本文引用格式：薛云飞.基于机器学习的汽车二氧化碳排放量预测研究[J].自动化与信息工程,2023,44(1):22-26;45.
XUE Yunfei. Research on prediction of automobile carbon dioxide emissions based on machine learning[J]. Automation & Information Engineering, 2023,44(1):22-26;45.

基于机器学习的汽车二氧化碳排放量预测研究

薛云飞
（重庆交通大学机电与车辆工程学院，重庆 400074）

摘要：针对汽车尾气排放物中二氧化碳（CO_2）的排放量测量设备价格昂贵且测量精度低的问题，进行基于机器学习的汽车二氧化碳排放量预测研究。首先，利用斯皮尔曼等级相关系数分析汽车特征之间的相关性，并过滤冗余特征；然后，利用随机森林算法筛选出影响 CO_2 排放量的 4 个核心特征；最后，分别基于线性回归、梯度提升树、XGBoost、支持向量机 4 种机器学习算法建立 CO_2 排放量的预测模型，并通过模型效果对比和网格搜索调参，确定最佳的预测模型为基于梯度提升树算法构建的模型。预测值和真实值的对比结果表明，基于梯度提升树算法构建的模型具有较高的预测精度，能有效预测不同汽车每公里的 CO_2 排放量。

关键词：机器学习；CO_2 排放量；斯皮尔曼等级相关系数；随机森林算法；预测模型

中图分类号：TP181　　**文献标志码：**A　　**文章编号：**1674-2605(2023)01-0004-06

DOI：10.3969/j.issn.1674-2605.2023.01.004

Research on Prediction of Automobile Carbon Dioxide Emissions Based on Machine Learning

XUE Yunfei
(School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China)

Abstract: Aiming at the problem of the high price and low measurement accuracy of the emission measurement equipment of carbon dioxide (CO_2) in automobile exhaust emissions, the research on the prediction of automobile carbon dioxide emissions based on machine learning is carried out. Firstly, the correlation between automobile features is analyzed by using Spearman rank correlation coefficient, and redundant features are filtered; Then, the random forest algorithm is used to screen out four core characteristics that affect the emission of CO_2; Finally, the prediction model of CO_2 emissions is established based on four machine learning algorithms, namely linear regression, gradient lifting tree, XGBoost and support vector machine, and the best prediction model is determined based on gradient lifting tree algorithm through model effect comparison and grid search parameter adjustment. The comparison between the predicted value and the real value shows that the model based on gradient lifting tree algorithm has high prediction accuracy and can effectively predict the CO_2 emissions per kilometer of different automobile.

Keywords: machine learning; CO_2 emissions; Spearman rank correlation coefficient; random forest algorithm; prediction model

0 引言

随着科技和经济的飞速发展，我国汽车保有量迅速增长，汽车尾气已成为我国大气污染物的主要来源之一[1]。汽车发动机工作时，燃油中的碳与氧结合生成的 CO_2 约占汽车尾气总排放量的 20%[2]。CO_2 会引发温室效应，影响全球气候变化，因此对汽车尾气中的 CO_2 排放量进行测量是非常必要的。通过测量得到规定条件下汽车的 CO_2 排放量，不仅可以确定汽车是否符合环保检测尾气标准，还可以为环境污染管理提供碳排放数据。

目前，测量汽车 CO_2 排放量的方法大都根据光学原理，利用 CO 和 CO_2 等气体对不同频率的红外光有

不同吸收率的特点进行测量。汽车尾气的测量设备主要有化学发光分析仪、可移动的四极质谱仪、新型非分光红外线（non-dispersive infrared, NDIR）设备和改进的氢火焰离子化检测器（flame ionization detector, FID）等。王刚等[3]针对轻型汽车设计一款便携式车载排放测试设备，依据非分光红外法原理测量汽车的CO_2排放量，稳态工况下的测量误差为2.54%。苏茂辉[4]利用NDIR分析仪来测量汽车尾气排放物中CO及CO_2的浓度，测量误差稳定在2.5%之内。隋修武等[5]采用一体化结构设计一套汽车排放瞬态工况法测量用气体流量分析仪，用于测量汽车尾气排放物中CO_2的浓度值及排放量，测量误差仅为0.93%。以上测量设备价格昂贵，动态响应差，只能满足CO_2浓度变化微小的工况。随着人工智能技术的快速发展，有些学者将其应用于汽车尾气排放量的测量，如李小颖等[6]基于神经网络建立汽车尾气排放物中CO的软测量模型，该模型可在没有汽车尾气排放物专用测量仪器时进行CO排放量的测量。受此启发，本文基于机器学习与数据挖掘技术，利用汽车行驶的信息数据来预测CO_2排放量。

1 数据描述及预处理

本文的研究数据来源于开放数据平台Kesci上的2022年加拿大汽车燃油消耗等级数据。该数据集有15个字段，共946条记录，每条记录包含唯一的汽车特征，数据集中的汽车特征信息如表1所示。

表1 汽车特征信息

变量名称	特征类型	特征含义
ModelYear	无用特征	车型年份
Make	离散	汽车品牌
Model	离散	车型类型
VehicleClass	离散	汽车类别
EngineSize(L)	连续	发动机容积
Cylinders	离散	气缸数
Transmission	离散	变速器
FuelType	离散	燃料类型
FuelConsumption(City(L/100 km))	连续	城市燃料消耗等级
FuelConsumption(Hwy(L/100 km))	连续	公路油耗等级
FuelConsumption(Comb(L/100 km))	连续	油耗综合等级
FuelConsumption(Comb(mpg))	连续	燃料消耗综合评级
CO_2Rating	离散	二氧化碳排放等级
SmogRating	离散	烟雾污染物排放等级
CO_2Emissions(g/km)	标签	二氧化碳排放量

在Python3.8环境中进行CO_2排放量预测的分析和建模，编辑器采用Spyder。将2022年加拿大汽车燃油消耗等级数据导入Python后，先删除无用特征ModelYear；再采用独热编码方式对5列字符型的离散型特征进行编码处理，以方便后续输入模型的分析。

2 特征选择

2.1 斯皮尔曼相关性分析

斯皮尔曼相关性分析作为一种常用的描述性分析方法，可检查特征间的相关性。当特征间的相关性过大时，可能引起模型不稳定，导致模型的鲁棒性较差[7]。2个特征的相关性可用相关系数的绝对值来表征。斯皮尔曼根据特征数据的位置顺序计算2个特征的相关

系数，不受数据本身影响，计算流程为：

1） 对2个特征X、Y排序；

2） 排序后的位置信息称为秩，秩的差记为d_i，d值的个数记为n；

3） 将d_i和n代入公式(1)，计算相关系数ρ_s：

$$\rho_s = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)} \tag{1}$$

因为特征中异常值的秩只出现在数据的头尾，所以斯皮尔曼相关系数降低了异常值对相关性的影响。2个特征之间的相关性等级如表2所示。

表2 2个特征之间的相关性等级

相关系数的绝对值	0.0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0
相关等级	极弱相关或无相关	弱相关	中等程度相关	强相关	极强相关

由表 2 可知：当 2 个特征的相关系数的绝对值在 0.8~1.0 之间时，说明 2 个特征呈极强相关；当 2 个特征的相关系数的绝对值大于 0.95 时，说明 2 个特征极度相似，近似呈线性关系。本文设定相关性阈值为 0.95，即 2 个特征的相关系数的绝对值大于 0.95 时，只保留其中 1 个。

利用斯皮尔曼相关性分析计算汽车特征之间的相关系数，并以热力图的形式将特征之间的相关系数可视化，如图 1 所示。

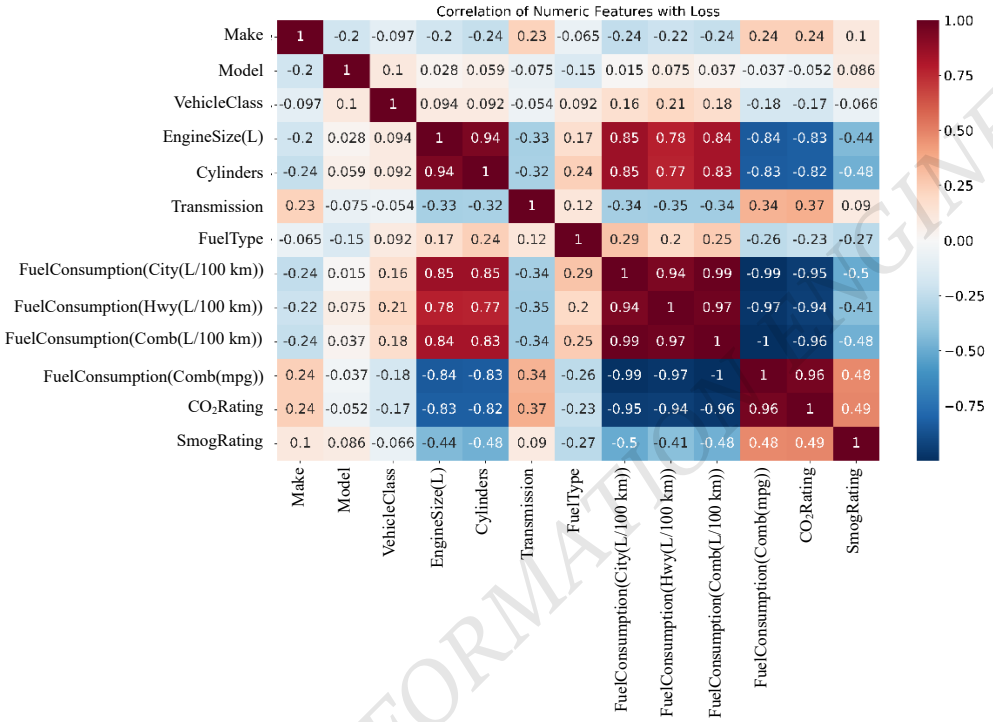

图 1　汽车特征之间的相关性

由图 1 可知，汽车不同特征之间共有 8 个相关系数的绝对值大于阈值 0.95。可删除 FuelConsumption (Comb(L/100 km))、CO_2Rating、FuelConsumption (Comb(mpg))这三列冗余特征。

2.2　基于随机森林算法的特征重要性评分

将删除冗余特征后的数据按 7∶3 的比例随机划分为训练集和测试集。其中，训练集数据有 662 个样本，测试集数据有 284 个样本。基于随机森林算法对斯皮尔曼相关性分析后的汽车特征进行重要性评分，只保留重要性评分较高的汽车特征来挖掘影响 CO_2 排放量的核心特征。对于回归问题，随机森林内部节点的特征按方差减少的标准来选择[8]。假设共有 n 个特征 X_1、X_2、$X_3 \ldots X_n$，它们的重要性评分用 VIM 表示，方差 Var 的计算公式为

$$Var_n = \sum_{i=1}^{S_n} (y_{ni} - x_n)^2 \tag{2}$$

式中：S_n 为节点 n 中训练集样本的个数，y_{ni} 为各个样本的值，x_n 为节点 n 中训练集样本的输出均值。

特征 X_j 在节点 n 的重要性，即节点 n 分枝前后的方差变化量为

$$VIM_{jn}^{(Var)} = Var_n - Var_l - Var_r \tag{3}$$

式中：Var_l、Var_r 分别为分枝后 2 个新节点的方差。

如果特征 X_j 在决策树 i 中出现的节点在集合 N 中，则 X_j 在第 i 棵树的重要性为

$$VIM_{ij}^{(Var)} = \sum\nolimits_{n \in N} VIM_{jn}^{Var} \tag{4}$$

假设随机森林有 r 棵树，则

$$VIM_j^{(Var)} = \sum_{i=1}^{r} VIM_{ij}^{Var} \tag{5}$$

将汽车特征的重要性评分由高到低排序，如图 2 所示。

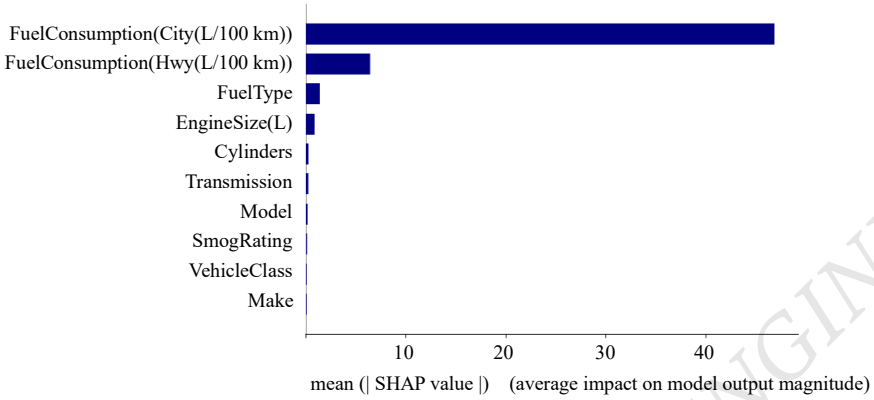

图 2 汽车特征重要性评分

由图 2 可知，FuelConsumption(City(L/100 km))特征与 CO_2 排放量的相关性最大；在 EngineSize(L)之后，汽车特征的重要性变得微乎其微。本文选择重要性评分较高的 4 个特征 FuelConsumption(City(L/100 km))、FuelConsumption(Hwy(L/100km))、FuelType、EngineSize(L)，即对 CO_2 排放量影响较大的特征进行建模。

3 模型构建

本文基于线性回归、梯度提升树、XGBoost、支持向量机 4 种机器学习算法分别建立汽车 CO_2 排放量的预测模型。

线性回归是利用线性回归方程的最小平方函数对一个或多个自变量和因变量之间的关系进行建模的一种回归分析[9]。

梯度提升树以决策树为基学习器，对于回归问题决策树是二叉回归树，其模型可表示为决策树的加法模型[10]，通过负梯度拟合的方式进行迭代，逐渐减小与样本真实值之间的残差。

XGBoost 作为梯度提升树的高效实现[11]，主要从算法本身、算法运行效率、算法健壮性 3 个方面做了优化，对每个弱学习器的建立过程做并行选择，找出合适的子树分裂特征和特征值。

支持向量机处理回归问题时，拟合训练的数学模型可表达为多维空间的某一曲管。如预测值与真实值的差值小于阈值，将不对此样本点作惩罚；若超出阈值，则计算惩罚量[12]。

在 Python3.8 环境中导入各个机器学习算法的模块，利用训练集的 662 个样本训练各模型，各模型的超参数为默认值；再将测试集 284 个样本的特征数据导入训练好的模型进行预测。

通过对比平均绝对误差（mean absolute error, MAE）、均方根误差（root mean square error, RMSE）、平均百分比误差（mean absolute percentage error, MAPE）、拟合优度（R-squared, R^2）4 个回归性能评估指标，分析模型在测试集上的效果。4 个回归性能评估指标的计算公式分别为

$$M_{\mathrm{AE}} = \frac{1}{n}\sum_{i=1}^{n}\left|y_{\mathrm{T}i} - y_{\mathrm{P}i}\right| \tag{6}$$

$$R_{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|y_{\mathrm{T}i} - y_{\mathrm{P}i}\right|^2} \tag{7}$$

$$M_{\mathrm{APE}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_{\mathrm{T}i} - y_{\mathrm{P}i}\right|}{\left|y_{\mathrm{T}i}\right|} \tag{8}$$

$$R^2=\frac{\sum_{i=1}^{n}\left(y_{Pi}-y_{mean}\right)^2}{\sum_{i=1}^{n}\left(y_{Ti}-y_{mean}\right)^2} \tag{9}$$

式中：n 为测试集样本的数量，y_{Ti} 为测试集样本的真实值，y_{Pi} 为对应样本的预测值，y_{mean} 为测试集样本真实值的均值。

4 个模型在测试集上的回归性能评估指标如表 3 所示。

表 3　4 个模型在测试集上的回归性能评估指标

	M_{AE}	R_{MSE}	M_{APE}	R^2
支持向量机	8.46	17.13	3.96%	0.68
线性回归	6.55	11.39	2.71%	0.88
XGBoost	2.36	6.39	0.87%	0.94
梯度提升树	1.80	4.93	0.71%	0.96

由表 3 可知，基于梯度提升树算法构建的 CO_2 排放量预测模型的 4 个回归性能评估指标均明显优于其他模型。

对基于梯度提升树算法构建的模型进行网格搜索调参。因为树的棵数 *n_estimators* 和最大深度 *max_depth* 超参数对模型效果的影响较大，所以主要对这 2 个超参数进行调节。调参时，设置 *n_estimators* 的范围为 10~600，步长为 10；*max_depth* 的范围为 1~16，步长为 1。以 R_{MSE} 作为调参目标，网格搜索不同参数组合时，该模型在测试集上的 R_{MSE} 如图 3 所示。

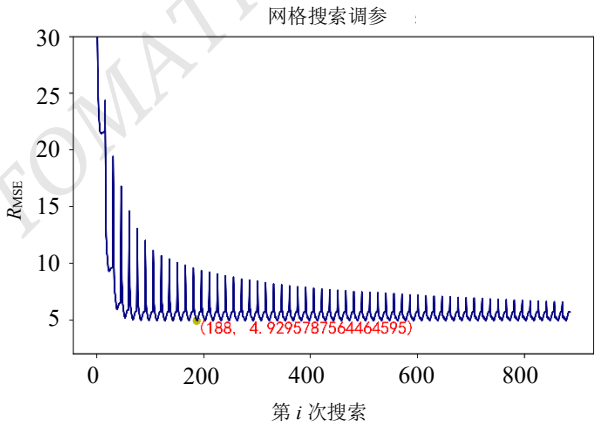

图 3　网格搜索调参结果

由图 3 可以看出，网格搜索在第 188 次超参数组合时，R_{MSE} 最小，此时对应的 *n_estimators* 为 220，*max_depth* 为 8。基于梯度提升树算法构建的模型调参前后的预测结果对比如表 4 所示。

表 4　基于梯度提升树算法构建的模型调参前后的预测结果对比

	M_{AE}	R_{MSE}	M_{APE}	R^2
调参前	2.16	5.24	0.86%	0.96
调参后	1.80	4.93	0.71%	0.96

由表 4 可知，模型调参后，测试集上的 M_{AE}，R_{MSE} 和 M_{APE} 均有一定程度的减小；可认为当 *n_estimators* 为 220，*max_depth* 为 8，其他参数为默认值时，基于梯度提升树算法构建的模型就是本文 CO_2 排放量预测的最佳模型。

为了直观查看样本预测值和真实值的情况，利用折线将预测值和真实值可视化。基于梯度提升树算法构建的模型预测值和真实值的对比折线图如图 4 所示。

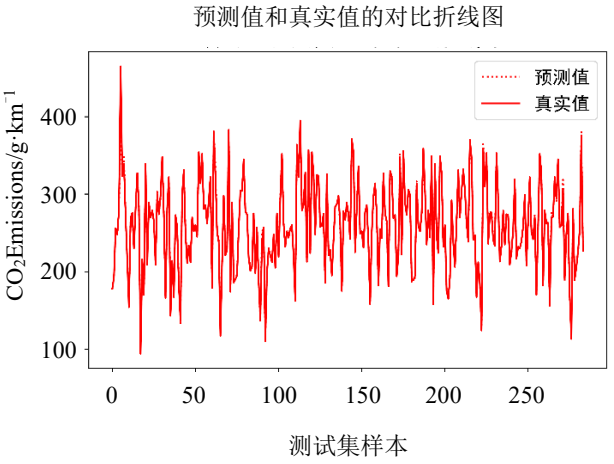

图 4　基于梯度提升树算法构建的模型预测值和真实值的对比折线图

由图 4 可以看出，只有少部分样本的预测值和真实值存在较小误差，绝大部分样本都能准确预测，模型预测效果较优。

（下转第 45 页）