DOI:10.19789/j.1004-9398.2022.02.003
文献引用:惠姣姣.基于主成分分析和两阶段聚类分析的汽车行驶工况评价模型[J].首都师范大学学报(自然科学版),2022,43(2):14-19.
HUI J J. Evaluation model of vehicle driving condition based on principal component analysis and two-stage cluster analysis[J]. Journal of Capital Normal University(Natural Science Edition),2022,43(2):14-19.

基于主成分分析和两阶段聚类分析的汽车行驶工况评价模型

惠　姣　姣

(云南大学数学与统计学院，云南 昆明　650091)

摘要:为了准确评价汽车行驶工况,促进汽车节能减排.本文在修正异常数据的基础上,提取了1 743条运动学片段,并构建了一种包括平均速度、发动机转速比和扭矩百分比等18个特征参数的汽车运动特征评价体系.利用基于主成分分析和两阶段聚类分析的汽车行驶工况评价模型,将运动学片段的样本数据聚为低速工况、中速工况和高速工况3类.聚类结果的准确率为98.970%,说明本文所建的评价模型是合理有效的.

关键词:汽车行驶工况;主成分分析;两阶段聚类分析;PTV评价模型
中图分类号:U491.1

Evaluation model of vehicle driving condition based on principal component analysis and two-stage cluster analysis

HUI Jiaojiao

(School of Mathematics and Statistics, Yunnan University, Kunming Yunnan　650091)

Abstract: To accurately evaluate the driving conditions of vehicles and promote the energy saving and emission reduction of vehicles. On the basis of correcting the abnormal data, 1 743 kinematics segments are extracted, and an evaluation system of vehicle motion characteristics including 18 characteristic parameters such as average speed, engine speed ratio and torque percentage is constructed. Finally, a vehicle driving condition evaluation model based on principal component analysis and two-stage cluster analysis was used to cluster the sample data of 1 743 kinematic segments into three categories: low-speed operating conditions, medium-speed operating conditions, and high-speed operating conditions. The accuracy of the clustering results is 98.970%, which shows that the evaluation model built in this paper is reasonable and effective.

Keywords: vehicle driving condition; principal component analysis; two-stage cluster analysis; PTV evaluation model
CLC: U491.1

收稿日期:2021-02-21

0　引　言

近年来,随着汽车保有量的快速增长,我国的交通状况得到了很大改善,研究表明,以新标欧洲循环测试(new European driving cycle,NEDC)工况为基准所优化标定的汽车,实际油耗与法规认证结果偏差较大[1].多年来,这一问题始终没有得到纠正,这无形中影响了政府的公信力.造成实际油耗与法规认证结果偏差较大的原因,一方面是我国长期以NEDC工况为基准优化汽车的各项特征参数,随着我国道路及交通状况的改变,NEDC工况越来越不符合我国实际情况[2];另一方面,我国地域辽阔,各个城市发展程度、交通状况及气候条件不同,因此各个城市的汽车行驶工况也存在显著差异.在这种条件下,构建符合某个城市实际情况的汽车行驶工况显得尤为重要.

汽车行驶工况的确定,是汽车行业一项重要的基础技术,是车辆能耗、排放测试方法和限值标准的基础,也是汽车各项特征参数优化的主要基准.该技术的实施,需调查车辆的实际行驶状况,并运用统计分析方法,定量描述在特定交通环境下(如:高速公路、城市道路等),小轿车、公交车和重型车等车辆的行驶状况[3].对于汽车行驶工况的构建,刘明哲[4]利用基于主成分分析(principal component analysis,PCA)的分层聚类方法,将汽车行驶工况凝聚为均速工况、加速工况、减速工况和怠速工况4类;李耀华等[5]通过马尔可夫蒙特卡洛方法,构建了西安市城市公交行驶工况;刘宇等[6]利用基于PCA的*K*-means均值聚类,对海湾大学城的汽车行驶工况与交通安全进行分析;徐婷等[7]首先利用马尔科夫蒙特卡洛模拟法,构建行驶工况,并利用基于PCA的*K*-means均值聚类,将运动学片段聚为加速路段、匀速路段和道路线形复杂路段3类.

已有文献都是利用常见的聚类分析方法来处理数据,但当样本量很大时,聚类结果的精确程度会有所下降.由于本文数据样本量较大,因此在对汽车行驶工况进行分析时,采用改进的聚类分析方法——两阶段聚类分析(two-stage cluster analysis,TSCA).本文首先选取与汽车行驶工况有关的18个特征参数,并利用PCA将其综合成若干个主成分,在此基础上再进行TSCA,以期建立基于PCA和TSCA的汽车行驶工况评价模型,即PTV评价模型.

1　数据预处理

汽车行驶工况的评价,需建立在运动学片段的基础上进行,而本文获取到的数据是GPS记录的实时数据,且数据中存在大量异常值.因此,首先要对数据进行预处理,并在此基础上将实时样本数据划分为运动学片段.运动学片段是指汽车从怠速状态开始至下一个怠速状态开始之间的车速区间[8],其提取应在时间连续并修正异常值的基础上进行.

本文使用的数据为某辆车时间段为2017年12月18日13:42:13—24日13:37:49 GPS设备所采集的数据,总样本量为185 725条.汽车在行驶过程中,由于各种客观因素,可能会导致采集到的数据含有异常值,对异常值进行如下处理:

(1)由于隧道、高层建筑覆盖等多种因素,导致检测车速的GPS信号丢失,而造成记录时间出现中断.对于时间不连续问题,利用Rstudio1.4.1103软件对数据进行插补,并按时间顺序重新排列[9].

(2)汽车的加速度和减速度超出理论值.通常,普通轿车速度由0增加至100 km/h的时间应>7 s,紧急刹车时最大减速度应为7~8 km/h[10].经过计算,当车速为0~100 km/h时,加速度应≤3.968 3 m/s^2,若>3.968 3 m/s^2,则按边界值3.968 3 m/s^{2}来修正.

(3)怠速时间>180 s.怠速即汽车停止运动,但发动机保持最低转速运转的连续过程.通常,怠速时间>180 s为异常情况,此时,将该部分划分为2条运动学片段即可.

(4)由于车主停车不熄火等候,或停车熄火了,但采集数据设备仍在运行等不当操作,从而产生异常停车数据.对于此类异常值,使用Rstudio1.4.1103软件对等候时间进行判断,若≤180 s,按怠速处理;若>180 s,则视为长时间停车,从而分成2条运动学片段.

(5)由于堵车等多种外在因素,导致的车速长时间过低(最高车速<10 km/h).本文将断续低速行驶时间以180 s为界限,若堵车时间>180 s,则属于异常情况,将其划分为2条运动学片段;若≤180 s,则按照怠速的情况处理.

对185 725条原始数据经过上述5种情形的异常数据处理,共填充15 277条数据,经处理后,时间上变为连续的,数据量变为201 002条.根据运动学片段和怠速的定义,应该以GPS车速为0作为分界

点,划分运动学片段.本文采用Rstudio1.4.1103软件中的base包将201 002条数据划分为1 743条运动学片段.该车一段汽车行驶工况曲线如图1所示,其包含7条运动学片段.

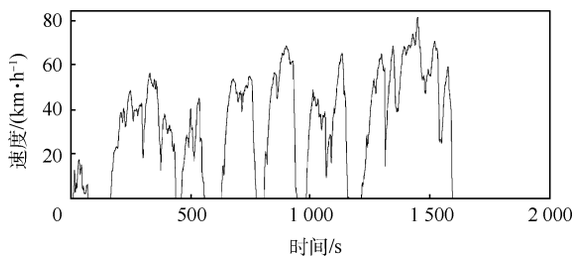


图1 汽车行驶工况曲线

2 PTV评价模型的构建

2.1 汽车运动特征评价体系

在评估汽车行驶工况之前,首先构建一个合理的汽车运动特征评价体系.经过查阅参考文献,评估体系中应包含18个特征参数,分别为平均速度、平均行驶速度、怠速时间比、平均加速度、平均减速度、加速时间比、减速时间比、速度标准差、加速度标准差[11]、经度、纬度、发动机转速比、扭矩百分比、瞬时油耗、油门踏板开度、空燃比、发动机负荷百分比和进气流量[12].其中,平均速度、平均行驶速度、怠速时间比、平均加速度、平均减速度、加速时间比、减速时间比、速度标准差和加速度标准差这9个特征参数的数据需根据运动学片段的划分计算得到.

为方便表述,将1 743条运动学片段看作1 743条样本数据,并将18个特征参数分别记为x_i(i=1,2,3,…,18).从数据形式来看,各特征参数的量纲不统一,在建立模型之前,需对数据进行标准化处理.对于正向因子,采取$(x_i - x_{max})/(x_{max} - x_{min})$进行标准化;对于负向因子,采取$(x_{max} - x_i)/(x_{max} - x_{min})$进行标准化[13].

2.2 PCA

由于特征参数过多,在建模时为了避免"维数灾难",首先对1 743条样本数据的18个特征参数,进行PCA,将18个特征参数综合成18个主成分,以达到降维的目的[14].PCA的总方差解释如表1所示.一般情况下,主成分的累积方差贡献率达到80.000%时,便认为对应的主成分具有良好的解释原始信息的能力[15].前5个主成分的方差贡献率达到82.156%,即前5个主成分解释了82.156%的原始信息,这说明前5个主成分能较好地解释18个特征参数的信息.

18个主成分的碎石图如图2所示.第1个主成

表1 总方差解释

主成分	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差/%	累积/%	合计	方差/%	累积/%	合计	方差/%	累积/%
1	8.234	45.744	45.744	8.234	45.744	45.744	5.547	30.819	30.819
2	2.047	11.372	57.116	2.047	11.372	57.116	3.048	16.932	47.752
3	1.896	10.536	67.652	1.896	10.536	67.652	2.665	14.805	62.557
4	1.564	8.687	76.339	1.564	8.687	76.339	2.001	11.114	73.670
5	1.047	5.817	82.156	1.047	5.817	82.156	1.527	8.486	82.156
6	0.800	4.447	86.603	—	—	—	—	—	—
7	0.675	3.752	90.355	—	—	—	—	—	—
8	0.529	2.941	93.296	—	—	—	—	—	—
9	0.472	2.625	95.920	—	—	—	—	—	—
10	0.279	1.552	97.473	—	—	—	—	—	—
11	0.205	1.140	98.613	—	—	—	—	—	—
12	0.093	0.516	99.129	—	—	—	—	—	—
13	0.063	0.348	99.477	—	—	—	—	—	—
14	0.035	0.194	99.671	—	—	—	—	—	—
15	0.029	0.160	99.831	—	—	—	—	—	—
16	0.012	0.067	99.898	—	—	—	—	—	—
17	0.011	0.059	99.957	—	—	—	—	—	—
18	0.008	0.043	100.000	—	—	—	—	—	—

注:—指无数据.

分到第2个主成分的曲线非常陡峭;在第5个主成分之后,特征值降到<1,且曲线较为平缓.这说明在利用PCA处理汽车运动特征评估体系中的18个特征参数时,降维效果很好,前5个主成分能够较好地代替18个特征参数的信息.

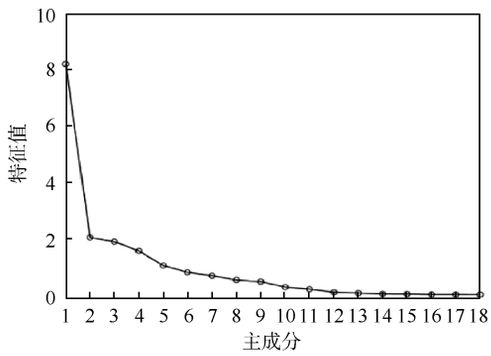


图2　主成分分析(PCA)碎石

前5个主成分与18个特征参数的相关关系如表2所示.

选取的5个主成分记为$F_1,F_2,\cdots,F_5$,则5个主成分与各特征参数之间的关系为

$$\begin{cases} F_1 = 0.881x_1 + 0.930x_2 + \cdots + 0.770x_{18}, \\ F_2 = 0.281x_1 + 0.089x_2 + \cdots + 0.294x_{18}, \\ \quad\vdots \\ F_5 = 0.062x_1 + 0.090x_2 + \cdots + 0.068x_{18}. \end{cases}$$

由表2可以看出:第1主成分主要解释了平均行驶速度、速度标准差、平均速度、发动机转速、空燃比、进气流量、油门踏板开度和加速度标准差8个特征参数的信息;第2主成分主要解释了加速时间比、减速时间比和怠速时间比3个特征参数的信息;第3主成分主要解释了发动机负荷百分比、扭矩时间比和瞬时油耗3个特征参数的信息;第4主成分主要解释了经度和纬度2个特征参数的信息;第5主成分主要解释了平均加速度和平均减速度2个特征参数的信息.这说明前5个主成分能较好地解释18个特征参数的信息.

2.3　TSCA

聚类分析是一种数据规约技术,可以把大量的观测值规约为若干个类,该技术被广泛应用于生物行为科学、市场以及医学研究中.聚类分析的一般步骤为:选择合适的变量、缩放数据、寻找异常点、计算距离、选择聚类算法、获得一种或多种聚类方法、确定类的数目、获得最终的聚类解决方案、结果可视化、解读类和验证结果[16].最常见的聚类方法是层次聚类和划分聚类.考虑到数据样本量较大,本文采用TSCA对1 743条样本数据进行聚类分析.针对数据量较大的利用层次方法的平衡迭代规约和聚类(balance iterative reducing and clustering using hierarchies,BIRCH)算法进行准聚类.BIRCH算法基于CF树生长的思想,对传统层次聚类法加以改进,其实质是层次聚类方法与其他聚类方法相

表2　前5个主成分与18个特征参数的相关关系

特征参数	主成分1	主成分2	主成分3	主成分4	主成分5
平均速度	0.881	0.281	0.293	−0.001	0.062
平均行驶速度	0.930	0.089	0.245	0.005	0.090
怠速时间比	−0.454	−0.743	−0.169	0.021	0.016
平均加速度	0.088	−0.032	−0.028	0.021	0.839
平均减速度	−0.044	−0.006	0.037	0.036	−0.821
加速时间比	0.236	0.938	0.111	0.026	0.011
减速时间比	−0.040	−0.930	−0.115	−0.013	0.013
速度标准差	0.888	0.064	0.180	0.015	0.101
加速度标准差	0.520	0.214	0.144	0.015	0.242
经度	−0.006	0.013	−0.038	0.996	−0.013
纬度	−0.010	0.004	−0.031	0.996	−0.004
发动机转速	0.816	0.425	0.272	−0.033	0.068
扭矩时间比	0.459	0.184	0.831	−0.039	−0.004
瞬时油耗	−0.192	−0.403	−0.698	−0.009	0.105
空燃比	0.780	0.380	0.369	−0.065	0.136
油门踏板开度	−0.592	0.056	−0.022	−0.004	0.170
发动机负荷百分比	0.332	−0.029	0.878	−0.059	−0.017
进气流量	0.770	0.294	0.499	−0.048	0.068

结合的多阶段聚类[17],准聚类过程如表3所示.可知,BIRCH算法将1 743条样本数据初步凝聚成15个子类.其中,第1个子类的施瓦兹的贝叶斯判别准则(Schwarz's Bayesian criterion,BIC)值为1 744.221,在15个子类中数值最大,该子类包含的信息量最多.第5个子类的BIC值为1 139.408,在15个子类中数值最小,则该子类包含的信息量最少.

表3 准聚类过程

子类	BIC值	ΔBIC	BIC变化率	距离测量比
1	1 744.221	—	—	—
2	1 388.628	−355.593	1.000	2.036
3	1 248.579	−140.048	0.394	1.336
4	1 160.887	−87.692	0.247	1.740
5	1 139.408	−21.480	0.060	1.614
6	1 151.979	12.572	−0.035	1.183
7	1 173.138	21.159	−0.060	1.063
8	1 197.066	23.928	−0.067	1.080
9	1 224.272	27.206	−0.077	1.078
10	1 254.432	30.159	−0.085	1.256
11	1 292.312	37.880	−0.107	1.007
12	1 330.410	38.099	−0.107	1.040
13	1 369.657	39.246	−0.110	1.221
14	1 414.109	44.452	−0.125	1.110
15	1 460.906	46.797	−0.132	1.020

注:BIC为施瓦兹的贝叶斯判别准则;ΔBIC表示BIC的变化量;—指无数据.

以15个子类为对象,利用凝聚法(agglomerative hierarchical clustering method,AHCM),逐个凝聚子类,直到获得期望的子类数量[18].15个子类凝聚为3类,分别为子类1、2和3,对应BIC值分别为2 243.811、1 958.271和1 566.429,均>1 000.000,这说明最终凝聚成的3个类包含了足够的信息量[12].根据汽车行驶工况将这3个类分别记为:低速工况、中速工况和高速工况.1 743条样本数据中,含有11条离群值,对其进行去噪处理后,利用TSCA方法进行聚类汇总,对应结果列于表4.可知低速工况、中速工况和高速工况的聚类数量分别为873、578和281条.

表4 聚类汇总表

类别	数量/条	分占比/%	总占比/%
低速工况	873	50.1	50.1
中速工况	578	33.2	33.2
高速工况	281	16.1	16.1
离群值	11	0.6	0.6
总数	1 743	100.0	100.0

为了直观地表示TSCA的聚类效果,其质量图如图3所示.粉色区域代表聚类效果较差,黄色区域代表聚类效果中等,绿色区域代表聚类效果较好,蓝色数据条已经超过0.5,落入绿色区域,处于聚类质量图的"good"阶段,说明本文建立的基于PCA和TSCA的PTV评价模型效果较好.

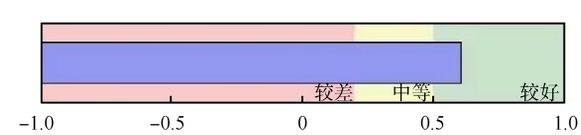
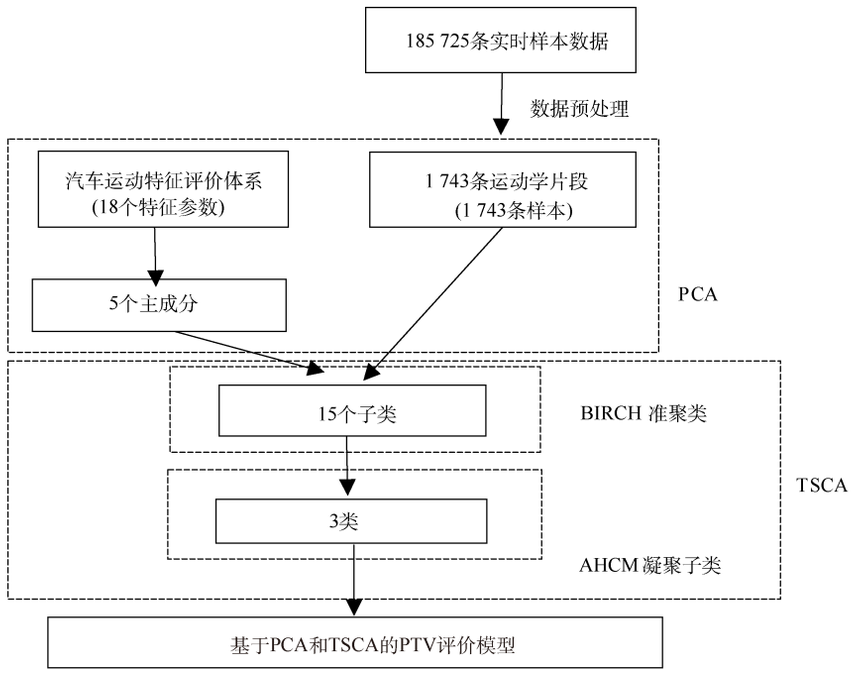


图3 聚类质量

综上,本文构建的PTV评价模型的流程如下:实际中,速度<30 km/h时,为低速工况;速度在30~80 km/h时,为中速工况;速度≥80 km/h时,为高速工况.首先,对该车辆的1 743条样本数据按实际情况划分为低速、中速和高速工况,作为实际结果;其次,基于PCA和TSCA的PTV评价模型,将1 743条样本数据聚类为低速、中速和高速工况,作为聚类结果.将实际结果与聚类结果进行对比分析,结果显示,本文构建的基于PCA和TSCA的PTV评价模型的准确率为98.970%,这一数据足以显示本文所建模型的合理性.



注:TSCA为两阶段聚类分析;PTV为汽车行驶工况;BIRCH为平衡迭代规约和聚类;AHCM为凝聚法.

图4　基于PCA和TSCA的PTV评价模型的流程

3　结束语

本文提出了基于PCA和TSCA的PTV评价模型,利用PCA将18个特征参数综合成5个主成分,且这5个主成分的方差贡献率达82.160%,在此基础上,利用TSCA评估某个运动学片段的工况属于低速工况、中速工况还是高速工况.将聚类结果与实际结果进行对比表明,本文建立的基于PCA和TSCA的PTV评价模型拟合效果良好,可用于汽车行驶工况的研究.

参 考 文 献

[1] 王昊,章桐,宋珂.上海市行驶工况构建与研究[J].内燃机与配件,2018(13):35-38.

[2] 李阿午.太原市轻型车行驶工况构建与排放特性的研究[D].太原:太原理工大学,2018.

[3] 许凌,鲁杨,张蓉,等.汽车怠速熄火时间的调查与分析[J].汽车实用技术,2018,44(11):65-69.

[4] 刘明哲.基于运动学片段和分层聚类的汽车行驶工况构建[J].内蒙古农业大学学报(自然科学版),2021,42(2):73-78.

[5] 李耀华,任田园,邵攀登,等.基于马尔可夫链的西安市城市公交工况构建[J].中国科技论文,2019,14(2):121-128.

[6] 刘宇,黄俊革,卢思同,等.海湾大学城汽车行驶工况与交通安全分析[J].公路与汽运,2021,202:24-28.

[7] 徐婷,崔世超,刘明,等.高原山区乘用车行驶工况构建方法研究[J].公路交通科技,2021,38(2):118-124.

[8] 野晨晨.沈阳市乘用车城市道路行驶工况研究[D].长春:吉林大学,2018.

[9] 汤银才.R语言与统计分析[M].北京:高等教育出版社,2008.

[10] DEMUYNCK J, BOSTEELS D, PAEPE M D, et al. Recommendations for the new WLTP cycle based on an analysis of vehicle emission measurements on NEDC and CADC[J]. Energy Policy,2012,49(1):234-242.

[11] 田宇.太原市轻型车道路行驶工况构建与分析[D].太原:太原理工大学,2018.

[12] 张宏,姚延钢,杨晓勤.城市道路轻型汽车行驶工况构建[J].西南交通大学学报,2019,54(6):1139-1154.

[13] 石琴,仇多洋,吴冰,等.基于粒子群优化支持向量机算法的行驶工况识别及应用[J].中国机械工程,2018,29(15):1875-1883.

[14] 何晓群.多元统计分析[M].北京:人民大学出版社,2019.

[15] 吴喜之.复杂数据统计方法:基于R的应用[M].北京:中国人民大学出版社,2015.

[16] 钟森鸣,黄坚,吴嘉健,等.新能源汽车动态工况EMI测试系统构架设计与关键技术分析[J].中国测试,2017,43(8):76-79.

[17] 姜平,石琴,陈无畏,等.基于小波分析的城市道路行驶工况构建的研究[J].汽车工程,2011,33(1):70-73+51.

(责任编辑:马田田)