DOI:10.19651/j.cnki.emt.2107317

基于 Python 的汽车运行油耗预测模型的构建*

黄 赫 储江伟 艾曦峰 李 红
(东北林业大学 交通学院 哈尔滨 150040)

摘 要: 运用 Python 语言对 OBD 采集的车辆运行数据搭建油耗预测模型。以车速 v,发动机转速 n,进气管绝对压力 P,节气门位置 TP,冷却液温度 CT,负荷率 L,怠速时间 IT 及加速度 a 等作为自变量,百公里油耗作为因变量,用 SelectKbest 函数将参数与因变量相关性强度进行排序并做简要分析,用基于 Tensorflow 的多层感知机(MLP)神经网络模型以及支持向量机(SVM)多元线性回归模型同时对油耗进行预测。支持向量机模型 RMSE 为 0.088,MAE 为 0.56;Tensorflow 神经网络模型 RMSE 为 0.132,MAE 为 0.70。结论说明模型比较可靠,可为进一步分析汽车油耗与车辆运行状态参数之间的关系提供理论依据。

关键词: OBD;油耗;支持向量机(SVM);神经网络;多元回归
中图分类号: U121 **文献标识码:** A **国家标准学科分类代码:** 580.99

Construction of vehicle fuel consumption forecast model based on Python

Huang He Chu Jiangwei Ai Xifeng Li Hong
(School of Traffic and Transportation, Northeast Forestry University, Harbin 150040, China)

Abstract: Using Python language, the fuel consumption prediction model is built based on the vehicle operation data collected by OBD. Taking vehicle running state parameters such as vehicle speed v, engine speed n, intake pipe absolute pressure P, throttle position TP, coolant temperature CT, load rate L, idle time IT, acceleration a as independent variables and 100 km fuel consumption as dependent variables, the correlation intensity between parameters and dependent variables is sorted by SelectKbest function and briefly analyzed. The (MLP) neural network model of multilayer perceptron based on Tensorflow and the multiple linear regression model of support vector machine (SVM) are used to predict the fuel consumption at the same time. Support vector machine model RMSE is 0.088, MAE is 0.56, Tensorflow neural network model RMSE is 0.132, MAE is 0.70. The results show that the two models are accurate in the prediction of fuel consumption, which can provide a theoretical basis for further elucidating the relationship between vehicle fuel consumption and vehicle running state parameters.

Keywords: OBD;fuel consumption;support vector machine (SVM);neural network;multiple regression

0 引 言

随着时代的进步,科技的发展,道路上的汽车也是越来越多,但同时汽车的发展也给我们带来了诸多负面的影响,例如尾气对环境的破坏,还有对能源的消耗。因此对油耗进行多方面的研究对节约资源是有意义的。许癸驹等[1]将 BP 神经网络应用到 ODFM 系统识别。杨亚联等[2]运用 DP(dynamic programing)算法在特定工况下对 EVT 构型进行仿真,采用 GRNN 神经网络对仿真所得油耗进行建模并与之进行对比验证。李洪亮等[3]针对车辆运行油耗监测的难点设计了以 STC89C52 为核心的车载油耗监测系统。Kanarachos 等[4]运用 PPMCC(pearson product-moment correlation coefficient)对影响燃油消耗高的驾驶行为指标滤波,再由这些高油耗指标生成聚合模型。Yamashita 等[5]以城市道路为基础建立了车辆平峰期和高峰期的油耗模型,在不同路况背景下的油耗模型以及速度和加速度不相同的条件下的车辆瞬时油耗模型。Capraz 等[6]分别采用神经网络,多元线性回归以及支持向量机 3 种油耗预测模型对车辆瞬时燃油消耗以及总燃油消耗进行预测。宋大凤等[7]提出的以理论分析为基础的预测模型,可以快速地为混合动力系统开发提供准确的燃油经济性及成本分析结果。马荣影等[8]运用 Python 语言对车辆运行参数对燃油

收稿日期:2021-07-19
* 基金项目:中央高校基本科研业务费专项资金(2572020AW49)项目资助

消耗的影响进行建模分析。马荣影等[9]采用 K-means 聚类方法对车辆行驶状态参数进行一系列的分析。金辉等[10]由于稳态油耗的稳态模型使得车辆在瞬态条件运行时的油耗预测与真实油耗有较大误差，基于稳态油耗模型再引入瞬态修正模块，构建瞬态油耗模型 BIT-TFCM，并用实际油耗对 BIT-TFCM 做验证。张金辉等[11]通过最小二乘法对所构建的车辆行驶瞬态油耗模型的参数进行估计，引入加权因子，进一步提高模型准确性，并加以验证。孙凤英等[12]从环境因素等诸多角度对我国车辆运行油耗进行一系列分析并给出相应看法与措施。马健军等[13]对车辆在市郊道路工况下所采集的 OBD 数据进行建立油耗回归模型并进行误差分析。姜平等[14]采用主成分分析的方法对所采取油耗等数据进行处理，并对传统 BP 神经网络油耗模型以及主成分分析与神经网络融合的油耗模型进行比较分析。冯莲[15]根据移动端的传感器数据，对加速度传感器感知的车辆加速度存在的随机噪声和重力分量问题搭建了道路坡度估计和加速度修正模型，又以道路坡度等为输入建立油耗预测模型。陈俐等[16]采取神经网络与时间节点融合的全新策略对公路货运费用进行预测。王立宇等[17]以反向神经网络为基础对油耗以及排放等进行了预测并进行验证。

由于 Python 语言的简单易操作，并且拥有多个功能性的类库，所以在处理数据以及机器学习领域很是便利。本文以 OBD 采集的 17 000 余条车辆实时运行状态数据为基础，提取与油耗相关的特征参数，再基于特征参数分别建立支持向量机模型和神经网络模型；通过对两个模型的误差分析，验证模型的准确性和可靠性，为进一步说明车辆油耗与车辆运行状态之间的关系提供理论依据。

1　基于 OBD 的汽车运行状态数据获取及预处理

1.1　汽车运行状态数据采集

本文数据样本来自冬季低温工况下通过 OBD 获得的车辆多次短时运行的数据集，相关参数有行驶车速 v，进气管绝对压力 P，节气门位置 TP，负荷率 L，怠速时间 IT，发动机转速 n，冷却液温度 CT，加速度 a 等。各参数详解见文献[18]。原始数据如表 1 所示。

表 1　由 OBD 采集数据示例

序号	IT/s	a/(m·s^{-1})	L/%	CT/℃	n/(r·min^{-1})	v/(km·h^{-1})	TP/%	P/kPa	FC/(L/100 km)
1	0	0	0	12.00	0	0	7.04	98.00	0
2	0	0	0	12.00	0	0	7.04	98.00	0
…	…	…	…	…	…	…	…	…	…
17 020	102.00	−3.45	27.63	85.00	1 798.00	29.00	5.32	66.00	8.88

1.2　数据预处理

1)运行状态数据的基本统计量

本文使用 Pandas 对数据进行处理，用 read_csv 函数读取所采数据集。使用 describe()函数对各参数做基本统计，包括 Q(参数总量)，M(参数平均值)，Std(参数标准差)，Max(参数最大值)，Min(参数最小值)，Uq(参数上四分位数)，Dq(参数下四分位数)等。数据基本统计如表 2 所示。

表 2　对 OBD 采集数据的统计分析结果

参数	IT/s	a/(m·s^{-1})	CT/℃	P/kPa	n/(r·min^{-1})	v/(km·h^{-1})	L/%	TP/%	FC/(L/100 km)
Q	17 020	17 020	17 020	17 020	17 020	17 020	17 020	17 020	17 020
M	165.0	−0.01	71.23	48.85	1 202.27	27.40	35.61	4.74	16.82
Std	112.0	0.53	19.76	25.70	391.33	24.05	23.31	2.80	9.82
Min	0	−2.78	−17.0	20	0	0	0	0	8.72
Uq	76	−0.27	64.0	27	887	5.00	17.65	2.75	11.7
$Median$	151.0	0	77.0	38	1 227	22	27.84	3.53	13.59
Dq	263.0	0.28	86.0	65	1 460	47	46.67	6.67	17.22
Max	511	2.22	94.0	100.00	3 378	94	100.00	17.65	50

据表 2 统计可知各参数总量均为 17 020，无空缺值且无异常值。但数据的量值范围差异过大且有符号正负不统一等情况，需对数据做归一化处理，解决数据量纲不同的问题。

2)数据归一化

归一化公式如式(1)所示。

$$x_i = (x_i - x_{\min})/x_{\max} - x_{\min} \tag{1}$$

通过式(1)，将样本数据均归一化到[0,1]之间，使各

参数的样本数据具有统一性,提高模型精准度。归一化后　　数据如表 3 所示。

表 3　采集数据归一化处理结果部分示例

IT/s	a/(m·s^{-1})	CT/℃	P/kPa	n/(r·min^{-1})	v/(km·h^{-1})	L/%	TP/%	FC/(L/100 km)
0.05	0.62	0.21	0.23	0.43	0.12	0.33	0.31	0.70
0.05	0.62	0.21	0.23	0.43	0.12	0.33	0.31	0.70
0.05	0.62	0.24	0.24	0.34	0.13	0.33	0.31	0.70
…	…	…	…	…	…	…	…	…
0.05	0.49	0.24	0.26	0.36	0.13	0.33	0.31	0.68
0.05	0.49	0.24	0.27	0.39	0.13	0.34	0.31	0.68

2　汽车运行状态对油耗影响的分析及特征参数提取

2.1　汽车运行状态对油耗影响的分析

根据所采集的数据做出车辆运行速度区间分布比例,如图 1 所示。由图 2 可看出,车辆行驶速度在 0～20 km/h 的状态占近 50%,可看出车辆行驶的平均车速低且运行时间较长;而发动机还处于低负荷率状态,导致油耗较高;此外,车辆频繁的加减速行驶使一部分能量用于克服加速阻力或转化为制动时的热能,也会导致油耗增加。

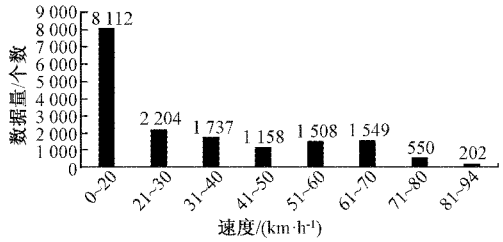


图 1　速度区间分布

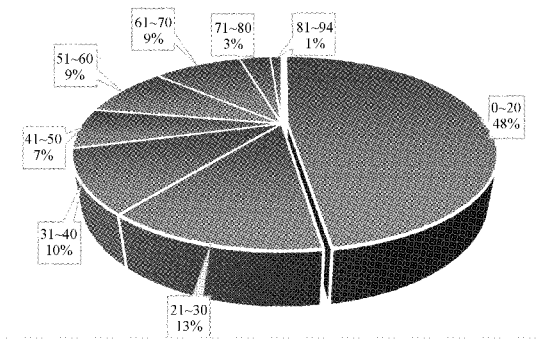


图 2　速度区间所占百分比

由于实际交通状况等因素(红绿灯及交通拥堵等)影响,导致在行驶过程中有多次的加减速,这样的行驶状态同样也会使运行油耗增加。

车辆在运行的过程中发动机冷却液的温度变化曲线,如图 3 所示。由于实验是在冬季低温工况下进行,车辆运行时多是冷车启动或低温启动,发动机的机油粘稠度高,导致发动机运转阻力大幅度增加;发动机工作温度低,导致燃油雾化差,因此,汽车冷车启动或低温启动时,发动机喷油量增加,直至温度上升到正常行车温。

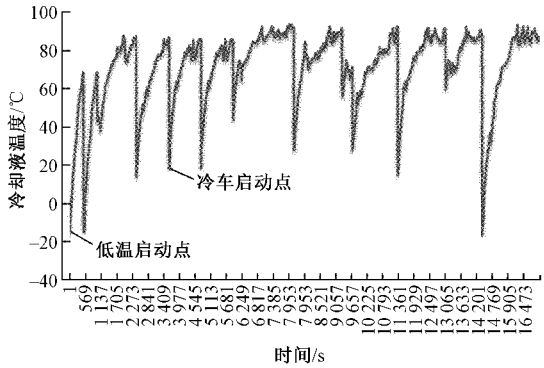


图 3　车辆发动机冷却液温度拟合曲线

2.2　特征参数提取

汽车运行油耗不仅与汽车运行状态参数有关,发动机工作状态对其也有一定的影响,所以各个参数对油耗的影响程度必定是不同的。

为减少模型计算量,需要进行特征参数选择。调用 Python 特征选择库中的 SelectKbest 函数选出对油耗影响最大的参数。通过设定一个阈值 n,选出 n 个与油耗最相关的参数(即最佳特征参数)。分别设置 $n=1,2,\cdots,8$,则可以得出各参数与油耗相关性强弱顺序,各参数对油耗影响强度排序从大到小为:1 冷却液温度、2 车速、3 节气门位置、4 发动机转速、5 怠速时间、6 负荷率、7 加速度、8 进气管绝对压力。

再调用 score 函数,计算 n 取各个数值所对应的模型得分。由图 4 可知,对于 MLP 模型当参数个数为 6 时得分最高,所以建模时选取与油耗相关性排名前 6 的参数能达到最佳效果。

3　基于神经网络的汽车运行油耗模型构建及验证

3.1　汽车运行油耗模型构建

1)基于支持向量机的油耗模型

支持向量机(support vector machines,SVM),通过寻求结构化风险最小,从而提高泛化能力,进而在统计样本

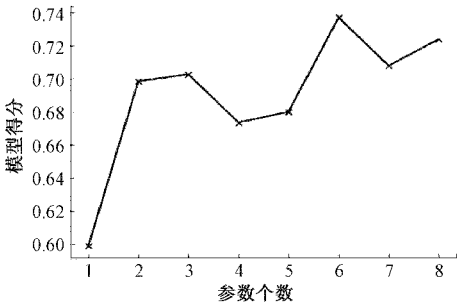


图 4　参数个数与所对应模型得分

量较小的时候也能获得较好的效果。

给定一个训练样本 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$y_i\in R$，期望学成一个回归模型，使得 $f(x)$ 与 y 最大限度的接近，ω 和 b 是待定模型参数。

假设 $f(x)$ 与 y 之间最大可以有 ε 的误差，即只有 $f(x)$ 和 y 之间的差值绝对值大于 ε 时才计算。随即 SVR 问题便转化为：

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}l_\varepsilon(f(x_i),y_i) \tag{2}$$

其中，C 为正则化常数，l_ε 为不敏感损失(insensitive loss)函数。

$$l_\varepsilon(z)=\begin{cases}0, & |z|\leqslant\varepsilon\\ |z|-\varepsilon, & \text{其他}\end{cases} \tag{3}$$

引入松弛变量 ξ_i 和 $\hat{\xi}_i$ 可将公式改为：

$$\begin{aligned}&\min_{w,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i,\hat{\xi}_i)\\&f(x_i)-y_i\leqslant\varepsilon+\xi_i,\\&y_i-f(x_i)\leqslant\varepsilon+\hat{\xi}_i,\\&\xi_i\geqslant 0,\hat{\xi}_i\geqslant 0,i=1,2,\cdots,m\end{aligned} \tag{4}$$

引入拉格朗日乘子 μ_i：

$$L(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu})=\min_{w,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\xi_i\mu_i-\sum_{i=1}^{m}\hat{\xi}_i\hat{\mu}_i+\sum_{i=1}^{m}\alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)+\sum_{i=1}^{m}\hat{\alpha}_i(yi-f(xi)-\varepsilon-\hat{\xi}_i) \tag{5}$$

再令 $L(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},\mu,\hat{\mu})$ 对 $w,b,\xi,\hat{\xi}$ 的偏导为 0 可得：

$$w=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i \tag{6}$$

上述公式均需满足：

$$\begin{aligned}&\alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)=0,\\&\hat{\alpha}_i(y_i-f(x_i)-\varepsilon-\hat{\xi}_i)=0,\\&\alpha_i\hat{\alpha}_i=0,\xi_i\hat{\xi}_i=0,\\&(C-\alpha_i)\xi_i=0,(C-\hat{\alpha}_i)\hat{\xi}_i=0\end{aligned} \tag{7}$$

SVR 的解形如：

$$f(x)=\sum_{i\ 1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^{\mathrm{T}}x+b \tag{8}$$

综上可求出：

$$b=y_i+\varepsilon-\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^{\mathrm{T}}x \tag{9}$$

考虑到特征映射形式：

$$w=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^{\mathrm{T}}x\phi(x) \tag{10}$$

SVR 可表示为：

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)k(x_i^{\mathrm{T}}x)+b \tag{11}$$

其中，$k(x_i^{\mathrm{T}}x)=\phi(x_i)^{\mathrm{T}}\phi(x_j)$ 为核函数。

在本文中 SVM 模型里的 Kernel 运算核使用的是 rbf(高斯核函数)，C 惩罚因子参数为 1，投影超平面维度 degree 为 3，其余参数使用学习库中默认的参数。

2)Tensorflow 框架下的神经网络油耗模型

Tensorflow 框架下的多层感知机网络(multiple layer perceptron network，MLP)是一种如今普遍运用的人工神经网络，有较强的自主学习以及泛化能力。它由多个连接层组成，连接层通常为输入层、隐含层和输出层。如图 5 所示，是一个具有 2 个隐含层的 MLP 结构，样本在模型中由输入层向输出层传播，m 维输入数据通过神经网络的处理被映射成 n 维输出数据，其核心是通过误差逆向传播修正各神经元节点的权重值 w 和阈值 b。

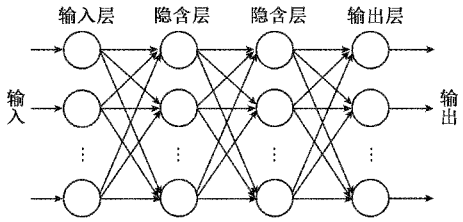


图 5　MLP 结构

本文 MLP 油耗模型的输出为平均油耗，该模型输入层神经元数量为输入变量的数量 6，隐藏层层数为 1，隐藏层使用神经元数量为 12，输出参数个数为 1，采用的 softmax 函数(式(12)为 K 个线性函数的 Softmax 函数的复合)。Batchsize 为 50，一共训练的迭代次数 epoch 为 50 次，训练集与测试集比例 8∶2，其中训练集的 10%用来做验证。

$$P(y=j)=\frac{e^{x^{\mathrm{T}}}W_j}{\sum_{k-1}^{K}e^{x^{\mathrm{T}}W_k}} \tag{12}$$

3.2　油耗预测模型的结果验证及分析

油耗预测模型能否达到实际要求的效果，需要进行模型的验证以及对预测偏差的分析计算。验证能通过并且偏差也足够小，才可以将油耗预测模型实际应用。

调用 metrics. mean_squared_error 函数计算模型均方

根误差(RMSE):是观测值与真实值误差的平方和与观测次数 m 比值的平方根,评价观测值同真实值的误差。调用 metrics. mean_absolute_erroe 函数计算模型的平均绝对误差(MAE):预测值和观测值之间绝对误差的平均值。模型验证误差如表 4 所示,由表 4 可看出 MLP 油耗模型的 RMSE 为 0.132,MAE 为 0.70;SVR 模型的 RMSE 为 0.088,MAE 为 0.56;可看出两模型可靠性都比较强,SVR 模型更稳定一些,精准性更高一些。

表 4　模型误差验证

误差	RMSE	MAE
MLP	0.132	0.70
SVR	0.088	0.56

基于机器学习的模型在训练时,迭代次数可以改变模型训练效果的,迭代次数越高,相应的效果也就越好,但花费的时间也越长。截取使用 Tensorboard 所绘出的 loss 曲线如图 6 所示。由图 6 可知模型在迭代二十几次的时候误差就已经趋于稳定,由此可知本模型训练时间短,模型训练速度快。

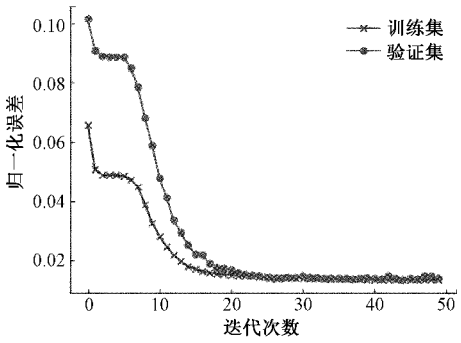
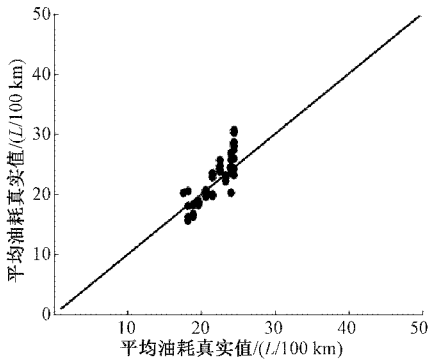
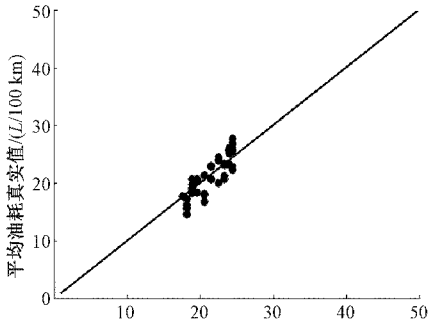


图 6　Tensorflow 神经网络误差曲线

根据两个模型分别绘出的部分油耗预测值与实际油耗值的偏差情况如图 7(a)、(b)所示;点到斜线的垂直距离越大表示误差越大,由图可看出两模型预测效果都不错,SVR 油耗模型向中线收敛程度更高一些,模型相对更精准一些。



(a) MLP回归结果比较



(b) SVR回归结果比较

图 7　真实值与预测值

4　结　　论

本文通过在车辆上安装 OBD,采集油耗以及相关车辆运行状态参数,结合机器学习对数据进行归一化等处理,并使用神经网络以及支持向量机方法建立油耗预测模型,并对两种方法进行了验证及分析,本文所建立模型可为实时油耗预测相关研究提供参考依据。但在实际中车辆自身配置、装载情况以及驾驶技术等对汽车运行时的油耗均有大小不等的影响。根据本文研究得出如下结论:

1)各参数对油耗影响强度,冷却液温度是最强的,其次是车速,进气管绝对压力对其影响相对最小。

2)通过汽车行驶速度的曲线以及行驶特点(多低速以及频繁加减速)进行分析,分析得出本实验数据预测的油耗会比正常工况下高很多。再通过预测值与真实值的拟合曲线整体来看,模型油耗变化态势与真实值的变化趋势拟合度较高,进一步验证了模型的稳定性。

3)基于 Python 语言所分别构建的支持向量机模型 RMSE 为 0.088,MAE 为 0.56;Tensorflow 神经网络模型 RMSE 为 0.132,MAE 为 0.70。相比之下 MLP 模型训练速度更快一些,SVR 模型精确性更好一些,总体上看误差均比较理想,可以证明模型的可行性以及可靠性。

参考文献

[1]　许癸驹,钱雅楠,代红英. 基于人工神经网络的 OFDM 系统信道辨识与补偿[J]. 电子测量技术,2021,44(8):120-124.

[2]　杨亚联,邓淇元,刘强寿,等. 混合动力耦合系统神经网络油耗模型构建[J]. 重庆大学学报,2019,42(7):1-9.

[3]　李洪亮,储江伟. 基于 STC89C52 的车载油耗实时监测系统设计[J]. 森林工程,2014,30(1):137-140.

[4]　KANARACHOS S, MATHEW J E, FITZPATRICK M. Instantaneous vehicle fuel consumption estimation using smartphones and recurrent neural networks[J]. Expert Systems with Applications, 2019, 120: 436-447.

[5]　YAMASHITA R J, YAO H H, HUNG S W, et al. Accessing and constructing driving data to develop fuel

consumption forecast model[J]. IOP Conference Series Earth and Environmental Science,2018,113(1):012217.
[6] CAPRAZ A G, OZEL P, SEVKLI M, et al. Fuel consumption models applied to automobiles using real-time data: A comparison of statistical models[J]. Procedia Computer Science,2016,83:774-781.
[7] 宋大凤,吴西涛,曾小华,等.基于理论油耗模型的轻混重卡全生命周期成本分析[J].吉林大学学报(工学版),2018,48(5):1313-1323.
[8] 马荣影,韩锐,艾曦锋,等.基于 Python 的汽车油耗多参数回归模型构建方法[J].公路交通科技,2020,37(6):145-150.
[9] 马荣影,储江伟,艾曦锋,等.基于 python 的汽车运行状态参数对油耗影响的聚类[J].交通科技与经济,2020,22(3):42-44,74.
[10] 金辉,周敏,李世杰.基于瞬态修正的车辆燃油消耗模型建模[J].北京理工大学学报,2017,37(5):473-477,484.
[11] 张金辉,李克强,徐彪,等.基于最小二乘法的车辆瞬态燃油消耗估计[J].汽车工程,2018,40(10):1151-1157.
[12] 孙风英,金世勇.影响汽车运行燃料消耗量的多元分析[J].森林工程,2009,25(4):62-64.
[13] 马健军,黄玉华,龙志军,等.市郊道路工况车辆油耗模型研究[J].南方农机,2017,48(8):12-13.
[14] 姜平,石琴.基于行驶工况特征的汽车燃油消耗的预测[J].汽车工程,2014,36(6):643-647.
[15] 冯莲.基于移动终端传感器数据的汽车行驶油耗估计方法[D].重庆:重庆大学,2019.
[16] 陈俐,方叶祥,甘平,等.基于大数据分析的多维公路货运价格预测问题研究[J].交通科技与经济,2021,23(4):65-72.
[17] 王立宇,滕勤,庄远.基于 BP 神经网络的喷水汽油机性能预测[J].内燃机与动力装置,2021,38(3):25-30.
[18] 周晓飞.别克世纪轿车发动机电控系统数据流及诊断分析[J].汽车维修,2009(6):45-47.

作者简介

黄赫,硕士研究生,主要研究方向为汽车技术状态监测与性能仿真。
E-mail:1145073697@qq.com

储江伟(通信作者),博士,教授,主要研究方向为汽车运行品质控制理论与方法。
E-mail:cjw_62@163.com