_i, 1) + \cdots + f(x_i, n) \tag{5.33}$$

其中$f(x_i, n)$就是样本i第n个特征的 Shapley value 值，y_{base}是数据集的平均模型输出。

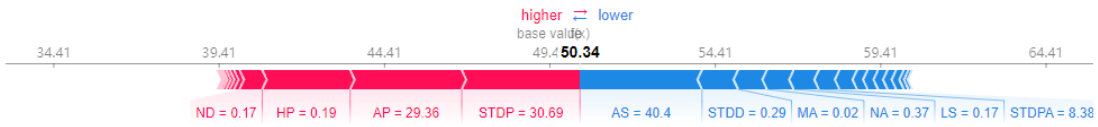

图 5.12 Shapley value 局部特征图

可以看出，此段行程的踏板开度整体标准差、平均踏板开度以及高油门踏板开度的 Shap 值拉高了模型的预测结果，使其最终预测结果高于平均预测结果，呈现出高油耗的水平，因此在驾驶员结束这段行程后，后台便可通过这份数据分析报告提醒驾驶员在下次行驶时注意这几个指标代表的驾驶行为。

（2）宏观策略

通过对 Shapley value 特征鸟瞰图的分析，我们可以总结出一些有利于降低油耗的驾驶行为建议：

（a）通过第五章的实验结果可以发现，油门踏板开度较高的猛踩油门型驾驶风格普遍有更高的油耗，同时我们可以明显发现油门踏板的开度会显著影响油耗。因此驾驶员应平稳控制车速的增加和减少，避免猛踩油门的操作；

（b）当车速控制在安全车速以内时，平均速度与燃油消耗量呈负相关关系。因此，驾驶员应当在保持安全车速的同时，尽可能提高行驶速度。

（c）观察三种驾驶风格下油耗和驾驶行为之间的关系，可以发现加速度对于油耗的影响普遍较小，因此驾驶员应控制踩油门的力度，使货车能够稳步地提速。

5.5 本章小结

本章首先介绍了一些关于燃油经济性数据的预处理过程；接着详细介绍了两种集成学习模型的训练和学习过程，并对两种模型进行了性能比较，选择了表现更优的 XGBoost 算法作为本次实验仿真的最终算法；最后利用真实的驾驶数据进行了实验仿真，并结合 Shapley value 分析影响油耗的主要因子以及单个行程段的驾驶行为分析。

6 总结和展望

本文在现有研究基础上，提出了驾驶风格的分类识别模型，通过构建驾驶行为指标，可以有效识别驾驶员的驾驶风格，模型的识别准确率达到了 95%以上。接着基于不同驾驶风格构建了油耗预测模型，考虑到本文所选数据具有多维且非线性的特点，因此选择了预测效果较好的集成学习模型，三种驾驶风格的油耗预测模型的R^2介于 70%-80%之间，同时选择采用 Shapley value 的方法解决模型可解释性较差的缺点，实现了高准确度高透明度的模型构建。

本文的主要工作如下：

（1）数据预处理和指标构建。本文预先获取的数据来自各类车载设备，存在数据不同步、数据有缺失、数据有异常等问题，因此需要事先对原始数据进行预测处理；对处理完的数据进行分析，获取了 28 个有关驾驶行为的指标，用于后续的实验研究。

（2）提出了基于聚类和支持向量机的驾驶风格分类识别模型。考虑到提取的指标数据维度过大，处理起来比较费时，因此选用主成分分析的方法将 28 个指标降维成 6 个主成分；然后利用 PSO-Kmeans 将 1252 段行程数据聚类为 3 类，根据每一类的数据统计分析，分别将这三类驾驶风格定义为猛踩油门型、高速行驶型和频繁变速型；基于聚类结果给原始数据打上驾驶风格标签，再根据这些带有标签的数据训练一个支持向量机三分类模型，通过验证集的验证，模型预测准确率高达 96.5%，可用于实际工作中对不同驾驶风格的识别。

（3）提出基于集成学习的油耗预测模型。本文根据不同驾驶风格构建了不同类别的油耗预测模型，同时比较了随机森林 RF 和极端梯度增强 XGBoost 两个算法在油耗预测时的准确率，对比之后选择了性能更好的 XGBoost 算法，算法的准确率在 70-80%之间，最后运用 Shapley value 进行了特征分析，发现平均油门踏板开度和平均速度是影响油耗最重要的特征，同时三类驾驶风格呈现的特征重要性图也不同，反映不同驾驶风格下驾驶行为对于油耗影响的差异性。

由于本人知识水平有限，虽然在现有研究的基础上提出了驾驶风格分类模型和油耗预测模型，并得到了一定的实验结论，但考虑到不管是驾驶风格还是油耗都是一个复杂的研究对象，需要全面的考虑才能获取更具有现实意义的结论，因此本文的不足之处主要体现在以下两点：

（1）本文在第二章提及了驾驶风格是一个受外界影响的变量，但文章在讨论驾驶风格时只关注了驾驶员的驾驶行为，忽略了行驶工况、天气等外部因素的

影响。

（2）本文构建的油耗预测模型考虑的维度不够丰富，后续可以再进行开拓性的研究。

参考文献

[1] Xu J, Saleh M, Hatzopoulou M. A machine learning approach capturing the effects of driving behaviour and driver characteristics on trip-level emissions[J]. Atmospheric Environment, 2020, 224: 117311.

[2] Shang R, Zhang Y, Shen Z J M. Analyzing the Effects of Road Type and Rainy Weather on Fuel Consumption and Emissions: A Mesoscopic Model Based on Big Traffic Data[J]. IEEE Access, 2021, 9: 62298-62315.

[3] Xu J, Tu R, Ahmed U, et al. An eco-score system incorporating driving behavior, vehicle characteristics, and traffic conditions[J]. Transportation Research Part D: Transport and Environment, 2021, 95: 102866.

[4] Walnum H J, Simonsen M. Does driving behavior matter? An analysis of fuel consumption data from heavy-duty trucks[J]. Transportation research part D: transport and environment, 2015, 36: 107-120.

[5] Martinez C M, Heucke M, Wang F Y, et al. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(3): 666-676.

[6] Li Y, Tang G, Du J, et al. Multilayer perceptron method to estimate real-world fuel consumption rate of light duty vehicles[J]. IEEE Access, 2019, 7: 63395-63402.

[7] 冯闪.汽车运行油耗的影响因素浅析[J].内燃机与配件,2018(22):33-35.

[8] Ben-Chaim M, Shmerling E, Kuperman A. Analytic modeling of vehicle fuel consumption[J]. Energies, 2013, 6(1): 117-127.

[9] Benajes J, García A, Monsalve-Serrano J, et al. Fuel consumption and engine-out emissions estimations of a light-duty engine running in dual-mode RCCI/CDC with different fuels and driving cycles[J]. Energy, 2018, 157: 19-30.

[10] 张登,祖春胜,赵超超.主成分分析与神经网络结合的燃油消耗预测[J].农业装备与车辆工程,2015,53(06):47-52.

[11] 张贤彪.汽车驾驶油耗影响因素及节油策略分析[J].时代汽车,2020(09):18-19.

[12] 郑天雷,金约夫,张晓龙.重型重型卡车关键参数对油耗影响的模拟分析[J].天津科技,2015,42(04):24-27.

[13] Kamal M A S, Mukai M, Murata J, et al. Ecological vehicle control on roads with up-down slopes[J]. IEEE Transactions on Intelligent TransportationSystems, 2011, 12(3): 783-794.

[14] Jianqiang W, Qianwen Y U, Shengbo L I, et al. Eco speed optimization based on real-time information of road gradient[J]. Journal of Automotive Safety and Energy, 2014, 5(03): 257.

[15] Faria M V, Duarte G O, Varella R A, et al. How do road grade, road type and driving aggressiveness impact vehicle fuel consumption? Assessing potential fuel savings in Lisbon, Portugal[J]. Transportation Research Part D:Transport and Environment, 2019, 72: 148-161.

[16] Zhang S, Wu Y, Liu H, et al. Real-world fuel consumption and CO2 (carbon dioxide) emissions by driving conditions for light-duty passenger vehicles in China[J]. Energy, 2014, 69: 247-257.

[17] Barth M, Boriboonsomsin K. Energy and emissions impacts of a freeway-based dynamic eco-driving system[J]. Transportation Research Part D: Transport and Environment, 2009, 14(6): 400-410.

[18] Carrese S, Gemma A, La Spada S. Impacts of driving behaviours, slope and vehicle load factor on bus fuel consumption and emissions: a real casestudy in the city of Rome[J]. Procedia-Social and Behavioral Sciences, 2013,87: 211-221.

[19] Zheng F, Li J, Van Zuylen H, et al. Influence of driver characteristics on emissions and fuel consumption[J]. Transportation Research Procedia, 2017,27: 624-631.

[20] Díaz-Ramirez J, Giraldo-Peralta N, Flórez-Ceron D, et al. Eco-driving key factors that influence fuel consumption in heavy-truck fleets: A Colombian case[J]. Transportation Research Part D: Transport and Environment, 2017, 56: 258-270.

[21] Berry I M. The effects of driving style and vehicle performance on the real-world fuel consumption of US light-duty vehicles[D]. Massachusetts Institute of Technology, 2010.

[22] Huang Y, Ng E C Y, Zhou J L, et al. Eco-driving technology for sustainable road transport: A review[J]. Renewable and Sustainable Energy Reviews,2018, 93: 596-609.

[23] Fafoutellis P, Mantouka E G, Vlahogianni E I. Eco-driving and its impactson fuel efficiency: An overview of technologies and data-driven methods[J].Sustainability, 2020, 13(1): 226.

[24]张隅希,段宗涛,朱依水,王路阳,周祎,郭宇.机动车油耗模型研究综述[J].计算机工程与应用,2021,57(24):14-26.

[25]Zhou M, Jin H, Wang W. A review of vehicle fuel consumption models toevaluate eco-driving and eco-routing[J]. Transportation Research Part D: Transport and Environment, 2016, 49: 203-218.

[26]Cachón L, Pucher E. Fuel consumption simulation model of a CNG vehicle based on real-world emission measurement[C]//8th International Conference on Engines for Automobiles. 2007 (2007-24-0114).

[27]Heywood J B. Internal combustion engine fundamentals[M]. McGraw-Hill Education, 2018.

[28]Perrotta F, Parry T, Neves L C. Application of machine learning for fuel consumption modelling of trucks[C]//2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017: 3810-3815.

[29]Chen Y, Zhu L, Gonder J, et al. Data-driven fuel consumption estimation: A multivariate adaptive regression spline approach[J]. Transportation Research Part C: Emerging Technologies, 2017, 83: 134-145.

[30]Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436.

[31]Wysocki O, Deka L, Elizondo D, et al. Heavy duty vehicle fuel consumption modelling based on exploitation data by using artificial neural networks[C]//International Work-Conference on Artificial Neural Networks. Springer, Cham, 2019: 794-805.

[32]Xu Z, Wei T, Easa S, et al. Modeling relationship between truck fuel consumption and driving behavior using data from internet of vehicles[J]. Computer-Aided Civil and Infrastructure Engineering, 2018, 33(3): 209-219.

[33]Kanarachos S, Mathew J, Fitzpatrick M E. Instantaneous vehicle fuel consumption estimation using smartphones and recurrent neural networks[J]. Expert Systems with Applications, 2019, 120: 436-447.

[34]Yao Y, Zhao X, Liu C, et al. Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones[J].Journal of Advanced Transportation, 2020, 2020: 1-11.

[35]Bousonville T , Dirichs M, Thilo Krüger. Estimating truck fuel consumption with machine learning using telematics, topology and weather data[C]// 2019 International Conference on Industrial Engineering and Systems

Management (IESM). 2019.

[36] Wang Q, Zhang R, Lv S , et al. Open-pit mine truck fuel consumption pattern and application based on multi-dimensional features and XGBoost[J]. Sustainable Energy Technologies and Assessments, 2021, 43:100977.

[37] Pelkmans L, Debal P, Hood T, et al. Development of a simulation tool to calculate fuel consumption and emissions of vehicles operating in dynamic conditions[R]. SAE Technical Paper, 2004.

[38] Mohammadnazar A, Arvin R, Khattak A J. Classifying travelers' driving style using basic safety messages generated by connected vehicles: application of unsupervised machine learning[J]. Transportation research part C: emerging technologies, 2021, 122: 102917.

[39] Johnson D A, Trivedi M M. Driving style recognition using a smartphone as a sensor platform[C]//2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). Ieee, 2011: 1609-1615.

[40] Corti A, Ongini C, Tanelli M, et al. Quantitative driving style estimation for energy-oriented applications in road vehicles[C]//2013 IEEE InternationalConference on Systems, Man, and Cybernetics. IEEE, 2013: 3710-3715.

[41] Wang R, Lukic S M. Review of driving conditions prediction and driving style recognition based control algorithms for hybrid electric vehicles[C]//2011 IEEE Vehicle power and propulsion conference. IEEE, 2011: 1-7.

[42] Huang X, Tan Y, He X. An intelligent multifeature statistical approach forthe discrimination of driving conditions of a hybrid electric vehicle[J]. IEEETransactions on Intelligent Transportation Systems, 2010, 12(2): 453-465.

[43] Manzoni V, Corti A, De Luca P, et al. Driving style estimation via inertialmeasurements[C]//13th International IEEE Conference on Intelligent Transportation Systems. IEEE, 2010: 777-782.

[44] Qu T, Chen H, Cao D, et al. Switching-based stochastic model predictive control approach for modeling driver steering skill[J]. IEEE Transactions onIntelligent Transportation Systems, 2014, 16(1): 365-375.

[45] Dörr D, Grabengiesser D, Gauterin F. Online driving style recognition using fuzzy logic[C]//17th international IEEE conference on intelligent transportation systems (ITSC). IEEE, 2014: 1021-1026.

[46] Gilman E, Keskinarkaus A, Tamminen S, et al. Personalised assistance for

fuel-efficient driving[J]. Transportation Research Part C: Emerging Technologies, 2015, 58: 681-705.

[47] Murphey Y L, Milton R, Kiliaris L. Driver's style classification using jerk analysis[C]//2009 IEEE workshop on computational intelligence in vehicles and vehicular systems. IEEE, 2009: 23-28.

[48] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30.

[49] 彭娜. 基于行车状态数据的激进型驾驶行为检测[D].长安大学,2019.

[50] 程颖,张佳乐,张少君,郭继孚,张达.大型货运车辆生态驾驶及节油潜力评估[J].交通运输系统工程与信息,2020,20(06):253-258.

[51] Montazeri-Gh M, Fotouhi A, Naderpour A. Driving patterns clustering based on driving features analysis[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2011, 225(6): 1301-1317.

[52] Augustynowicz A. Preliminary classification of driving style with objective rank method[J]. International journal of automotive technology, 2009, 10(5):607-610.

[53] Wickramanayake S, Bandara H M N D. Fuel consumption prediction of fleet vehicles using machine learning: A comparative study[C]//2016 Moratuwa Engineering Research Conference (MERCon). IEEE, 2016: 90-95.

[54] 赵晓华,姚莹,伍毅平,陈晨,荣建.基于主成分分析与 BP 神经元网络的驾驶能耗组合预测模型研究[J].交通运输系统工程与信息,2016,16(05):185-191+204.

[55] Marafie Z, Lin K J, Wang D, et al. AutoCoach: Driving Behavior Management Using Intelligent IoT Services[C]//2019 IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA). IEEE, 2019: 103-110.

[56] Deng L, Yu D. Deep learning for signal and information processing[J]. Microsoft research monograph, 2013.

[57] Bakhit P, Said D, Radwan L. Impact of acceleration aggressiveness on fuelconsumption using comprehensive power based fuel consumption model[J]. Civil and Environmental Research, 2015, 7(3): 148-157.

[58] Li L, Wang X, Song J. Fuel consumption optimization for smart hybrid electric vehicle during a car-following process[J]. Mechanical Systems and Signal Processing, 2017, 87: 17-29.

[59] 胡浈,张瑞新,刘鑫,冯读康,张秋涵.基于大数据的矿用卡车驾驶风格识别算法研究[J].软件,2021,42(03):19-21+64.

[60]基于驾驶风格分类的卡车油耗预测[D].长安大学,2020.

[61]范艺璇,阚秀,曹乐,沈颉.改进 PSO-K-means 算法在汽车行驶工况估计中的应用[J].智能计算机与应用,2021,11(07):80-85+90.