基于用户大数据的工况分类与构建方法

柳子强　王晓旭　王大志

（中国第一汽车股份有限公司研发总院，长春 130013）

【摘要】为提高可靠性试验所用工况与用户实际驾驶行为的相关性，提出一种基于用户大数据的用户工况分类及试验工况构建方法，通过主成分分析法对大量数据进行降维处理，利用K均值聚类方法对数据进行分类并选取代表片段，采用马尔可夫方法对数据进行排序，最终构建出符合用户实际驾驶情况的可靠性试验工况，并根据该计算流程对某大数据平台下20位用户的行驶数据进行实际计算，与目前普遍采用的工况相比，数据来源角度能更准确地描述用户实际驾驶状态。

关键词：大数据　主成分分析　K均值聚类　马尔可夫过程

中图分类号：U467.3　　**文献标志码：**A　　**DOI：** 10.20104/j.cnki.1674-6546.20230100

A Working Condition Classification and Construction Method Based on User Big Data

Liu Ziqiang, Wang Xiaoxu, Wang Dazhi

(Global R&D Center, China FAW Corporation Limited, Changchun 130013)

【**Abstract**】To improve the correlation between reliability test condition and user actual driving behavior, this paper proposed a method of user condition classification and test condition construction based on user big data. In this method, the principal component analysis method was used to reduce the dimensionality of a large amount of data, and the K-means clustering method was used to classify the data and select representative segments, sort the data through the Markov method, and finally construct a reliability test condition conforming to the actual driving situation of the user. Driving data from twenty users under a big data platform was calculated according to this calculation process. Compared with the condition commonly used at present, the method from perspective of data source can describe user actual driving state more accurately.

Key words: Big data, Principal component analysis, K-means clustering, Markov process

【引用格式】柳子强, 王晓旭, 王大志. 基于用户大数据的工况分类与构建方法[J]. 汽车工程师, 2023(7): 37-43.

LIU Z Q, WANG X X, WANG D Z. A Working Condition Classification and Construction Method Based on User Big Data[J]. Automotive Engineer, 2023(7): 37-43.

1　前言

目前，可靠性试验中采用的城市汽车行驶工况通常为欧洲城市循环工况，该工况加减速过程简单，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[1]。

使用用户大数据可以更准确地描述用户实际驾驶状态，有效提高试验工况与用户驾驶行为的关联性。本文通过大数据的提取和处理，并结合主成分分析、K均值聚类、马尔可夫排序等数据处理方法，实现对用户各种驾驶行为下典型工况的构建，以进一步提高试验结果与用户实际驾驶状态的一致性。

2　数据处理

2.1　数据信息

本文采用的大数据主要来自某企业某车型用户大数据平台，基于用户实际驾驶车辆定期反馈的数据进行研究，用户的选取主要参照以下原则：选取该车型用户车辆总行驶里程由高到低排序的前

10个城市，下载相应用户某自然年全年的行驶数据作为分析对象。前10个城市所有用户的总行驶里程如表1所示。

表1　某车型用户总行驶里程由高到低排序的前10个城市

城市名称	总里程/km
东莞市	301 474.8
上海城区	276 864.3
深圳市	260 965.0
佛山市	223 405.4
青岛市	217 580.7
广州市	212 662.1
杭州市	198 604.4
北京城区	187 081.8
中山市	178 809.4
成都市	177 604.7

2.2　数据整理

进行数据提取和聚合后，需要对大数据片段进行划分，从而形成多个完整的行驶工况。选取具有代表性的行驶工况，将这些工况的速度-时间数据按一定的时间周期划分，可得到多个运行片段[2]，片段划分遵循如下原则：

a. 将连续2个速度为0的时间段划分为一个运行片段，某个片段的速度-时间曲线如图1所示。

b. 计算划分后的运行片段的最大速度，若最大速度不超过5 km/h，则将此片段视为无效片段，并从片段集合中去除。

c. 对剩余的有效运行片段进行编号。

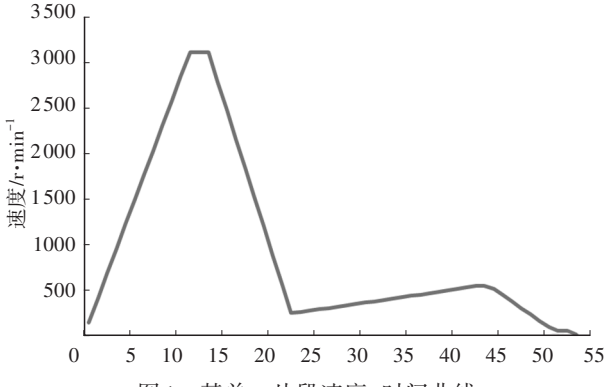

图1　某单一片段速度-时间曲线

本文最终将获得的数据划分为104 357个运行片段。

2.3　特征参数的选取和计算

从对驾驶特点的描述角度，本文选取了18个特征参数，如表2所示。

表2　工况特征参数

序号	特征参数
1	最大速度 V_{max}
2	平均速度 V_{m}
3	速度标准差 V_{sd}
4	最大加速度 a_{max}
5	最大减速度 a_{min}
6	加速度标准差 a_{sd}
7	加速段平均加速度 a_{amean}
8	减速段平均减速度 a_{dmean}
9	扭矩标准差 T_{rq_sd}
10	平均正扭矩 T_{rq_pmean}
11	平均负扭矩 T_{rq_nmean}
12	扭矩增加时最大波动量 $T_{rq\_range\_max}$
13	扭矩减小时最大波动量 $T_{rq\_range\_min}$
14	扭矩波动标准差 $T_{rq\_range\_sd}$
15	加速时间比例 P_{ta}
16	减速时间比例 P_{td}
17	总时间 T
18	总行驶里程 S

对每一个驾驶循环工况进行18维特征参数计算，最终形成带有特征参数信息的片段集合，称为特征参数矩阵。

3　基于主成分分析法的数据分析

本文采用主成分分析法将高维且具有一定相关性的复杂特征指标转化为低维的多个互不相关的主成分，并保留原始特征参数中的大量信息[3]，为后续的分类和分析工作提供基础。

3.1　主成分分析过程

假设共有 n 个运行片段，每个运行片段有 p 个特征参数，将计算出的所有运行片段的特征参数矩阵记为 $\boldsymbol{X}$。设随机变量 X_i 的均值为 $\boldsymbol{\mu}_i$、协方差矩阵为 $\boldsymbol{\Sigma}_i$，则通过主成分分析，对 p 个特征参数进行线性变换，生成的综合指标即为主成分，记为 y_1、y_2、……y_p。其中特征参数矩阵 $\boldsymbol{X}$ 可以表示为：

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p)' \qquad (1)$$

式中，$x_{ij}(i=1,2,\cdots,n, j=1,2,\cdots,p)$为特征参数矩阵中的特征参数。

经过变换后主成分的表达式为：

$$\begin{cases} y_1 = \boldsymbol{l}_{11}x_1 + \boldsymbol{l}_{21}x_2 + \cdots + \boldsymbol{l}_{p1}x_p = l_1'\boldsymbol{X} \\ y_2 = \boldsymbol{l}_{12}x_1 + \boldsymbol{l}_{22}x_2 + \cdots + \boldsymbol{l}_{p2}x_p = l_2'\boldsymbol{X} \\ \vdots \\ y_p = \boldsymbol{l}_{1p}x_1 + \boldsymbol{l}_{2p}x_2 + \cdots + \boldsymbol{l}_{pp}x_p = l_p'\boldsymbol{X} \end{cases} \tag{2}$$

式中，x_1~x_p为原始特征参数；$l_{ij}(i,j=1,2,\cdots,p)$为主成分变换后的参数矩阵；l_1'~l_p'为各主成分中特征参数的载荷系数。

根据相关系数矩阵或协方差矩阵求解各主成分y_p，其中方差及协方差计算公式分别为：

$$var(y_j)=l_j'\sum \boldsymbol{l}_j,\ j=1,2,\cdots,p \tag{3}$$

$$cov(y_j,y_k)=l_j'\sum \boldsymbol{l}_k,\ j,k=1,2,\cdots,p \tag{4}$$

式中，$\boldsymbol{l}_j$为协方差参数矩阵。

各特征参数的量纲不同，会引起各变量的分散程度差异较大。在通过协方差矩阵求解特征值与对应的特征向量前，为消除量纲不同带来的不合理影响，常对各原始变量进行标准化处理。标准化处理后的变量为：

$$z_j = \frac{x_j - E(x_j)}{\sqrt{var(x_j)}}, j = 1,2,\cdots,p \tag{5}$$

式中，$E(x_j)$为$\boldsymbol{X}$向量的平均值。

向量$\boldsymbol{Z}=(z_1,z_2,\cdots,z_p)^{\mathrm{T}}$的协方差矩阵为相关系数矩阵$\boldsymbol{\rho}=[\boldsymbol{\rho}(x_i,x_j)]_{p\times p}$。主成分分析后主成分的协方差矩阵为对角矩阵$\boldsymbol{\Lambda}$，其对角线元素为相关系数矩阵$\boldsymbol{\rho}$的特征根$\lambda_1$、$\lambda_2$、……、$\lambda_p$。其中相关系数矩阵$\boldsymbol{\rho}(x_j,y_i)=\sqrt{\lambda_i}\boldsymbol{l}_{ij}$，第$i$个主成分$y_i$与原始特征参数变量$x_j$间的相关系数称为因子载荷量，它反映原始特征参数变量与主成分之间的密切程度，其绝对值越接近1，说明关系越密切。

各主成分的总累积贡献量可以用r'表示：

$$tr(\boldsymbol{\sum})=tr(\boldsymbol{\Lambda})=r=\sum_{j=1}^{r}\lambda_j \tag{6}$$

根据式(6)，第j个主成分的贡献率为λ_j/r'，则前m个主成分的累积方差贡献率为$\sum_{i=1}^{m}\lambda_i/r'$，当累积贡献率达到80%~90%时，提取前$m$个主成分，可以代替原始变量中大部分特征信息量。

当特征参数变量x_1~x_p在某个主成分上的载荷系数近似时，对其主成分的解释较为困难，可以通过因子分析中方差最大化正交旋转法来实现对因子载荷矩阵的旋转，使每个变量在主成分上方差最大化，载荷矩阵每行或每列的元素平方向0和1两级分化，因子载荷越接近1说明相关性越强[4]，以此实现对因子载荷系数的解释。

假设共有n个特征参数，主成分因子数量为m，若每次进行2个因子的旋转计算，共有$m(m-1)/2$种旋转方法，将此视为一次循环，直到因子载荷矩阵中各列的方差和最大且收敛。假设载荷矩阵为$\boldsymbol{A}$：

$$\boldsymbol{A}=(a_{ij}),\ i=1,2,\cdots,n;\ j=1,2 \tag{7}$$

令正交矩阵为$\boldsymbol{Q}$：

$$\boldsymbol{Q}=\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \tag{8}$$

式中，φ为矩阵正交后的旋转角度。

旋转后的因子载荷矩阵$\boldsymbol{B}$为：

$$\boldsymbol{B}=\boldsymbol{AQ}=(b_{ij}),\ i=1,2,\cdots,n;\ j=1,2 \tag{9}$$

为使旋转后的因子载荷矩阵各列方差总和$V=V_1+V_2$最大，V_j的计算公式为：

$$V_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b_{ij}^2}{h_i^2}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{b_{ij}^2}{h_i^2}\right)^2,\ j=1,2 \tag{10}$$

式中，$h_i^2=a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2$为$m$个因子对第$i$个变量方差的贡献度。

以上计算过程均可通过SPSS软件进行，最终得到主成分分析的成分矩阵，将成分矩阵与原始的特征参数矩阵相乘，即可得到主成分分析结果。

3.2 主成分分析结果

通过主成分分析，可以得到降维后的结果，其总方差解释如表3所示。

表3 总方差解释

成分	初始特征值		
	贡献值	方差百分比/%	累积百分比/%
1	7.652	42.513	42.513
2	3.977	22.095	64.608
3	1.379	7.661	72.269
4	1.240	6.889	80.158
5	1.004	5.580	84.738
6	0.910	5.058	89.796
7	0.665	3.694	93.490
8	0.372	2.064	95.554
9	0.239	1.330	96.884
10	0.218	1.214	98.098
11	0.142	0.789	98.887
12	0.088	0.489	99.376
13	0.044	0.246	99.622
14	0.033	0.182	99.804
15	0.020	0.109	99.913
16	0.013	0.073	99.986
17	0.002	0.013	99.999
18	0.000	0.001	100.000

其中方差百分比即为此主成分所包含的原始信息的比例,累积百分比80%以上的成分即可覆盖大部分的原始信息。本文中,前4个主成分累积贡献量为80.158%,因此可将原始数据的特征参数矩阵由18维降低到4维。

4 基于K均值聚类方法的分类过程

作为经典的数据挖掘方法,K均值聚类方法是一种无监督的学习方法[5]。在不明确运行片段分为哪些典型工况时,运行K均值聚类算法给定聚类数量K,即为K类典型工况。按照点与点之间的距离,将每个点分到距离最近的类簇中心所代表的类别中,所有样本点分配完成后重新计算该类簇中所有样本点的平均值,即为新的聚类中心点,之后继续迭代分类,直至类簇中心点的变化很小或达到给定的迭代次数[6]。

4.1 K均值聚类分析过程

K均值聚类需要事先指定K值,因此需要一种可以确定最优K值的方法,目前主要有4种方法,即方差比准则、大卫-博丁指数、轮廓系数法、手肘法。

在实际分析中,分类数量不应出现较大值,因此可在小范围内采用穷举法,然后根据判定式进行最优解的确认,本文对分类数在2~10范围内进行穷举,综合考虑4种方法的计算结果,确定K=4。

4.2 K均值聚类结果

本文使用SPSS软件对主成分分析后形成的4维特征参数矩阵进行聚类,取K=4,最大循环次数为100次,部分结果如表4所示。

表4 K均值聚类部分结果

聚类类别	到聚类中心的距离	片段编号
4	0.160	57
4	0.214	51
4	0.233	34
4	0.271	115
4	0.272	78
4	0.325	81
4	0.342	102

在结果中,每个运行片段都有一个对应的聚类编号,以及该片段到聚类中心的距离,与聚类中心距离越近,代表这一片段越能反映其所在分类的特征。因此,选取4类中与聚类中心最近的4个片段,即为这4类各自的代表片段,结果如表5所示。

表5 代表片段编号和距离

片段编号	分类	距离
61500	1	0.253 90
75115	2	0.176 39
26069	3	0.154 32
23821	4	0.107 85

由于代表片段为与聚类中心距离最近的片段,因此在定义分类的含义时,可以直接观察代表片段的特征对分类进行解释。综合考虑各类代表片段的信息后,对各分类的特征解释如下:

a. 分类1。城郊工况,主要典型特征为车速处于中高速,有较为频繁的加减速,加速度较大,其典型片段如图2所示。

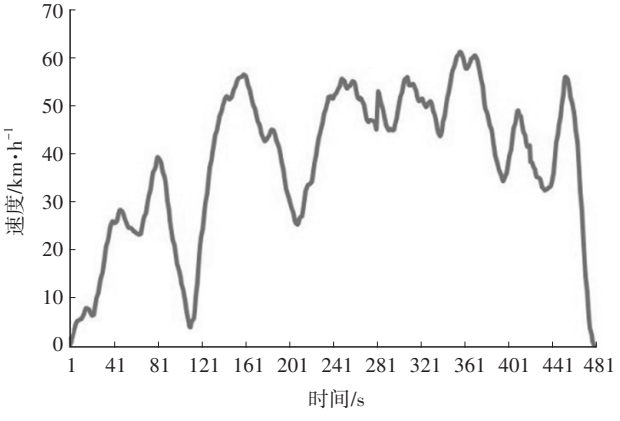

图2 城郊工况速度曲线

b. 分类2。高速工况,主要典型特征为车速长时处于高速段,减速不剧烈,加速度也较小,其典型片段如图3所示。

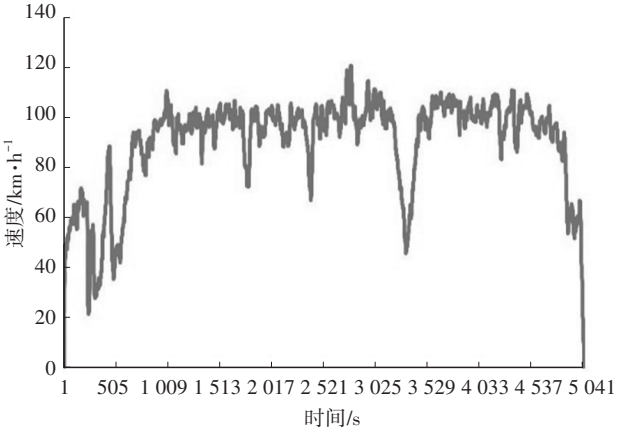

图3 高速工况速度曲线

c. 分类3。城市拥堵工况,主要典型特征为车速处于低速段,加速度较小,其典型片段如图4所示。

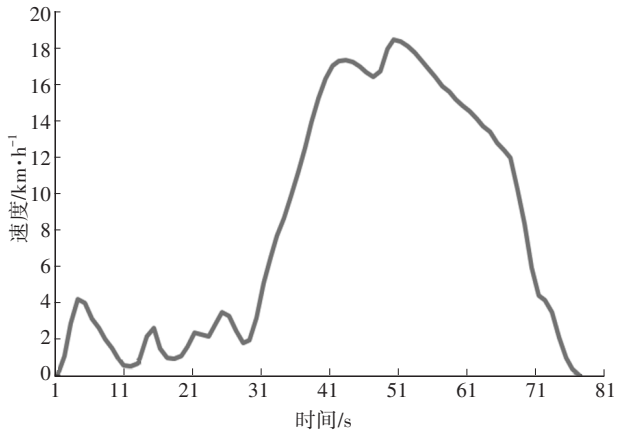

图4　城市拥堵工况速度曲线

d. 分类4。城市快速路，主要典型特征为速度处于中速，有一定加减速，加速度较大，其典型片段如图5所示。

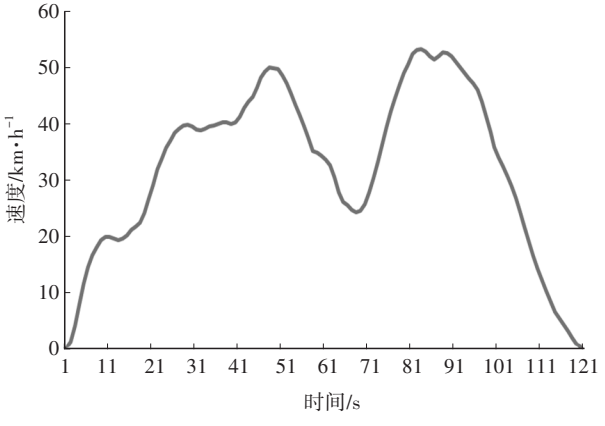

图5　城市快速路工况速度曲线

5　基于马尔可夫链的片段排序过程

5.1　试验片段的选取

选取各分类中用户行驶过程损伤较大的片段作为试验工况，以便缩短试验工况的时间，利用20位用户行驶大数据全过程的平均总损伤与试验工况的损伤进行协同计算，最终得出各片段循环次数如表6所示，为了便于计算，循环次数进行取整。

表6　各试验片段及循环次数

分类	试验片段编号	时长/s	循环次数/次
1	74837	194	800
2	20295	1 140	600
3	103898	27.5	15 000
4	41222	140.5	12 000

试验片段的转速和转矩变化情况如图6所示。

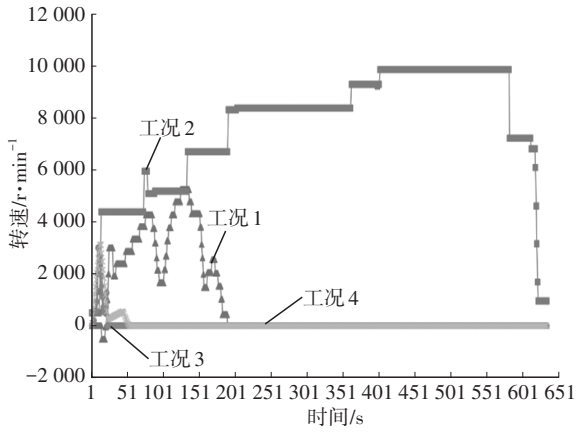

图6　试验片段转速变化情况

5.2　马尔可夫过程

准确复现用户实际使用条件下的载荷作用效果是电驱动系统可靠性评价的核心，其本质是通过合理的工况组合顺序使得各部件性能同步退化到全生命周期设定目标。本文根据马尔可夫原理分析原始工况片段之间的顺序关系。

马尔可夫链实际上是一组离散随机变量的集合[7]，具体指对概率空间(Ω,F,P)内以一维可数集为指数集的随机变量集合$X=\{X_n:n>0\}$，假设随机变量的取值范围均在可数集内，$X=s_i$, $s_i\in s$，且随机变量的条件概率满足如下关系：

$$P(X_{t+1}|X_t,\cdots,X_1)=P(X_{t+1}|X_t) \tag{11}$$

将式(1)中的X称为马尔可夫链，对于一个固定的马尔可夫链模型，式(11)表明，随着马尔可夫链的增长，链中事件参数的分布不变。基于马尔可夫链的这种性质，通过模型时间转移矩阵构建循环工况能够代表整个原始数据中用户实际行驶工况[8]。

采用最大似然估计法计算各类典型工况间的状态转移矩阵，马尔可夫过程可以通过贝叶斯公式求得稳态概率：

$$P(Z_0,Z_1,\cdots,Z_t)=P(Z_0)\prod_{\tau}P(Z_\tau|Z_{\tau-1}) \tag{12}$$

式中，$P(Z_0)$为Z_0事件的先验概率；$P(Z_\tau|Z_{\tau-1})$为在$Z_{\tau-1}$事件发生的前提下，Z_τ事件发生的概率。

重复上述过程，N次重复观察的公式为：

$$P(Z_0,Z_1,\cdots,Z_T|N)=P(Z_0)\prod_{r,s}P_{rs}^{Nrs} \tag{13}$$

通过极大似然函数，假设从工况r转移到工况s，可得出各工况间的状态转移概率方程：

$$P_{rs}=N_{rs}/\sum_{s}N_{rs} \tag{14}$$

式中，P_{rs}为当前状态工况r转移到下个时刻状态工况s的概率，r,s=1,2,3,4；N_{rs}为工况r转移到工况s的

频次[9]。

将聚类后的4种典型工况特征作为马尔可夫过程的4个状态，构成状态空间，根据每个片段所属的工况类别构建马尔可夫链模型。对于一个固定的马尔可夫过程，通过各工况间的状态转移概率方程，即可计算各工况状态转移概率矩阵[10]。

本文使用nCode进行辅助计算，马尔可夫转移概率矩阵实际计算结果如表7所示。根据表7，最终的工况顺序为工况3–工况4–工况2–工况1。

表7　马尔可夫转移概率矩阵

前序工况	后序工况次数/次			
	工况1	工况2	工况3	工况4
工况4	3 170	5 059	10 100	0
工况3	424	1 056	0	10 100
工况2	11	0	1 019	5 097
工况1	0	11	461	3 133

6　工况结果整理

根据以上计算结果，本文构建的工况为：试验工况3运行150次，试验工况4运行100次，试验工况2运行6次，试验工况1运行8次；以上工况循环100次。

按运行顺序，各工况示意如图7~图10所示。

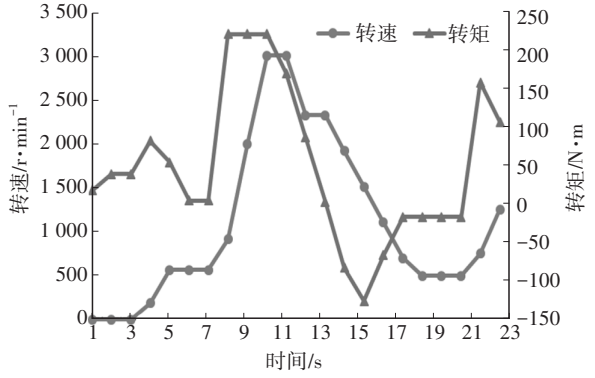

图7　工况3

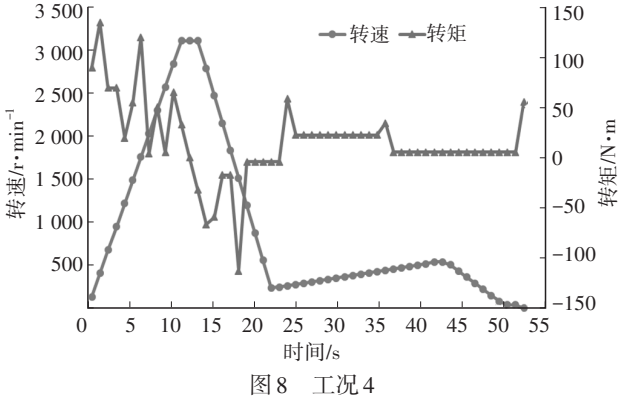

图8　工况4

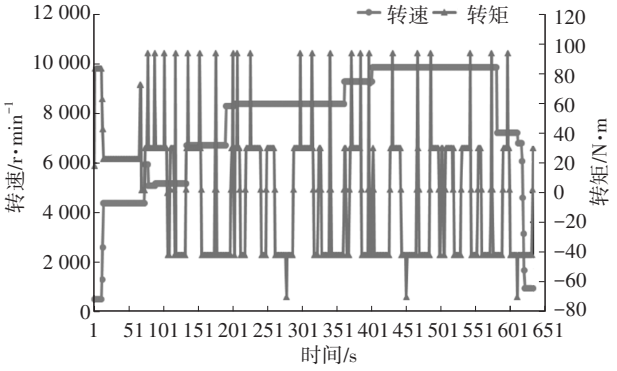

图9　工况2

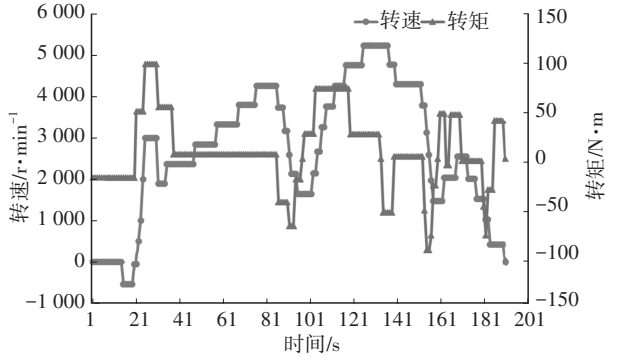

图10　工况1

目前，电驱系统普遍使用的循环耐久试验工况包括定转速定转矩运行，不同转速、转矩下的持续工况循环运行，以及交变工况的循环运行。采用本文提出的方法构建的工况与当前普遍采用的工况相比，具有以下优点：

a. 工况由用户数据提取和处理获得，且各工况有对应的含义，能更准确地反映用户实际驾驶时的工况信息。

b. 由于采用了各工况中大数据计算得到的损伤值偏大的用户数据作为代表片段，通过本文方法计算出的工况实际用时将少于目前工况用时，有利于试验周期的压缩。

7　结束语

本文通过对提取后的用户大数据进行主成分分析、K均值聚类、马尔可夫排序，构建了基于用户大数据的可靠性试验工况，与当前采用的试验参考工况相比，更贴合用户实际驾驶状态。

参　考　文　献

[1] 石琴, 仇多洋, 吴靖. 基于主成分分析和FCM聚类的行驶工况研究[J]. 环境科学研究, 2012, 25(1): 70–76.
SHI Q, QIU D Y, WU J. Research on Driving Cycles Based on Principal Component Analysis and Fuzzy C-Means

Clustering[J]. Research of Environmental Sciences, 2012, 25(1): 70-76.

[2] 徐宗煌, 林瑶. 基于主成分和聚类分析的汽车行驶工况构建[J]. 宁夏大学学报(自然科学版), 2021, 42(3): 270-276.
XU Z H, LIN Y. Construction of Vehicle Driving Cycle Based on Principal Component Analysis and Cluster Analysis[J]. Journal of Ningxia University (Natural Science Edition), 2021, 42(3): 270-276.

[3] 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[J]. 汽车工程, 2011, 33(3): 256-261.
SHI Q, ZHENG Y B, JIANG P. A Research on Driving Cycle of City Roads Based on Microtrips[J]. Automotive Engineering, 2011, 33(3): 256-261.

[4] 段宇帅. 基于主成分分析与K-means聚类的汽车行驶工况构建[J]. 软件导刊, 2022, 21(5): 175-180.
DUAN Y. Construction of Vehicle Driving Cycle Based on Principal Component Analysis and K-Means Clustering[J]. Software Guide, 2022, 21(5): 175-180.

[5] SEN M K, DASGUPTA U. Hyperrelations and Generalized Hypergraphs[J]. International Journal of Machine Learning and Cybernetics, 2013, 4(5): 565-574.

[6] 尹安东, 赵韩, 周斌, 等. 基于行驶工况识别的纯电动汽车续驶里程估算[J]. 汽车工程, 2014, 36(11): 1310-1315.
YIN A D, ZHAO H, ZHOU B, et al. Driving Range Estimation for Battery Electric Vehicles Based on Driving Cycle Identification[J]. Automotive Engineering, 2014, 36(11): 1310-1315.

[7] 姜平, 石琴, 陈无畏. 聚类和马尔科夫方法结合的城市汽车行驶工况构建[J]. 中国机械工程, 2010, 21(23): 2893-2897.
JIANG P, SHI Q, CHEN W W. Driving Cycle Construction Method of City Motors Based on Clustering Method and Markov Process[J]. China Mechanical Engineering, 2010, 21(23): 2893-2897.

[8] 曹骞, 李君, 刘宇, 等. 基于大数据和马尔科夫链的行驶工况构建[J]. 东北大学学报(自然科学版), 2019, 40(1): 77-81.
CAO Q, LI J, LIU Y, et al. Construction of Driving Cycle Based on Big Data and Markov Chain[J]. Journal of Northeastern University (Natural Science), 2019, 40(1): 77-81.

[9] 龚文轩. 乘用车行驶工况模型构建方法研究: 以福州市为例[D]. 南昌: 南昌航空大学, 2021.
GONG W X. Research on Construction Method of Typical Driving Cycle of Passenger Car: A Case Study of Fuzhou City[D]. Nanchang: Nanchang Hangkong University, 2021.

[10] 曹骞, 李君, 刘宇, 等. 基于马尔科夫链的长春市乘用车行驶工况构建[J]. 吉林大学学报(工学版), 2018, 48(5): 1366-1373.
CAO Q, LI J, LIU Y, et al. Construction of Driving Cycle Based on Markov Chain for Passenger Car in Changchun City[J]. Journal of Jilin University (Engineering and Technology Edition), 2018, 48(5): 1366-1373.

（责任编辑　王　一）
修改稿收到日期为4月6日。