

地铁车辆车轮轮缘异常磨损检测中的箱线图算法和改进孤立森林算法

习佳星 沈 钢 许承焯

(同济大学铁道与城市轨道交通研究院, 201804, 上海//第一作者, 硕士研究生)

摘 要 对某地铁列车48个车轮的实测廓形监控数据进行了批量分析,以轮缘磨损面积变化速率和轮缘根部滚动圆半径差之半变化速率为分析指标,分别采用箱线图算法和改进孤立森林算法进行磨损异常检测。箱线图算法具备较好的抗干扰能力,可得到对各指标单独检测的客观统计结果;改进孤立森林算法运行效率得到提高。经对比,2种算法得到的磨损异常检测结论较一致,验证了两种算法的可行性。

关键词 地铁车辆; 轮缘磨损检测; 孤立森林算法; 箱线图算法

中图分类号 U270.331+.1

DOI:10.16037/j.1007-869x.2022.12.023

Boxplot Algorithm and Improved Isolated Forest Algorithm in Metro Vehicle Wheel Flange Abnormal Wear Detection

XI Jiaxing, SHEN Gang, XU Chengzhuo

Abstract The measured profile monitoring data of 48 wheels of a metro train is analyzed in batches. Taking the change rate of the flange wear area and the half change rate of the rolling circle radius difference at the flange root as analysis indexes, the boxplot algorithm and the improved isolated forest algorithm are used to detect the wear anomaly respectively. The boxplot algorithm has good anti-interference ability and can obtain objective statistical results of independent detection of each index. While the operation efficiency of the improved isolated forest algorithm is elevated. By comparison, the wear anomaly detection results obtained by the two algorithms are consistent, verifying the feasibility of both.

Key words metro vehicle; flange wear detection; isolated forest algorithm; boxplot algorithm

Author's address Institute of Rail Transit, Tongji University, 201804, Shanghai, China

轮缘是影响列车轮对导向和防止脱轨的关键部位。轮缘是否存在异常磨损对线路运营安全至关重要。当前对轮缘异常磨损的研究较少:文献

[1]对深圳轨道交通9号线轮缘严重磨损问题进行研究,通过比较左右侧车轮轮缘磨损量,发现轮对明显磨损不均匀;文献[2]对上海轨道交通4号线列车运营期内的轮缘万km磨损量进行计算,发现其显著高于与其部分共线的3号线列车车轮轮缘磨损量;文献[3]对广州轨道交通3号线频繁镟轮的现象进行分析,发现一段范围内轮缘厚度磨损速率远超出正常值。而轮缘异常磨损研究尚缺少对轮缘磨损监控数据挖掘分析的有效方法。

轮缘磨损异常值检测常采用统计学方法、基于距离的方法和基于树的方法等。箱线图法是一种统计学方法,对数据分布类型没有限制,抗干扰性好,其计算结果相对客观^[4]。孤立森林法是一种基于树的方法,没有利用距离或密度测量,具有简单、高效的优点^[5]。本文基于某地铁自动化采集设备采集的整列列车48个车轮的实测廓形数据,分别采用箱线图算法和孤立森林算法进行轮缘磨损异常检测。

1 轮缘磨损检测指标的选择

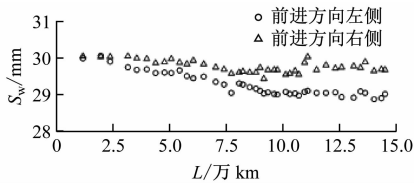
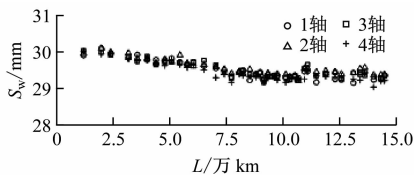
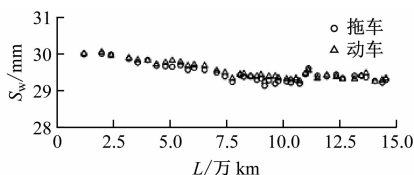
1.1 实测廓形数据分析

基于某地铁列车的实测廓形数据,对轮缘厚度 S_w 按左右、轴位及动拖车分别进行计算分析,得到镟修后列车运行里程 L 为1.2万km至14.5万km时, S_w 在各维度下的磨损情况,如图1—图3所示。

由图1—图3可知,该列车轮缘存在偏磨,但不同轴位、不同车辆的磨损差异不明显。可见,仅依靠轮缘厚度难以判断轮缘是否存在异常磨损。

1.2 基于轮缘磨损规律选择指标

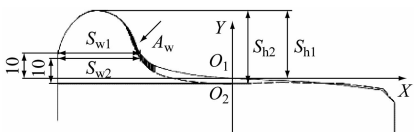
轮缘根部是轮对自导向能力的关键部位,其磨损是导致轮缘磨损的起因。如果轮缘根部发生磨损,其轮径差减小会导致轮对自导向能力减弱,当车辆通过曲线时为防止脱轨,轮缘会与钢轨内轨面

图1 左、右侧的车轮 S_w 平均值变化Fig. 1 Change of S_w average value of left and right wheels图2 不同轴位的车轮 S_w 平均值变化Fig. 2 Change of S_w average value with different axle positions图3 动车及拖车的车轮 S_w 平均值变化Fig. 3 Change of S_w average value of power car and trailer wheels

接触并产生轮缘侧磨。由此,本研究基于轮缘根部和侧面磨耗规律选择指标。

轮缘根部发生磨耗后,在相同的横移量下轮对所能产生的轮径差减小,自导向能力减弱,车轮轮缘区段会与钢轨产生磨耗。因此,本文拟定以轮缘根部的轮径差变动量来衡量轮缘根部的磨耗。

现阶段,相关单位将 S_w 作为轮缘侧磨衡量指标。 S_w 测量以名义滚动圆接触点作为基准。当踏面名义滚动圆处存在磨耗时,可能出现图4所示情况,即磨耗廓形 S_{w2} 与新廓形 S_{w1} 相等。此外,当名义滚动圆处磨耗速率大于轮缘侧磨速率时,可能表现出轮缘“假增厚”。因此,本文选用磨耗面积 A_w 来衡量轮缘侧磨(图4阴影)。将磨耗廓形与参考廓形作对比, A_w 能直接表示轮缘的侧磨量。

注: S_{h1} 表示新廓形的轮缘高度, S_{h2} 表示磨耗廓形的轮缘高度。图4 S_w 测量与 A_w Fig. 4 S_w measurement and A_w

由上述轮缘磨耗规律,本研究以轮缘根部轮径差之半(横移量为12 mm)变化速率 R_f 和轮缘磨耗面积(法向磨耗面积,且距轮背横向距离为20~30 mm区段)变化速率 R_a 为指标,对轮缘磨耗进行检测辨识。

2 基于箱线图算法的磨耗检测

2.1 箱线图相关概念

文献[6]于1977年提出经典的箱线图理论。箱线图主要由最小值、下四分位数 Q_1 、中位数 Q_2 、上四分位数 Q_3 和最大值5个数值点组成, Q_1 与 Q_3 的差值为四分位距 I_{QR} 。本研究将样本数据中大于 $Q_3 + 1.5 I_{QR}$ 或小于 $Q_1 - 1.5 I_{QR}$ 的值定义为异常值。

2.2 算法流程及结果分析

针对本研究中的监控数据分别计算在连续里程区段内各车轮轮缘磨耗速率的变化情况,识别出存在异常磨耗的车轮及其里程区段。以 R_a 为例,假设整列车各车轮轮缘侧磨面积为 $A_{w1}, A_{w2}, A_{w3}, \dots, A_{wn}$, n 为列车车轮样本总数。用 $A_{wi,1}, A_{wi,2}, A_{wi,3}, \dots, A_{wi,t}$ 表示第 i 个车轮轮缘磨耗面积 A_{wi} 下的 t 个有序趋势值,设 l 为时间序列观察窗口长度($l < t$), ΔL 为观察窗口长度内的列车运行里程, $L_{p,t}$ 为 t 时刻各分位数所在样本数据中的位置。箱线图检测识别算法的具体流程如下:

1) 计算得到各车轮在观察窗口长度 l 内的轮缘磨耗面积变化值 $\Delta A_{wi,t}$ 和 t 时刻磨耗速率 $R_{A_{wi,t}}$:

$$\Delta A_{wi,t} = A_{wi,t} - A_{wi,t-1}, i = 1, 2, 3, \dots, n \quad (1)$$

$$R_{A_{wi,t}} = \Delta A_{wi,t} / \Delta L, i = 1, 2, 3, \dots, n \quad (2)$$

2) 计算得到 t 时刻样本数据内各分位数所处位置 $L_{p,t}$ 及数值 $Q_{p,t}$:

3) 计算得到 t 时刻样本数据内的 R_a 最大值 $Q_{u,t}$:

4) 判断 $A_{wi,t} \leq Q_{u,t}$ 是否成立,若不成立则标记为异常值。

5) 更改时刻值 $t = t + l$,并重复1)~4),完成对各个时刻测量值的异常状态辨识。

图5为箱线图检测识别算法的流程图。

根据箱线图算法,将 R_a 换成 R_f ,同样可以计算并观察连续里程区段内 R_f 的波动大小及异常状况。

本文分别计算 L 为0~10.7万km的车轮 R_a 和 R_f 的监控数据分析结果。

选取典型的存在异常磨耗和正常磨耗车轮的监控结果,如图6及图7所示。从图6a)中可知,当

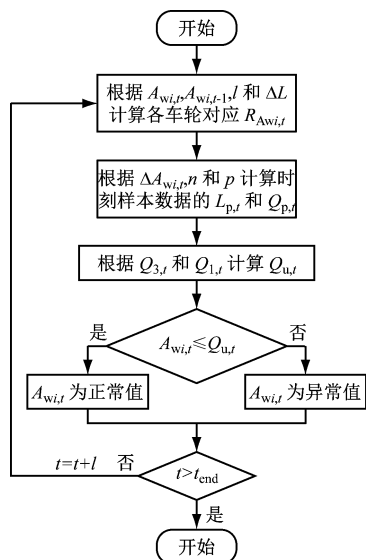
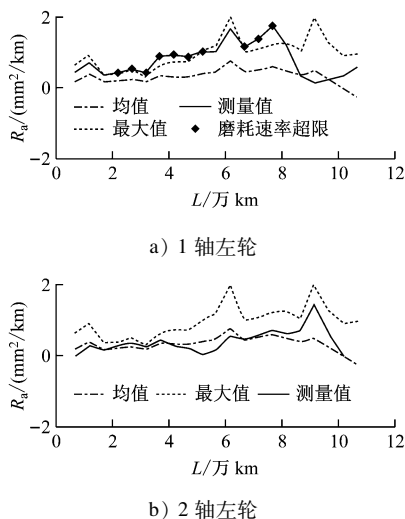


图 5 箱线图检测识别算法

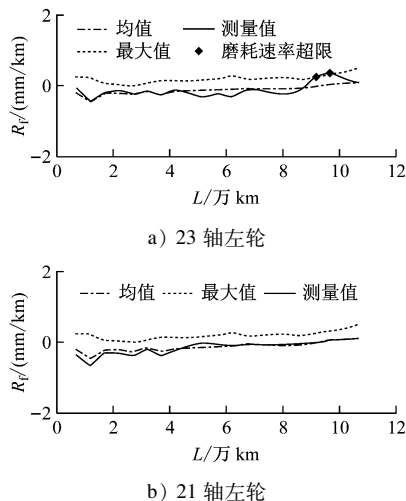
Fig. 5 Diagram of boxplot detection and recognition algorithm

图 6 R_a 检测识别Fig. 6 Detection and identification of R_a

L 为 1.7 万~5.7 万 km 及 6.2 万~8.2 万 km 时,该车轮 R_a 超出最大值,识别出该车轮存在异常磨耗。同理,由图 7 a) 可识别,当 L 为 8.7 万~10.2 万 km 时,该车轮存在异常磨耗。对比图 6 b) 及图 7 b) 中的正常磨耗车轮可见,各车轮轮缘磨耗速率基本在其磨耗速率均值曲线附近小幅度波动。

由图 6 及图 7 可知,基于箱线图的轮缘磨耗检测识别方法能辨识出车轮存在异常磨耗的状况,客观地得到了各指标的检测识别结果。

L 为 1.7 万~10.7 万 km 时,箱线图算法有效辨识出的轮缘异常磨耗情况统计结果见表 1。

图 7 R_f 检测识别Fig. 7 Detection and identification of R_f

$L = 10.5$ 万 km 时,整列车的轮缘法向磨耗面积均值、轮缘根部的轮径差之半均值分别为 8.18 mm^2 、 0.95 mm ;表 1 中 1 号、2 号、28 号、34 号、41 号、42 号、46 号和 47 号车轮的轮缘法向磨耗面积分别为 19.71 mm^2 、 14.43 mm^2 、 8.94 mm^2 、 13.77 mm^2 、 9.57 mm^2 、 9.58 mm^2 、 8.64 mm^2 和 8.30 mm^2 ,3 号、23 号、34 号、38 号、41 号、42 号、43 号、47 号和 48 号车轮的轮缘根部轮径差之半值分别为 2.41 mm 、 2.64 mm 、 2.58 mm 、 1.72 mm 、 2.46 mm 、 1.34 mm 、 1.68 mm 、 1.37 mm 和 1.31 mm ,都明显高于平均水平,表现出磨耗异常。这与实际磨耗情况基本一致。

3 基于孤立森林算法的磨耗检测

3.1 孤立森林相关概念

周志华教授等于 2008 年在第八届 IEEE 数据挖掘国际会议上提出孤立森林理论^[7],提出异常数据可基于路径长度被检测出来。二叉搜索树的平均路径长度 $c(n)$ 为:

$$c(n) = 2h(n-1) - 2(n-1)/n \quad (3)$$

其中 n 为样本个数; $h(i)$ 为调和数,该值可被估计为 $h(i) = \ln(i) + 0.577\ 215\ 664\ 9$ (Euler 常数)。 $c(n)$ 用来标准化样本 x 的路径长度 $h(x)$ 。

异常分值 $s(x, n)$ 用来判断数据异常的程度,定义如下:

$$s(x, n) = 2^{-E(h(x))/c(n)} \quad (4)$$

其中 $E(h(x))$ 为样本 x 在一群孤立树中的路径长度的期望。

表 1 箱线图算法的磨耗异常检测计算结果

Tab. 1 Wear anomaly detection calculation results of boxplot algorithm

指标	车轮号	$L/\text{万 km}$	$Q_{u,i}$ 最大值
R_a	1	1.7 ~ 5.7	0.62
	1	6.2 ~ 8.2	1.22
	2	2.7 ~ 5.7	0.69
	28	8.2 ~ 9.2	1.37
	34	8.7 ~ 10.2	1.47
	41	7.2 ~ 8.2	1.17
	42	1.7 ~ 4.7	0.50
	46	4.7 ~ 5.7	0.95
R_f	47	4.7 ~ 5.7	0.95
	3	8.7 ~ 9.7	0.251
	23	8.7 ~ 10.2	0.286
	34	8.7 ~ 9.7	0.251
	38	4.7 ~ 5.7	0.166
	41	10.2 ~ 10.7	0.434
	42	10.2 ~ 10.7	0.434
	43	10.2 ~ 10.7	0.434
	47	10.2 ~ 10.7	0.434
	48	10.2 ~ 10.7	0.434

3.2 孤立森林算法的改进

孤立森林算法在构建孤立树的过程中,存在分割数据随机性较强的问题。对此,本文改进了孤立森林算法:先分析采样数据,判断此样本集是否适合构造孤立树,以避免随机选择的根节点中包含较多没有离群点的样本集;随后,在构造孤立树时,用特定的切割点将孤立树分成左右子树。

对于数据集 m ,随机选择 j 个样本点作为孤立树根节点样本,再随机选择其中一维作为切割属性。由数据的分布特性可知,样本点中超过上界值的概率很低。因此,若有数据点落在此区域外,则所选根节点样本中包含异常点的可能性很大。如果样本集中的最大值 $\max(j)$ 大于这个上界值,则将 j 样本放入树的根节点,否则构建为 1 棵空树。

在第一次选择切割点时,取根节点样本数据中相应切割属性下的上界值作为切割点。在下一个子空间选择切割点时,则以该子样本数据最大值 z_{\max} 与最小值 z_{\min} 之间的黄金分割点作为切割点。递归上述过程直到当前子树只包含 1 个数据点或达到最大限制的树高。此树定义为孤立树。

孤立森林算法的改进去除了可能含有干扰属性的孤立树,加快了迭代,提高了运行效率及稳定性。

3.3 改进算法流程及结果分析

结合 R_a 和 R_f 综合计算各车轮轮缘磨耗速率的异常得分 $s(j,m)$ 。具体流程如下:

步骤 1:初始化孤立森林。设置孤立树分叉的最大限制高度。

步骤 2:初始化生成孤立树算法参数。输入数据集 m ,该数据集存入的是镟修后 10.7 万 km 里程范围内 48 个车轮的轮缘磨耗速率值,具有 R_a 和 R_f 2 个指标维度。生成孤立树的总数即是数据集的采样次数为 T ,采样大小为 j 。样本中的 2 个指标维度即代表孤立树分叉过程中的 2 种切割属性。

步骤 3:判断当前磨耗速率样本数据中的最大值 $\max(j) >$ 上界值(由样本数据的分布特性确定)是否成立。若不成立,则舍弃该样本集,重新选取 1 个样本集。

步骤 4:随机选择 1 个指标维度作为切割属性,选择相应上界值作为初次切割点 p ,在选取的指标维度下对磨耗速率样本集中的数据进行比较,磨耗速率 $\geq p$ 的放在右子树,磨耗速率 $< p$ 的放在左子树。

步骤 5:判断样本集的分割是否达到最大限制高度或者当前子样本中是否只有 1 个磨耗速率数据点。若是,则完成对样本集的 1 次计算,并且开始构建下一个样本集。

步骤 6:递归构造孤立子树,即不断分叉。左子树(即磨耗速率数据的子样本集)选择数据最大值与最小值之间的 0.618 比例处(即 $z_{\min} + 0.618 \times (z_{\max} - z_{\min})$)为切割点;右子树相应选择 0.382 比例处($z_{\min} + 0.382 \times (z_{\max} - z_{\min})$)作为切割点。选择黄金分割点可减小切割随机性,能让孤立树在生成子空间时每次都切割为与样本父节点均等比例大小的左右子树^[5];并且 2 个分割点比例之和为 1,可使得构造子树时迭代速度加快,且具有一定稳定性^[7]。

步骤 7:重复步骤 6,完成每棵孤立树的构建。

步骤 8:重复步骤 3—步骤 7,遍历所有孤立树。设定检测阈值 S_m ,计算得到 $s(j,m)$,将磨耗速率得分超过阈值的数据点确定为异常点。

图 8 为改进后的孤立森林算法流程图。采用改进后的孤立森林算法对 $L < 10.7$ 万 km 的列车车轮

轮缘磨耗情况进行计算。

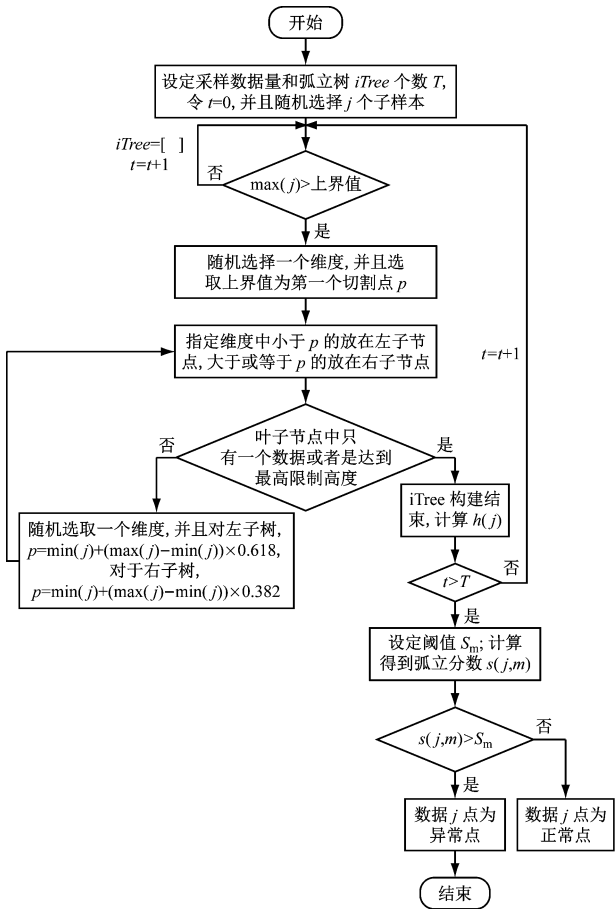


图 8 改进后的孤立森林检测识别算法

Fig. 8 Improved isolated forest detection algorithm

表 2 为 $L < 10.7$ 万 km, 选用不同 S_m 时, 改进后的孤立森林算法对存在异常磨耗的车轮及其里程区段的辨识结果。和表 1 对照的误报情况见表 3。其中误报车轮数是将正常磨耗车轮监测为异常磨耗车轮的个数, 总误报次数为误报车轮数与非连续区段对应误报次数乘积的总和。

由表 3 可知, 改进后的孤立森林算法所得异常检测结论与箱线图算法所得结论较一致, 两种算法检测出来的存在异常磨耗的车轮及其里程区段结果基本一致, 验证了两种算法应用于地铁车轮轮缘磨耗异常检测的可行性。

由表 2 及表 3 还可看出: 当 S_m 为 $0.820 \sim 0.831$ 时, 能辨识出存在异常磨耗的车轮及其里程区段, 存在 $1 \sim 3$ 个车轮误报的情况; 当 S_m 减小至 0.800 时, 总误报次数相对增加; 当 S_m 为 $0.832 \sim 0.930$ 时, 出现不能辨识出车轮存在异常磨耗的状况。可见, 过高的阈值降低了算法辨识的准确性。

当 $S_m = 0.831$ 时, L 分别为 1.97 万 km、 4.01 万

表 2 不同 S_m 下的异常辨识结果

Tab. 2 Abnormal identification results with different S_m

车轮号	$L/\text{万 km}$	辨识结果						
		$S_m = 0.800$	$S_m = 0.820$	$S_m = 0.825$	$S_m = 0.831$	$S_m = 0.832$	$S_m = 0.875$	$S_m = 0.930$
1	1.7 ~ 5.7	✓	✓	✓	✓	✓	✓	✓
1	6.2 ~ 8.2	✓	✓	✓	✓	✓	✓	✓
2	2.7 ~ 5.7	✓	✓	✓	✓	✓	✓	✓
3	8.7 ~ 9.7	✓	✓	✓	✓	✓	✓	✓
23	8.7 ~ 10.2	✓	✓	✓	✓	✓	✓	✓
28	8.2 ~ 9.2	✓	✓	✓	✓	✓	✓	✓
34	8.7 ~ 10.2	✓	✓	✓	✓	✓	✓	✓
38	4.7 ~ 5.7	✓	✓	✓	✓	✓	×	×
41	7.2 ~ 8.2	✓	✓	✓	✓	✓	✓	✓
41	10.2 ~ 10.7	✓	✓	✓	✓	✓	✓	✓
42	1.7 ~ 4.7	✓	✓	✓	✓	✓	✓	✓
42	10.2 ~ 10.7	✓	✓	✓	✓	×	×	×
43	10.2 ~ 10.7	✓	✓	✓	✓	✓	✓	✓
46	4.7 ~ 5.7	✓	✓	✓	✓	✓	✓	✓
47	4.7 ~ 5.7	✓	✓	✓	✓	✓	✓	✓
47	10.2 ~ 10.7	✓	✓	✓	✓	✓	✓	×
48	10.2 ~ 10.7	✓	✓	✓	✓	✓	✓	✓

注: ✓表示能检测出某个区段里程内某个车轮磨耗异常, 并且经过对比, 异常检测结果是一致的; ×表示未能检测出磨耗异常。

表 3 不同 S_m 下的异常辨识误报情况统计结果

Tab. 3 Statistical results of abnormal identification false alarms with different S_m

S_m	误报车轮数	总误报次数	误报车轮占车轮总数比/%
0.800	4	4	8.33
0.820	3	3	6.25
0.825	2	2	4.17
0.831	1	1	2.08
0.832	1	1	2.08
0.875	1	1	2.08
0.930	1	1	2.08

km 和 9.35 万 km 时的轮缘磨耗检测识别结果如图 9 所示。由图 9 a) 可知, $S_m = 0.831$, $L = 1.97$ 万 km 时, 1 号、42 号轮被辨识为异常点, 且明显与同列车其余位置的车轮区分开来; 图 9 b) 中的孤立分数结果显示, 1 号、42 号轮的 $s(j, m)$ 值高于 S_m 。分析图 9 c) — 图 9 f), 同样可得类似结论。这说明改进后的孤立森林算法能够较好地辨识出存在异常磨耗的车轮及其里程区段, 得到相应的综合检测识别结果。

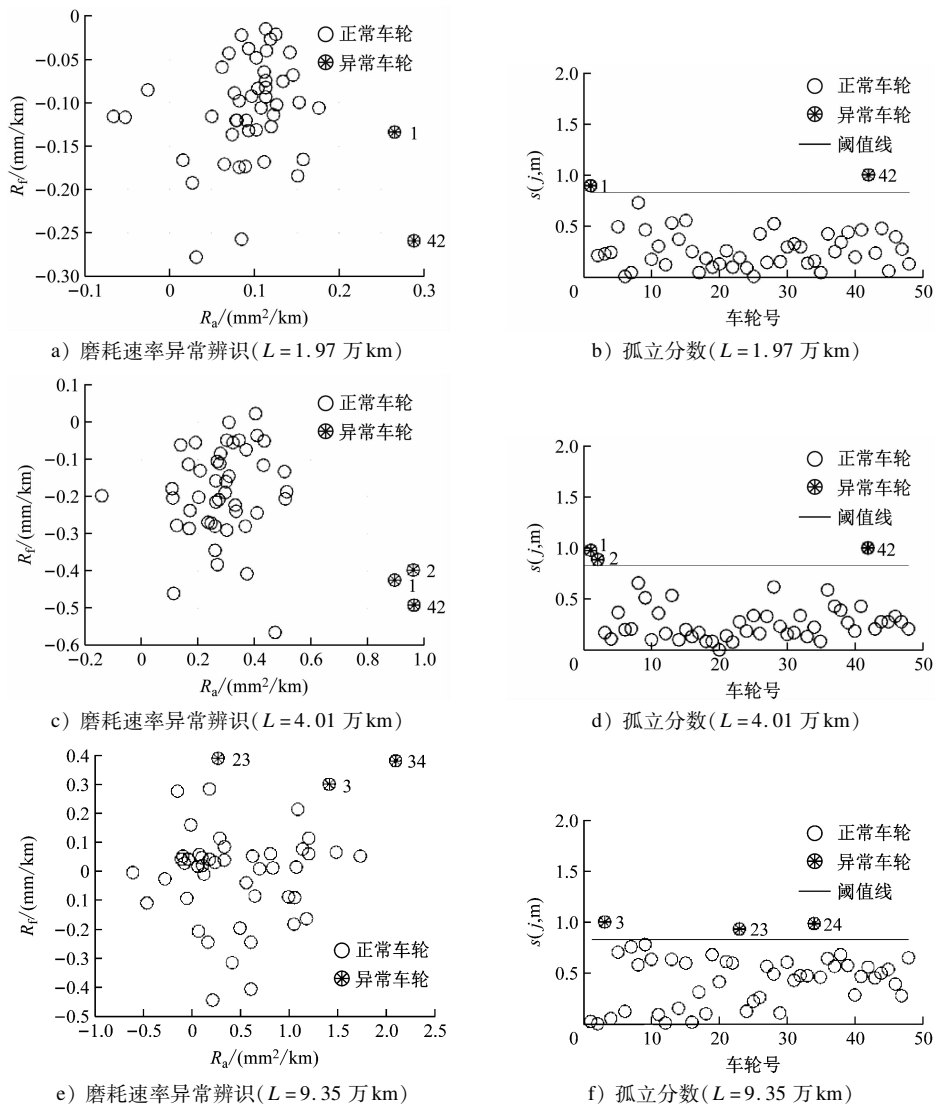


图9 轮缘磨损异常辨识
Fig. 9 Abnormal identification of flange wear

4 结语

本文基于某地铁列车48个车轮的实测廓形数据,结合轮缘磨损面积变化速率和轮缘根部轮径差变化速率,对大量实测数据进行了有效分析和处理,提出了两种能检测和辨识出存在异常磨损的车轮及对应里程区段的数据挖掘的方法。

箱线图法具有好的抗干扰能力,对数据类型没有限制。本文通过箱线图算法完成了对指标的单独检测,得到客观的统计结果,能很好地辨识出异常磨损的车轮及其对应里程。

本文对孤立森林算法中生成孤立树及分割成左右子树的过程进行了优化改进,提高了计算效率。通

过改进后的孤立森林算法完成了对指标的综合检测。经比较,改进后的孤立森林算法与箱线图法可以得到较一致的磨损异常检测结论,验证了两种方法的可行性。此外,在改进后的孤立森林算法中,阈值设为0.820~0.831时的检测可靠性最好。

参考文献

- [1] 李涛,刘志远,赵卓,等.城市轨道交通车辆车轮轮缘严重磨损分析[J].城市轨道交通研究,2018(11):74.
LI Tao, LIU Zhiyuan, ZHAO Zhuo, et al. Analysis on seriously wheel flange wear of urban rail transit vehicle[J]. Urban Mass Transit, 2018(11):74.

(下转第137页)