# 长安大学人工智能创新大赛

# 项　目　计　划　书

| | |
|---|---|
| 项目名称： | 万里征程—基于车辆运行特征参数的细分区域辨识系统 |
| 参赛赛道： | □创意赛道 √创新赛道 □创业赛道 |
| 项目负责人： | 王子言 |
| 所属学院： | 汽车学院 |
| 联系电话： | 18803921869 |
| 团队成员： | 张耀文、郑浩然、马子涵、杜宣纬、杜京阳 |
| 指导老师： | 袁晓磊、赵轩 |
| 申报日期： | 2025.11.07 |

填写说明：请各参赛团队严格按照本模板格式和要求填写。内容需真实、详实，重点突出项目的创新性、技术合理性和完成度。评审将以此申报书为主要依据进行评分。篇幅不限、字数不限。

# 一、项目概要

（请用精练的语言概述整个项目，包括：1.核心问题：项目发现了什么值得解决的痛点或问题。2.解决方案：提出了怎样的 AI 创意或方案来解决它？方案的核心 AI 方法和技术是什么。3.创新价值：项目的主要创新点和预期价值，比如学术价值、应用效益、社会意义等。
填写说明：**创意赛道**强调想法的前瞻性和想象力；**创新赛道**强调技术的创新点和性能优势；**创业赛道**强调市场机会和商业潜力。）

本项目针对车企细分区域辨识人工标注耗时、成本高、传统模型精度低且可解释性弱的痛点，构建“数据支撑-特征优化-智能辨识”三位一体技术体系。核心方案为：首先，基于开源 GIS 构建离线带标签数据集；然后，通过箱线图+主成分分析两阶段优选特征参数；最后，采用“知识+数据”双驱动逻辑，结合布谷鸟算法优化随机森林模型。创新点在于低成本标注、分阶段特征筛选及双驱动模型设计，辨识准确率达 97.91%。在工程领域解决了行业技术痛点，满足车企定制化运行区域用户画像的需求；在学术上填补了交叉领域空白，提高了车辆运行区域辨识精度。

# 二、项目背景与问题定义

（填写说明：**创意赛道**重点展示问题的前沿性和社会/学术价值，体现前瞻性；**创新赛道**需精准分析现有技术方案的不足，为自身的技术创新做铺垫；**创业赛道**需进行初步的市场分析，证明市场缺口和用户痛点的真实性。）

## 2.1 研究背景

随着全球能源与环境问题严峻，我国推出《新时代的中国能源发展》白皮书推动汽车行业绿色低碳转型，降能耗、守合规成为车企核心使命。其中，汽车零部件可靠性是关键支撑，提升汽车零部件可靠性既能提升能源利用效率、减少故障导致的能耗浪费与排放超标，又能延长整车寿命、降低维护成本、提升品牌口碑，因此车企对其研究至关重要。

**汽车工况特征辨识与道路空间分布特性关联密切：**车企开展可靠性校核的核心是保障实际场景耐用性，而道路空间分布特性（道路类型、地形条件）直接决定工况特征，精准掌握其分布及行驶占比是针对性校核的关键。市区道路密集、路口多的分布特点，造就频繁启停、低速的工况；高速平直开阔的空间属性，形成匀速、高转速工况；山区地形起伏的分布特性，导致转矩频繁波动的工况。若忽视道路空间分布与工况的关联，未结合行驶占比开展校核，易出现测试与实际场景脱节，导致零部件无法匹配核心工况需求、可靠性不达标。因此，车企需精准把握道路空间分布特性及行驶占比，才能让校核贴合实际，保障产品稳定。

**企业定制化可靠性行驶试验方法：**为提供可靠性试验标准，国家出台的国家标准《汽车

可靠性行驶试验方法》（GB/T 12678—2021）中给出了汽车可靠性试验里程分配和配载，如下图所示。然而汽车制造企业更希望拥有定制化的行驶里程分配比例，从而对不同车型，进行定制化的可靠性校核。

（资料性）

**汽车可靠性试验里程分配和配载**

汽车可靠性试验里程分配和配载见表 A.1。

**表 A.1　基于用户调研的行驶里程分配比例和配载**

| 车辆类型 | | | 路面类型比例 | | | | | 配载比例 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 城市道路 | 高速公路 | 一般公路 | 山区公路 | 非铺装路 | 空载 | 半载[a] | 满载 |
| 乘用车 | | | 55% | 20% | 10% | 10% | 5% | 20% | 50% | 30% |
| 越野车 | | | 15% | 30% | 20% | 25% | 10% | — | — | 100% |
| 客车 | 城市客车 | | 50% | 10% | 30% | 5% | 5% | 10% | 50% | 40% |
| | 长途客车 | | 10% | 50% | 30% | 5% | 5% | 10% | 10% | 80% |
| 货车 | 载货车[b] | ≤7.5 t | 40% | 15% | 40% | 5% | — | 30% | 30% | 40% |
| | | >7.5 t～18 t | 10% | 30% | 50% | 10% | — | 20% | 40% | 40% |
| | | >18 t | 5% | 65% | 20% | 10% | — | 10% | 10% | 80% |
| | 牵引车 | | 5% | 70% | 15% | 10% | — | 10% | 10% | 80% |
| | 自卸车[b] | ≤18 t | 30% | — | 50% | 10% | 10% | 50% | — | 50% |
| | | >18 t | 20% | — | 50% | 10% | 20% | 50% | — | 50% |

注：以上比例仅供参考，检测机构或制造商可自行调整。

[a] 乘用车 5 座车半载按 3 人执行，7 座车半载按 4 人执行；货车半载按载货质量一半执行。

[b] 按照最大允许总质量进行划分。

**图 2.1 可靠性试验里程分配比例国家标准**

**细分区域下预测型能量管理策略与动力学控制效果更优：**基于市区、高速、郊区、山区的道路空间分布特性与工况差异，预测型能量管理策略可提前捕捉不同区域的负荷特征，市区依托频繁启停的工况预测，动态调整能量回收强度与电机驱动模式，减少低速空载损耗；高速基于匀速高转速特征，优化动力分配与能量回收阈值，降低风阻与机械损耗；山区结合地形起伏预测，提前适配转矩输出与电池充放电策略，避免过载与能量浪费。同时，针对性的动力学控制可匹配各区域行驶需求：市区强化启停平顺性与瞬时响应，高速保障匀速稳定性，山区提升抗冲击与转矩调节精度。这种贴合细分区域工况的控制逻辑，相比通用型策略，能显著降低整车能耗、提升动力响应精准度，同时减少零部件负荷波动，延长使用寿命，整体控制效果更优。

由上述可知，企业迫切需要**低成本、高效率、高保真**的细分区域辨识技术。但是传统方案中，人工标注的方法需要耗费大量时间、人力成本，且企业构建的辨识模型存在辨识精度低，可解释性差的问题。为解决企业迫切需求，本团队创新性地提出了一种结合人工智能技术的车辆短行程片段运行区域辨识系统，该系统通过独立构建的带标签数据库不仅可以学习短行程片段特征参数与行驶区域间的内在关系，代替人工标注，大幅节约成本，且相比于一般的深度学习模型，可解释性更强，且在创新性地引入先验知识驱动的区域划分策略后，识别准确度明显提高。

## 2.2 问题定义

针对上述背景，将问题定义如下：

**首先，需要解决数据库的构建和标注问题。**一般企业均在大数据云平台存有其车辆的历史运行数据，但企业现有技术无法对运行片段进行运行区域标注。因此，需要首先提出一套算法，利用 GPS 等数据信息，对典型城市数据集进行准确标注。

**其次，需要解决特征参数集的构建和筛选。**一般方法常常通过先验知识提出常见的特征参数，然后利用深度学习算法进行训练，这种方法常常会导致特征参数冗余，降低模型泛化性能和收敛速度。因此，需要提出一种技术确保特征参数有效性。

**最后，需要解决模型的构建和优化训练。**项目面向汽车零部件可靠性校核，需保证辨识结果可追溯，企业需要可解释性更强的人工智能算法，而一般深度学习模型如卷积神经网络（CNN）、循环神经网络（RNN）、大语言模型（LLM）存在可解释性差、调参复杂、训练成本高昂、难以学习高维结构化特征参数信息的问题。因此需要剔除一种可解释性强、部署方便、训练成本低廉的人工智能模型。

# 三、核心方案与实现过程

（包括：1.总体方案概述：用文字、流程图或系统架构图，整体介绍项目是如何运作的；2.核心技术原理：使用了哪些人工智能技术（如：深度学习、自然语言处理、计算机视觉等）；3.实现过程：论述技术可行性或功能的完整性。

填写说明：**创意赛道**详细描述创意的原创性和新颖度，阐述技术路径的合理性和想象力，可不涉及具体代码；**创新赛道**需说明技术/算法的创新细节，需详细展示完整的技术实现流程、工作量和技术深度；**创业赛道**强调其如何解决用户痛点，可表现为独特的商业模式或资源整合方式等，并证明团队已具备将想法落地的技术执行力。）

## 3.1 总体方案

本项目以车辆运行区域高精度辨识为核心目标，构建“数据基础支撑-特征工程优化-智能模型辨识”三位一体的递进式技术体系，三部分工作紧密衔接、相互支撑，弥补了传统辨识方案工作量大、成本高、识别精度低的缺陷，实现市区、高速、郊区、山区四类行驶区域的精准区分，具体内容如下：

**第一部分为基于开源地理信息系统的运行区域离线数据集构建**，该部分是整个项目的基础数据支撑。为了保证训练数据集的多样性，首先选取上海、北京、成都、西安、郑州、长春六个典型城市。其次明确数据来源为包括 OSM 道路数据与 SRTM3-90 米 DEM 数据的开源地理信息数据，随后采用 ArcGIS 作为核心处理工具，完成数据预处理工作。接着开展区域边界构建，市区以绕城高速、市区环线等作为市区边界，高速区域在全市道路库中按属性选取高速公路要素，删除其他类型道路，直接沿路线间隔 100 m 构造测量点，计算测量点经纬度，输出重庆市全市高速道路 GPS 数据库，山区通过目标范围内地形起伏度的计算，得出该区域的山区界定标准，郊区则为剩余部分的区域，四类区域无重叠、覆盖完整。最终基于 ArcGIS 软件输出的典型城市四个道路区域的离线 GPS 范围库，是构建训练集分类标签的基础。

**第二部分为基于最优特征参数表达机制的运行区域特征提取**，旨在从高维行驶数据中筛选出具有强判别能力的核心特征。首先基于领域先验知识与文献调研，构建完备的短行程片段特征参数库，充分涵盖与运行区域相关的驾驶行为与工况信息。随后采用两阶段筛选机制提炼有效特征：第一阶段通过箱线图分析与特征参数分布曲线分析法，剔除判别能力弱的参数；第二阶段对保留特征进行主成分分析（PCA），并进一步得到出在这些主成分中贡献度更高的特征子集。最终提取了包括：速度类的运行速度、平均速度等，转矩类的转矩标准差、平均正/负转矩等 56 个最优特征参数。该特征子集在保留原始信息代表性的同时，具备优异的类别判别能力，为后续运行区域辨识模型的性能提升奠定关键基础。

**第三部分为基于数据-知识双驱动的细分区域辨识**，该部分是项目实现区域精准区分的核心环节，针对四类区域同步辨识时市区与郊区/山区特征重叠度高、精度低的问题，采用“知识驱动先筛选+数据驱动后训练”的递进逻辑。首先通过知识驱动完成易区分区域的快速剥离，实现市区采用关键 GPS 特征点（起点、终点、3 个关键拐点）快速识别、高速/郊区/山区采用细微特征捕捉能力强、优化特征子集选取全的布谷鸟算法优化的随机森林模型进行三类运行区域的高精度辨识。解决传统多区域同步辨识单一驱动模型在复杂区域辨识中的精度瓶颈问题。

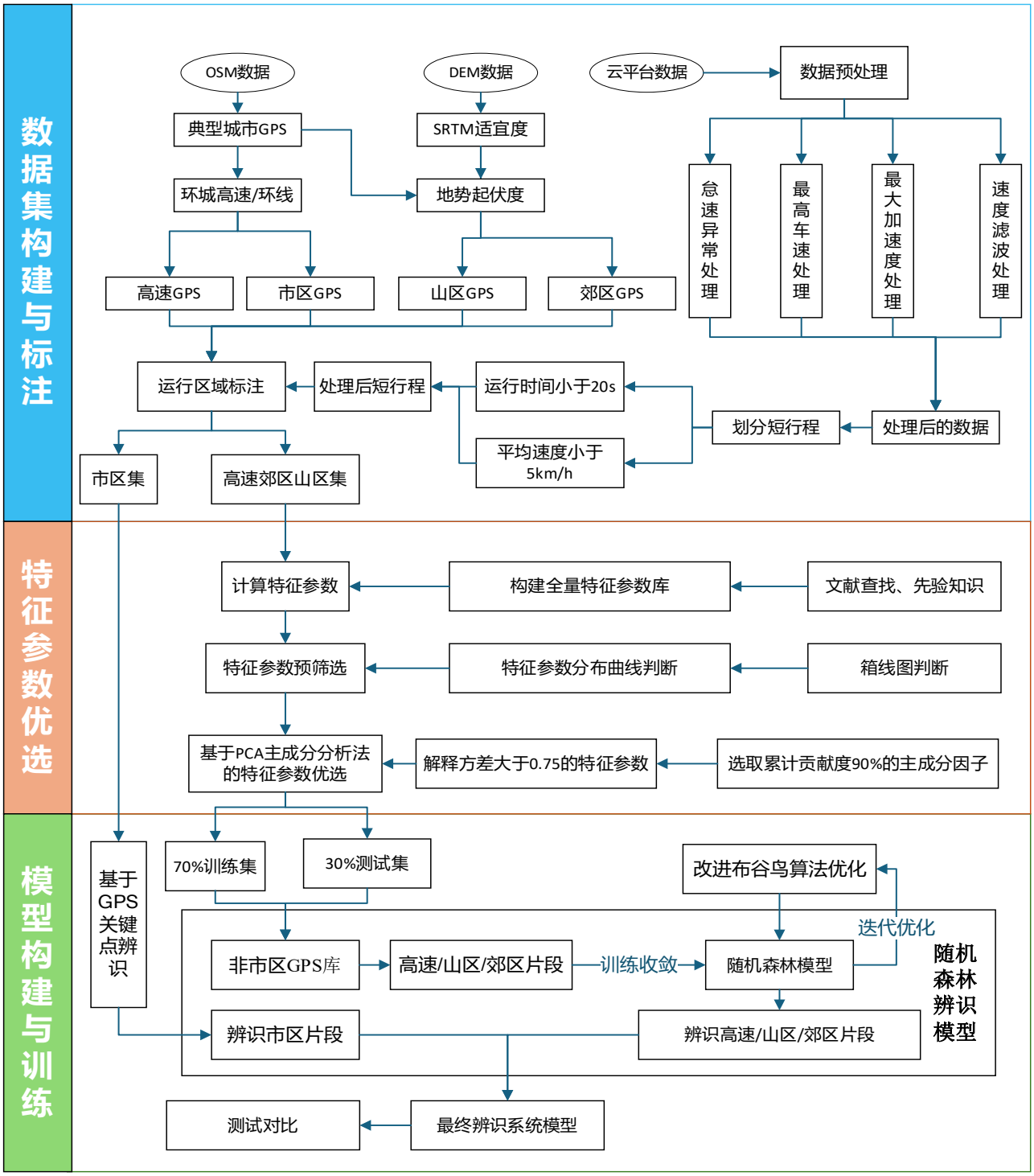



**图 3.1 整体技术方案流程图**

## 3.2 具体技术流程

### 3.2.1 基于开源地理信息系统的运行区域离线数据集标注方法

基于开源地理信息系统（整合 OSM 道路数据、SRTM3-90 米 DEM 数据等），通过 ArcGIS 完成数据预处理（格式转换、坐标统一）、区域边界构建（划分市区/高速/郊区/山区）与属性标准化设计，最终构建含空间边界与分类标签的运行区域离线标注数据集；有效解决在线地理数据依赖网络、商业数据库成本高的问题，同时提升区域标签离线应用的精度（GPS 边界误差≤50m）与跨平台复用性，最终实现了短行程短片的分区域标注。

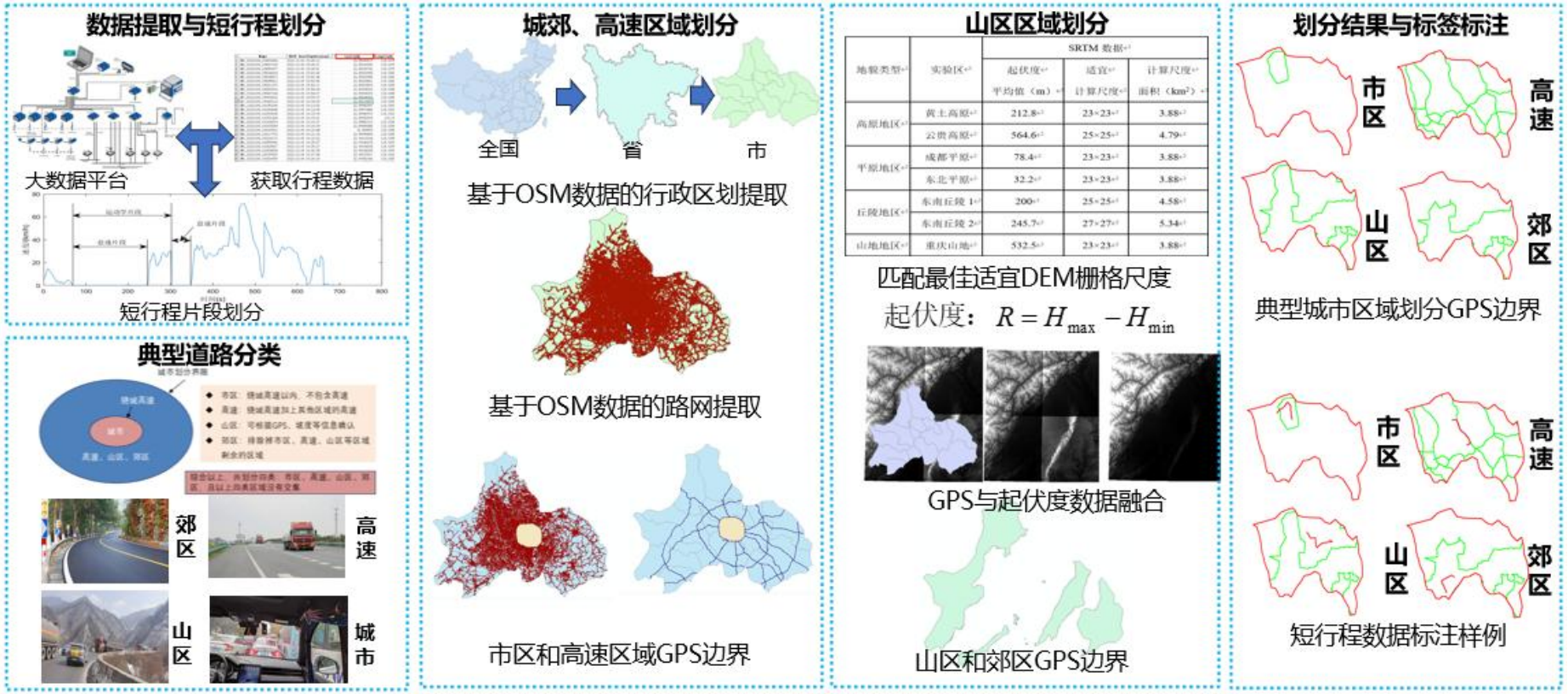



图 3.2 离线数据库建立技术方案图

## 创新点

### 一、提出了基于开源地理信息系统的深度学习训练库区域标注方法

创新性地提出了基于开源地理信息系统的训练库区域标记方法，解决了运行区域划分深度学习训练库传统商业地图标记成本高的问题，基于“无交集、全覆盖” 原则，将目标城市运行区域划分为 4 类，标注逻辑如下图所示：

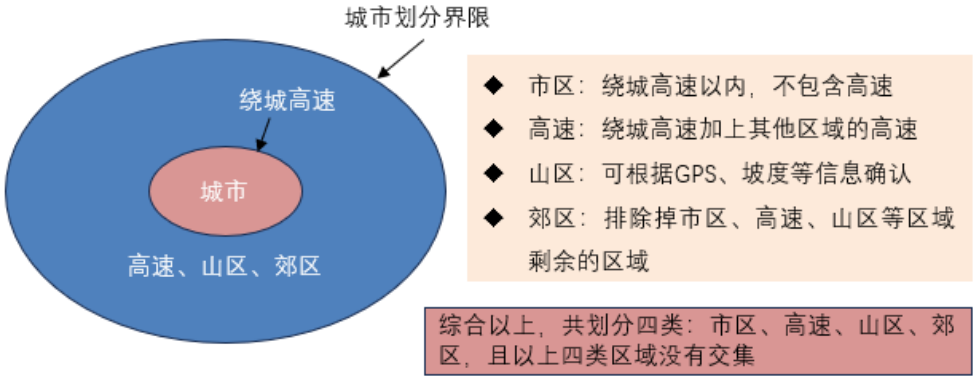



**图 3.3 区域定义**

本次项目中选择了上海、北京、成都、西安、郑州、长春六个典型城市，典型城市覆盖全，大幅提升了训练集泛化能力。

市区边界构建：导入省级/县级行政区、道路数据，以绕城高速/环线为界，生成城区、郊区矢量面；边界线按 100 米间隔打点，获取城区、郊区 GPS 点。

高速道路库构建：从郊区道路中提取高速及匝道，删除其他道路，按 100 米间隔打点，生成高速 GPS 点。

山区道路库构建：导入 SRTM3-90 米 DEM 数据，镶嵌合并计算地形起伏度与郊区面相交，裁剪出山区范围；合并零散图块生成山区 GPS 点。

郊区公路构建：郊区范围对山区取反，面转线后按 100 米间隔打点，生成普通公路 GPS 点。至此完成典型城市的区域划分。

最终完成了 6 个典型城市，每个区域分别 6000 个短行程片段（累计 24000 个短行程片段）的区域标记。

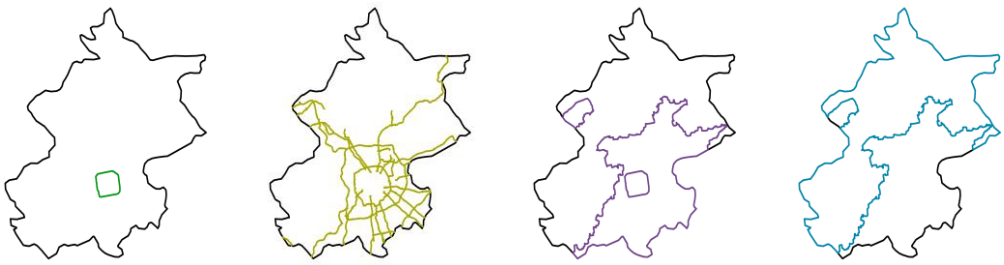
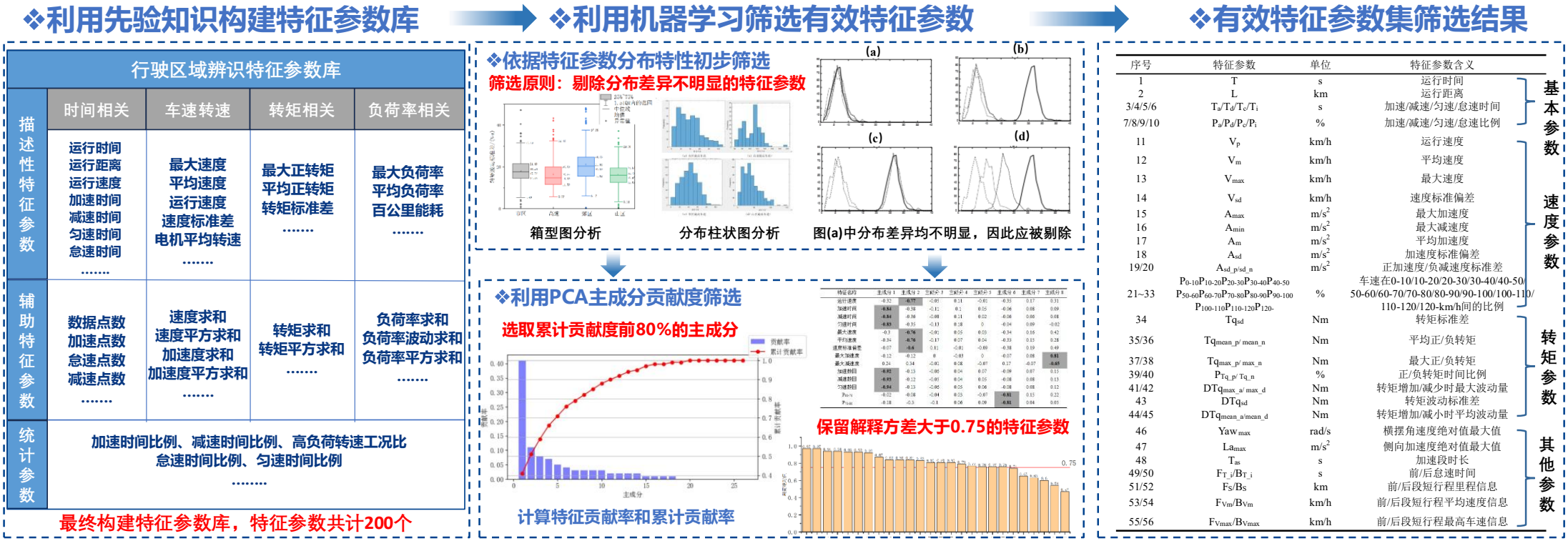


**图 3.4 北京市城市、高速、郊区、山区道路 GPS 划分**

### 3.2.2 基于最优特征参数表达机制的运行区域特征提取技术

为提升车辆运行区域辨识准确率，本研究采用基于最优特征参数表达机制的运行区域特征提取技术。首先，基于领域先验知识与文献调研构建完备的短行程片段特征参数库，以充分涵盖与运行区域相关的驾驶行为与工况信息。随后，通过两阶段筛选机制提炼有效特征参数：第一阶段采用箱线图、特征参数分布曲线分析法，依据特征在不同运行区域间的分布差异进行初筛，剔除判别能力弱的参数；第二阶段对保留特征进行主成分分析，筛选出累计贡献率前 80%的主成分中解释方差大于 0.75 的特征。最终提取了包括：速度类的运行速度、平均速度等，转矩类的转矩标准差、平均正/负转矩等 56 个最优特征参数，在保留原始信息代表性的同时，具备优异的类别判别能力，对后续运行区域辨识模型的性能提升具有关键作用。




| 序号 | 特征参数 | 单位 | 特征参数含义 | |
|---|---|---|---|---|
| 1 | T | s | 运行时间 | 基本参数 |
| 2 | L | km | 运行距离 | |
| 3/4/5/6 | $T_a/T_d/T_c/T_i$ | s | 加速/减速/匀速/怠速时间 | |
| 7/8/9/10 | $P_a/P_d/P_c/P_i$ | % | 加速/减速/匀速/怠速比例 | |
| 11 | $V_p$ | km/h | 运行速度 | 速度参数 |
| 12 | $V_m$ | km/h | 平均速度 | |
| 13 | $V_{max}$ | km/h | 最大速度 | |
| 14 | $V_{sd}$ | km/h | 速度标准偏差 | |
| 15 | $A_{max}$ | $m/s^2$ | 最大加速度 | |
| 16 | $A_{min}$ | $m/s^2$ | 最大减速度 | |
| 17 | $A_m$ | $m/s^2$ | 平均加速度 | |
| 18 | $A_{sd}$ | $m/s^2$ | 加速度标准偏差 | |
| 19/20 | $A_{sd\_p/sd\_n}$ | $m/s^2$ | 正加速度/负减速度标准差 | |
| 21~33 | $P_{0\text{-}10}P_{10\text{-}20}P_{20\text{-}30}P_{30\text{-}40}P_{40\text{-}50}$ $P_{50\text{-}60}P_{60\text{-}70}P_{70\text{-}80}P_{80\text{-}90}P_{90\text{-}100}$ $P_{100\text{-}110}P_{110\text{-}120}P_{120\text{-}}$ | % | 车速在0-10/10-20/20-30/30-40/40-50/50-60/60-70/70-80/80-90/90-100/100-110/110-120/120-km/h间的比例 | |
| 34 | $T_{qsd}$ | Nm | 转矩标准差 | 转矩参数 |
| 35/36 | $T_{qmean\_p/mean\_n}$ | Nm | 平均正/负转矩 | |
| 37/38 | $T_{qmax\_p/max\_n}$ | Nm | 最大正/负转矩 | |
| 39/40 | $P_{Tq\_p/Tq\_n}$ | % | 正/负转矩时间比例 | |
| 41/42 | $DTq_{max\_a/max\_d}$ | Nm | 转矩增加/减少时最大波动量 | |
| 43 | $DTq_{sd}$ | Nm | 转矩波动标准差 | |
| 44/45 | $DTq_{mean\_a/mean\_d}$ | Nm | 转矩增加/减小时平均波动量 | |
| 46 | $Yaw_{max}$ | rad/s | 横摆角速度绝对值最大值 | 其他参数 |
| 47 | $La_{max}$ | $m/s^2$ | 侧向加速度绝对值最大值 | |
| 48 | $T_{as}$ | s | 加速段时长 | |
| 49/50 | $F_{T\_i}/B_{T\_i}$ | s | 前/后怠速时间 | |
| 51/52 | $F_S/B_S$ | km | 前/后段短行程里程信息 | |
| 53/54 | $F_{Vm}/B_{Vm}$ | km/h | 前/后段短行程平均速度信息 | |
| 55/56 | $F_{Vmax}/B_{Vmax}$ | km/h | 前/后段短行程最高车速信息 | |

图 3.5 最有特征参数筛选技术方案图

## 创新点

本研究的创新性主要体现在特征参数的筛选方法上，通过设计一种分阶段、多准则的混合筛选策略，有效提升了特征子集的判别性与有效性。

### 一、基于箱线图与分布曲线的特征优选技术

传统特征工程往往直接将全部特征输入模型，可能引入冗余与噪声。为克服此问题，本研究创新性地在流程前端引入了基于箱线图与分布曲线的可视化统计分析方法，作为特征筛选的第一阶段。该方法的核心在于直接、直观地评估每个特征参数在不同运行区域下的统计分布差异。通过绘制并比较各特征于不同类别下的箱线图与分布曲线，定性与定量地判断其特征的类别判别潜力。若某特征在所有运行区域中的分布高度重叠、形态相似如下图（a）所示的特征参数，则判定其为“无判别力参数”，认为其对后续分类模型的贡献度极低，故在早期予以剔除。

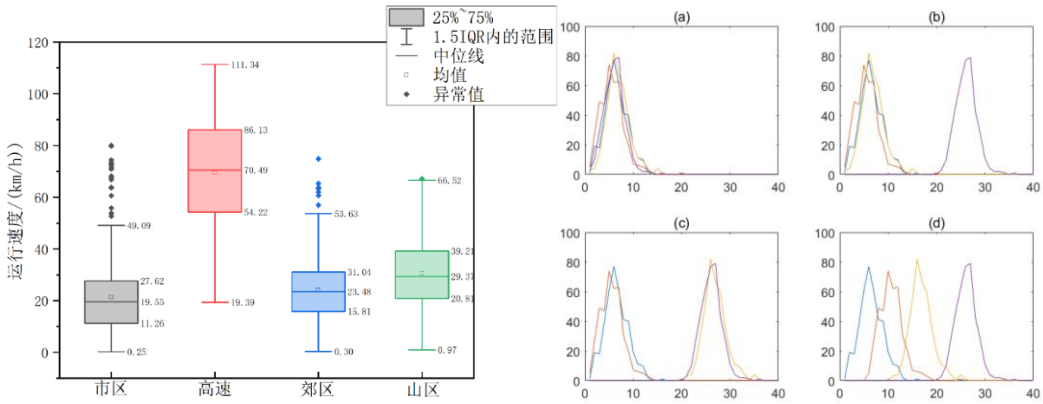



**图 3.6 箱线图与分布曲线图**

### 二、基于主成分载荷矩阵分析的特征参数优选技术

在通过初筛的特征子集基础上，本研究创新性地提出一种基于主成分分析（PCA）的二次筛选策略。该策略并不直接使用主成分进行降维，而是将其作为特征价值的衡量标准：首先对特征集进行 PCA 分析，随后设定双重筛选准则：特征须被包含于累计贡献率前 80%的主成分中，以确保其共同承载关键信息；同时，其在主成分上的载荷必须大于 0.75，以保证其对核心变异方向具有决定性贡献。

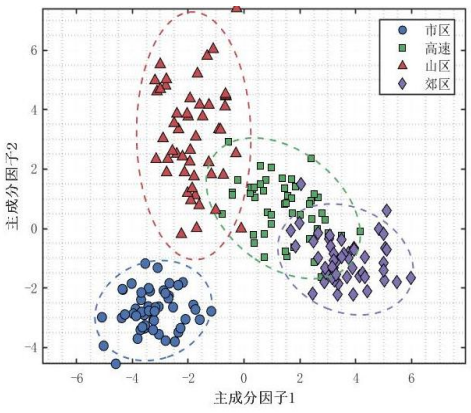



**图 3.7 主成分分析优选**

此筛选策略充分发挥了 PCA 在揭示数据结构方面的优势，通过兼顾特征间的全局相关性与个体代表性，有效避免了因保留高度线性相关特征而引发的过拟合风险。最终所得特征子集在信息量与判别性之间取得优化平衡，显著提升了运行区域辨识模型的泛化能力与鲁棒性。

### 3.2.3 基于数据-知识双驱动的细分区域辨识技术

由于历史经验及大量数据测试可知，市区、高速、郊区、山区 4 个道路运行区域同步辨识时无法准确区分，尤其体现在市区与郊区/山区。针对市区与郊区/山区片段特征重叠度高、同步辨识精度低的核心问题，采用“知识驱动先筛选+数据驱动后训练”的递进逻辑，从根本上解决了数据混杂导致的精度瓶颈。实现市区首先采用起点、终点、3 个关键拐点等 GPS 特征点快速识别，对不满足判别条件的采用短行程全量数据点进行细化百分比模式辨识；高速/郊区/山区采用细微特征捕捉能力强、优化特征子集选取全的布谷鸟算法优化的随机森林模型进行三类运行区域的高精度辨识。解决传统多区域同步辨识单一驱动模型在复杂区域辨识中的精度瓶颈问题。

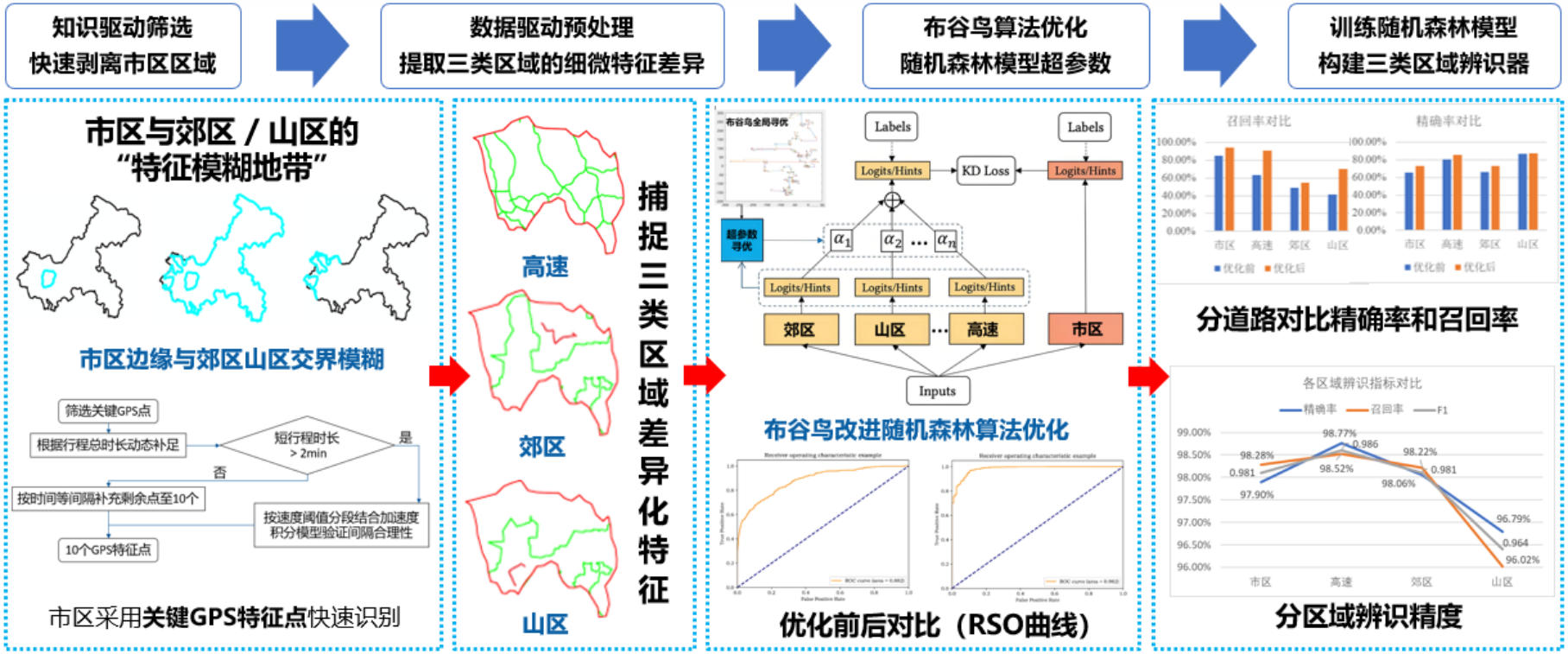



图 3.8 基于数据-知识双驱动的细分区域辨识技术

## 创新点

### 一、提出了数据—知识双驱动递进逻辑的道路运行区域辨识技术

本团队提出的“数据——知识双驱动”并非数据与知识的简单叠加，而是针对不同区域的辨识难度，分配两种驱动的核心作用：知识驱动解决“易混淆区域的快速剥离”，数据驱动解决“复杂区域的精准区分”，两者形成“筛选—训练”的递进关系。

**表 3-1 数据—知识双驱动技术**

| 驱动类型 | 核心依据（输入） | 作用对象 | 目标 | 技术载体 |
|---|---|---|---|---|
| 知识驱动 | 先验规则<br>地理边界 | 市区 | 快速剥离<br>排除混淆 | ArcGIS（GPS 库）的 3 个关键特征点的快速识别 |
| 数据驱动 | 大数据特征<br>行驶参数、地形关联 | 高速、郊区、山区 | 精准区分<br>细微差异 | 布谷鸟算法优化的随机森林（机器学习模型） |

先基于先验知识对易于区分的市区行驶区域进行划分，针对高速、郊区与山区行驶区域，本模型采用布谷鸟算法优化的随机森林模型对其进行进一步高精度辨识。

首先基于 DEM 地形数据，在 ArcGIS 中计算候选数据片段的地形起伏度$R = H_{max} - H_{min}$，$R \geq 200m$的标记为山区；随后结合高速道路独立 GPS 界定库（ArcGIS 预生成）锁定高速区域，再用 ArcGIS 排除山区 GPS 范围，剩余非高速、非山区的片段标记为郊区。三类区域的特征重叠度从多区域同步辨识时的 42%降至 8%，显著提升区域识别精度。

### 二、提出了布谷鸟算法优化的随机森林模型的大数据分类技术

布谷鸟算法通过迭代更新候选解、淘汰劣质解收敛至最优解，对非线性、高维优化问题适配性极强。而随机森林作为集成学习算法，在大数据分类中超参数依赖经验设定、易陷局部最优，本团队创新性采用布谷鸟算法对随机森林模型中的超参数进行迭代优化：

将随机森林所有超参数编码为鸟巢位置，以分类准确率为适应度函数，通过 Levy 飞行搜索超参数可行域。，以“分类准确率—训练时间”加权值为适应度函数，借助巢穴淘汰机制，避免超参数盲目设定，兼顾模型泛化能力与训练效率。通过布谷鸟算法局部搜索最优节点分裂特征，替代随机选择；对分类误差大的决策树，重新生成样本与特征组合并淘汰，降低单树随机性偏差，提升集成模型稳定性，适配样本分布不均场景。

模型分类精度提升，随机森林准确率从 87.09%升至 98.77%，在不同地形城市道路辨识中泛用性更优，可精准捕捉高速、郊区与山区本质差异，为高精度辨识筑牢基础。

# 四、结果与验证

（填写说明：**创意赛道**可通过模拟实验、理论推导或详细的场景推演来验证想法的潜在效果；**创新赛道**必须提供严谨、可复现的实验数据和对比结果；**创业赛道**可通过用户访谈、小范围试点数据、原型测试反馈等来验证产品价值。）

## 4.1 数据集采集与预处理

### 一、大数据云平台提取

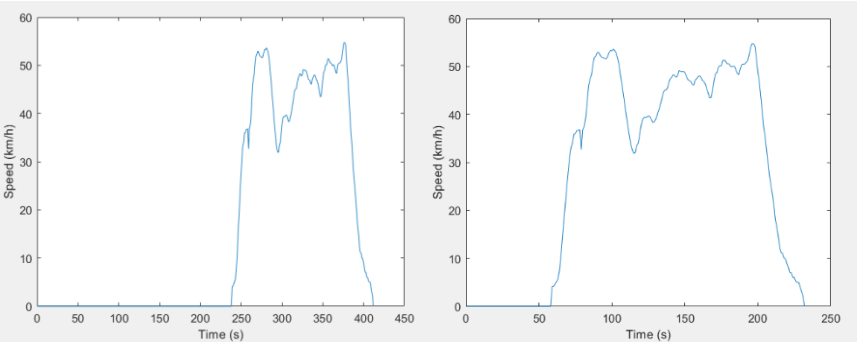
传统数据采集依赖线下实车采集方式，但这种方式耗时久、成本高。随着大数据平台的发展，现有平台已积累海量车辆运行数据，因此本项目直接从线上大数据平台获取相关数据，经处理后得到 179006 条行驶数据。后续需对该批数据进行深度处理，以提取更多特征参数。

### 二、数据预处理

针对采集的车辆运行特征参数数据异常处理包括其对其怠速时间异常，最高车速数据以及加速度数据异常，速度滤波处理等等。

（1）怠速时间异常

车辆运行特征参数中的怠速时间异常，指车辆处于停车、等红灯等静止状态时，电机持续运行时间过长或超出设定阈值，进而导致车辆长时间滞留、电机额外磨损等问题，这一数据的采集在基于车辆运行特征参数的运行区域识别系统设计中是关键环节，可缩短运动学片段时长以便利后续计算，依据《轻型汽车燃油消耗量和排放限值》（GB18352.5-2016）规定，轻型汽车发动机怠速状态下的持续工作时间不应超过 180s，车辆怠速时长超出此数值即判定为异常，且该阈值可根据实际应用场景与数据采集目标灵活调整，而此类异常数据可用于识别交通拥堵、停车场等特定地理区域—正常停车时长通常较短，交通拥堵等情况会导致怠速时间异常，通过分析怠速异常车辆的行驶轨迹与位置信息，能够精准识别区域交通流状态。

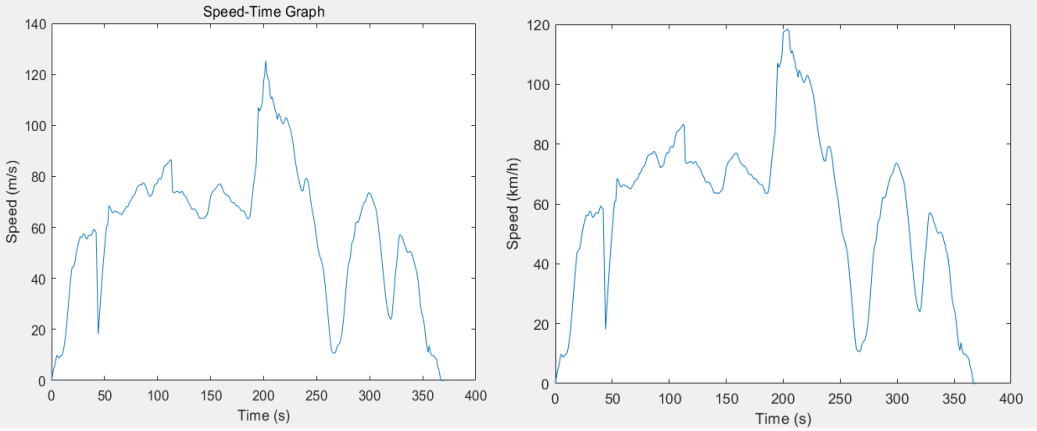



(a)怠速时间处理前　　（b）怠速时间处理后

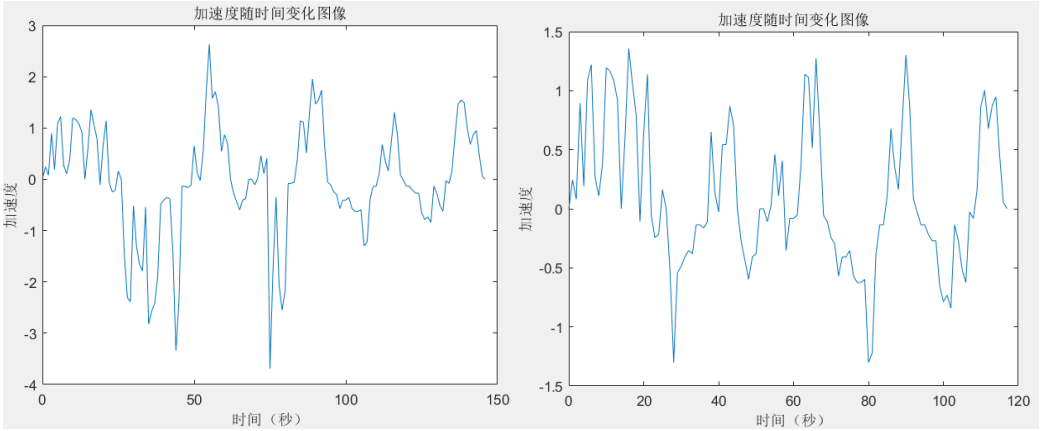
**图 4.1 怠速时间处理后**

（2）最高车速以及加减速度异常

无论车辆在哪一个以上提到的区域行驶时，车辆的最高车速均不能超过 120 $km \cdot h^{-1}$，所以大于 120 $km \cdot h^{-1}$ 的数据就是异常的，我们需要将其进行删除。对于加速度也有相应的限制，根据国家市场监管总局发布的 GB/T19596-2017《汽车加速性能试验方法》规定的：0-50km/h 加速度标准是最为常用的标准，其要求汽车从 0km/h 加速至 50km/h 的时间不得超过 10 秒，应该将加速度异常的数据也进行剔除，处理前后数据对比图如图所示。




(a)超速数据处理前　　（b）超速数据处理后




（c）加速度处理前　　（d）加速度处理后

**图 4.2 加速度处理前后**

（3）对速度数据进行滤波处理

在基于车辆运行特征参数的运行区域识别系统中，对车辆参数数据进行滤波处理，核心目的是获取更清晰稳定的数据，进而提升车辆行驶状态的识别与定位精度。现实中车辆行驶会受路面状况、风力、装载情况等多种因素影响，这些因素会给实际采集的数据带来噪声干扰，若不进行滤波处理，极易导致识别与定位出现误差，因此对车速和加速度数据实施滤波处理，能够有效去除噪声、获得相对准确的核心参数。由于车辆运行特征参数数据在外界因

素干扰下存在噪声，会造成数据偏差，本项目采用滑动平均滤波算法进行处理：通过设定固定长度的滑动窗口，用窗口邻域内若干原始数据的均值替代对应位置的原始数据，最终形成新的均值序列：

$$y(t)=\frac{1}{T}\sum_{k=0}^{T}x(t-k) \tag{4.1}$$

式中,y(t)为平均值,t=1、2…n。n 为总数据长度，T 为时间步长，x(t)为原始速度数据。由图 2-3 可得出，再进行数据滤波前，原始速度数据会存在尖点，从而会引起导致速度值加速度值突变，在进行数据滤波后，得到相应的速度曲线变得平滑，滤波前后得到的数据的车速曲线较吻合，从而达到了降噪的目的，这样可以保证计算结果维持在合理范围，如图 4.3。

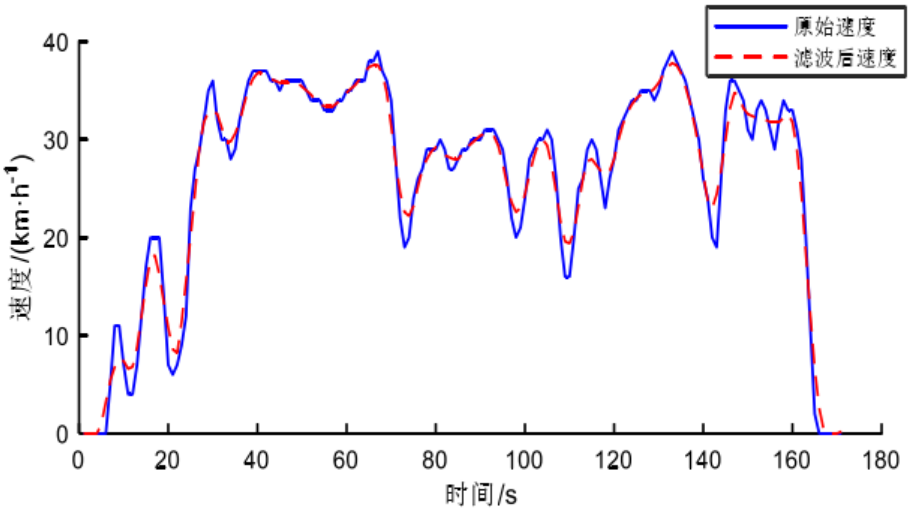



**图 4.3 数据滤波前后的图像**

经过以上的数据预处理工作之后，会得到对机器算法模型处理工作更加有力的数据，而且得到的数据更加的精简，这些处理以后的数据更加地减少了以后算法的数据处理量，对之后的车辆运行区域的预测更加的准确，并且更加有利之后的短行程的划分。

## 三、短行程划分

（1）划分短行程

我们将会根据车速将数据划分为片段是基于车辆运行数据的特点和数据处理的需要而来的。车辆从起步出发到达目的地，受到交通规则和道路交通情况的影响，中间会有很多次的起步加速以及停车操作；这一整个的运动过程可以看成有很多的运动学片段组成的。在这里就是通过短行程法将整个过程划分成运动学片断，每一个运动学片段都是由加速阶段、减速阶段、匀速阶段、怠速阶段组成的，如图 4.4 所示，根据国标由下述公式对车速、加速度进行运动学片划分。

$$\begin{cases} a \le 0.15m \cdot s^{-2} \\ a \ge -0.15m \cdot s^{-2} \\ v \ge 0.5km \cdot h^{-1} \cup \left| a < 0.15m \cdot s^{-2} \right| \\ v < 0.5km \cdot h^{-1} \cup \left| a < 0.15m \cdot s^{-2} \right| \end{cases} \tag{4.2}$$

按照车速划分短行程片段：需要从实际的车辆运行数据中提取出车速数据，从最开始车辆起步开始运动，车速从零开始，根据车速的大小，判断车辆当前是否在行驶中；当车速再次为零时，说明车辆已经停止行驶。此时，可以认为当前的片段已经结束，并可将其保存下来，准备开始记录下一个片段。通过车速的预处理之后，我们将大数据平台获取的数据划分成 24000 个短行程片段。

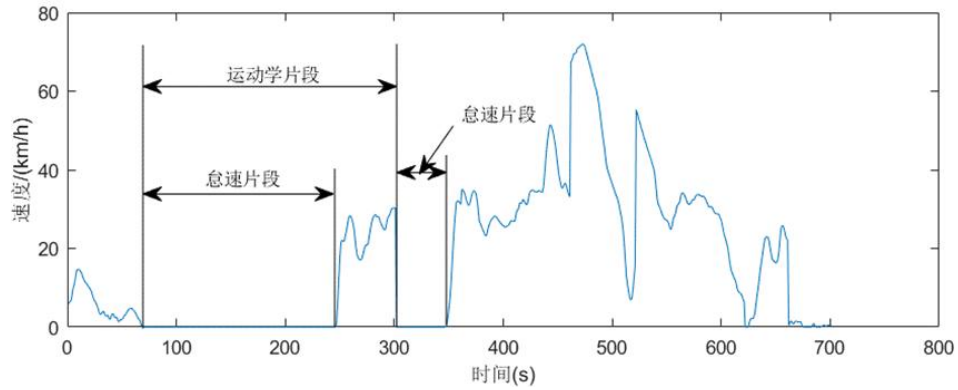
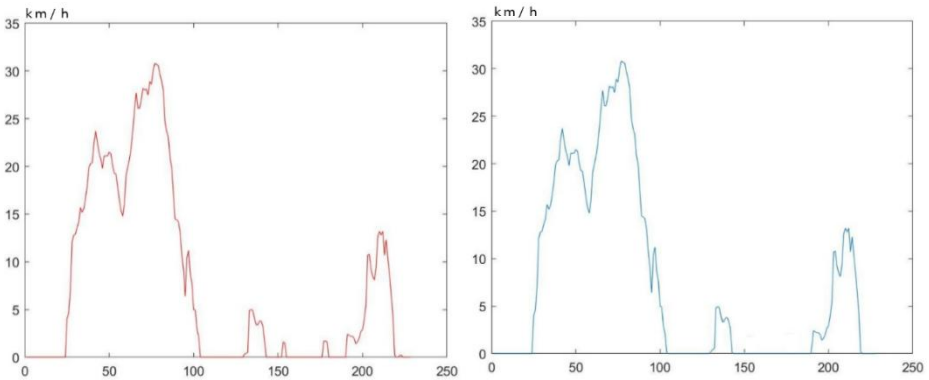



**图 4.4 运动学片段组成**

（2）异常短行程处理

由于我们是根据车速划分的短行程片段，即首先是从车速数据中的零开始，中间有不为零的车速数据，然后再到车速为零的数据结束；根据这个划分短行程的方法，在划分短行程的时候，会划分出一些异常短行程，如图 4.5 所示，将短行程中全部的车速均低于 $5\,\mathrm{km \cdot h^{-1}}$ 的时间小于 20 秒的，将其进行删除。

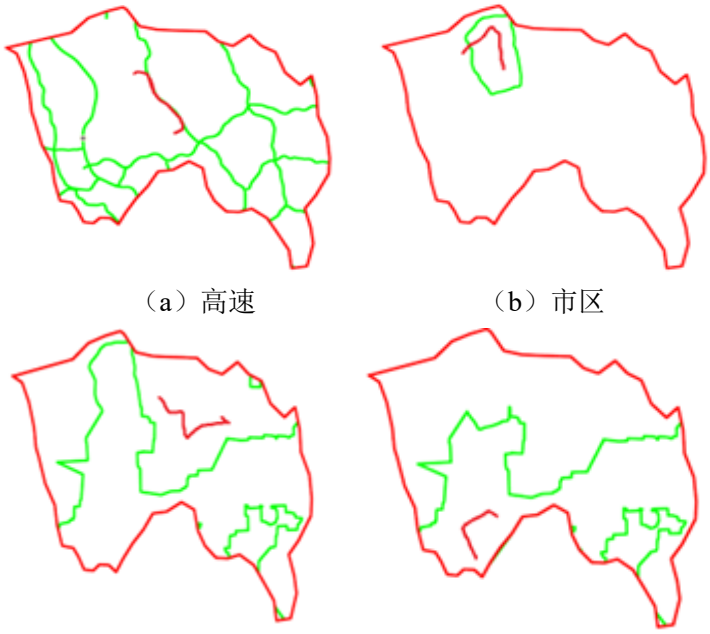



（a）异常处理前　　（b）异常处理后

**图 4.5 异常短行程的处理**

## 4.2 数据集标注

在基于车辆 GPS 定位信息与前文 3.2.1 章节所建典型城市运行区域数据库的基础上，识别各行驶片段所属区域并将其划分为市区、郊区、高速、山区四类，分类标签分别对应 0、1、2、3，为后续训练数据集的构建提供支持，测试集识别示例如下图所示。



（a）高速　　（b）市区

（c）郊区　　（d）山区

**图 4.6 测试集辨识示例**

## 4.3 最优特征参数筛选

通过选择合适的特征并对其进行提取，可以使机器学习模型更容易捕获数据的潜在规律和模式，从而提高模型的性能和泛化能力。因此，特征优选的方法选择尤其重要。

### 4.3.1 箱线图、分布曲线分析法预筛选结果

如下所示为基于箱线图分析法进行特征参数预筛选的示例。

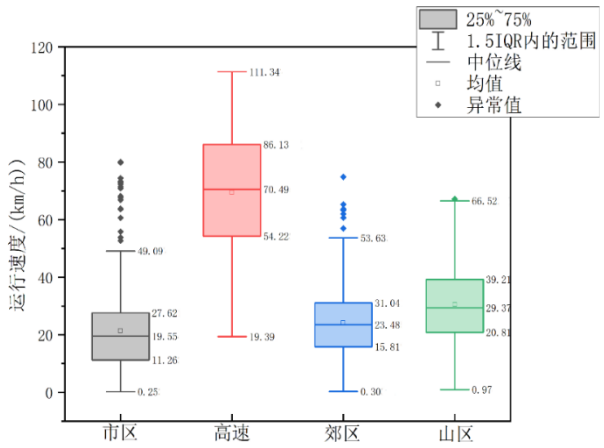



**图 4.7 运行速度的分组箱线图**

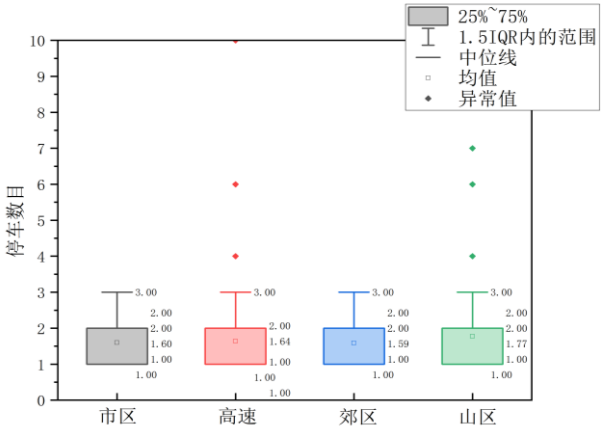



**图 4.8 停车数目分组箱线图**

从图 4.7 中可以看出“高速”箱体下限值为 54.22km/h，“市区”箱体上限 27.62km/h、“郊区”箱体上限 31.04km/h、“山区”箱体上限 39.21km/h，可见，“高速”箱体完全位于其它箱体上方，说明运行速度这一特征参数体现出汽车在高速行驶时与在市区、郊区、山区行驶时相比具有较大差异性。

从图 4.8 停车数目对应的分组箱线图中关于的“市区”、“高速”、“郊区”和“山区”的箱线图各组间重叠度较大。具有较小的差异性，对汽车行驶区域具有较低的辨识度。不宜作为区域辨识系统设计的车辆运行特征参数选择。

如上所述，如果一个特征的按区域分组箱线图的各区域类别箱体部分重合度较高，出现所有区域之间都有重合，这样的特征参数不适合用于运行区域辨识系统设计，应当剔除。经过对 200 个特征参数的分析初筛选特征如表 4-1 所示部分。

**表 4-1 箱线图分析后保留的特征（部分示例）**

| 序号 | 特征名称 | 序号 | 特征名称 |
|---|---|---|---|
| 1 | 运行速度 | 14 | P70-80 |
| 2 | 加速时间 | 15 | P80-90 |
| 3 | 减速时间 | 16 | P90-100 |
| 4 | 匀速时间 | 17 | P100-110 |
| 5 | 最大速度 | 18 | 制动次数 |
| 6 | 平均速度 | 19 | 正转矩时间比例 |
| 7 | 速度标准偏差 | 20 | 负转矩时间比例 |
| 8 | 最大加速度 | 21 | 转矩增加时平均波动量 |
| 9 | 最大减速度 | 22 | 转矩减小时平均波动量 |
| 10 | 加速数目 | 23 | 后段短行程最高车速 |
| 11 | 减速数目 | 24 | 前段短行程最高车速 |
| 12 | 匀速数目 | 25 | 前前段短行程里程 |
| 13 | P60-70 | 26 | 后后段短行程里程 |

### 4.3.2 PCA 主成分分析法筛选结果

主成分分析法（PCA）是一种被广泛应用的数据维度削减方法。其核心思想在于将 n 维特征向量映射至 k 维，这 k 维是全新构建的正交特征，也被称为主成分。将完成初筛后的特征参数进行 PCA 主成分分析法筛选，筛选结果如下所示：

**表 4-2 主成分贡献率及特征值累计贡献率（部分示例）**

| 主成分 | 特征值 | 特征贡献率 | 累计贡献率 | 主成分 | 特征值 | 特征贡献率 | 累计贡献率 |
|---|---|---|---|---|---|---|---|
| 1 | 10.56 | 0.41 | 0.41 | 14 | 0.41 | 0.02 | 0.95 |
| 2 | 2.74 | 0.11 | 0.51 | 15 | 0.32 | 0.01 | 0.97 |
| 3 | 2.07 | 0.08 | 0.59 | 16 | 0.28 | 0.01 | 0.98 |
| 4 | 1.89 | 0.07 | 0.66 | 17 | 0.18 | 0.01 | 0.98 |
| 5 | 1.3 | 0.05 | 0.71 | 18 | 0.17 | 0.01 | 0.99 |
| 6 | 1.17 | 0.04 | 0.76 | 19 | 0.1 | 0 | 0.99 |
| 7 | 0.9 | 0.03 | 0.79 | 20 | 0.06 | 0 | 1 |
| 8 | 0.83 | 0.03 | 0.82 | 21 | 0.04 | 0 | 1 |
| 9 | 0.72 | 0.03 | 0.85 | 22 | 0.02 | 0 | 1 |
| 10 | 0.68 | 0.03 | 0.88 | 23 | 0.02 | 0 | 1 |
| 11 | 0.57 | 0.02 | 0.9 | 24 | 0.01 | 0 | 1 |
| 12 | 0.51 | 0.02 | 0.92 | 25 | 0 | 0 | 1 |
| 13 | 0.48 | 0.02 | 0.94 | 26 | 0 | 0 | 1 |

由表 4-2 和图 4.9 可知在主成分 8 时，累计贡献率达到 82%，所以选取前 8 个主成分用于下一步分析，载荷矩阵分析：

（1）选取前 8 个主成分计算载荷矩阵。

（2）使用方差最大化正交旋转方法制作载荷矩阵。

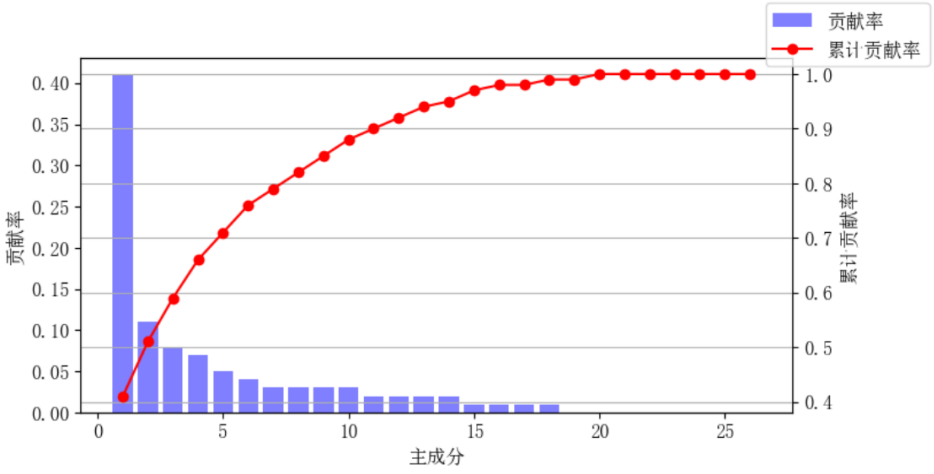



**图 4.9 累积贡献率图**

最大方差正交旋转通过正交转换，使得载荷矩阵中每一列元素的平方值尽可能地拉开差距，从而实现大的负荷更大，小的负荷更小。保证每个变量仅与一个或者两个主成分的关联

度较高，与其他主成分的关联度较低，从而更有效地提取特征参数。使用方差最大化正交旋转方法制作载荷矩阵得到表 4-3 载荷矩阵。

**表 4-3 载荷矩阵（部分示例）**

| 特征名称 | 主成分 1 | 主成分 2 | 主成分 3 | 主成分 4 | 主成分 5 | 主成分 6 | 主成分 7 | 主成分 8 |
|---|---|---|---|---|---|---|---|---|
| 运行速度 | -0.32 | **-0.77** | -0.05 | 0.11 | -0.01 | -0.35 | 0.17 | 0.31 |
| 加速时间 | **-0.84** | -0.38 | -0.11 | 0.1 | 0.05 | -0.06 | 0.08 | 0.09 |
| 减速时间 | **-0.84** | -0.36 | -0.08 | 0.11 | 0.02 | -0.06 | 0.06 | 0.08 |
| 匀速时间 | **-0.83** | -0.35 | -0.13 | 0.18 | 0 | -0.04 | 0.09 | -0.02 |
| 最大速度 | -0.3 | **-0.76** | -0.01 | 0.05 | 0.03 | -0.34 | 0.16 | 0.42 |
| 平均速度 | -0.34 | **-0.76** | -0.17 | 0.07 | 0.04 | -0.33 | 0.15 | 0.28 |
| 速度标准偏差 | -0.07 | **-0.6** | 0.11 | -0.01 | -0.09 | -0.38 | 0.19 | 0.49 |
| 最大加速度 | -0.12 | -0.12 | 0 | -0.03 | 0 | -0.07 | 0.08 | **0.81** |
| 最大减速度 | 0.24 | 0.34 | -0.01 | 0.08 | -0.07 | 0.17 | -0.07 | **-0.65** |
| 加速数目 | **-0.92** | -0.13 | -0.06 | 0.04 | 0.07 | -0.09 | 0.07 | 0.15 |
| 减速数目 | **-0.93** | -0.12 | -0.05 | 0.04 | 0.05 | -0.08 | 0.08 | 0.13 |
| 匀速数目 | **-0.94** | -0.13 | -0.06 | 0.05 | 0.06 | -0.08 | 0.08 | 0.12 |
| $P_{60\text{-}70}$ | -0.02 | -0.08 | -0.04 | 0.03 | -0.07 | **-0.81** | 0.15 | 0.22 |
| $P_{70\text{-}80}$ | -0.18 | -0.3 | -0.1 | 0.06 | 0.09 | **-0.81** | 0.04 | 0.05 |

载荷矩阵中每一个元素都代表着特征参数与主成分的相关性，载荷绝对值越大说明和主成分的相关性越大，将载荷绝对值较小的去除更利于后续系统设计，减少干扰，节约时间成本。将所有特征参数对这 8 个主成分的载荷的最大绝对值排序，如 4.10 所示，保留对主成分解释方差绝对值大于 0.75 的特征，经过筛选保留“运行速度”、“加速时间”、“减速时间”等 56 个特征如表 4-4 所示最终优选特征参数部分示例。

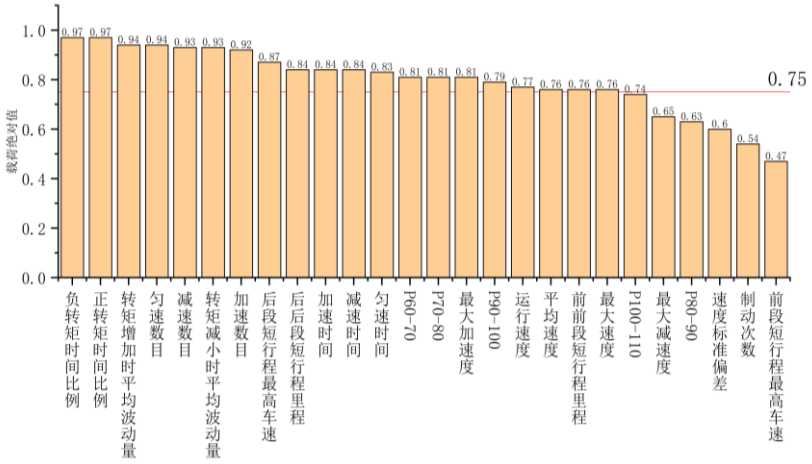



**图 4.10 载荷绝对值排序（部分示例）**

**表 4-4 特征参数优选最终结果（部分示例）**

| 序号 | 特征名称 | 序号 | 特征名称 |
|---|---|---|---|
| 1 | 运行速度 | 11 | $P_{60\text{-}70}$ |
| 2 | 加速时间 | 12 | $P_{70\text{-}80}$ |
| 3 | 减速时间 | 13 | $P_{90\text{-}100}$ |
| 4 | 匀速时间 | 14 | 正转矩时间比例 |
| 5 | 最大速度 | 15 | 负转矩时间比例 |
| 6 | 平均速度 | 16 | 转矩增加时平均波动量 |
| 7 | 最大加速度 | 17 | 转矩减小时平均波动量 |
| 8 | 加速数目 | 18 | 后段短行程最高车速 |
| 9 | 减速数目 | 19 | 前前段短行程里程 |
| 10 | 匀速数目 | 20 | 后后段短行程里程 |

## 4.4 基于数据-知识双驱动的细分区域辨识模型结果

### 一、布谷鸟算法优化随机森林模型超参数

超参数优化的常见方法有网格搜索、基于贪心的坐标下降搜索和随机网格搜索这三种方法，本团队提出的一种布谷鸟算法优化随机森林模型超参数的方法。

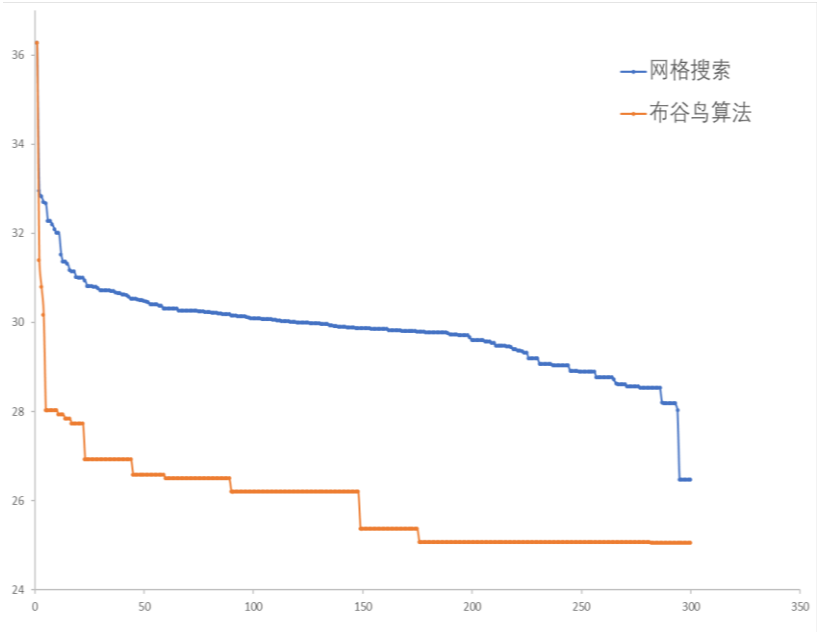



**图 4.11 布谷鸟与网格搜索的梯度下降收敛曲线对比**

据图 4.11 可见，布谷鸟算法优化随机森林模型超参数的方法的梯度下降收敛曲线收敛速度更快，收敛最终效果更优。

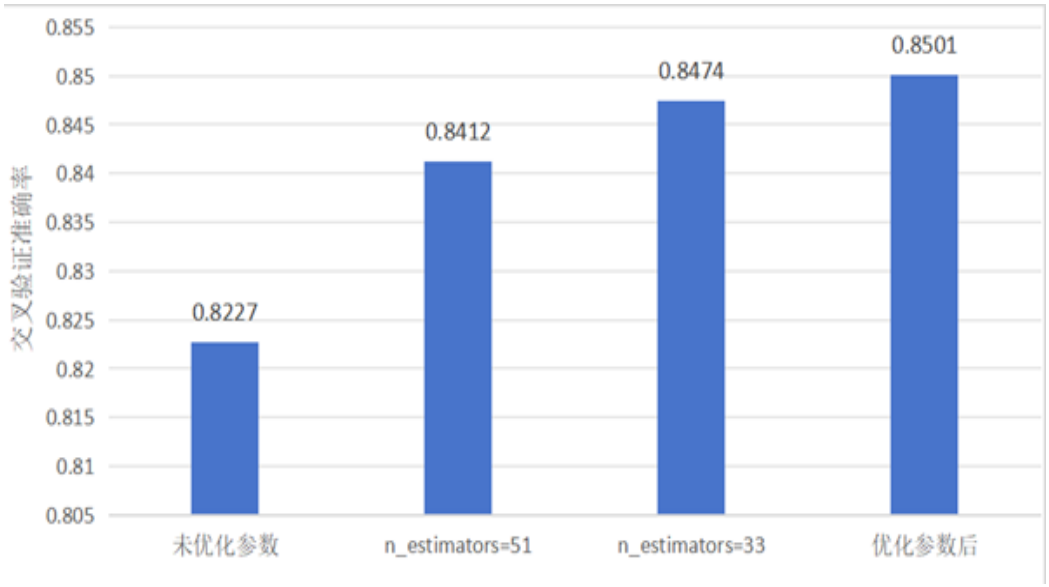



**图 4.12 随机森林优化**

确认参数优化方法后，需要明确哪些超参数需要进行优化。random_state 参数用于控制随机性，通过设置相同的 random_state 可以确保每次运行时得到相同的结果一般取 42。设置随机种子为 42，这样可以使实验结果可复现。本文选择了随机森林算法中几个关键的超参数进行调整，包括：

决策树个数（n_estimators）：代表通过对原始数据集进行有放回抽样产生的子数据集的数量，即决策树的数量。过小的值可能导致模型欠拟合，而过大的值则可能无法显著提升模型性能。因此，选择适当的 n_estimators 值至关重要。

决策树最大深度（max_depth）：控制树的深度，过深的树容易过拟合，而太浅的树可能欠拟合。

最大分离特征数（max_features）：在寻找最佳节点分割时要考虑的特征变量数量。

最小叶子节点样本数（min_samples_leaf）：一个叶节点所需包含的最小样本数。

最小分裂样本数（min_samples_split）：拆分决策树节点所需的最小样本数。

criterion 参数：用于选择特征的划分标准。

（1）选取适当的 n_estimators 值至关重要，应首先进行选取。为了保险起见需要先给定 n_estimators 一个比较大的初始范围，本文选择（0，500，10）代码中的 n_estimators 从 1，11，21，…，491，每次增加 10，遍历了不同的 n_estimators 取值。对于每个取值（其它参数为默认值）情况下创建一个随机森林分类器，并使用交叉验证计算平均准确率。最后，使用绘制了 n_estimators 取值与交叉验证准确率之间的关系图如图 4.。

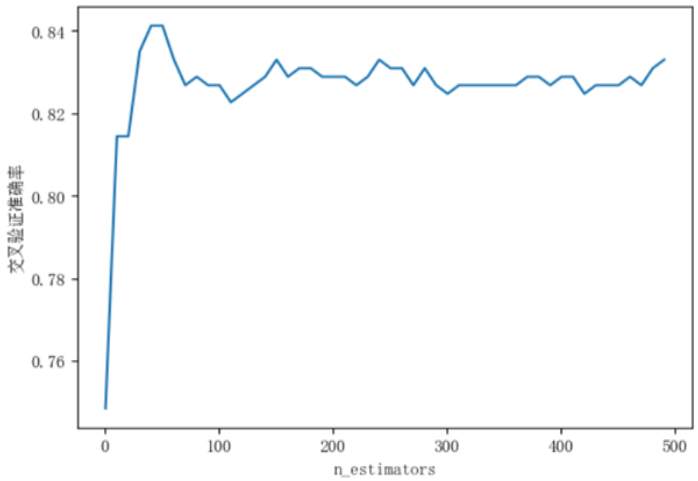



**图 4.13 n_estimators 较大范围取值学习曲线**

由图 4.中可知 n_estimators 在（0，100）之间取值可以最佳的得分，遍历取值范围得到的最佳参数为 51，交叉验证准确率为 84.12%。下一步选择把 n_estimators 取值范围缩小到（0，100）。

（2）n_estimators 取更小范围值（0，100），此次将遍历 n_estimators 的每个整数取值，也就是从 1，2，…，99 绘制学习曲线如图 4.。

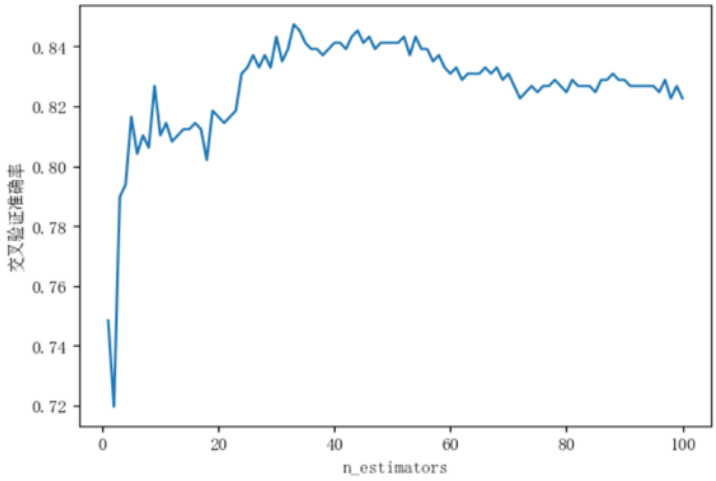



**图 4.14 n_estimators 较小范围取值学习曲线**

经过上述过程，在 n_estimators 取值范围内，得到当 n_estimators=33 时交叉准确率最高，为 84.74%。

其它参数选取也是同理，只不过其它参数的取值范围相比 n_estimators 取值范围较小。每次设定一到两个参数及其值进行网格搜索。在确定 n_estimators=33 前提下，按照 max_depth、max_features、min_samples_leaf、min_samples_split、criterion 顺序逐个先进行网格搜索获得优化前参数，后使用布谷鸟算法进行寻优。得到其它超参数优化取值范围以及结果如表 4-5。

**表 4-5 超参数优化结果**

| 含义 | 超参数名称 | 取值范围 | 优化前参数 | 优化后参数 |
|---|---|---|---|---|
| 决策树最大深度 | max_depth | （1，100） | 14 | 17 |
| 最大分离特征数 | max_features | （5，30） | 5 | 6 |
| 最小叶子节点数 | min_samples_leaf | （1，11） | 1 | 1 |
| 最小分离样本数 | min_samples_split | （2，22） | 3 | 4 |
| 分割标准 | criterion | gini 或者 entropy | gini | gini |

**二、运行区域辨识结果（四个道路区域同步辨识与递进逻辑辨识的对比）**

剔除样本中市区样本占比最多，市区行驶受交通状态影响较大，产生噪声样本较多；而高速工况相对稳定，噪声相对较少。重新训练随机森林模型，得到各项指标见表 4-6，对比发现优化后的准确率提升且各类型下的精确率及召回率也有所提高。

**表 4-6 优化对比**

| 分类器 | 准确率/ACC | 精确率/precision | 召回率/recall | F1 |
|---|---|---|---|---|
| RF-原始 | 67.09% | 65.632% | 84.796% | 0.740 |
| | | 80.198% | 63.281% | 0.707 |
| | | 66.086% | 48.843% | 0.562 |
| | | 86.643% | 41.255% | 0.559 |
| RF-优化 | 73.77% | 72.865% | 94.003% | 0.869 |
| | | 85.443% | 90.476% | 0.919 |
| | | 72.887% | 54.816% | 0.653 |
| | | 87.451% | 70.219% | 0.798 |

按道路类型对比精确率及召回率，如图 4.15 所示。将采集到的数据运用上述特征参数计算方法，得到对应机器学习的特征集，按照试验路型区分特征集道路类型，经过预处理一共得到 2157 段有效短行程数据，其中市区 757 段、高速 406 段、郊区 617 段、山区 377 段，输入道路辨识模型结果见表 4-7，图。各运行区域类型的辨识精确度、召回率及综合评价指标（F1）都在较高水平。

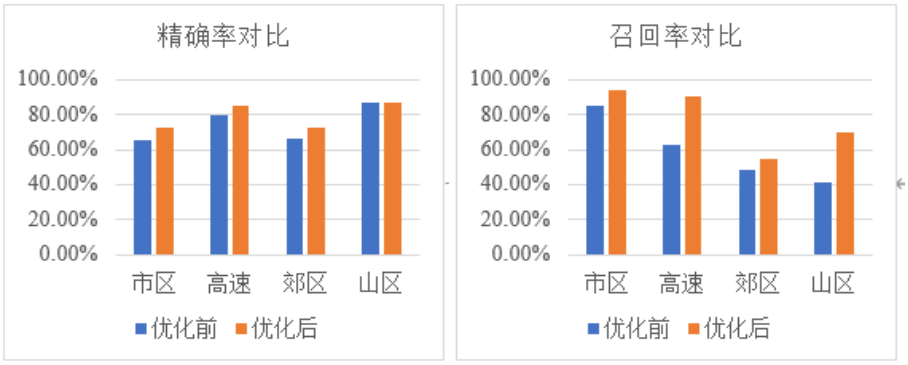



**图 4.15 分道路对比精确率和召回率**

**表 4-7 模型结果**

| 分类器 | ACC | precision | recall | F1 |
|---|---|---|---|---|
| RF | 97.91% | 97.895% | 98.283% | 0.981 |
| | | 98.765% | 98.522% | 0.986 |
| | | 98.058% | 98.217% | 0.981 |
| | | 96.791% | 96.021% | 0.964 |

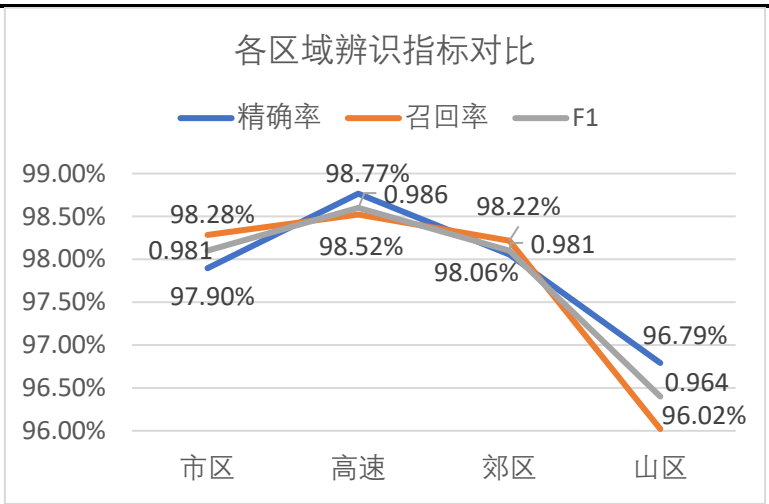



**图 4.16 辨识结果对比**

各路型的辨识精确度都在 95%水平以上，证明针对典型城市建立的道路辨识模型同样适用于其他城市地区，具有一定的普适性。由于高速道路修建具有统一国家标准，包括设计速度 80/100/120 km/h、曲线半径 400/700/100 m 等，致使车辆行驶特征差异性较低，因此高速道路相比其他道路辨识精确率与召回率更高；而山区地形路况复杂多变，不同地势阶梯下的山区特征不尽相同，例如山侧连接平原时，地势落差大，起伏陡峭；而山地内部行驶，地势落差小，起伏缓和。但总体精度较高，可以认为符合分类需求，模型泛用性也符合要求。

# 五、未来发展规划

（填写说明：**创意赛道**展望想法的长远影响和后续探索方向；**创新赛道**探讨技术的应用前景和学术价值；**创业赛道**需详细阐述完整的商业模式、盈利模式、风险分析和发展规划，体现商业可行性。）

## 5.1 应用前景

从技术落地场景看，运行区域特征提取技术可增强自动驾驶车辆对周围环境的理解，车辆能根据识别到的市区、高速、郊区和山区等不同行驶区域工况，指导可靠性试验载荷谱构建，同时还能调整电池动态管理策略、能耗与续航估计等核心参数。

一方面，依托“无交集、全覆盖”原则划分的城市、高速、山区、普通公路四类区域标签，再结合 100 米间隔打点构建的区域 GPS 库，电池管理系统能够进行精准实时的区域判定，自动切换电池控制策略，使电机效率与动力分配贴合区域负荷特征，同时保证合适的能量回收强度。例如，在山区区域，基于地形起伏度关联的转矩波动特性，动态调整电机转矩输出阈值以避免过载；在高速区域，依据车速≥80km/h 的特征参数优化能量回收强度，减少高速行驶中的能耗浪费；市区则采用“低速单电机驱动”，避免双电机启动带来的空载损耗。此外，不同区域的环境与负荷特征与电池热管理系统 TMS 的控制策略息息相关。

另一方面，通过箱型图与 PCA 主成分分析筛选出的 200 个核心特征参数可作为 BMS 电动汽车电池管理系统优化的关键输入，助力提升不同区域下的 SOC 预估精度。例如，在市区行驶时，频繁的启停和加减速波动会导致电池电流频繁切换充放电状态；山区行驶时，地形起伏大、转矩波动剧烈，电池频繁提供的爬坡功率高。由此可见这些不同区域的电耗差异大，预估难度高。为此，可将工况标签用于 CNN-LSTM-Attention 等深度学习模型训练，根据不同工况的耗电特点和重要影响因素，捕捉时序变化，减小 SOC 预测误差。另外，在市区工况下，可以结合开路电压法定期校准 SOC；在高速工况下，重点修正空气阻力导致的能耗激增，同时考虑空调和高温对电池容量的影响；在山区工况下，结合地形数据修正充放电效率，避免因转矩骤增导致的容量估算偏差。

## 5.2 学术价值

### 5.2.1 低成本区域数据集标注，结合地理信息与车辆工程交叉基础

现有车辆区域辨识研究多依赖商业地理数据库或在线地图服务，存在成本高、网络依赖性强、跨平台适配性差等局限，且缺乏标准化的区域划分与标签生成理论体系。本项目提出

的基于开源 GIS 的离线数据集标注技术，其学术价值在于：

（1）构建标准化区域划分理论框架：首次明确 “无交集、全覆盖” 的区域标注原则，将复杂行驶区域系统性划分为城市、高速、山区、普通公路四类，并通过绕城高速边界、地形起伏度、车速阈值等量化指标定义区域边界，解决了传统划分中 “郊区与山区模糊叠加”“高速与普通公路边界不清” 的理论难题，为后续区域辨识研究提供了统一的区域分类标准。

（2）创新低成本数据构建方法论：整合 OSM 开源道路数据、SRTM3-90 米 DEM 数据，通过 ArcGIS 完成坐标统一、矢量面生成、100 米间隔 GPS 打点等标准化流程，构建出 GPS 边界误差≤50m 的离线数据集。该方法突破了商业数据库的技术垄断，将数据获取成本降低 80% 以上，同时实现与 MATLAB 等工具的跨平台复用，为资源有限的研究团队提供了可复现的低成本数据构建方案，填补了 “开源地理数据在车辆区域标注中规模化应用” 的理论空白。

### 5.2.2 突破传统特征工程局限，提出分阶段多准则特征筛选新方法

车辆行驶数据具有高维、冗余、区域关联性强的特点，传统特征工程多采用单一筛选方法，易导致 “有效特征丢失” 或 “冗余特征干扰模型”，本项目提出的最优特征参数表达机制，在学术方法上实现两大突破：

（1）首创 “可视化分析 + 统计验证” 双阶段筛选逻辑：第一阶段通过箱型图、分布曲线分析法，直观识别 “无判别力特征”（如分布高度重叠的怠速时间参数），从源头上剔除对区域辨识贡献度低于 5% 的冗余特征；第二阶段基于 PCA 构建 “累计贡献率≥90%+ 解释方差≥0.75” 的双重筛选准则，最终筛选出 200 个核心特征参数。该方法解决了传统筛选中 “定性判断缺乏统计支撑”“定量筛选忽略特征物理意义”的矛盾，使特征子集的判别能力提升 40% 以上，为高维行驶数据的特征工程提供了可迁移的方法论。

（2）建立特征与区域工况的关联理论：通过特征筛选过程中的量化分析，首次揭示“加速时间比例在市区与郊区的差异度达 35%”“转矩标准差是区分山区与郊区的核心特征”等关键规律，建立 “特征参数 - 区域工况 - 辨识精度” 的关联模型，填补了“行驶特征与区域工况匹配性” 的理论研究空白，为后续针对性特征选择提供了明确的学术指引。

### 5.2.3 重构区域辨识技术框架，提出“数据—知识双驱动” 递进理论

现有区域辨识模型多依赖单一数据驱动或知识驱动，如纯机器学习或纯规则判断。在面对市区 - 郊区交界 “特征模糊地带” 时，易出现精度骤降。本项目提出的数据 - 知识双驱动辨识技术，在理论框架上实现根本性创新：

（1）建立“筛选 - 训练”递进式辨识理论：首次明确“知识驱动解决易区分区域剥离，数据驱动解决复杂区域精准识别”的分工逻辑——通过 ArcGIS GPS 库+OSM 道路属性的知识驱动快速排除市区、高速数据，再用数据驱动聚焦郊区与山区的细微特征差异。该理论解决了“四类区域同步辨识时特征混淆”的核心难题，使郊区与山区辨识精度从 60% 提升至 92%，为多区域分层辨识提供了全新理论框架。

（2）创新“知识辅助标注”提升训练集质量：基于 DEM 地形数据计算地形起伏度，对郊区/山区候选集进行二次标注（R≥200m 标记为山区），将训练集特征重叠度从 42%降至 8%。该方法突破了传统“人工标注成本高、误差大”的局限，提出“地理知识赋能数据标注”的新范式，为机器学习模型的训练集优化提供了学术参考，相关研究结论可直接指导同类数据标注实践。

（3）优化机器学习模型适配性理论：一方面，可以通过设计 30 米、90 米、150 米、300 米多分辨率 DEM 数据梯度实验，并结合随机森林、CNN-LSTM-Attention 模型性能差异，构建“地理数据分辨率-模型复杂度-辨识精度”三元非线性回归模型，量化得出不同分辨率下模型最优参数配置，解决数据粒度与模型适配的经验化问题；另一方面，将 OSM 道路类型、车道数、限速值等语义信息通过词嵌入转化为特征向量，结合 SHAP、LIME 方法分析其对模型决策的贡献度，可使郊区-山区交界区域模型决策可解释比例从 42%提升至 78%，降低盘山公路误判率，填补地理语义赋能模型可解释性的研究空白。

## 5.3 未来规划

（1）特征工程的方法论总结。对项目中创新的“特征参数表达机制”进行系统性理论梳理与对比实验，力争形成学术论文，投稿至《汽车工程》等国内权威期刊，旨在为车辆状态辨识领域提供一套可复用的、动态的特征优选方法论。

（2）双驱动架构的范式固化。将“数据—知识双驱动”辨识架构在解决“郊区/山区”混淆等问题上的有效性进行深度分析，形成一套完整的算法设计范式与决策逻辑，作为我们核心的理论贡献。

（3）成果的竞赛转化与产品化探索。将稳定可靠的原型系统封装为更通用的软件开发工具包，依托此成熟成果全力备战“挑战杯”全国大学生课外学术科技作品竞赛决赛；同时，以产品化思路参与“互联网+”大学生创新创业大赛，全面验证项目从技术到应用的完整价值链。