doi: 10.12068/j.issn.1005-3026.2019.01.015

基于大数据和马尔科夫链的行驶工况构建

曹 骞, 李 君, 刘 宇, 曲大为
(吉林大学 汽车仿真与控制国家重点实验室, 吉林 长春 130022)

摘 要: 为提高代表行驶工况的准确性,对行驶工况构建算法进行了研究. 在沈阳市选取10辆乘用车并采用自主驾驶方式收集行驶数据,组建了大样本数据库. 首先根据傅里叶变换对原始数据进行了降噪滤波,然后采用改进的Kneser-Ney平滑方法计算状态转移概率矩阵,提出了基于马尔科夫链的行驶工况构建算法,最后开发了沈阳市乘用车代表行驶工况,并将其与数据库总体特征进行对比. 结果表明,构建工况与数据库总体的平均偏差为2.46%,所有特征参数偏差均在10%以内,验证了算法的有效性.

关 键 词: 行驶工况;乘用车;大样本;马尔科夫链;算法
中图分类号: U 491.2 **文献标志码**: A **文章编号**: 1005-3026(2019)01-0077-05

Construction of Driving Cycle Based on Big Data and Markov Chain

CAO Qian, LI Jun, LIU Yu, QU Da-wei
(State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun 130022, China. Corresponding author: LIU Yu, E-mail: liuyu1981@jlu.edu.cn)

Abstract: In order to improve the accuracy of typical driving cycle, the constructing algorithm for typical driving cycle was studied. 10 passenger cars in Shenyang City were selected to collect the driving data by autonomous driving, and the big sample database was established. Firstly, the raw data was filtered for noise reduction by using Fourier transform method. Secondly, the modified Kneser-Ney smoothing method was applied to compute the state transfer probability matrix, and the driving cycle constructing algorithm based on Markov chain was proposed. Finally, the typical driving cycle for passenger cars in Shenyang City was constructed and compared with the overall characteristics of the database. The results showed that the average deviation between the constructed cycle and the database population is 2.46%, the deviation values of all the characteristic parameters are within 10%, and the validness of the proposed algorithm is thus verified.

Key words: driving cycle; passenger car; big sample; Markov chain; algorithm

代表行驶工况是反映某一地区车辆行驶特征的速度-时间曲线,它可作为测试标准用于对污染物排放、燃油经济性等指标进行评价,相关的检测结果对于车辆研发具有重要的参考价值. 研究指出,地理特征、驾驶风格以及道路等级等因素会显著影响代表行驶工况的准确性[1]. 为了解决传统行驶工况代表性不足的问题,有研究者以城市行驶特征为基础构建了独立的代表工况[2-4],通过与传统行驶工况对比,证实独立开发的城市工况能够更充分地反映当地的行驶特征,进而提高检测结果的准确性.

行驶工况构建的常用方法有短行程法[5]和马尔科夫链法[6-7]. 前者往往先对短行程样本分类,然后从各类中选择匹配程度最高的样本组合成完整工况,在这一过程中,可选用的聚类方法有K均值聚类法[8]、判别分析法[9]等. 不过K均值聚类法对于分布复杂的样本往往陷于局部最优解;判别分析法则需要事先建立模型样本,对于大

收稿日期: 2018-03-29
基金项目: 中国新能源汽车产品检测工况研究和开发项目(工信部 工装函[2015]615号)
作者简介: 曹 骞(1986-),男,山东济南人,吉林大学博士研究生; 李君(1961-),男,吉林长春人,吉林大学教授,博士生导师.

样本的情况,聚类数量往往并不确定,因而难以选择合适的模型样本,从而导致聚类效果不理想. 另外,短行程法受限于单个样本的长度,难以精确控制代表工况的时长. 马尔科夫链法不依赖短行程样本,可随机生成指定时长的代表工况,同时避免因样本聚类误差带来的不利影响,因而构建工况能够获得较高的精度. 马尔科夫链法的关键是建立能够反映实际行驶规律的状态转移概率矩阵,如果数据量较小或者状态概率运算不准确,那么得到的状态转移概率矩阵就会出现较大误差,从而影响最终的工况构建效果.

综上所述,为了充分反映实际行驶特征,提高代表行驶工况的准确性,本文首先采集了沈阳市乘用车行驶数据,建立了大样本数据库. 然后将行驶数据转换为行驶状态,采用改进的 Kneser – Ney 平滑方法计算行驶状态转移概率矩阵,提出了基于马尔科夫链的行驶工况构建算法,基于 MATLAB 平台构建了沈阳市乘用车代表行驶工况. 最后通过与数据库的统计结果进行对比,验证了算法的有效性.

1 行驶数据的采集与处理

1.1 行驶数据采集

为了兼顾不同驾驶习惯以及道路等级对行驶特征的影响,本文在沈阳市选择了 10 位驾驶员的轻型乘用车作为行驶数据的收集对象,采用自主驾驶的方式收集行驶数据,即事先不指定具体路线和行驶时段,由驾驶员根据日常需要安排行驶行为. 由此可以实现对不同行驶区域和行驶时段的充分覆盖.

本文利用车载终端采集行驶数据,数据采集频率设定为 1 Hz. 数据采集终端实时记录车辆的方位、车速、时间等信息,将其保存在内部存储器中,然后通过无线网络发送至管理平台数据库以便对数据调取和分析. 为了使统计结果尽可能充分反映实际行驶特征,体现大数据的特点,本文在沈阳市进行了为期一年的行驶数据采集,最终得到 11 002 604 条行驶数据,据此组建了大样本数据库并用于构建代表工况.

1.2 行驶数据滤波处理

由于车辆振动以及数据采集设备本身噪声的影响,行驶数据往往会出现异常波动,如果不对此进行处理,容易降低特征参数统计结果的准确性. 为此本文首先采用傅里叶变换对行驶数据进行滤波处理,纠正行驶数据偏差,然后再将其用于工况构建与分析.

傅里叶变换由式(1)和式(2)两部分组成,其中式(1)是傅里叶正变换表达式,即根据时域信号 $f(x)$ 计算得到频域信号 $F(u)$;式(2)则是傅里叶逆变换表达式,即通过频域信号 $F(u)$ 还原为时域信号 $f(x)$.

$$F(u) = \sum_{x=0}^{N-1} f(x)\, e^{-j2\pi ux/N}, \tag{1}$$

$$f(x) = \frac{1}{N}\sum_{u=0}^{N-1} F(u)\, e^{j2\pi ux/N}. \tag{2}$$

式中:x 和 u 分别为离散时域变量和频域变量;N 为离散信号的数量.

因为噪声通常可以视为高频信号,可以首先通过傅里叶正变换将时域信号转换到频域;然后把高频信号的值调整为 0 或者某一较小的值,而低频信号的值保持不变;最后再由傅里叶逆变换将频域信号还原到时域,由此便可以实现信号的平滑处理,达到降噪的目的.

图 1 是针对某一行驶片段的滤波前后速度曲线对比结果. 从图中可以看到,在滤波之前,原始速度曲线存在许多“尖峰”突变,由这些突变计算得到的加速度值往往会出现失真,而在滤波之后,速度曲线变得更加平滑,“尖峰”明显减少,这样可以保证计算结果维持在合理范围,由此可见,采用傅里叶变换滤波的效果较为理想.

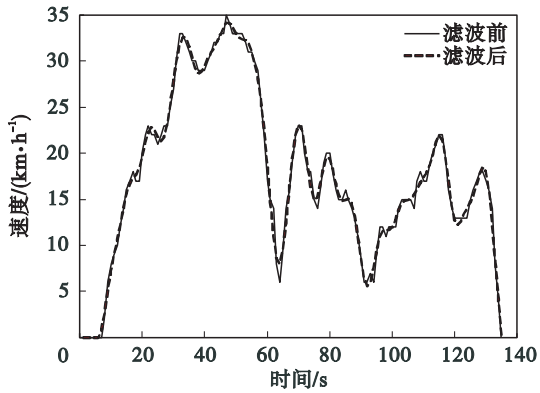

图 1 速度数据滤波结果

Fig. 1 Result of speed data filtering

2 状态转移概率矩阵

基于马尔科夫链的行驶工况构建方法是将车速变化视为一个随机过程,如果将互不重叠的速度区间视为不同的行驶状态,那么每一个速度值都可以依据所属区间转换为相应的行驶状态,并且前后两个相邻的行驶状态存在因果关系,也就是说当前行驶状态会影响下一行驶状态. 状态转

移概率就是对这一影响的反映，它表征的是从某一状态转移到另一状态的概率.

将状态转移概率进行有序排列，便可以建立如式(3)所示的状态转移概率矩阵 $\boldsymbol{P}$. 其中，$P_{ab}(a,b=1,2,\cdots,M)$ 表示从当前状态 a 转移到下一状态 b 的概率，ab 就是两两组合的二元状态，M 为行驶状态的总数. 利用该矩阵可以控制行驶状态的随机变化，并由此生成一组行驶状态序列，根据车速与速度区间之间的对应关系，可以将行驶状态序列转换为速度序列，该速度序列就是最终得到的行驶工况.

$$\boldsymbol{P}=(P_{ab})_{M\times M}=\begin{bmatrix}P_{11} & P_{12} & \cdots & P_{1M}\\ P_{21} & P_{22} & \cdots & P_{2M}\\ \vdots & \vdots & & \vdots\\ P_{M1} & P_{M2} & \cdots & P_{MM}\end{bmatrix}. \tag{3}$$

本文将按速度区间定义行驶状态，规定每一个行驶状态对应的速度区间长度为 10 km/h，从怠速开始将车速划分到对应的行驶状态，即将在区间[0 10)内的车速值作为状态 1，位于区间[10 20)内的车速值作为状态 2，以此类推，直至状态区间涵盖所有的速度范围. 将车速数据转换为行驶状态之后，再统计两两组合的行驶状态之间的转移概率，最终得到状态转移概率矩阵.

状态转移概率通常采用极大似然估计法求得，即将两两组合的二元状态的样本数与当前状态的样本数之比作为状态转移概率. 如果状态的样本数较多，统计结果接近真实情况，那么采用极大似然法计算状态转移概率的精确度较高；如果状态的样本数较少尤其为 0 时，统计结果将会偏离现实情况，那么利用该方法估算的状态转移概率就不准确，如果继续用于工况构建则会影响工况精度. 为此，本文采用改进的 Kneser－Ney 平滑方法[10-11]估计状态转移概率，该方法对状态转移概率进行平滑处理，避免出现概率值为 0 的情况，具有较高的可信度. 该方法的具体计算如式(4)～式(9)所示.

$$P_{ab}=\frac{c(ab)-D\langle c(ab)\rangle}{\sum_b c(ab)}+\gamma(a)p(b). \tag{4}$$

$$D\langle c(ab)\rangle=\begin{cases}0, & c(ab)=0;\\ D_1, & c(ab)=1;\\ D_2, & c(ab)=2;\\ D_{3+}, & c(ab)\geqslant 3.\end{cases} \tag{5}$$

其中：

$$\left.\begin{aligned}D_1&=1-2Y\frac{n_2}{n_1},\\ D_2&=2-3Y\frac{n_3}{n_2},\\ D_{3+}&=3-4Y\frac{n_4}{n_3},\end{aligned}\right\} \tag{6}$$

$$Y=\frac{n_1}{n_1+2n_2}; \tag{7}$$

$$\gamma(a)=\frac{D_1N_1(a\cdot)+D_2N_2(a\cdot)+D_{3+}N_{3+}(a\cdot)}{\sum_b c(ab)}; \tag{8}$$

$$p(b)=\frac{N_{1+}(\cdot b)}{N_{1+}(\cdot\cdot)}. \tag{9}$$

式中：P_{ab}表示相邻两个状态 ab 的转移概率，a 表示当前状态，b 表示下一状态，其含义与式(3)中一致；$c(ab)$表示状态组合 ab 的样本个数；参数 D 是根据状态组合 ab 的样本个数计算的折扣系数；n_1，n_2，n_3 和 n_4 分别是样本数为 1，2，3 和 4 时的二元状态 ab 的组合个数.

系数 γ 与当前状态 a 有关，其中，$N_1(a\cdot)$表示与当前状态 a 组合的下一状态的个数，并且这些二元状态组合的样本数都是 1；$N_2(a\cdot)$和 $N_{3+}(a\cdot)$的含义与 $N_1(a\cdot)$类似，但前者二元状态组合的样本数为 2，而后者的样本数必须不小于 3.

参数 $p(b)$为下一状态 b 的估计概率，其中，$N_{1+}(\cdot b)$表示与下一状态 b 组合的当前状态的个数，并且二元状态组合的样本数不小于 1；$N_{1+}(\cdot\cdot)$表示样本数大于等于 1 的所有二元状态组合的个数.

结合行驶状态的定义，本文首先将车速数据集转换为行驶状态集，然后根据式(4)～式(9)计算状态转移概率，最终得到状态转移概率矩阵 $\boldsymbol{P}$，根据 $\boldsymbol{P}$ 绘制出如图 2 所示的状态转移概率直方图. 观察图 2 可以发现，在对角线上的概率值最大，说明车辆行驶倾向于维持稳定车速，在对角线两侧的概率值显著下降，说明车辆在不同行驶状态之间的转换频率较低. 因为车辆在行驶过程中，车速变化一般较为平缓，很少发生大幅度突变，所以图 2 所示的状态转移概率矩阵符合车速的一般规律，计算结果是合理的.

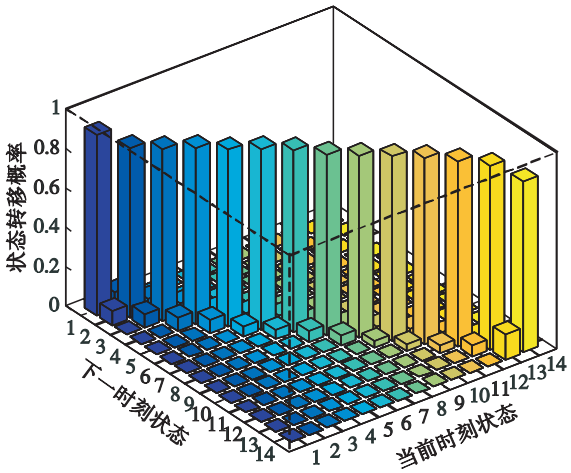

图 2 状态转移概率矩阵直方图

Fig. 2 Histogram of state transfer probability matrix

3 构建行驶工况

结合状态转移概率矩阵的计算,本文利用 MATLAB 软件按图 3 所示流程编制程序并构建代表行驶工况,具体实施步骤如下:

1) 首先将当前状态 a 设置为 1,在(0 1)区间内生成服从均匀分布随机数 r,按式(10)确定下一行驶状态 b. 如此再从下一状态开始重复上述计算过程,得到一组行驶状态序列,该序列数须满足指定时长并且最后一个状态再次返回到初始状态 1.

$$b = \underset{k\in[1,M]}{\operatorname{argmin}}\left(k \mid \sum_{g=1}^{k} p_{ag} \geqslant r\right). \tag{10}$$

式中:p_{ag} 表示从状态 a 转移到状态 g 的概率;k 表示下一状态的序号;M 是状态总数.

2) 将行驶状态序列转换为候选工况,计算式为

$$v_s = [(s-1)+r]\cdot \Delta d. \tag{11}$$

式中:v_s 表示与状态 s 对应的速度值;r 是在[0 1)区间内服从均匀分布的随机数;Δd 是状态对应的速度区间长度.

3) 计算候选工况与样本数据库的特征参数平均绝对偏差,用于计算的特征参数如表 1 所示. 如果偏差值在 5% 以内,则视为合格工况,将其作为代表工况输出,否则返回步骤 1)再次构建候选工况. 计算平均绝对偏差的公式为

$$w = \frac{\sum_{m=1}^{n} | C_m - G_m |}{n}. \tag{12}$$

其中:n 为特征参数个数;C_m 和 G_m 分别为候选工况和样本数据库总体的第 m 个特征参数值;w 是计算得到的平均绝对偏差值.

表 1 选择的特征参数

Table 1 Selected characteristic parameters

特征参数	符号	单位
平均行驶速度	v_{sp}	$km\cdot h^{-1}$
速度标准差	v_{sd}	$km\cdot h^{-1}$
加速度标准差	a_{sd}	$m\cdot s^{-2}$
加速段平均加速度	a_{ac}	$m\cdot s^{-2}$
减速段平均减速度	a_{de}	$m\cdot s^{-2}$
加速比例	P_a	%
减速比例	P_d	%
匀速比例	P_c	%
怠速比例	P_i	%

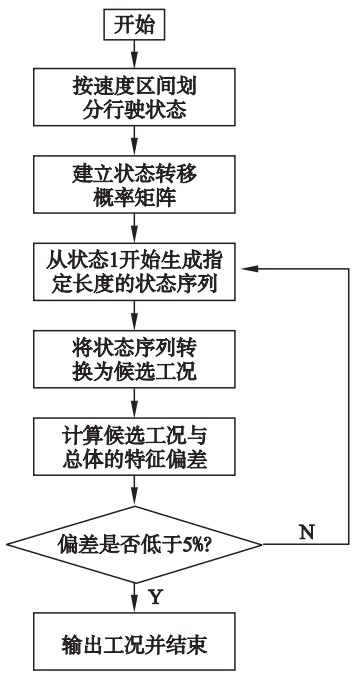

图 3 代表行驶工况构建流程

Fig. 3 Flow chart of typical driving cycle construction

本文拟定代表工况的时长为 1 800 s,根据以上步骤最终得到沈阳市乘用车代表行驶工况,如图 4 所示. 从图 4 可以看到,构建工况由若干短行驶片段组成,这些行驶片段可以分为低速、中速和高速 3 个类别,说明采用马尔科夫链法可以使构建工况呈现不同的分类特征.

表 2 所示为构建的代表工况与样本数据库的统计特征对比. 根据表 2 计算的平均百分比偏差为 2.46% . 特征参数的百分比偏差大部分小于 5% ,匀速比例和减速比例的偏差值略大,但也在 10% 以内. 因此,从统计结果看,构建工况能够准确反映样本总体的特征,本文所提出的构建算法能够获得较高的统计精度.

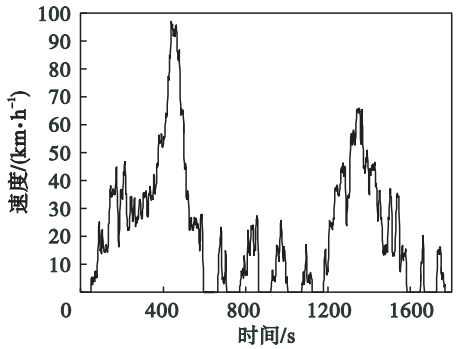

图 4　沈阳市乘用车代表行驶工况

Fig. 4　Typical driving cycle for passenger cars in Shenyang City

表 2　代表工况与样本数据库对比

Table 2　Comparison between typical driving cycle and sample database

参数	样本数据库	工况	偏差/%
v_{sp}	28.64	29.74	3.84
v_{sd}	20.57	20.75	0.88
a_{sd}	0.447 5	0.446 9	0.13
a_{ac}	0.447 8	0.447 1	0.16
a_{de}	−0.490 4	−0.493 0	0.53
P_a	24.14	25.00	3.56
P_d	20.85	21.94	5.23
P_c	25.65	23.67	7.72
P_i	29.36	29.39	0.10

4　结　　论

1)本文以马尔科夫链为基础,采用改进的 Kneser－Ney 平滑方法估计状态转移概率,提出了代表行驶工况的构建算法. 该算法可为车辆行驶工况的构建提供理论参考.

2)组建了沈阳市乘用车行驶数据库,据此构建了代表行驶工况. 将该工况与数据库总体特征进行对比,结果表明两者的平均百分比偏差为 2.46% ,特征参数偏差值均在 10% 以内,说明构建工况能够准确反映总体行驶特征,验证了工况构建算法的有效性.

参考文献:

[1] Wang Q D, Huo H, He K B, et al. Characterization of vehicle driving patterns and development of driving cycles in Chinese cities[J]. *Transportation Research Part D: Transport and Environment*, 2008, 13(5): 289－297.

[2] Ho S H, Wong Y D, Chang V W C. Developing Singapore driving cycle for passenger cars to estimate fuel consumption and vehicular emissions[J]. *Atmospheric Environment*, 2014, 97: 353－362.

[3] Fotouhi A, Montazeri-Gh M. Tehran driving cycle development using the *k*-means clustering method [J]. *Scientia Iranica*, 2013, 20(2): 286－293.

[4] 郑殿宇,吴晓刚,陈汉,等. 哈尔滨城区乘用车行驶工况的构建[J]. 公路交通科技,2017,34(4):101－107.
(Zheng Dian-yu, Wu Xiao-gang, Chen Han, et al. Construction of driving conditions of Harbin urban passenger cars[J]. *Journal of Highway and Transportation Research and Development*, 2017, 34(4): 101－107.)

[5] Nutramon T, Supachart C. Influence of driving cycles on exhaust emissions and fuel consumption of gasoline passenger car in Bangkok [J]. *Journal of Environmental Sciences*, 2009, 21(5): 604－611.

[6] Brady J, O′Mahony M. Development of a driving cycle to evaluate the energy economy of electric vehicle in urban areas [J]. *Applied Energy*, 2016, 177: 165－178.

[7] Balau A E, Kooijman D, Rodarte I V, et al. Stochastic real-World drive cycle generation based on a two stage markov chain approach[J]. *SAE International Journal of Materials & Manufacturing*, 2015, 8(2): 390－397.

[8] Montazeri-Gh M, Fotouhi A. Traffic condition recognition using the *k*-means clustering method[J]. *Scientia Iranica*, 2011, 18(4): 930－937.

[9] Jing Z C, Wang G L, Zhang S P, et al. Building Tianjin driving cycle based on linear discriminant analysis [J]. *Transportation Research Part D: Transport and Environment*, 2017, 53: 78－87.

[10] Bishop J D K, Axon C J, McCulloch M D. A robust, data-driven methodology for real-world driving cycle development [J]. *Transportation Research Part D: Transport and Environment*, 2012, 17(5): 389－397.

[11] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling[J]. *Computer Speech and Language*, 1999, 13(4): 359－394.