文章编号：2095-2163(2021)07-0080-07 中图分类号：TP391 文献标志码：A

改进 PSO-K-means 算法在汽车行驶工况估计中的应用

范艺璇，阚 秀，曹 乐，沈 颉
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：针对城市道路上轻型车的行驶工况问题，分析福建省莆田市某实际道路采集的行驶数据和道路交通运行特征，对实采数据进行清洗并划分成运动学片段，根据车辆运行机制和运动学片段统计分布特点，采用 PCA 方法对特征参数进行降维处理，设计改进的 PSO-K-means 算法构建车辆行驶工况，并从 10 个主要特征参数角度与实际工况进行对比，结果表明所构建工况能够准确反映车辆在实际道路上的行驶特征，说明使用改进 PSO-K-means 算法构建轻型车行驶工况的合理性和有效性。

关键词：PCA 分析；数据清洗；改进 PSO-K-means 算法；行驶工况

Application of improved PSO-K-means algorithm in the estimation of driving cycle

FAN Yixuan, KAN Xiu, CAO Le, SHEN Jie
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 Aiming at the driving cycle of light vehicles on urban roads, the driving data and road traffic operation characteristics collected from a real road in Putian City, Fujian Province are analyzed. The collected data are cleaned and divided into kinematic segments. According to the vehicle operation mechanism and the statistical distribution characteristics of kinematic segments, PCA method is used to reduce the dimension of the characteristic parameters, and the improved PSO-K-means algorithm is designed to construct vehicle driving cycle. The paper compares the constructed driving cycle and actual driving cycle from the perspective of 10 main characteristic parameters. The results show that the constructed driving cycle can accurately reflect the driving characteristics of the vehicle on the actual road, which shows the rationality and effectiveness of using the improved PSO-K-means algorithm to construct the driving cycle of light vehicles.

【Key words】 PCA analysis; data cleaning; improved PSO-K-means algorithm; driving cycle

0 引 言

近年来，随着乘用车保有量的迅猛增长，道路交通、能源消耗和排放污染等一系列问题随之出现，行驶工况作为衡量车辆能耗、排放测试和行驶特征的重要标准，其构建问题一直受到相关领域学者的广泛关注[1-5]。由于各城市发展背景和环境不同，采用统一的行驶工况标准进行汽车能耗/排放等认证显然不合适，因此，依据不同城市的实际汽车行驶数据，构建反映实际道路行驶工况具有重要的研究意义。

为适应不同地区的车辆行驶特征和道路条件，现有行驶工况研究大多针对具体的地区展开。刘燕[6]应用 K-means 聚类方法研究了具有山地道路特性的重庆市行驶工况。高建平等人[7]采用主成分分析和改进的模糊聚类(FCM)方法构建了符合郑州市交通特征的行驶工况。Amirjamshidi 等人[8]运用多目标遗传(MOGA)算法构建了多伦多市卡车的行驶工况，并进行了车辆排放试验。宋怡帆[9]使用改进的 AP 聚类方法针对深圳市的轻型车进行行驶工况分析。刘子谭等人[10]从估计区间的角度改进 K-means 聚类方法，并研究了广州市的轻型车行驶工况。

本文基于莆田市某型号汽车的行驶数据，利用改进的 PSO-K-means 算法构建了适应该地区该车型的行驶工况，论文的具体内容结构如图 1 所示。第 2 节介绍了基于改进的 PSO-K-means 算法的流程。第 3 节阐述了数据清洗的过程和运动学片段的划分。第 4 节根据运动学片段分布特点和车辆行驶特征，提取典型特征参数，通过 PCA 对典型特征降维，得到 4 个主要成分。第 5 节基于改进的 PSO-K-means算法，构建汽车行驶工况，并结合车辆实际运行情况，评估所构建行驶工况的合理性。

作者简介：范艺璇(1997-)，女，硕士研究生，主要研究方向：数据处理；阚 秀(1983-)，女，博士，副教授，主要研究方向：智能控制、路径规划、网络化系统建模等；曹 乐(1986-)，男，博士，讲师，主要研究方向：惯性传感器、组合导航技术、先进传感技术等；沈 颉(1993-)，男，硕士，主要研究方向：智能控制。

通讯作者：阚 秀 Email：xiu.kan@ sues.edu.cn

收稿日期：2021-04-08

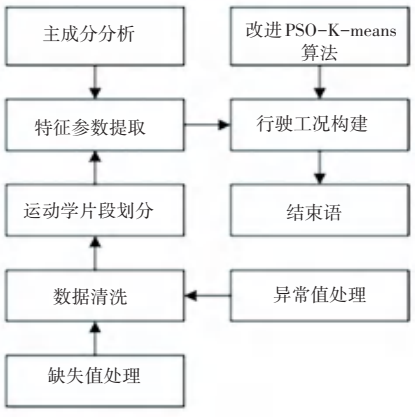

图 1　本文框架结构图

Fig. 1　The frame structure of this paper

1　改进 PSO-K-means 算法

粒子群优化算法[11]（PSO）是一种进化计算技术，具有易实现、收敛快和精度高等优点，且对初始值要求不高，而 K-means 聚类方法具有聚类效果好但对初始中心点敏感的特点，本文将 PSO 算法和 K-means方法结合，使得改进后的 PSO-K-means 算法实现对行驶工况的精确快速估计。PSO-K-means算法的流程如下所示：

（1）初始化粒子群：随机生成 m 个粒子，每个粒子的位置由 k 个样本的 d 个特征信息决定，即初始聚类中心位置。

（2）利用适应度函数计算每个粒子的个体极值和全局最优值的适应度值，适应度定义如下：

$$f(x)=\sum_{j=1}^{k}\sum_{S_i\in C_j}\|S_i-Z_j\| \tag{1}$$

其中，C_j 为 k 个聚类中心对应的 k 个类别；S_i 为类 C_j 中的其他所有点；Z_j 为聚类中心。初始化粒子速度 $v_i(t)$，计算个体适应值，确定个体极值位置 $xBest_i$ 和种群达到的全局最优位置 $xgBest$。

（3）设置最大迭代次数 $t_{\max}$，当前迭代次数 $t=1$。设置判断粒子群收敛速度的适应度方差阈值为 θ，方差 σ^2 计算公式如下：

$$\sigma^2=1/m\sum_{i=1}^{m}\left[f(x_i)-f_{avg}\right]^2 \tag{2}$$

其中，$f(x_i)$ 为粒子 i 的适应度值，f_{avg} 为所有粒子的适应度均值。

（4）根据每个粒子的个体极值位置 $xBest_i$ 和全局最优位置 $xgBest$，按以下公式更新粒子的速度与位置信息：

$$v_i(t)=\omega(t)*v_i(t-1)+c_1*\rho_1(xBest_i-x_i(t))+c_2*\rho_2(xgBest-x_i(t)) \tag{3}$$

$$X_i(t+1)=X_i(t)+H_0(1-t/t_{\max})V_i(t+1) \tag{4}$$

其中，$x_i(t)$ 为第 i 个粒子所在的位置；$v_i(t)$ 为第 i 个粒子的速度；c_1，c_2 分别为惯性因子和约束因子 ρ_1 和 ρ_2 为取值[0,1]区间的随机数；$\omega(t)$ 为惯性权重。

针对理想 PSO 算法中前期全局搜索强后期局部搜索强的特点，对 $\omega(t)$ 值采用如下公式刻画的自适应操作[12]：

$$\omega(t)=\omega_{\max}-(\omega_{\max}-\omega_{\min})t/t_{\max} \tag{5}$$

其中，$\omega_{\max}$ 为最大惯性权重，$\omega_{\min}$ 为最小惯性权重。

（5）判断当前迭代次数 t 是否等于最大迭代次数 $t_{\max}$，如果 $t=t_{\max}$ 则输出适应度值最小的粒子为 k 个聚类中心；如果 $t<t_{\max}$ 且 $\sigma^2\leqslant\theta$，则输出适应度值最小的粒子为 k 个聚类中心。若 $\sigma^2>\theta$，继续重复（4）、（5）过程。

（6）计算种群中每个个体与以上步骤中得到的聚类中心之间的距离，按照如下公式计算个体 a 与个体 b 第 h 个特征之间的距离：

$$D_{ab}(2)=\left(\sum_{h=1}^{n}|x_{ah}-x_{bh}|^2\right)^{\frac{1}{2}} \tag{6}$$

将每个样本归为距离最近的中心点，更新每个数据簇的中心点。

（7）重复步骤（6）直至聚类中心不发生变化，算法结束。

2　数据清洗与运动学片段提取

行驶数据来自于车联网管理平台数据库，车辆通过无线传输设备将车载传感器数据信息发送至车联网管理平台数据库，由于 GPS 信号丢失、环境因素或传感器老化等因素会造成数据部分丢失、不连续和异常等现象，为尽可能真实地还原车辆实际行驶状况，首先要对原始数据进行清洗，本文通过对汽车行驶时相应参数变化的分析，对原始数据的丢失或异常部分进行插值拟合、替换和剔除等清洗处理操作，具体清洗处理流程如图 2 所示。

2.1　缺失数据值处理

（1）若信号丢失前车速>10 km/h，且 GPS 车速不为 0，采用如下插值方法将丢失数据补齐，此时需用到的公式为：

$$x_i,\ldots,x_{i+n-1}=\begin{cases}x_i,\ldots,x_{i+n-1} & n=0\\ \text{三次样条插值} & 0<n\leqslant 100\\ \text{傅里叶插值} & 100<n\leqslant 300\\ 0 & n>300\end{cases}\quad i=0,1,\ldots,k \tag{7}$$

其中，n 为丢失数据点，数据点的间隔以 s 为单位。

(2)若信号丢失前车速<10 km/h，则视为异常，将该信号缺失段的数据点删除。

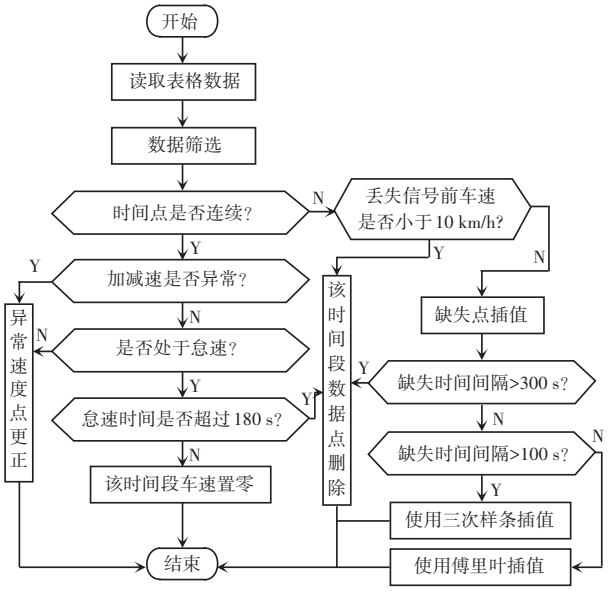

图 2 数据清洗流程图

Fig. 2 Data cleaning flow chart

2.2 异常数据值处理

(1)存在汽车加、减速异常的数据(此型号轻型车一般情况下：0~100 km/h 的加速度时间大于 7 s，紧急刹车最大减速度在 7.8~8 m/s^2)，因此针对 2.1 节中已经插补后的数据值的情况，通过双树复小波算法，查找加减速异常值，然后对异常值进行筛选和剔除。

将行驶工况看作一个随时间变化的离散小波信号，基于双树复小波变换[13]，默认汽车加速状态下的加速度为平均加速度，刹车状态下的最大减速度为瞬时减速度。以 2017-12-18 18:01:50 至 2017-12-18 18:08:29 中 400 组数据为例，选取时刻记为 $t_i(i=1,2,\cdots,400)$。并截取其时间 — 车速图像，设 t_i 时刻速度 v_i 数据异常，通过小波分析将异常点筛选，并按如下公式得到更正点 v_i'，数学公式可写为：

$$v_i' = \frac{\sum_{a=1}^{n} v_{i+a} + \sum_{a=1}^{n} v_{i-a}}{2n} \tag{8}$$

其中，v_{i+a} 表示 t_i 时刻前 a 个点的速度；v_{i-a} 表示 t_i 时刻后 a 个点的速度；n 为数据点数。

图 3 为一段含异常点的时间-速度图，虚线框处速度和加速度值出现异常，按照上述处理方式，可以得到更正后的时间-速度图如图 4 所示。

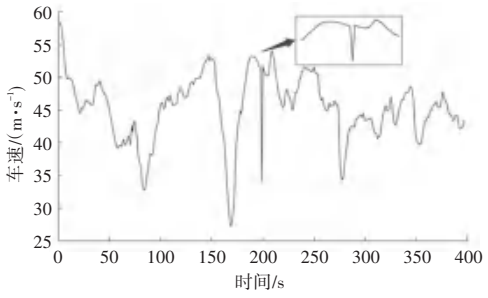

图 3 含异常值时间-速度图

Fig. 3 Time-velocity diagram with outliers

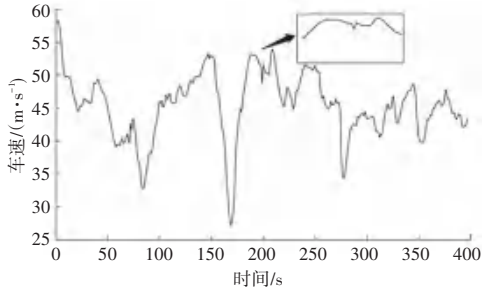

图 4 修正后时间-速度图

Fig. 4 Corrected time-velocity diagram

(2)调查表明福建省交通信号红灯持续时间一般不大于 180 s，因此设定车辆的最长怠速时间为 180 s。对于车辆处于怠速且怠速时间超过 180 s 的时间段以及发动机转速为 0 但采集设备仍运行的情况下的数据点进行删除，对于怠速时间在 180 s 之内的数据段车速置为 0。将车速跳变的地方用连线表示出来，其密集程度表示车速数据的连贯性。

经过 2.1 节和 2.2 节对原始数据清洗处理后，处理前后数据如图 5 和图 6 所示，具体就是车速密连贯性图，序列号为数据的编号，但是时间并非连续的，所以纵轴的尺度较之横轴大。图 5 中，颜色越深处表示清洗处理前数据缺失量越大。由图 6 可以看出，清洗处理后数据较为均匀，能够反映真实的行驶状况，为后续构建合理的行驶工况提供依据。

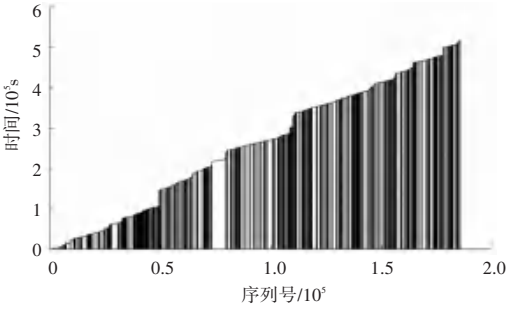

图 5 原数据

Fig. 5 Original data

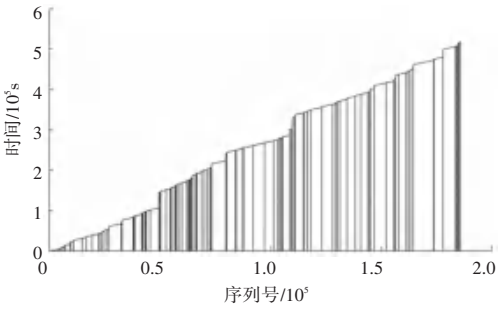

图 6　清洗处理后数据

Fig. 6　Data after cleaning

2.3　运动学片段的提取

运动学片段是指汽车从一个怠速状态开始至下一个怠速状态开始之间的车速区间，且一个标准的运动学片段需要包括加速状态、减速状态、巡航/匀速状态和怠速状态[14]。提取步骤为：将车速较慢且时间不长的片段进行降噪处理，将片段时间小于 20 s的剔除，遍历所有数据点，遇到速度为 0 的点即记录该位置为起始点，当速度从非 0 点跳至 0 的时刻，记该位置为结束点，结束点与起始点之间的时间片段大于20 s则保留为运动学片段，重复此过程操作，具体运动学片段提取算法流程如图 7 所示。

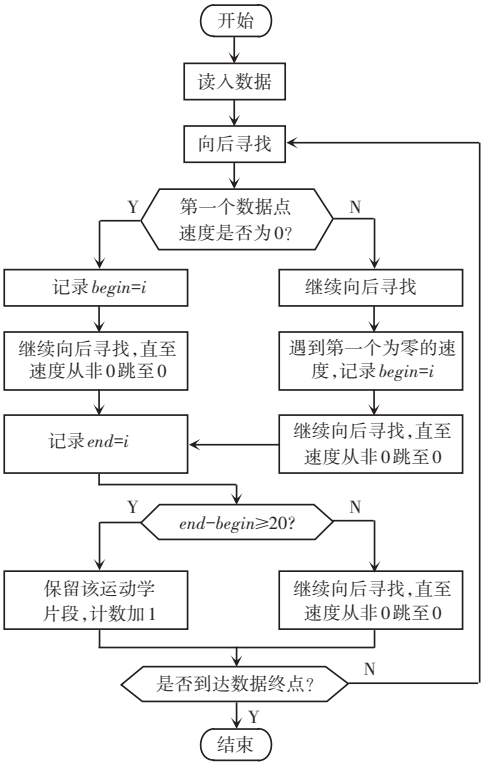

图 7　运动学片段提取流程图

Fig. 7　Flow chart of kinematic fragment extraction

基于所给行驶数据，按照上述步骤提取出 3 408 个运动学片段。

3　特征参数

3.1　提取有效特征参数

分析车辆行驶机制和运动学片段分布特点，选取 10 个主要特征参数，见表 1。

表 1　特征参数选取

Tab. 1　Feature parameter selection

序号	符号	描述	公式
1	V_m	平均速度	$V_m = \dfrac{S}{T}$
2	V_{mr}	平均行驶速度	$V_{mr} = \dfrac{S}{(T - T_i)}$
3	A_m	平均加速度	$A_m = \dfrac{sum\{a_i \mid a_i \geqslant 0.1\}}{T_a}$，$i = 1,2,\ldots,k-1$
4	D_m	平均减速度	$D_m = \dfrac{\mid sum\{a_i \mid a_i \leqslant -0.1\} \mid}{T_d}$，$i = 1,2,\ldots,k-1$
5	P_i	怠速时间比	$P_i = \dfrac{T_i}{T}$
6	P_a	加速时间比	$P_a = \dfrac{T_a}{T}$
7	P_d	减速时间比	$P_d = \dfrac{T_d}{T}$
8	V_s	速度标准差	$V_s = \sqrt{\dfrac{1}{k-1}\sum_{j=1}^{k}(V_j - V_m)^2}$，$j = 1,2,\ldots,k$
9	A_s	加速度标准差	$A_s = \sqrt{\dfrac{1}{k-1}\sum_{j=1}^{k}a_j{}^2}$，$j = 1,2,\ldots,k$
10	$V_{\max}$	最大速度	$V_{\max} = \max\{V_j, j = 1,2,\ldots,k\}$

表 1 中，$S = \sum_{i=1}^{k} V_i, j = 1,2,3,\ldots,k$，是该运动学片段所有数据点速度的总和，$T$ 代表该运动学片段的总点数，T_i 为速度为 0 的数据点的总个数，T_a 为该运动学片段中加速度不小于 0.1 m/s² 的总点数，T_d 为该运动学片段中加速度小于 − 0.1 m/s² 的总点数。

3.2　PCA 降维处理

上述过程选取的 10 个特征参数间存在一定的相关性，PCA 方法在保持数据信息的前提下，将特征参数进行组合，形成新的相互独立的参数，降低估计行驶工况的计算复杂度[15]。根据处理后的3 408 个运动学片段和选取的 10 个特征参数，可构成如下

运动学特征值参数矩阵：

$$\boldsymbol{Z}_{m\times n}=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix} \tag{9}$$

$$m = 3\ 408,\ n = 10$$

其中，z_{ij} 为第 i 个运动学片段的第 j 个特征参数。将特征参数矩阵的每一列进行标准化处理，得到矩阵 $\boldsymbol{X}$，即：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{10}$$

其中，$x_{ij}=\dfrac{(z_{ij}-\bar{z_j})}{s_j}$，$\bar{z_j}=\dfrac{1}{m}\sum_{i=1}^{m}z_{ij}$，

$s_j^2=\dfrac{1}{m-1}\sum_{i=1}^{m}(z_{ij}-\bar{z_j})^2$，$i=1,2,\cdots,m$，$j=1,2,\cdots,n$。

矩阵 $\boldsymbol{X}$ 的协方差矩阵为：

$$\boldsymbol{\Sigma}=\begin{bmatrix} s_1^2 & cov(1,2) & \cdots & cov(1,10) \\ cov(2,1) & s_2^2 & \cdots & cov(2,10) \\ \vdots & \vdots & \ddots & \vdots \\ cov(10,1) & cov(10,2) & \cdots & s_{10}^2 \end{bmatrix} \tag{11}$$

其中，$s_x^2=cov(x,x)$，$cov(x,y)=cov(y,x)=\dfrac{1}{m-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$。

运动学特征值参数矩阵 $\boldsymbol{Z}_{m\times n}$ 对应相关矩阵 $\boldsymbol{R}$ 为：

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

其中，$r_{xy}=\dfrac{cov(x,y)}{s_x s_y}$。此时相关矩阵 $\boldsymbol{R}$ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_{10}\geqslant 0$，特征值对应的标准正交化特征向量为：

$$[\boldsymbol{e}_1\quad\boldsymbol{e}_2\quad\boldsymbol{e}_3\quad\cdots\quad\boldsymbol{e}_n]=\begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nn} \end{bmatrix} \tag{13}$$

设 $\dfrac{\lambda_i}{\sum_{j=1}^{10}\lambda_j}$ 为第 i 个主成分的贡献率，$\dfrac{\sum_{r=1}^{l}\lambda_r}{\sum_{j=1}^{10}\lambda_j}$ 为前 r 个成分的累计贡献率，经验表明累计贡献率大于 80%的成分为工程上所需求的主成分。统计结果见表 2。分析表 2，发现前三个主成分的特征值均大于 1，所以选择前三个主成分作为特征参数数据的代表，由于第四个主成分的累计贡献率为 81.99%，超过了一般工程应用需求的 80%，故最终选用 4 个主成分。

表 2 各主成分方差贡献率和累计方差贡献率

Tab. 2 Variance contribution rate and cumulative variance contribution rate of each principal component

主成分	特征值	贡献率/%	累计贡献率/%
M_1	6.659	47.830	47.83
M_2	2.485	17.560	65.39
M_3	1.330	9.500	74.88
M_4	0.998	7.113	81.99

各个主成分与特征参数对应特征值之间的相关关系见表 3。

表 3 主成分相关系数表

Tab. 3 Principal component correlation coefficient

特征参数	第一主成分	第二主成分	第三主成分	第四主成分
平均速度	0.805	0.490	0.193	0.041
平均行驶速度	0.788	0.447	0.364	0.104
平均加速度	0.489	−0.664	0.332	0.149
平均减速度	−0.309	0.766	−0.018	−0.013
怠速时间比	−0.777	−0.261	0.428	0.090
加速时间比	0.671	0.210	−0.618	−0.158
减速时间比	−0.795	0.453	0.170	0.064
速度标准差	0.794	0.308	0.379	0.119
加速度标准差	0.764	−0.540	−0.055	−0.028
最大速度	−0.400	0.019	−0.304	0.951

特征参数所对应的主成分上的相关系数绝对值越大，该成分与这些特征参数的相关性就越高，对表 3 中各特征参数与 4 个主成分的相关系数进一步分析可知：

(1)第一主成分与减速时间比、加速度标准差、速度标准差、平均速度、平均行驶速度这几个特征参数的载荷系数最高，因此主要代表减速时间比、加速度标准差、速度标准差、平均速度和平均行驶速度的特征值信息。

(2)第二主成分与平均加速度、平均减速度的载荷系数绝对值都超过了 0.6，相关性较高，因此主

要代表平均加速度、平均减速度。

(3)第三主成分与加速时间比、怠速时间比的载荷系数的绝对值较大,因此主要代表加速时间比、怠速时间比。

(4)第四主成分与最大速度的载荷系数非常高,因此主要代表最大速度的特征值。

4　行驶工况估计

根据城市交通状况,可将车辆行驶状态分为 3 类:

(1)拥堵行驶工况:交通状况拥堵,车辆行驶速度缓慢,车辆需经常启停。

(2)稳态流动行驶工况:没有拥堵,车流数目较多,平均行驶速度较低。

(3)畅通行驶工况:路面交通状况良好,车流数目较少,怠速状态少。

将其特征参数降维后的3 408个运动学片段进行分类。依据经验设定初始 K 值为 3,把所有的运动学片段划分成上述 3 种状态,得到拥堵行驶工况的数目有 426 个,稳态流动行驶工况的数目有2 130 个,畅通行驶工况的数目有 852 个。

应用改进 PSO-K-means 算法,将低速工况、高速工况、中速工况进行连接,合成持续时间1 289 s的道路行驶工况,构建成如图 8 所示的由八段数据组成的汽车行驶工况曲线。

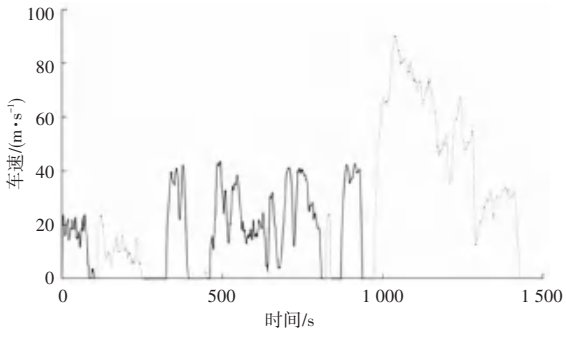

图 8　汽车行驶工况曲线

Fig. 8　Vehicle driving cycle curve

实际工况总速度占比和构建工况总速度占比基本吻合,表 4 给出构建工况与实际工况中各项参数值,可以看出对应参数差距很小,说明所估计的行驶工况科学合理。图 9 为构建工况和实际工况相关雷达图,表明实际工况和构建工况在特征参数中相关性较高,进一步说明所估计行驶工况的合理性和有效性。

表 4　汽车工况分析表

Tab. 4　Analysis of vehicle driving cycle

参数	构建工况	实际工况	所占比	参数	构建工况	实际工况	所占比
平均速度 v_m	27.73	26.87	0.032	加速时间比 P_a	0.30	0.29	0.035
平均行驶速度 v_t	31.78	31.09	0.022	减速时间比 P_d	0.49	0.47	0.042
平均加速度 A_{am}	1.30	1.57	0.170	速度标准差 V_s	14.5	10.3	0.407
平均减速度 D_m	1.35	1.49	0.090	加速标准差 A_s	2.21	1.73	0.277
怠速时间比 P_i	0.15	0.13	0.150	最大速度 v_{max}	90.4	104	0.130

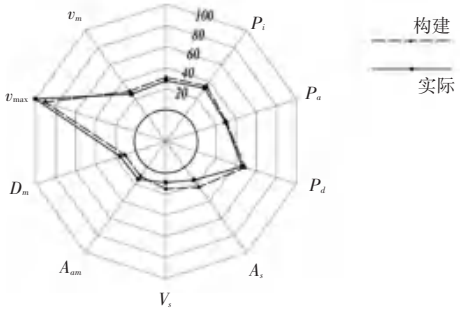

图 9　构建工况与实际工况相关雷达图

Fig. 9　Radar chart related to constructed driving cycle and actual driving cycle

5　结束语

本文根据福建省莆田市某型号轻型车的行驶数据,研究了其在实际道路上的行驶工况估计问题。根据行驶道路特征和数据采集传输原理,清洗原始数据并进行运动学片段划分,分析车辆运行机制和运动学片段分布特点,提取主要特征参数并使用 PCA 方法降维处理,利用改进的 PSO-K-means 算法估计车辆行驶工况,并从 10 个主要特征参数角度对比构建工况与实际工况,数据显示各项特征参数值占比相近,进一步说明所估计行驶工况的科学性和有效性。

参考文献

[1] ANDRE M. Driving cycles development: Characterization of the methods[J]. SAE S pecial Publications, 1996, 1201(12): 312,322.

[2] LEE T C, JUDGE G G, ZELLNER A. Estimating the parameters of the Markov probability model from aggregate time series data [J]. Journal of the American Statistical Association, 1970, 66 (335):653.

(下转第 90 页)