基于K-均值聚类分析的城市道路汽车行驶工况构建方法研究*

彭育辉[1,2]　杨辉宝[1]　李孟良[2]　乔学齐[3]

（1.福州大学，福州　350116；2. 中国汽车技术研究中心，天津 300300；3.厦门金龙旅行车有限公司，厦门 361006）

【摘要】以实时采集的乘用车行驶数据为数据源，进行了城市道路汽车行驶工况构建方法的研究。分别运用运动学片段分析法、主成分分析法和K均值聚类分析法对实测数据进行降维和分类，提出以Silhouette函数实现对聚类结果的筛选，以减少人为选择的误差，并根据聚类中心的大小筛选所需运动学片段构建候选工况。在目标代表工况的遴选方面，提出了综合6个特征参数和最大SAFD差异值的评价标准。最后通过试验验证了该行驶工况构建方法的有效性和精确性。

主题词：汽车行驶工况　运动学片段　主成分分析　K-均值聚类分析

中图分类号：U467.1+1　文献标识码：A　文章编号：1000-3703(2017)11-0013-06

Research on the Construction Method of Driving Cycle for the City Car Based on K-Means Cluster Analysis

Peng Yuhui[1,2], Yang Huibao[1], Li Mengliang[2], Qiao Xueqi[3]

(1. Fuzhou University, Fuzhou, 350116; 2. China Automotive Technology and Research Center, Tianjin 300300; 3. Xiamen Golden Dragon Wagon Bus Co., Ltd, 361006)

【Abstract】The construction method of city road driving cycle was studied with the real-time collected passenger vehicle driving data as data source. The kinematics fragment analysis, principal component analysis and K-means clustering analysis were applied separately for dimensionality reduction of the measured parameters and classification, then Silhouette Function was proposed to screen the clustering results, to reduce artificial selection error, and the required kinematic fragment candidate cycle was selected according to size of the clustering center. In term of choice of typical driving cycle, an evaluation criterion that integrated six characteristic parameters and the maximum SAFD difference value was proposed. Finally test proved validity and accuracy of the construction method of the driving cycle.

Key words: Driving cycle, Kinematic fragment, Principal component analysis, K-means clustering analysis

1　前言

汽车行驶工况(Driving Cycle)又称车辆测试循环，是描述典型车辆行驶的速度-时间曲线，用于确定车辆污染物排放量、燃油消耗量、新车型技术开发和评估以及测定交通控制的风险等，是汽车工业一项共性的核心技术[1]。目前，中国以及大多数国家是参照欧洲的汽车行驶工况[2]制定并实施本国的排放法规。中国城市汽车行驶工况采用的是ECE(欧洲城市工况)工况，其加减速过程简单，与实际车辆行驶状态差距较大，已不能满足车辆开发测试的需要。针对此问题，我国学者已经先后建立了符合本地区的汽车工况[1,3-5]，而在构建行驶工况时，主要采用主成分分析法与聚类分析法相结合的研究方法，其中K-均值聚类(K-Means Cluster)分析法在构建城市汽车行驶工况中应用较多，但在对K-均值聚类分析结果筛选时，普遍通过对比各类对应的汽车行驶状态时间分布比例，人为判断聚类结果是否合理，对聚类结果的合理性存在影响。

为此，本文以实时采集的乘用车行驶数据为研究数据源，基于K-均值聚类方法对汽车行驶数据进行分析，

*基金项目：工业和信息化部"中国新能源汽车产品检测工况研究开发"项目资助(FZU201600201603)。

提出一种以Silhouette函数筛选聚类结果,并根据聚类结果构建汽车行驶工况的方法,减少了人为选择K-均值聚类结果时存在的误差。

2 数据采集与前期处理

2.1 数据采集

数据采集在厦门市城市道路上进行,为使采集的数据真实反映城市道路汽车的行驶特点,利用5辆私家车按照各自目的紧跟行驶车流的方式采集数据,时间为早上7:00至晚上20:00,共15天。车辆行驶路线如图1所示,涵盖海沧区的霞飞路、马青路、阳光路、海沧大桥以及湖里区湖里大道和仙岳路,体现较为典型的厦门岛内(湖里区)与岛外(海沧区)的道路交通情况。

图1 试验车辆行驶路线

数据采集车载终端与车辆的OBD接口相连接,如图2所示。在汽车行驶过程中,以1 Hz的频率采集包括车速、发动机转速、扭矩百分比、瞬时油耗、进气歧管温度、发动机负荷、进气歧管压力等参数。通过信息化数据平台对车辆进行实时监测,确保采集数据的连续性和正确性。

(a)数据采集车载终端 (b)车载终端安装效果

图2 数据采集车载终端及安装效果

2.2 数据前期处理

采集数据中存在汽车振动导致GPS漂移引起的部分加速度大于2.5 $m\cdot s^{-2}$的数据点,为减少后续数据分析的误差,前期处理过程中需要将这些数据异常点剔出;为保证工况合成结果具有代表性,选用10^5 s左右的数据点进行汽车行驶工况的构建分析。经前处理后数据源的SAFD(Speed Acceleration Frequency Distribution)分布如图3所示,部分速度-时间曲线和相对应的加速度-时间曲线见图4。

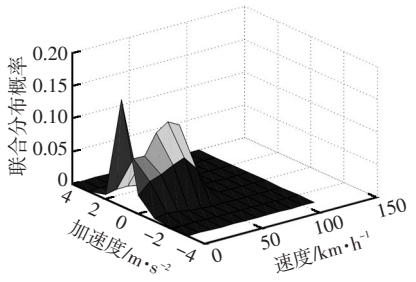
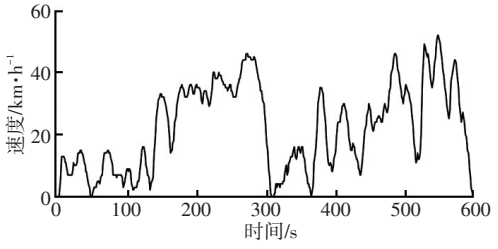
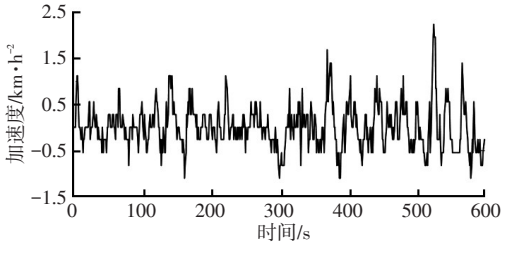


图3 数据源的SAFD分布图



(a)速度-时间曲线



(b)加速度-时间曲线

图4 部分速度-时间曲线和相对应的加速度-时间曲线

3 数据分析

3.1 运动学片段分析

汽车在行驶过程中,由于受到道路交通条件的限制,存在多次怠速、加速、巡航和减速的状态。为此将从1个怠速开始到下1个怠速开始定义为1个运动学片段。

汽车行驶状态定义[6]为:怠速为发动机正在工作但车速为0的连续过程;加速为汽车加速度大于0.1 m/s^2的连续过程;巡航为汽车加速度的绝对值小于0.1 m/s^2但车速不为0的连续过程;减速为汽车加速度小于−0.1 m/s^2的连续过程。根据以上定义,可将汽车行驶过程看作如图5所示的多个运动学片段的组合,通过对每个运动学片段的行驶特征参数进行分析,得出该车辆的行驶特征。

在汽车行驶过程中,需要选用一些基本的特征参数反映每个运动学片段的行驶特征。本文选用如表1所列的15个特征参数描述运动学片段。将数据源分割成297个运动学片段并计算各运动学片段的15个特征参数,部分分析结果见表2。

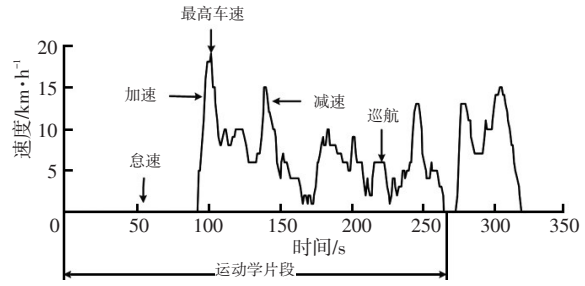


图5　运动学片段

表1　用于描述运动学片段的15个特征参数

特征参数	定义
T/s	片段持续时间
S/km	行驶距离
V_a/km·h^{-1}	平均速度
V_x/km·h^{-1}	平均行驶速度
a_{max}/m·s^{-2}	最大加速度
a_{min}/m·s^{-2}	最大减速度
T_i	怠速时间比例
T_a	加速时间比例
T_d	减速时间比例
T_C	巡航时间比例
V_{max}/km·h^{-1}	最高车速
V_{std}/km·h^{-1}	速度标准偏差
a_a/m·s^{-2}	平均加速度
a_{std}/m·s^{-2}	加速度标准偏差
a_d/m·s^{-2}	平均减速度

表2　运动学片段特征参数值

序号	T/s	S/km	v_a/km·h^{-1}	···	T_i/s	T_a/s	T_C/s	v_{std}	a_{std}
1	51	0.121	9.456	···	5	10	18	4.533	0.372
2	267	0.327	6.811	···	94	41	75	4.426	0.293
3	101	0.001	0.562	···	93	4	0	0.259	0.103
···	···	···	···	···	···	···	···	···	···
151	21	0.004	0.191	···	11	4	2	1.077	0.176
152	57	0.054	3.438	···	6	19	14	2.096	0.389
153	204	1.708	30.15	···	59	40	23	24.159	0.605
···	···	···	···	···	···	···	···	···	···
295	365	4.971	49.03	···	31	90	124	21.187	0.368
296	312	2.36	27.231	···	3	113	71	13.649	0.495
297	121	0.596	17.752	···	5	45	27	7.224	0.561

3.2　特征参数标准化

基于每个运动学片段中汽车行驶速度的变化可计算相应运动学片段的15个特征参数。由于各特征参数的量纲不统一会引起各变量取值的分散程度差异较大,导致后续进行降维分析和聚类分析过程中优先照顾方差较大的变量,使处理结果的稳定性差,为此,在对特征参数进行降维处理前进行特征参数标准化:

$$A_{m\times n}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix} \tag{1}$$

式中,a_{ij}(i=1,2,…;j=1,2,…,n)是第i个运动学片段的第j个参数。

对矩阵(1)进行标准化得到标准化矩阵X:

$$X_{mn}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & & x_{mn} \end{bmatrix} \tag{2}$$

$$x_{ij}=a_{ij}-\bar{a}_j\big/s_j^2 \tag{3}$$

$$\bar{a}_j=\frac{1}{m}\sum_{i=1}^{m}a_{ij} \tag{4}$$

$$s_j^2=\frac{1}{m-1}\sum_{i=1}^{m}\left(a_{ij}-\bar{a}_j\right) \tag{5}$$

3.3　特征参数降维

利用特征参数对运动学片段进行描述的过程中,为保证运动学片段的特征信息不丢失,选取的15个特征参数必然存在所反映的信息重叠,为此采用主成分分析法对特征参数进行降维处理。

由矩阵(2)计算协方差矩阵Σ:

$$\sum=\begin{bmatrix} s_j^2 & \mathrm{cov}(1,2) & & \mathrm{cov}(1,n) \\ \mathrm{cov}(1,2) & s_j^2 & & \mathrm{cov}(2,n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}(n,1) & \mathrm{cov}(n,2) & & s_n^2 \end{bmatrix} \tag{6}$$

$$s_x^2=\mathrm{cov}(x,x) \tag{7}$$

$$\mathrm{cov}(x,y)=\mathrm{cov}(y,x)=\frac{1}{m-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right) \tag{8}$$

由矩阵(1)得相关矩阵R:

$$R=\frac{1}{m-1}\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & & r_{mn} \end{bmatrix} \tag{9}$$

$$r_{xy}=\frac{\mathrm{cov}(x,y)}{s_x s_y} \tag{10}$$

用λ_i表示矩阵R的特征参数$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n\geqslant0$,计算出相应的正交化特征参数向量为:

$$\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}=\begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & & e_{nn} \end{bmatrix} \tag{11}$$

设 $\lambda_1 \Big/ \sum_{j=1}^{n}\lambda_j$ 为第 i 个主成分的贡献率，贡献率越大所表达的信息越多，前 r 个特征参数的贡献率之和为 $\sum_{r=1}^{l}\lambda_r \Big/ \sum_{j=1}^{n}\lambda_j$；理论上 $\sum_{r=1}^{l}\lambda_r \Big/ \sum_{j=1}^{n}\lambda_j \geqslant 80\%$，前 r 个特征参数就能够满足工程需要，即每个运动学片段有 r 个主成分。

通过对297个运动学片段的特征参数进行主成分分析，得到如表3所列的主成分贡献率及累积贡献率。

表3　主成分贡献率及累积贡献率

成分	特征值	贡献率/%	累积贡献率/%
1	8.855	59.030	59.030
2	2.975	19.833	78.863
3	1.131	7.541	86.404
4	0.870	5.800	92.205
5	0.419	2.792	94.996
6	0.277	1.844	96.841
7	0.170	1.134	97.974
8	0.106	0.708	98.682
9	0.083	0.554	99.236
10	0.059	0.396	99.632
11	0.029	0.193	99.825
12	0.013	0.084	99.908
13	0.009	0.058	99.967
14	0.005	0.033	100.000
15	-1.010×10^{-13}	-1.068×10^{-13}	100.000

理论上，进行主成分分析时，选取累积贡献率大于80%的主成分进行后续的聚类分析，虽然会有一定的信息损失，但对最终的结果影响较小[7]。由表3可知，前8个主成分累积贡献率达到98.682%，已经包含15个特征参数的所有信息，其中前3个主成分的累积贡献率之和已经达到86.404%，并且这3个主成分的特征值都大于1，因此，用前3个主成分就能反映出15个特征参数的大部分信息。

某个参数在某个主成分上的载荷系数的绝对值越大，则说明该参数与这个主成分的相关系数越大。表4为主成分载荷矩阵，由表4可知，第1主成分M1主要反映减速时间比例、行驶距离、片段持续时间、加速时间比例、巡航时间比例、平均速度；第2主成分M2主要反映加速度标准偏差、平均加速度、平均减速度；第3主成分M3主要反映怠速时间比例。根据主成分分析结果，选取第1、第2和第3主成分进一步分析。

表4　主成分载荷矩阵

特征参数	M1	M2	M3
减速时间比例 T_d	0.963	0.133	−0.068
行驶距离 S	0.962	0.069	−0.019
片段持续时间 T	0.957	0.036	0.168
加速时间比例 T_a	0.950	0.124	−0.107
巡航时间比例 T_C	0.923	−0.069	−0.108
平均速度 V_a	0.853	0.414	−0.090
平均行驶速度 V_x	0.797	0.497	0.043
最高车速 $V_{\max}$	0.793	0.544	0.005
速度标准偏差 V_{std}	0.675	0.613	0.046
加速度标准偏差 a_{std}	0.186	0.896	−0.332
平均加速度 a_a	0.042	0.883	0.077
平均减速度 a_d	0.182	−0.807	−0.049
最大加速度 $a_{\max}$	0.488	0.771	−0.058
最大减速度 $a_{\min}$	−0.333	−0.649	0.118
怠速时间比例 T_i	−0.017	−0.090	0.987

3.4　K-均值聚类分析及结果选择

K-均值聚类分析法是根据对所研究问题的了解程度确定分类数，这样即可在每一类中选择1个有代表性的样品作为聚点（初始聚点），其它样品根据与该点的“亲疏程度”进行分类，这种“亲疏程度”可以用欧氏距离（Euclidian Distance）、欧氏平方距离（Squared Euclidian Distance）、明可夫斯基距离（Minkowski Distance）、切比雪夫距离（Chebyshev Distance）和布洛克距离（Block Distance）5种片段差异程度表示。

在K-均值聚类分析过程中，参考相关文献[3,4]采用欧氏平方距离表示各类中的“亲疏程度”。其定义如下：设 x_{ik} 表示第 i 个片段的第 k 个变量，每个片段定义了 p 个变量，d_{ik} 表示不同片段之间的距离，则欧氏平方距离为：

$$d_{ij}=\sum_{k=1}^{P}\left(x_{ik}-x_{jk}\right)^2 \tag{12}$$

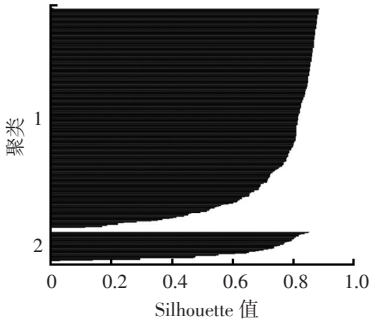
基于特征参数的第1、第2和第3主成分，运用K-均值聚类分析方法对所有的运动学片段进行分析，得到分为两类、3类和4类的聚类分析结果。根据K-均值聚类分析结果，运用Silhouette函数绘制轮廓图，从轮廓图上判断每个运动学片段的分类是否合理。其中Silhouette函数的定义[8]如下：

$$L(i)=\frac{\min(b)-a}{\max\left[a,\min(\mathrm{b})\right]} \tag{13}$$

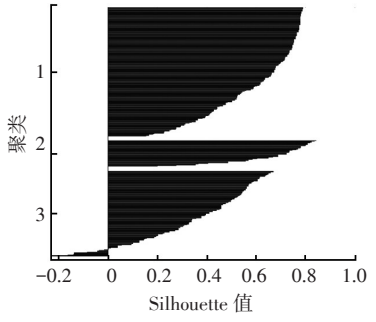
式中，a 为第 i 个运动学片段与同类运动学片段之间的平均距离；b 为一个向量，其元素是第 i 个运动学片段与不同类的各运动学片段间的平均距离。

式(13)中,$L(i)$的取值范围为[-1,1],$L(i)$值越大,说明第i个运动学片段分类越合理;当$L(i)$<0时,说明第i个运动学片段的分类不合理,应该还有比目前更好的分类。

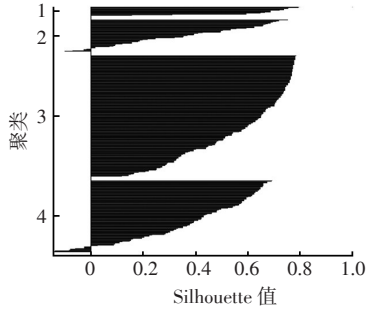
K-均值聚类各分类结果的Silhouette函数值的轮廓见图6。由图6a可看出,当分两类时,各类的Silhouette函数值均大于0,说明分类时各类型已经被很好地区分;由图6b可看出,分3类时,第3种类型的Silhouette函数值出现少量负值;由图6c可看出,分4类时,第1种类型和第4种类型的Silhouette函数值出现少量负值的情况,说明聚类分析分为3类和4类时存在未被很好区分的运动学片段。根据分析结果,选取分为两类的结果作为K-均值聚类分析的最终结果。



(a)分两类



(b)分3类



(c)分4类

图6　不同分类时Silhouette函数值的轮廓

根据Silhouette函数值的选择结果,对分为两类时各类型中行驶状态所占的时间比例进行统计,结果如图7所示。由图7可看出,类型1中怠速时间比例所占比重最大,达到37%,其它行驶状态的时间比例在20%左右,反映了汽车在城市拥挤道路的行驶过程;类型2中怠速时间比例所占比重不到5%,加减速时间比例和巡航时间比例分布较均匀,反映了汽车在较为顺畅道路行驶的过程。

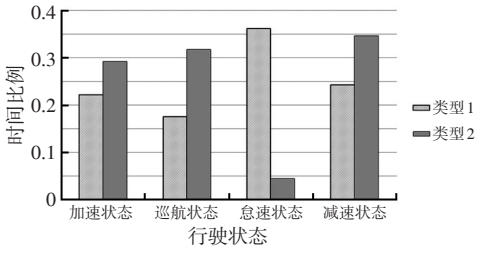


图7　各类型行驶状态时间比例

根据厦门市实际道路交通情况可知,厦门岛内湖里区交通状况复杂,道路拥挤,因此,第1种类型主要反映试验车辆在厦门岛内的行驶状态;厦门岛外海沧区道路较为顺畅,因此,第2种类型主要反映试验车辆在厦门海沧区的行驶过程。

4　行驶工况的构建

4.1　工况合成方法

基于K-均值聚类分析结果,已知各分类中所包含的运动学片段与其聚类中心的大小,由式(12)可知,当d_{ij}越小时,说明该运动学片段越能反映该类中的行驶特点。因此,本文的工况合成方法为:在每一种类型中,根据聚类中心的大小,由小到大分别筛选10个候选运动学片段;当某个类型中所包含的运动学片段不足10个时,将整个类别中的运动学片段作为候选的运动学片段;以各类的候选运动学片段为基础,从中随机筛选运动学片段组合成≥1 200 s的候选汽车行驶工况。

4.2　代表行驶工况的确定

为了从大量的候选工况中选取具有代表性的汽车行驶工况,国外学者采用最小性能值PV(Performance Value)作为筛选代表性工况的标准[9]。本文提出将加速时间比例、巡航时间比例、怠速时间比例、平均速度、平均加速度、加速度标准偏差以及最大SAFD差异值等7项作为主要筛选参数。通过仿真模拟计算,确定前6项与数据源的相对误差控制在±5%范围内,同时最大SAFD差异值的误差控制在±5%作为代表工况的筛选标准。

4.3　工况合成结果分析

根据前述工况合成方法和筛选标准,运用MATLAB软件合成如图8所示的1 514 s的汽车代表行驶工况,其SAFD分布如图9所示。

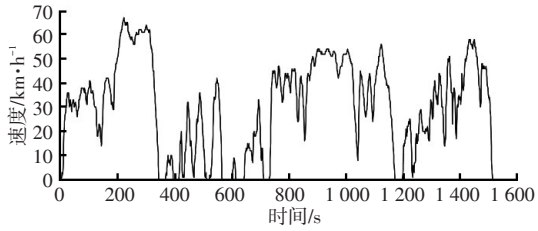


图8　代表行驶工况的速度-时间曲线

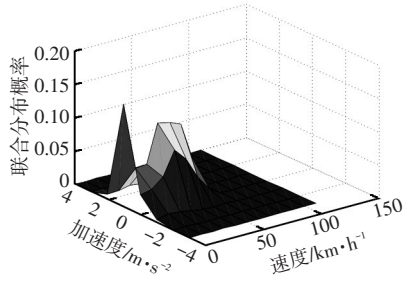


图9　代表行驶工况的SAFD分布图

将构建的代表行驶工况与实际采集数据源的特征参数进行对比分析,结果如表5所列,表5中主要特征参数的相对误差都维持在±5%的范围内。进一步分析代表行驶工况与数据源的SAFD差异值,如图10所示,由图10可看出,代表工况和数据源对应点的SAFD差异值在±2%的范围内。通过以上分析表明,所构建的代表工况可以反映试验车辆的整体行驶特征。

表5　特征参数差异率对比

特征参数	试验数据	代表行驶工况	差异率/%
加速时间比例T_a/%	28.8	29.7	3.1
巡航时间比例T_c/%	27.8	26.8	−3.6
怠速时间比例T_i/%	10.4	10.5	0.9
平均速度V_a/km·h^{-1}	30.103	30.663	1.9
平均加速度a_a/m·s^{-2}	0.530	0.543	2.5
加速度标准偏差a_{std}/m·s^{-2}	0.372	0.388	4.3

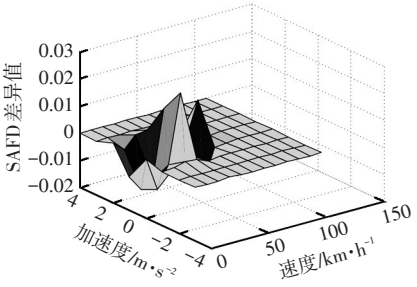


图10　数据源与代表工况的SAFD差异值

依据工况构建过程中的筛选准则,将合成的厦门市乘用车行驶工况与国外主要标准循环工况进行对比,结果如表6所列。由表6可知,合成的厦门市乘用车行驶工况中,加速时间比例和巡航时间比例与美国城市工况(UDDS)较为接近,平均速度和平均加速度与UDDS工况较接近,而平均加速度与日本汽车测试工况(J10-15)和欧盟的NEDC循环工况最为接近。但厦门工况中怠速时间比例低于美国城市工况,与欧盟提出的轻型车测试循环工况(WLTC)相比,各参数都存在较大差距。

表6　厦门市合成行驶工况与国外主要标准工况的比较

工况	T/s	T_a	T_i	T_C	V_a/km·h^{-1}	a_a/m·s^{-2}
厦门工况	1 514	0.297	0.105	0.268	30.66	0.543
UDDS	1 369	0.294	0.178	0.273	31.51	0.500
WLTC	1 800	0.335	0.242	0.142	22.11	0.392
J10—15	660	0.261	0.326	0.195	22.68	0.540
NEDC	1 184	0.229	0.239	0.374	33.21	0.540

5　结束语

基于厦门市乘用车典型交通线路实时采集的行驶数据,分别运用运动学片段分析法、主成分分析法和K-均值聚类分析法对实测数据降维和分类。提出以Silhouette函数实现对聚类结果的筛选,减少人为选择的误差,并根据聚类中心的大小筛选所需运动学片段构建候选工况。在目标代表工况的遴选标准方面,提出综合控制6个特征参数和SAFD最大差异值误差的方法,并通过实例验证了所提工况构建方法的有效性和精确性。通过将所构建的厦门工况与国外常用标准循环工况进行对比表明,所构建的厦门市工况与UDDS工况较为接近,但是怠速时间比例低于UDDS工况。

参　考　文　献

1　石琴,郑与波,姜平. 基于运动学片段的城市道路行驶工况的研究. 汽车工程,2011,33(3):256~261.
2　张建伟,李孟良,艾国和,等. 车辆行驶工况与特征的研究. 汽车工程,2005,27(2):220~224,245.
3　蔡锷,李阳阳,李春明,等. 基于K-均值聚类算法的西安市汽车行驶工况合成技术研究. 汽车技术,2015(8):33~36.
4　胡志远,秦艳,谭丕强,等. 基于大样本的上海市乘用车行驶工况构建. 同济大学学报(自然科学版),2015,43(10):1523~1527.
5　李孟良,张建伟,张富,等. 中国城市乘用车实际行驶工况的研究. 汽车工程,2006,28(6):554~557,529.
6　杨延相,蔡晓林,杜青,等. 天津市道路汽车行驶工况的研究. 汽车工程,2002,24(3):200~204.
7　范金城,梅长林. 数据分析. 北京:科学出版社,2010.
8　张德丰. MATLAB概率与数理统计分析. 北京:机械工业出版社,2010.
9　Jie Lin, Niemeier D A. Estimating Regional Air Quality VehicleEmission Inventories: Constructing Robust Driving Cycles. SAE.2002.

(责任编辑　文　楫)

修改稿收到日期为2017年3月5日。