大学生论文检测系统

文本复制检测报告单(全文标明引文)

№:ADBD2023R_20230601185022472820056026 检测时间：2023-06-01 18:50:22

篇名： 基于梯度提升回归树的纯电动汽车剩余续驶里程预测

作者： 陈文翼 (11910434;计算机科学与工程系;计算机科学与技术)

指导教师： 宋轩

检测机构： 南方科技大学

提交论文IP： 110.***.***.***

文件名： 毕业论文-陈文翼.pdf

检测系统： 大学生论文检测系统

检测类型： 大学生论文

检测范围： 中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
大学生论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
源代码库
CNKI大成编客-原创作品库
机构自建比对库

时间范围： 1900-01-01至2023-06-01

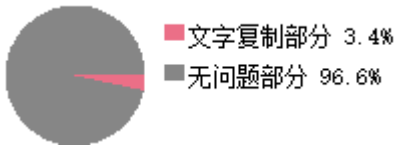
检测结果

去除本人文献复制比： 3.4% **跨语言检测结果：** -

去除引用文献复制比： **2.1%** **总文字复制比：** **3.4%**

单篇最大文字复制比： **1.2%（基于机器学习的纯电动汽车的行驶里程预测研究）**

重复字数：	[522]	总段落数：	[2]
总字数：	[15547]	疑似段落数：	[2]
单篇最大重复字数：	[189]	前部重合字数：	[171]
疑似段落最大重合字数：	[491]	后部重合字数：	[351]
疑似段落最小重合字数：	[31]		



指标： ☐ 疑似剽窃观点 ☑ 疑似剽窃文字表述 ☐ 疑似整体剽窃 ☐ 过度引用

相似表格：0 相似公式：没有数据 疑似文字的图片：0

4.6%	(491)	4.6% (491)	基于梯度提升回归树的纯电动汽车剩余续驶里程预测_第1部分 (总10625字)
0.6%	(31)	0.6% (31)	基于梯度提升回归树的纯电动汽车剩余续驶里程预测_第2部分 (总4922字)

（注释： 无问题部分 文字复制部分 引用部分）

指导教师审查结果

指导教师：宋轩
审阅结果：
审阅意见：指导老师未填写审阅意见

1. 基于梯度提升回归树的纯电动汽车剩余续驶里程预测_第1部分

总字数：**10625**

相似文献列表

去除本人文献复制比：4.6%(491) 文字复制比：4.6%(491) 疑似剽窃观点（0）

1	基于机器学习的纯电动汽车的行驶里程预测研究 高航(导师：毕军) -《北京交通大学硕士论文》- 2018-03-01	1.8%（189） 是否引证：是
2	基于改进符号回归算法和XGBoost算法的剩余续驶里程预测 田晟;甘志恒;吕清; -《广西师范大学学报(自然科学版)（优先出版）》- 2021-06-23 1	1.7%（183） 是否引证：否
3	基于文本嵌入的网信项目预评估模型设计与实现 唐周华(导师：鱼滨) -《西安电子科技大学硕士论文》- 2020-04-01	0.8%（83） 是否引证：否
4	基于标签和评分联合学习的协同过滤推荐算法研究 - 道客巴巴 -《互联网文档资源（https://www.doc88.co）》- 2020	0.5%（54） 是否引证：否
5	基于用户信任和偏好信息的推荐算法研究 徐玲玲(导师：刘晓红) -《山东理工大学硕士论文》- 2020-03-20	0.4%（44） 是否引证：否
6	中国氯碱行业氢能利用现状及趋势分析 郑结斌; -《中国氯碱》- 2023-01-20	0.3%（30） 是否引证：否

原文内容

分类号编号

U D C 密级

本科生毕业设计（论文）

题目： 基于梯度提升回归树的纯电动汽车剩余续驶里程预测

姓名： 陈文翼

学号： 11910434

系别： 计算机科学与工程系

专业： 计算机科学与技术

指导教师： 宋轩副教授

2023 年 6 月 2 日

诚信承诺书

1. 本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，
独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2. 除文中已经注明引用的内容外，本论文不包含任何其他人或
集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭
他人研究成果和伪造相关数据等行为。

4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本
人承担相应的法律责任。

作者签名:

年月日基于梯度提升回归树的纯电动汽车剩余续驶里程预测陈文翼

（计算机科学与工程系指导教师：宋轩）

[摘要]：*面对日益严峻的能源问题和环境污染问题，契合绿色、低碳、可*

持续发展观念的电动汽车进入了产业的高速发展期。但在当今电池能量密度尚未取得较大提升、充电基础设施建设不够完善的情况下，“里程焦虑”现象渐现端倪，严重影响了电动汽车的用户体验与接受度。因此，基于真实历史运行数据进行剩余续驶里程的建模预测对电动汽车的推广、

普及、发展都具有重要意义。本文的主要研究内容如下：

（1）纯电动汽车运行数据的筛选与预处理。从原始数据集中筛选出纯电动汽车的运行数据，并进行预处理，构建剩余续驶里程字段。

（2）影响纯电动汽车剩余续驶里程的关键特征参数分析。从纯电动汽车的运行数据出发，运用斯皮尔曼相关系数等方法分析了剩余续驶里程与各特征参数以及各特征参数之间的相关关系，初步筛选出与纯电动汽车的剩余续驶里程相关性较强的关键特征参数。

（3）基于梯度提升回归树（GBRT）与递归特征消除算法（RFE）的剩余续驶里程建模预测。使用递归特征消除算法得到表征剩余续驶里程的关键特征参数，减少特征之间共线性对模型的影响；基于梯度提升回归树对纯电动汽车的剩余续驶里程进行建模预测，并分析、评估模型的

预测结果。

[关键词]：机器学习；剩余续驶里程；纯电动汽车； GBRT； RFE

I

[ABSTRACT]: The issues of energy and environmental pollution are becoming increasingly severe in today’s world. Electric vehicles that adhere to the green, low-carbon, and sustainable development concepts have entered a period of rapid development in the industry. However, the energy density of electric vehicle batteries has not yet achieved significant improvement and the construction of charging infrastructure is not sufficiently complete. Under these circumstances, the phenomenon of ”range anxiety” has gradually emerged, which seriously affects the user experience and acceptance of electric vehicles. Therefore, modeling and predicting the remaining driving range based on real historical operational data is of great significance for the promotion, popularization, and development of electric vehicles. The main research contents of this paper are as follows:

(1) Screening and preprocessing of running data of pure electric vehicles.

(2) Analysis of key characteristic parameters affecting the remaining driving range of pure electric vehicles.

(3) Modeling and prediction of remaining driving distance based on gradient lifting regression tree (GBRT) and recursive feature elimination algorithm (RFE).

[Key words]: machine learning；remaining driving range；battery electric vehicles；GBRT；RFE

II

目录

1. 引言

1.1 研究背景及意义

当今，我国正面临着日益严峻的能源和环境问题。一方面，我国的能源消耗量巨大，且能源结构较不合理，过度依赖煤炭等化石能源；另一方面，我国的环境污染问题也不容轻视，大气污染、水污染、土壤污染等问题频发[1]，这些问题不仅会影响人民的健康，也会威胁到社会的可持续发展。因此，契合绿色、低碳、可持续发展观念的电动汽车成为了未来发展的大势所趋。

电动汽车具有节能低碳的优点，以电池储能代替传统燃油，既可以减少二氧化碳等温室气体的排放，又可以降低对化石能源的依赖[2-3]，对缓解环境污染、保障国家能源安全都起到了积极作用。但是，我国在电动汽车发展过程中仍有许多问题需要解决，如电池能量密度较低导致的续航里程不足，充电基础设施建设不够完善影响用户出行等[4]，这些问题使得电动汽车的用户容易产生“里程焦虑”，直接影响了电动汽车的用户体验和推广普及。

因此，纯电动汽车剩余续驶里程的精准预测将对缓解“里程焦虑”、提高用户体验、进一步普及和推广纯电动汽车具有重大意义[5]，拥有很高的研究价值。

1.2 国内外研究现状

因为全球日益严峻的环境问题，以及多国政府的大力发展，纯电动汽车剩余续驶里程的精准预测问题一直受到国内外学者的关注。

Zhao 等人[6]提出了一种有效消除数据分布不平衡的策略，并基于真实历史驾驶数据，融合了 XGBoost 和 LightGBM 两种先进的机器学习算法对电动汽车的剩余续驶里程进行了预测。田晟等人[7]运用改进符号回归算法扩充了数据的纬度，自动生成了与原标签字段拥有较高相关性的新特征字段，并对 XGBoost 模型进行超参数调

优，使得剩余续驶里程的预测结果中，最大相对绝对误差下降了 4.9%，平均绝对误

差和均方根误差下降超过 20%。Lamantia 等人[8]提出了一种新的电动汽车剩余续驶里程的估计方法，该方法将气动阻力系数的实时估计和先进驾驶辅助系统 (ADAS) 或队列行驶的模式集成到电动汽车的物理模型中，提高了高速公路上电动汽车剩余续驶里程的估计精度。Ayevide 等人[9]提出了一种融合 NARX 和卡尔曼滤波（KF）的递归 1

动态网络，利用来自天气、车辆速度和道路状况的信息来预测电池输出功率和剩余续驶里程，并在不同环境条件下进行了驾驶实验，与传统的基于模型的预测方法相比，预测的准确度有了明显的提升。

1.3 剩余续驶里程定义

对于纯电动汽车，续驶里程（Dring Range）指全电续驶里程（All Electric Range），即纯电动汽车满电状态下（荷电状态 SOC 对应 100%），仅依靠电池能量行驶至电池能量耗尽时（荷电状态 SOC 对应 0%），车辆能够行驶的距离[10]。

行驶里程是指车辆从满电状态开始，按照一定的工况行驶至当前时刻时，车辆所行驶的距离[11]。

剩余续驶里程（Remaining Driving Range）是指从当前时刻开始，直至 SOC 降至设定的截止 SOC 时，车辆累计行驶的距离[12]。剩余续驶里程在纯电动汽车的行驶过程中需要进行实时预测，当车辆无法再行驶时，剩余续驶里程为 0[13]。

1.4 本文内容与结构安排

本文的主要研究内容为纯电动汽车的剩余续驶里程预测。选用纯电动汽车的真实历史运行数据，结合对剩余续驶里程的相关影响因素分析，基于有监督机器学习进行建模预测。*本文分为五个章节，每一章节的具体内容如下：*

*第一章：引言。本章主要介绍*了纯电动汽车剩余续驶里程预测问题的研究背景及

意义，对剩余续驶里程预测问题的国内外研究现状进行了简单分析，并解释了纯电动汽车的剩余续驶里程的定义。

第二章：数据预处理。本章主要介绍了数据来源、原始数据集情况及数据的预处

*理，*对数据中的异常值、缺失值进行处理，构造了剩余续驶里程字段，并对数据进行了切片与压缩。

第三章：剩余续驶里程的关键特征参数。本章主要介绍了影响纯电动汽车剩余续

驶里程的关键特征参数的初步筛选。从各特征参数与剩余续驶里程之间的关系以及各特征参数相互之间的关系两个方面着手，利用斯皮尔曼相关系数对剩余续驶里程的关键特征参数进行分析。

第四章：基于梯度提升与递归特征消除算法的剩余续驶里程预测。本章在第三章 2

的基础上，利用递归特征消除算法（RFE）筛选出了预测剩余续驶里程的关键特征参

数，基于梯度提升回归树（GBRT）对剩余续驶里程进行建模预测，并对预测结果进

行了分析评估。

第五章：总结。本章对论文所做的工作进行了总结，对论文中存在的不足进行了

分析，并给出了今后改进的方向。

2. 数据预处理

2.1 原始数据集

原始数据采集自上海 2000 辆不同型号、不同动力类型的汽车，其中包含了汽车的车型数据及其 2020 年 10-12 月期间的车辆运行数据、位置数据、电池极值数据、驱动电机数据、次行标签数据以及出险标签数据，数据采集间隔为 10s。基于该原始数据集进行了纯电动汽车剩余续驶里程预测的相关研究，具体的参数及其含义如表 1 所示。

表 1 原始数据集表名字段名称字段含义

车辆运行数据

time 数据采集时间

vehiclestatus 车辆状态

chargestatus 充电状态

runmodel 运行模式

speed 车速

summile 累积里程

sumvolt 总电压

sumcurren 总电流

SOC 荷电状态

DCDC DC-DC 状态

gearnum 档位

resistance 绝缘电阻

车辆位置数据

longitude 经度

latitude 纬度

电池极值数据

max_single_volt 电池单体电压最高值

min_single_volt 电池单体电压最低值

maxtmp 最高温度值

mintmp 最低温度值 3

续表 1

表名字段名称字段含义

驱动电机数据

sequence 驱动电机序号

rotatingspeed 驱动电机转速

torque 驱动电机转矩

次行标签数据

tripkind 次行类型

starttime 开始时采集时间

endtime 结束时采集时间出险标签数据 insurance 是否出险

车型数据

id 车辆代码

model_seq 车型代码

power_seq 车辆动力类型

vehicle_type 车辆类型

curb_weight 整备质量

wheel_base 轴距

device_power 电池额定能量

device_capacity 电池额定容量

车型数据记录了 2000 辆汽车的基本状态参数，可用于筛选纯电动汽车；其中的

id 字段作为车辆的唯一标识，可在数据加工处理过程中充当索引。车辆运行数据反映了车辆运行过程中的状态参数，可作为特征参数建模预测汽车的剩余续驶里程。车辆位置数据中包含了车辆实时的经度、纬度坐标信息，与本文所研究的汽车剩余续驶里程相关度较低，为减少计算资源的消耗，对经、纬度字段进行删除。电池极值数据反映了车辆电池的物理状态，将其中的数据字段作为关键特征参数进行保留。在驱动电机数据中，所有字段的内容均与字段描述中的取值范围不符，故对其进行删除，不参与建模的过程。次行标签数据的内容与第一类车辆运行数据的内容存在重复，故对其进行过滤。出险标签数据与本文所研究的汽车剩余续驶里程相关度较低，故对其进行删除，不参与后续数据处理、建模的过程。

2.2 数据预处理

根据原始数据集的探查结果，确定与本文所研究的汽车剩余续驶里程相关的、作

为特征变量保留的数据字段有：id、time、vehiclestatus、chargestatus、runmodel、speed、

summile、sumvolt、sumcurrent、SOC、DCDC、gearnum、resistance、max_singlevolt、 4

min_singlevolt、max_tmp 和 min_tmp 等 17 个字段。基于保留的数据字段进行数据预处理。

2.2.1 数据筛选

原始数据集中的车辆运行数据记录共计 147727213 条，车辆动力类型涵盖纯电动汽车、插电式混合动力汽车与增程式电动汽车。基于本文研究的是纯电动汽车剩余续驶里程，故按照车辆动力类型对车型数据记录及车辆运行数据记录进行筛选，共计得到 330 辆纯电动汽车的车型数据及运行数据。

筛选出纯电动汽车的车型数据后，按照车型代码（model_seq）统计每种纯电动汽车车型所包含的车辆数量，330 辆纯电动汽车共分为 96 种车型。不同车型所含车辆总数前 10 排序如图 1 所示。

图 1 不同车型所含车辆总数前 10 排序

由图 1 的统计结果表明，车辆 id 代码为 123114323 的车型包含 24 辆纯电动汽车，

车辆数量排名第一，故本文接下来所采用的数据均为该 24 辆纯电动汽车的车辆运行数据，共计 2842747 条。

为预测汽车的剩余续驶里程，需选取车辆行驶状态下的运行数据记录，故需筛选

出 vehiclestatus 对应值为 1（车辆启动）的数据。统计全部车辆运行记录中的 charges-

tatus 对应值的分布情况，chargestatus 对应值均为 1（停车充电）或 3（未充电）或 4（充电完成），未出现 chargestatus 为 2（行驶充电）的情况，说明该纯电动汽车车型不具有能量回收机制，故需筛选出 chargestatus 对应值为 3（未充电）的数据。故按照 5 vehiclestatus = 1 且 chargestatus = 3 的筛选条件对该车型 24 辆纯电动汽车的运行数据进行初步筛选。

2.2.2 数据异常值处理

数据异常值的筛选主要从两个维度进行：

（1）数据字段描述中规定值的有效范围。

（2）字段含义的逻辑关系。如 SOC 在单次放电过程中应随时间的增加而降低；同一辆车的 summileage 应随时间的增加而增加；同一条数据记录中，max_tmpval 应不小于 min_tmpval，max_single_voltageval 应不小于 min_single_voltageval 等。

在经过异常值筛选后，纯电动汽车原生运行数据中原来为异常值的地方会被None 替代，因而出现了数值缺失。

2.2.3 数据缺失值处理

经对筛选得到的纯电动汽车原生运行数据的探查得知，纯电动汽车的原生运行数据中不存在缺失值，故汽车运行数据中的缺失值均来源于异常值处理步骤。

对于部分特征字段缺失的运行数据行，因相邻两条数据的采样时间间隔为 10s，选取该条数据前后各 3 条采样数据，即 1 分钟内的数据，再以该 6 条采样数据的均值对缺失值进行填充。

对于整行数据记录的缺失，采取整行数据记录删除的方式进行缺失值的处理。

2.2.4 数据切片

将进行了异常值和缺失值处理的纯电动汽车运行数据按照车辆 id 划分为 24 个单独的数据集。对于 24 辆纯电动汽车中每一辆汽车的运行数据均进行切片处理。

当汽车处于启动状态时，处于放电状态，SOC 的值随时间的增加而减少。将纯电动汽车的运行数据中*相邻的 SOC 最高值和最低值之间的一系列数据作为一个放电过程，*按照单个放电过程对纯电动汽车的运行数据进行切片，将原本连续的数据集切割成多段的连续放电过程数据片段，并按照车辆 id 分组保存，共计得到 2539 个放电过程片段。 6

按照剩余续驶里程的定义，一个理想的放电过程应为 SOC 从 100% 下降到 0% 的过程，但这样的放电过程在 24 辆纯电动汽车的所有放电过程片段中均未出现。在 2539 个放电过程片段中，SOC 从 90% 下降至 10% 的片段不足 10 段，从 80% 下降至 20% 的片段有 18 段，故本文将基于 SOC 从 80% 下降到 20% 的放电片段开展纯电动汽车的剩余续驶里程的建模预测。

2.2.5 剩余续驶里程字段构造

由表 2 可知，在纯电动汽车的原始运行数据记录中不存在“剩余续驶里程”数据字段，故需按照剩余续驶里程的定义构造该字段。在一个放电过程中，累积里程 summile 随着时间不断增加，则 SOC 为 20% 时对应的 summile 值为 summile 的最大值，剩余续驶里程 remaining_mile 等于该最大值减去各条运行数据记录中的 summile值所得之差。

为减少数据冗余、降低数据处理与建模中的计算资源消耗，按照 1 分钟的时间间隔对数据进行压缩。纯电动车单个放电过程内运行数据压缩后的结果表 2 所示。

表 2 单个放电过程内运行数据

time speed

sum

volt

sum

current

max

single_volt

min

single_volt

max

tmp

min

tmp

SOC

remain

mile 2020-11-03

11:33:00

70.56 378.12 21.76 3.96 3.92 20.0 19.0 80.0 159.0 2020-11-03

11:34:00

77.83 376.93 24.65 3.99 3.93 20.0 19.0 80.0 157.0 2020-11-03

11:35:00

73.14 374.36 48.68 3.94 3.9 20.0 19.0 79.0 156.0 2020-11-03

16:57:00

64.62 335.52 29.10 3.56 3.50 24.0 22.0 20.0 2.0 2020-11-03

16:58:00

67.82 334.38 43.48 3.53 3.48 24.0 22.0 20.0 1.0 2020-11-03

16:59:00

83.30 334.30 32.40 3.53 3.48 24.0 22.0 20.0 0.0 7

3. 剩余续驶里程的关键特征参数

3.1 剩余续驶里程的相关影响因素分析

为分析纯电动汽车剩余续驶里程与运行数据中保留的特征变量之间的关系，本小节中选用车辆 id 为 2fd598ea458da6034fc947b786846750 的一辆纯电动汽车 2020 年

12 月 1 日 11:53 至 18:24 的一段预处理后的完整放电过程片段，分别绘制了特征变量

speed、SOC、sumvolt、sumcurrent、max_tmp、min_tmp、max_singlevolt、min_singlevolt

与字段 remaining_mile 的散点分布图，得到剩余续驶里程 remaining_mile 与各特征变量之间的关系，分析如下：

（1）speed 与 remaining_mile

图 2 speed 与 remaining_mile 的相关关系根据图 2 可知，剩余续驶里程 remaining_mile 随行驶速度 speed 的分布是发散的，无明显规律可循，因此二者之间不存在线性相关关系。因此，为减少模型训练过程中的计算资源消耗、提高数据质量，speed 将不会作为关键特征变量进行保留。

（2）SOC 与 remaining_mile

图 3 SOC 与 remaining_mile 的相关关系 8

图 3 反映了剩余续驶里程 remaining_mile 和 SOC 的关系，从图中可以看出，SOC

越大，剩余续驶里程 remaining_mile 也越大，二者呈明显的线性正相关关系。SOC 能够反映动力电池的剩余可用能量，故选用 SOC 作为表征剩余续驶里程 remaining_mile的关键特征变量之一。

（3）sumvolt 与 remaining_mile

图 4 sumvolt 与 remaining_mile 的相关关系根据图 4 可知，剩余续驶里程 remaining_mile 随着总电压 sumvolt 的下降而减少，总电压和剩余续驶里程总体呈现非线性的正相关关系。故选用总电压 sumvolt 作为表征剩余续驶里程 remaining_mile 的关键特征变量之一。

（4）sumcurrent 与 remaining_mile

图 5 sumcurrent 与 remaining_mile 的相关关系根据图 5 可知，剩余续驶里程 remaining_mile 随总电压 sumcurrent 的分布是发散的，无明显规律可循，因此二者之间不存在线性相关关系。因此，sumcurrent 将不会作为关键特征变量进行保留。 9

（5）max_tmp 与 remaining_mile

图 6 max_tmp 与 remaining_mile 的相关关系根据图 6 可知，剩余续驶里程 remaining_mile 与电池单体最高温度 max_tmp 之间大致呈非线性的负相关关系。由于电动汽车的电池管理系统（BMS）的温度调节作用，电池最高温度 max_tmp 集中分布在 18 至 19 摄氏度之间，没有持续升高。

（6）min_tmpval 与 remaining_mile

图 7 min_tmp 与 remaining_mile 的相关关系根据图 7 可知，剩余续驶里程 remaining_mile 与电池单体最低温度 min_tmp 之间大致呈非线性的负相关关系，且与电池单体最高温度 max_tmp 的分布趋势相近。

（7）max_singlevolt 与 remaining_mile 10

图 8 max_singlevolt 与 remaining_mile 的相关关系根据图 8 可知，剩余续驶里程 remaining_mile 随着电池单体电压最高值 max_singlevolt的下降而减少，二者之间呈现非线性的正相关关系。

（8）min_singlevolt 与 remaining_mile

图 9 min_singlevolt 与 remaining_mile 的相关关系根据图 9 可知，剩余续驶里程 remaining_mile 随着电池单体电压最低值 min_singlevolt

的下降而减少，二者之间呈现非线性的正相关关系。因为电动汽车的电池组是由多个

单体电池串联而成，并通过 BMS 进行管理和控制[11]，故 max_singlevolt、min_singlevolt、

sumvolt 三者的变化曲线相似。

3.2 特征间相关性分析

利用散点分布图能够较为直观地体现与目标变量随特征变量变化的趋势，但对非线性关系的描述仍存在一定的不足，只能体现变化趋势，缺少量化的分析。因此，选用斯皮尔曼等级相关系数（即 spearman 相关系数）对数据各特征变量之间的相关性进行量化分析。 11

3.2.1 斯皮尔曼等级相关系数的基本原理

斯皮尔曼等级相关系数是用于度量两个变量之间相关性的一种无参假设检验[14]。

因其无需对数据分布做出任何假设，斯皮尔曼等级相关系数适用于各种类型的数据，具有广泛的普适性。斯皮尔曼等级相关系数的计算步骤如下。

（1）假设两个变量分别为 X={x1,x

2,x

3,...,x

n}，Y={y

1,y

2,y

3,...,y

n}。将 X 与 Y 同时

按照升序或降序排列，得到对应的有序数列 A 与 B。将 X 内每个元素在 A 中的位置、

Y 内每个元素在 B 中的位置分别记录下来，形成两个新的数列 M={m1,m

2,m

3,...,m

n}

和 N={n1,n

2,n

3,...,n

n}，即为 X 与 Y 分别对应的秩次数列[15]。

（2）将集合 M 和 N 中的元素对应相减，得到秩次差数列 D={d1,d

2,d

3,...,d

n}, 其中

di=m

i-n

i, 1≤i≤n。

（3）将值代入斯皮尔曼等级相关系数公式：

ρ = 1 − 6

n

i=1d2i

n(n2− 1) (1)

ρ 的取值范围是 [-1,1]，其中-1 表示完全反比例关系，1 表示完全正比例关系，0

表示没有关系。

3.2.2 变量间相关性分析

为了准确描述各特征变量与剩余续驶里程之间的相关性，利用 3.1 中同样的纯电

动汽车运行数据进行斯皮尔曼等级相关性分析，分析结果如图 10 所示。在图 10 中，

两变量对应的方块颜色越浅，则二者之间的斯皮尔曼相关系数越接近于 1，表明两变量之间的相关性越高。 1

2

图 10 各特征字段斯皮尔曼相关系数为便于比较各特征变量与目标变量剩余续驶里程之间的相关性强弱，单独提取出各特征变量与剩余续驶里程之间的斯皮尔曼相关系数，按照斯皮尔曼相关系数大

小进行排序并可视化，结果如图 11 所示。

指　标

疑似剽窃文字表述

1. 第五章：总结。本章对论文所做的工作进行了总结，对论文中存在的不足进行了 分析，并给出了今后改进的方向。

2. 基于梯度提升回归树的纯电动汽车剩余续驶里程预测_第2部分

总字数：**4922**

相似文献列表

去除本人文献复制比：0.6%(31)　　文字复制比：0.6%(31)　　疑似剽窃观点（0）

1	基于Boosting算法的多因子量化选股实证研究 王一卓(导师：吴臻) -《山东大学硕士论文》- 2020-05-20	0.6%（31） 是否引证：否

原文内容

13

图 11 各特征变量与剩余续驶里程的斯皮尔曼相关系数由图 11 可知，speed、sumcurrent 与剩余续驶里程 remain_mile 之间的斯皮尔曼相关系数都接近于 0，表明 speed、sumcurrent 与 remain_mile 之间的相关性很低，与 3.1

中剩余续驶里程的相关影响因素分析的结果相符，故不选用这两个特征参与后续的

模型训练。

由图 10 可知，存在多组特征变量之间的斯皮尔曼相关系数较大，具有较强的相关性，如 max_tmp 与 min_tmp 之间的相关系数为 0.87，max_singlevolt 与 min_singlevolt

之间的相关系数为 0.97，这表明选取的部分特称变量之间具有一定的共线性，在特征信息上具有一定的重叠，在后续的模型训练中可选择其中之一来降低特称变量之间的共线性对模型训练的影响[16]。

4. 基于梯度提升与递归特征消除算法的剩余续驶里程预测

纯电动汽车的剩余续驶里程预测问题本质上是一个有监督的回归问题，本章基于纯电动汽车的历史运行数据，结合上一章中剩余续驶里程的关键特征参数分析，使用基于梯度提升的回归算法对纯电动汽车的剩余续驶里程进行预测研究。

4.1 梯度提升回归树（GBRT）的基本思想

GBRT 的基本思想是通过迭代训练一系列决策树来构建一个预测模型。在每次迭代中，用梯度信息对当前的加法模型进行更新，并使用贪心法找到局部最优解。在整 14

个过程中，损失函数逐渐减小，而模型的性能则逐渐提升[17-18]。GBRT 的计算步骤如下。

设训练集为 D = {(x1,y

1),(x

2,y

2),(x

3,y

3),...,(x

n,y

n)}，其中 x

i

是输入的特征，yi是输出的标签。假设已经训练了 t-1 棵决策树，则第 t 棵决策树的目标是拟合前 t-1 棵决策树的残差，设前 t-1 棵决策树的预测结果为：

Ft−1(xi) =

t−1

k=1

fk(xi) (2)

其中，第 t 棵决策树的预测结果为:

fk(xi) = yi− Ft−1(xi) (3)

最终的预测函数是所有决策树的集成：

Ft(x) = Ft−1(x) + αft(x) (4)

其中，α 表示学习率（learning_rate），通常取值为 0.01 到 0.1，用于控制决策树的权重，防止过拟合。

4.2 基于递归特征消除的特征筛选

由第三章可知，在纯电动汽车的行驶数据中，大部分的特征参数都与剩余续驶里程相关，即会对剩余续驶里程的预测产生影响，且特征参数之间也存在一定的相关

性。为了降低特征参数之间的相关性对模型训练的影响，引入递归特征消除（RFE，Recursive Feature Elimination）对特征参数进行筛选。

4.2.1 递归特征消除的基本思想

递归特征消除 (RFE) 是一种基于模型的特征选择算法，通过反复训练模型来剔除较为不重要的特征，直至保留的特征数量降至预设的特征数量[19]。RFE 算法的基本步骤如下：

（1）使用所有现存特征来训练模型，计算每个特征的重要性并排序

（2）剔除重要性最低的特征 15

（3）重复步骤（1）和步骤（2），直到保留的特征数量降至预设的特征数量[20]。

4.2.2 特征参数筛选

在预设特征数量为 5 的条件下，使用 RFE 算法进行特征参数的筛选，各特征的重要性等级及选用情况如表 3 所示，各特征的重要性如图 12 所示。

表 3 各特征的重要性等级及选用情况

特征名称排序是否选用

sumvoltage 1 True

SOC 1 True

max_tmpval 2 False

min_tmpval 1 True

max_battery_single_voltageval 1 True

min_battery_single_voltageval 1 True

图 12 各特征的重要性如表 2 及图 12 所示，筛选出 sumvolt、SOC、min_tmp、max_singlevolt、min_singlevolt五个特征来进行汽车剩余续驶里程的建模预测。

4.3 训练集与测试集

本文选用车型代码为 123114323 的 24 辆纯电动汽的 18 个 SOC 从 80% 降至 20%

的完整放电片段，其中第 1 至 14 号片段作为训练集，记为 S1至 S 14

；15 至 18 号片段作为测试集，记为1 5

至 S1 8 。 16

4.4 模型参数

梯度提升回归模型的参数如表 4 所示。

表 4 模型部分参数

参数名称参数含义

loss 损失函数

learning_rate 学习率，用于控制每次迭代的权重

n_estimators 用于控制迭代步数

max_depth 决策树最大深度

subsample 控制每次迭代所用数据集大小

min_samples_split 内部节点再划分所需最小样本数

min_samples_leaf 叶子节点最少样本数梯度提升回归模型的参数调节步骤如下：

（1）损失函数选用均方差。

（2）考虑到模型的复杂性，将 learning_rate 设置为 0.1；在 100 至 500 的区间内寻找表现最佳的 n_estimators 值。

（3）对决策树的结构进行调整，包括 max_depth、min_samples_split 与 min_samples_leaf等。

通过不断调整模型的众多参数，得到的一个表现最佳的参数组合如表 5 所示。

表 5 参数调整结果

参数名称参数值

loss squared_error

learning_rate 0.1

n_estimators 200

max_depth 3

subsample 1

min_samples_split 10

min_samples_leaf 10 17

4.5 预测结果及模型评估

4.5.1 模型评估指标

为了更好地描述模型预测结果与实际剩余续驶里程之间的差异、更好地反映模

型性能与优化方向，本文将使用四个不同指标来评估该梯度提升回归模型：

（1）最大误差（MEm

in

），表示预测值与实际值之间的最大差距。

（2）最小误差（MEm

ax

），表示预测值与实际值之间的最小差距。

（3）均方根误差（RMSE，Root Mean Square Error），又称标准误差，用于衡量预测值与真实值之间的误差大小。RMSE 的值越小，表示模型的预测精度越高。计算公

式如下：

RM SE = 1

n

n

i=1

(yi− ˆy

i)2(5)

（4）均值绝对误差（MAE，Mean Absolute Error），用于衡量预测值与实际值之间

的差异大小。计算公式如下：

M AE = 1

n

n

i=1

|yi− ˆy

i| (6)

（5）R-squared，用于评估模型拟合度的统计指标。R-squared 的取值范围为 0 到

1，越接近 1 表示模型的拟合效果越好。计算公式如下：

R − squared = 1 −

SSR

T SS (7)

其中 SSR（Sum of Squared Residuals）代表残差平方和，TSS（Total Sum of Squares）代表总平方和。

4.5.2 模型预测结果评估

按照最佳参数组合训练剩余续驶里程预测的梯度提升回归模型，绘制剩余续驶里程的预测值与真实值的对比图。

按照上述五个指标对模型进行评价，评价结果如表 5 所示。

18

图 13 S1 5

预测值与真实值的对比图图 14 S1 6

预测值与真实值的对比图图 15 S1 7

预测值与真实值的对比图图 16 S1 8

预测值与真实值的对比图

表 6 各指标评结果

MEm

in ME

max RMSE MAE R-squared

S1

5-0.837 -19.109 7.295 4.589 0.952

S1

61.205 -4.332 4.237 3.540 0.980

S1

7-1.604 -5.298 3.145 2.399 0.992

S1

8-1.822 10.566 4.983 3.725 0.986

由表 6 的各项数据可知，在 4 段作为测试集的纯电动汽车放电片段中，S1 5

较其

他三个片段的测试结果而言表现较差，最大误差为-19.109，均方根误差为 7.295。从

图 13 的曲线拟合情况来看，预测值与真实值仍存在一定差距。但整体而言，在基于

梯度提升回归树的剩余续驶里程预测结果中，4 段片段的 R-squared 均在 0.95 以上，

MAE 均在 5Km 以内且最小值为 2.399Km，说明该梯度提升回归树模型的拟合程度较

好，但也仍存在很大的改进空间。

5. 总结与展望

随着全球环境污染问题的日益加剧，传统燃油汽车的尾气排放已经对环境造成了严重的影响。为了缓解这种现状，新能源汽车产业得到了突飞猛进的发展，对精确 19

预测电动汽车剩余续驶里程的需求也与日俱增。针对这一问题，本文选取了同一个型号的 24 辆纯电动汽车的历史运行数据，结合机器学习算法，对纯电动汽车的剩余续驶里程进行了建模预测。本文完成的工作如下：

（1）汽车历史运行数据的筛选与预处理。从原始数据集中筛选出纯电动汽车的车型数据，在基于此筛选出纯电动汽车的历史运行数据，对其进行预处理，并构建剩余续驶里程字段。将数据按照按照 SOC 变化的范围进行切片，提取出每辆车的单次放电片段。

（2）影响纯电动汽车剩余续驶里程的关键特征参数分析。运用斯皮尔曼相关系数来量化分析剩余续驶里程与各特征参数之间，以及各特征参数相互之间的相关性，初步筛选出与剩余续驶里程相关性较高的特征参数。

（3）基于 RFE 的 GBRT 剩余续驶里程建模预测。通过 RFE 对上一步保留的特征向量进行二次筛选，找到表现最佳的特征变量组合。运用 GBRT 算法预测纯电动汽

车的剩余续驶里程，并对预测结果从最大误差、最小误差、均方根误差、均值绝对误

差、R-square 五个角度进行分析。

最后得到的预测结果中，四个测试片段的 MAE 均在 5Km，R-squared 均在 0.95

以上，说明该模型具有较好的拟合能力，但也存在着很多的不足之处：

（1）训练数据的规模较小。原始数据经过筛选和预处理之后，符合 SOC 从 80 降至 20 的放电片段只有 18 段，数据量整体不大。未来可以采集更多的数据或是使用算法自动生成更多数据，扩大数据规模，提高数据质量。

（2）特征参数的纬度较少，只选用了 5 个特征参数参与模型训练。未来可以尝试根据先验知识或算法构建更多特征参数，扩充数据维度。

（3）本文只使用了 GBRT 一种回归算法预测纯电动汽车剩余续驶里程，未来可以选用多种机器学习算法融合的方式，提高模型的性能。 20

参考文献

[1] 王韶华. 基于低碳经济的我国能源结构优化研究[D]. 哈尔滨工程大学, 2013.

[2] BI J, WANG Y, SAI Q, et al. Estimating remaining driving range of battery electric vehicles based on real-world data: A case study of Beijing, China[J]. Energy, 2019,169(FEB.15): 833-843.

[3] 郭文双, 申金升, 徐一飞. 电动汽车与燃油汽车的环境指标比较[J]. 交通环保, 2002, 23(2): 4.

[4] 张文亮, 武斌, 李武峰, 等. 我国纯电动汽车的发展方向及能源供给模式的探讨[J]. 电网技术, 2009(4): 5.

[5] 杨晓东, 吕叶林, 许可. 基于 GRU-NN 模型的电动汽车实时能耗预测方法[J]. 交通节能与环保, 2022(004): 018.

[6] ZHAO L, YAO W, WANG Y U, et al. Machine Learning-Based Method for Remaining Range Prediction of Electric Vehicles[J]. IEEE Access, 2020, 8: 212423-212441.

[7] 田晟, 甘志恒, 吕清. 基于改进符号回归和 XGBoost 算法的剩余续驶里程预测方法[J]. 广西师范大学学报: 自然科学版, 2022, 40(2): 10.

[8] LAMANTIA M, SU Z, CHEN P. Remaining Driving Range Estimation Framework for Electric Vehicles in Platooning Applications[C]//2021 American Control Conference (ACC). 2021.

[9] AYEVIDE F K, KELOUWANI S, AMAMOU A, et al. Estimation of a battery electric vehicle output power and remaining driving range under subfreezing conditions [J/OL]. Journal of Energy Storage, 2022, 55: 105554. https://www.sciencedirect.com/science/article/pii/S2352152X22015456. DOI: https://doi.org/10.1016/j.est.2022 .105554.

[10] FRANKE T, KREMS J F. Interacting with limited mobility resources: Psychological range levels in electric vehicle use[J]. Transportation Research Part A: Policy and Practice, 2013, 48(FEB.): 109-122.

[11] 高航. 基于机器学习的纯电动汽车的行驶里程预测研究[D]. 北京交通大学, 2018.

[12] 刘光明, 欧阳明高, 卢兰光, 等. 基于电池能量状态估计和车辆能耗预测的电动汽车续驶里程估计方法研究[J]. 汽车工程, 2014, 36(11): 9.

[13] 李中耀, 李达峰. 纯电动汽车剩余续驶里程计算方法研究[J]. 汽车零部件, 2021(004): 000.

[14] XIAOFEN J, YONGCUN G, YOURUI H, et al. 基于斯皮尔曼等级相关性的彩色图像椒盐噪点检测算法[J]. 中国科学技术大学学报, 2019(001): 049. 21

[15] 贾科, 杨哲, 魏超, 等. 基于斯皮尔曼等级相关系数的新能源送出线路纵联保护[J]. 电力系统自动化, 44(15): 103.

[16] 吕清. 基于数据的纯电动车剩余续驶里程建模与预测研究[D]. 华南理工大学, 2021.

[17] 龚越, 罗小芹, 王殿海, 等. 基于梯度提升回归树的城市道路行程时间预测[J]. 浙江大学学报：工学版, 2018, 52(3): 8.

[18] 韩启迪, 张小桐, 申维. 基于梯度提升决策树 (GBDT) 算法的岩性识别技术[J]. 矿物岩石地球化学通报, 2018, 37(6): 8.

[19] 吴辰文, 梁靖涵, 王伟, 等. 基于递归特征消除方法的随机森林算法[J]. 统计与决策, 2017(21): 4.

[20] 毛勇, 皮道映, 刘育明, 等. 基于加速的支持向量机回归特征消去方法的关键变量辨识[J]. 中国化学工程学报（英文版）, 2006, 014(001): 65-72. 22

致谢

提笔到这里，我的四年大学生活也即将随着论文的结束而画上一个句号。借此机会，我想向那些一直以来给予我鼓励、支持和帮助的人表达最诚挚的感谢。

首先，我要感谢我的导师宋轩老师、副指导老师余庆老师对我的指导和帮助，使我受益匪浅。同时，我也要感谢课题组的学长学姐和同学们，感谢创新实践的队友们，感谢一直以来的陪伴和帮助。

其次，我要感谢致新书院和计算机系的老师们，感谢一直以来的教导、关心和帮助。同时，还要感谢我的家人和朋友们，你们的鼓励和支持是我克服困难的最大动力，也是我遇到困难时最坚强的后盾。

最后，再次诚挚地感谢所有支持和帮助过我的人，尽管我无法一一列举，但与你们的相识是我最大的幸运，谢谢你们！ 23

说明： 1.总文字复制比：被检测论文总重合字数在总字数中所占的比例
2.去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例
3.去除本人文献复制比：去除作者本人文献后，计算出来的重合字数在总字数中所占的比例
4.单篇最大文字复制比：被检测文献与所有相似文献比对后，重合字数占总字数的比例最大的那一篇文献的文字复制比
5.指标是由系统根据《学术论文不端行为的界定标准》自动生成的
6.红色文字表示文字复制部分;绿色文字表示引用部分;棕灰色文字表示作者本人已发表文献部分
7.本报告单仅对您所选择比对资源范围内检测结果负责

amlc@cnki.net
check.cnki.net

http://check.cnki.net/