[文章编号] 1003－4684(2019)05-0067-05

基于机器学习的新能源汽车残值评估方法

张子蓬，郝世林

（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 针对现有汽车残值评估方法信息使用较少、严重依赖人工检测、误差大和无法解决新能源汽车残值评估的问题，基于梯度提升回归树模型，使用行驶里程、使用时间、功率、过户次数等多维特征数据，训练模型对残值进行预测。从选取惩罚项、树的深度、基学习器类型以及提取特征重要性方面优化模型。最后，使用 Stacking 模型集成算法对二阶多项式、XGBoost、LightGBM 模型进行集成。实验结果表明，使用 Stacking 集成后的模型可以根据当前车况数据自动计算残值，不需要人工检测，具有实时性，较其他方法有更高的准确度。

[关键词] 新能源汽车；机器学习；大数据；残值评估

[中图分类号] TP391.4 [文献标识码] A

新能源汽车产业作为我国七大战略性新兴产业之一，近年来产业规模发展迅猛。与之相对的是，如今消费者在购买新能源汽车时，很大一部分顾虑来自于汽车残值的不确定性。保险行业对纯电动汽车商业保险也不完善，这些问题在一定程度上阻碍了纯电动汽车的普及。因此，亟需建立一套新能源汽车残余价值评估方法。

目前常用的方法主要有重置成本法、现行市价法、清算价格法[1]。这些方法存在建模简单、车辆信息使用较少、严重依赖人工检测（成本高、耗时长）、小数据集下无法建模[2-3]和无法解决新能源电动车残值评估的问题。

1 新能源汽车残值评估方法

1.1 评估方法框架

新能源汽车的残值和车型、行驶里程、使用时间等车况信息密切相关。本方法使用机器学习方法学习车况数据和车辆残值之间的关系，实现对残值的评估。

通过网络爬虫技术采集新能源汽车数据，包括车况信息和残值。对数据进行预处理、特征工程，从车况信息中提取出影响车辆残值的关键因素。使用这些关键因素和对应残值数据训练模型。最后使用测试数据验证模型泛化性能，保存泛化性能最佳的模型。整个评估方法框架见图 1。

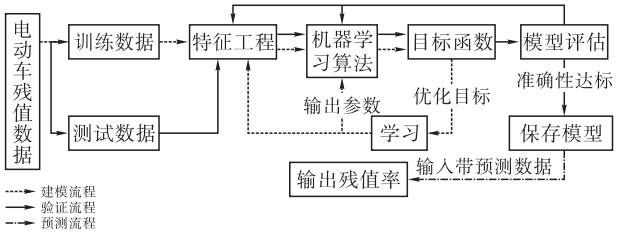


图 1 残值评估方法框架

1.2 采集新能源汽车数据

使用 scrapy 爬虫框架，编写 Python 脚本从 https://www.autoscout24.com/网站获取数据。数据格式见表 1。

表 1 数据格式

A	B	C	D	E	F	G	H
75000	43	65 kW	8490	White	4	5	1
55000	70	43 kW	8900	Black	5	5	1
47000	48	65 kW	8990	White	4	5	1

列名 A,B,C,D,E,F,G,H 分别代表：kilometers（行驶里程）、service_life（使用时长，单位月）、power（功率）、residual（残值单位：欧元）、body_color（颜色）、Nr_of_Doors（车门数）、Nr_of_Seats（车座数）、transfers（过户次数）

1.3 提升树模型的学习过程

本方法主要使用回归树模型建模。回归树模型可以解决高维非线性特征问题，能学习到数据中的非线性关系，能够很好的拟合数据，并且模型容易理解和阐述。

提升树模型[4-5]原理：采用加法模型和前向分布

[收稿日期] 2019－08－09
[基金项目] 国家自然科学基金青年基金项目(61603127)
[第一作者] 张子蓬(1968－)，男，山东武城人，理学博士，湖北工业大学副教授，研究方向为嵌入式系统，人工智能

算法构建多棵决策树。基于 Boosting 的思想，先学习前 $n-1$ 个学习器，然后基于前 $n-1$ 个学习器学习第 n 个学习器。最终这些决策树组合进行预测。

利用分割函数对数据集随机切分为训练集和测试集。其中测试集数据占总数据量的 20%。使用训练数据集训练模型，测试集验证模型的泛化性能。以 XGBoost[6-7] 模型为例，算法学习过程如下：训练数据集

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

其中 x_1 表示特征数据，y 表示残值率。

$\hat{y}_i=\sum_{k=1}^{K} f_k(x_i)$，$f_k\in F$ 。F 表示的是回归森林中的所有函数空间。$f_k(x_i)$ 表示的就是第 i 个样本在第 k 棵树中落在的叶子的权重。需要求的参数就是每棵树的结构和每片叶子的权重，可以设

$$\Phi=\{f_1,f_2,f_3,\cdots,f_k\}$$

损失函数取平方误差，目标函数表示为：

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n}\{2(\hat{y}_i^{(t-1)}-y_i)+f_t(x_i)^2\}+\Omega(f_t)+\mathrm{const}$$

$$\Omega(f_t)=\frac{1}{2}\lambda\sum_{j}^{T}\|w_j\|^2+\gamma T$$

提升树算法：

1)每一轮添加一棵树；

2)每一轮开始的时候，计算一阶导数和二阶导数

$$g_i=\partial_{\hat{y}t-1}\,l(y_i,\hat{y}^{t-1})\,,\ h_i=\partial^2_{\hat{y}t-1}\,l(y_i,\hat{y}^{t-1})$$

记 $G_j=\sum_{i\in I,j} g_i, H_j=\sum_{i\in I,j} h_i$

其中 I_j 表示第 j 个叶子里的样本集合。

3)统计所有分裂点信息。用贪婪的方式生长一棵树

$$\mathrm{Obj}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T$$

4)添加 $f_t(x)$ 到模型

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+\varepsilon f_t(x_i)$$

ε 是步伐大小，有助于防止过拟合。

基学习器，即每棵树的构建流程：

1)从深度为 0 的树开始；

2)对树的每个叶子节点，枚举所有的特征。对于每个特征，根据特征值对实例(样本)进行排序。对每个分裂点计算

$$\mathrm{Gain}=\frac{G_L^2}{H_L+\lambda}+\frac{G_L^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}-\gamma$$

确定增益最大特征和对应的分裂点。

3)添加 2)步骤找到的最优分裂点。添加这个分裂点后，计算损失函数值的变化。$\mathrm{Obj}_{\mathrm{split}}$ 为分裂后的损失函数值，$\mathrm{Obj}_{\mathrm{unsplit}}$ 为分裂之前损失函数值。如果 Gain1 大于设定值，选择分裂。否则，舍弃该分裂点

$$\mathrm{Obj}_{\mathrm{split}}=-\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}\right]+\gamma T_{\mathrm{split}}$$

$$\mathrm{Obj}_{\mathrm{unsplit}}=-\frac{1}{2}\left[\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]+\gamma T_{\mathrm{unsplit}}$$

$$\mathrm{Gain1}=\mathrm{Obj}_{\mathrm{unsplit}}-\mathrm{Obj}_{\mathrm{split}}=$$

$$\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma(T_{\mathrm{split}}-T_{\mathrm{unsplit}})\ ;$$

4)重复上述步骤 2)，3)。直到满足设定结束条件，停止分裂。回归树构建完成。

2 实验内容

2.1 数据预处理

缺失数据的特征有：body_color、Nr_of_Doors、Nr_of_Seats 和 transfers。

body_color、Nr_of_Doors、Nr_of_Seats 对应缺失值比例分别为：0.6%、0.6%、1.2%。由于这三个特征缺失值占比很小，取每个特征对应的众数进行填充。

过户次数缺失值比例约为 25.3%。采用逻辑回归(logistics regression)构建三分类模型(过户次数 T={0,1,2})。将数据分为数据集 A(过户次数不含缺失值)、B(过户次数含有缺失值)。使用 A 中的行驶里程和使用时间作为输入特征，过户次数为目标变量，训练三分类逻辑回归模型。对 B 的过户次数进行预测。最终验证集准确率为 98.1%。

对连续变量进行归一化和标准化。主要针对二阶多项式模型，树模型对此不要求。离散变量采用 onehot 编码。

归一化：

$$X_{\mathrm{norm}}=\frac{X-X_{\mathrm{minV}}}{X_{max}-X_{min}}$$

标准化

$$X'=\frac{X-\mu}{\sigma}$$

one-hot 编码：

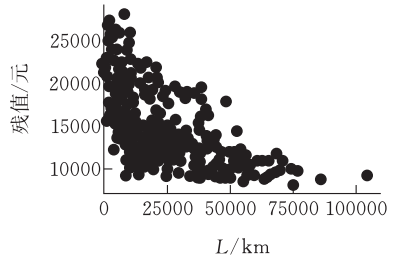
$$Z_j=\begin{cases}1 & \text{if } x \text{ is in category } j\\ 0\end{cases}$$

2.2 探索数据分布

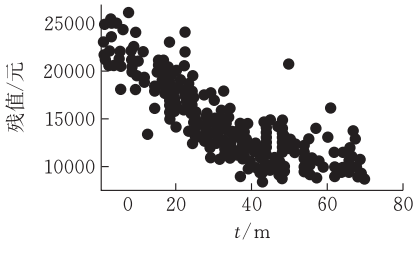
图 2 和图 4 以部分特征进行展示，其他特征均做同样处理。

图 2 显示行驶里程、使用时间和残值成强负相关。随着使用时间和里程的增加，新能源汽车残值逐渐减少。

图 3a 为原始图，以高斯分布为基准，残值分布有轻微的右拖尾。图 3b 是采用 Log 变化后的分布，变换后有效减轻了右摆尾。

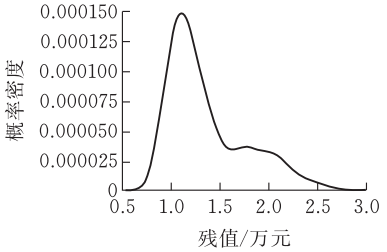


(a)行驶里程与残值关系

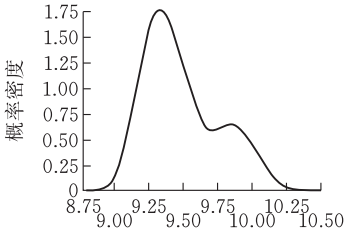


(b)使用时间与残值关系

图 2　行驶里程、使用时间和残值散点图



(a)原始残值概率密度分析

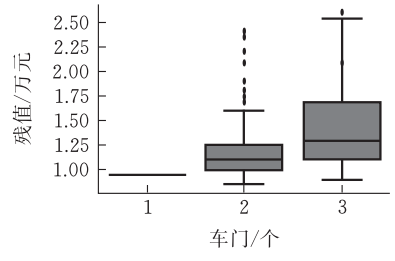


(b)残值对数变换后概率密度

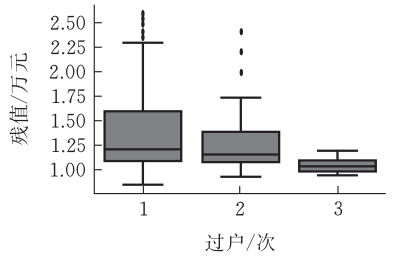
图 3　残值概率密度分布

图 4 显示，车门数为 4 和 5 的样本中残值有异常样本，存在残值偏高的情况。过户次数为 1 次和 2 次的样本中的残值存在偏高的情况。从数据集中删除这些异常点。

由图 5 可知，负相关最强的是行驶里程和使用时间。车身颜色、车座数跟残值的相关性较低，故删除这两个特征。



(a)车门数与残值的箱线图



(b)过户次数和残值的箱线图

图 4　车门数、过户次数和残值的箱线图

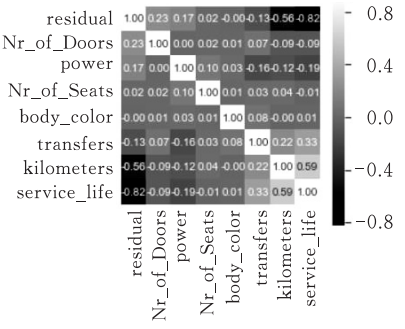


图 5　特征之间皮尔逊相关系数图

2.3　特征构造和特征选择

1)构造比率特征；根据数据实际含义构造特征。每月平均行驶里程 = 行驶里程/使用时间；行驶里程和使用时间的二阶特征。

2)添加统计特征；分别对每个分类特征分组求得连续变量每个分组的均值、方差、最大、最小值。对每个连续变量分组求离散变量的 nunique。

3)对类别特征使用 w2v 构建特征；将每一个样本的类别特征构造成一个句子，然后使用 word2vec 提取特征之间的相关关系。

以 LightGBM 为基分类器，使用 3 折递归特征消除法(RFECV)，筛选强相关特征。

2.4　构建模型和模型优化

步骤 2.3 筛选出的特征作为模型输入。使用 5 折交叉验证划分训练集训练模型。

使用二阶多项式、XGBoost、LightGBM 算法分别训练出 5 个模型。取每个算法对应的 5 个模型均值，作为预测结果。

对模型进行优化，通过模型评价指标对比得到。损失函数采用 Root Mean Square Error。二阶多项式最优模型基学习器为岭回归，其 L_2 范数系数为 0.343；XGBoost 最优模型使用 300 棵基学习器，树的深度为 5，L_1 正则项系数为 0.1；LightGBM 最优模型使用 1000 棵 gbdt 树，树的深度为 5，叶子数量为 31，min_data_in_leaf 为 10，min_child_samples 为 10，L_1 正则系数为 0.15。

XGBoost 和 LightGBM 模型训练结束后，可以打印出特征重要性列表。根据特征重要性可以对模

型训练过程做出优化:过滤重要性较低的特征,将剩余的特征放入模型训练,有效提升模型训练速度和提升拟合优度。

2.5 模型融合

通过比较模型之间的差异性,来获得最优模型组合。采用 Stacking[8] 算法将三个模型融合。二阶多项式、LightGBM 和 XGBoost 作为个体学习器,Catboost 作为次级学习器。Stacking 结合后的模型更稳定,过拟合风险更低。

3 实验结果

以下结果在 Python3.6 环境下,使用 jupyter noteboook 编写脚本完成。实验环境:Ubutun 16.04。使用主要库有:pandas、numpy、seaborn、sklearn、lightgbm、xgboost。使用机器学习方法和传统方法对比预测结果精度。

3.1 特征筛选结果

最终筛选出来的特征依次为:行驶里程、平均每月行驶里程、使用时间、不同功率下行驶里程的均值、车门数、不同车身颜色对应的 km 数的中位数、过户次数。

3.2 算法结果对比

结果使用 R^2 决定系数展示,即

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - y_{true})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

取值范围 0~1,越接近 1 说明预测值和真实值之间差距越小,拟合程度越高。

绘制模型学习曲线判断是否过拟合。以 LightGBM 的训练过程为例。记训练集的 R^2 为 train_{R^2},验证集 R^2 为 valid_{R^2},测试集 R^2 为 test_{R^2}。

由图 6 可知,在样本小于 50 时,train_{R^2} 和 valid_{R^2} 很接近,此时模型欠拟合。随着训练样本的增多,train_{R^2} 和 valid_{R^2} 都在逐渐增加,最终 train_{R^2} 收敛于 0.92。valid_{R^2} 收敛于 0.868。

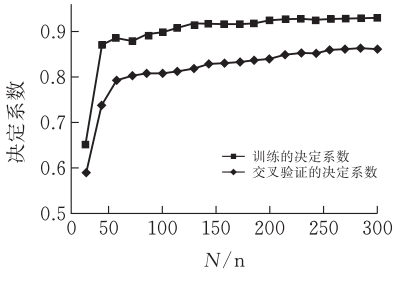


图 6 LightGBM 学习曲线图

传统方法以重置成本法和清算价格法为例,机器学习方法采用二阶多项式、LightGBM、XGBoost 三种算法。表 2 显示算法在验证集和测试集的预测精度。从测试集结果看,机器学习方法平均预测精度较传统方法平均提升 0.26。最后将二阶多项式、LightGBM、XGBoost 进行 Stacking 集成。结果较三个单独模型的平均值提升 0.029。

表 2 算法预测结果

	valid_{R^2}	test_{R^2}
重置成本法	0.656	0.626
清算价格法	0.591	0.556
二阶多项式	0.841	0.832
XGBoost	0.882	0.863
LightGBM	0.868	0.859
Stacking	0.896	0.881

4 总结

针对传统汽车残值评估方法存在的问题,介绍了基于机器学习的新能源汽车残值评估方法。本方法从特征选择、防止模型过拟合等方面优化模型,最终得到由 Stacking 算法对二阶多项式、LightGBM、XGBoost 集成模型。实验表明,该集成模型能够充分挖掘数据蕴含信息,有很好泛化性能,获得更高的准确度;具有很好的适用性和扩展性,并且具有很好的可解释性。

[参 考 文 献]

[1] 张弦.基于主题模型的车辆残值评估研究:[D].南京:南京大学 ,2018.

[2] 中永(苏州)信息技术有限公司. 一种二手车检测评价分析及残值评估系统:中国,109883725 [P],2019.06.14.

[3] 广东数鼎科技有限公司. 基于大数据的汽车残值预测模型及预测方法:中国,108154275[P],2017.12.29.

[4] Hu J, Min J. Automated detection of driver fatigue based on EEG signals using gradient boosting decision tree model[J]. Cognitive Neurodynamics, 2018, 12(12):1-10.

[5] Friedman J H. Greedy function approximation: A gradient boosting machine.[J]. Annals of Statistics, 2000, 29(5):1189-1232.

[6] Son J. Tracking-by-segmentation with online gradient boosting decision tree[C]// IEEE International Conference on Computer Vision, 2016.

[7] Chen T, Guestrin C. XGBoost: a scalable tree boosting system[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2016.

[8] Zhou Z H. Ensemble methods - foundations and algorithms[M]. Taylor & Francis,[s.n.],2012.

Research on Residual Value Evaluation Method of New Energy Vehicle Based on Machine Learning

ZHANG Zipeng, HAO Shilin

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: Aiming at the problem of less use of information, heavy reliance on manual detection, large error and inability to solve the problem of residual value evaluation of new energy vehicles in the existing vehicle residual value assessment method, a method for estimating the residual value of new energy vehicles based on machine learning is proposed. The method, based on the gradient boosting regression tree model, uses the multi-dimensional feature data such as mileage, time, power, and number of transfers, and the training model to predict the residual value. The model is optimized from the selection of penalty terms, the depth of the tree, the type of base learner, and the extraction of feature importance information. Finally, the second-order polynomial, XGBoost, and LightGBM models are integrated using a stacked model integration algorithm. The experimental results show that the model with stack integration can automatically calculate the residual value according to the current vehicle condition data, without manual detection, and has real-time performance, which has higher accuracy than other methods.

Keywords: new energy vehicles; machine learning; big data; residual value assessment

[责任编校: 张岩芳]

(上接第 66 页)

Text Feature Selection Based on Multi-strategy Improved Bat Algorithm

HOU Qiao,CHEN Hongwei

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: Feature selection is an important processing step of the text classification process. It is difficult to greatly improve the accuracy of text classification by traditional feature selection methods, when other classification processing and algorithms are set. Therefore, a new text feature selection method based on improved bat optimization is introduced. It uses traditional feature selection method to pre-select the original features, based on which Gaussian local perturbation and adaptive adjustment weights are used to improve the traditional bat group algorithm. The preference and classification accuracy of pre-selected features is used as the fitness of the individual in binary coding. The multi-strategy improved bat algorithm text feature selection algorithm MS-BA is proposed to realize the efficient solution of text feature selection optimization model. The results show that the classification accuracy of MS-BA is improved.

Keywords: feature selection; bat algorithm; text classification; multi-strategy improvement

[责任编校: 张岩芳]