计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):228−232 [doi: 10.15888/j.cnki.csa.007479]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 PSO_RF 双向特征选择和 LightGBM 设备故障检测①

韩金鹏[1,2], 李冬梅[2], 王　嵩[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
通讯作者: 韩金鹏, E-mail: hjp3701@163.com

摘　要: 仪器共享平台的发展提高了各高校的仪器设备的使用率. 但是在设备的使用过程中, 对设备的故障检测方面还没有得到改善. 针对上述问题, 本文收集了医用影像设备的相关数据, 采用 PSO_RF 的双向特征选择方法进行特征选择, 然后构建了基于 LightGBM (Light Gradient Boosting Machine) 的故障检测模型, 并将其应用于医用影像设备的故障检测中. 通过标准评价体系的建立及不同模型对故障诊断结果的对比, 相对于传统的机器学习算法, 该模型在故障检测的精确率、召回率、*F*1 值等评价指标上有较好的表现, 对于加快仪器故障点的发现以及提高仪器利用率具有积极推进作用.

关键词: 设备故障检测; 特征选择; LightGBM

引用格式: 韩金鹏,李冬梅,王嵩.基于 PSO_RF 双向特征选择和 LightGBM 设备故障检测.计算机系统应用,2020,29(7):228–232. http://www.c-s-a.org.cn/1003-3254/7479.html

Device Fault Detection Based on PSO_ RF Bidirectional Feature Selection and LightGBM

HAN Jin-Peng[1,2], LI Dong-Mei[2], WANG Song[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: The development of the instrument sharing platform has increased the utilization rate of instruments and equipment in various universities. However, during the use of the equipment, the fault detection of the equipment has not been improved. In view of the above problems, this study collected relevant data of medical imaging equipment, adopted the two-way feature selection method of PSO_RF for feature selection, then built a fault detection model based on LightGBM (Light Gradient Boosting Machine), and applied it to the fault detection of medical imaging equipment. Through the establishment of the standard evaluation system and the comparison of fault diagnosis results by different models, compared with the traditional machine learning algorithm, this model has a better performance in the accuracy rate, recall rate, *F*1 value and other evaluation indicators of fault detection, which has a positive role in accelerating the discovery of instrument fault points and improving the utilization rate of instruments.

Key words: equipment failure detection; feature selection; LightGBM

1　引言

随着科学技术的发展, 如何降低设备的维护维修成本成为非常关键的问题. 一些贵重精密仪器, 对于外部环境等因素十分敏感, 对于传统的人工故障检测, 存在维修周期长、费用贵、代价高的问题, 导致设备的利用率偏低. 仪器共享平台中存储了医用影像设备使

① 收稿时间: 2019-12-02; 修改时间: 2019-12-27; 采用时间: 2020-01-07; csa 在线出版时间: 2020-07-03

用时的各种信息, 但是这些并没有得到充分利用, 只是用于存档查询等功能, 如果利用系统中的设备信息, 找到仪器故障点以及引发仪器发生故障的原因, 那么就能为设备维护维修提供更有针对性的方向, 对缩短维修周期、减少维修费用以及提高仪器利用率具有积极意义.

关于故障检测问题, 为提升故障检测模型的性能, 许多专家学者尝试了多种方法. 文献[1,2]使用贝叶斯网络分别对电子设备、电流互感器进行故障检测, 有效地提高了故障检测的效率与准确率. 文献[3]使用基于 XGBoost 算法对变压器故障进行检测判别, 结果表明 XGBoost 分类效果最优. 文献[4,5]采用了 SVM 的故障检测模型分别用于轴承故障检测与钢丝绳诊断中, 验证该模型可以提高故障检测的准确率. 文献[6]使用改进粒子群算法优化 SVM 应用于煤层气单井系统故障诊断中, 验证了该算法的精度优于普通粒子群算法和遗传算法. 文献[7]中 P. Arpaia 等人使用隐马尔可夫模型检测流体机械故障, 该方法的准确性在 CERN 的螺杆压缩机上得到了验证. 上述文献大多单个分类器, 存在分类精度低、泛化能力弱等问题. 本文采用集成学习模型, 把单一的学习模型结合起来获得更准确的结果以及更好的泛化性能. 随机森林是一种集成学习模型, 具有很好的预测精度, 近年来已被广泛应用于特征选择领域. 特征选择算法有过滤式、包裹式及嵌入式, 考虑到本题的数据量及维度不大, 选择包裹式特征选择策略. 基于 Boosting 策略的 LightGBM 算法具有训练速度快、准确率高、支持并行学习等特点[8].

综上所述, 本文采用一种粒子群算法优化随机森林 (PSO_RF) 的双向特征选择和 LightGBM 的故障检测方法. 利用随机森林算法对原始数据进行双向特征选择, 然后将数据输入 LightGBM 模型中训练. 使用 LightGBM 不但大大加快训练速度, 还能保证不错的准确率. 将此模型应用于医学影像设备的故障原因检测中, 对加快检测故障点的发现具有重要意义.

2 算法描述

2.1 LightGBM

GBDT(梯度提升树), 通过多轮迭代, 来不断提高最终分类器的精度.

在 GBDT 到的迭代中, 假设我们前一轮得到的强学习器是$F_{t-1}(x)$, 损失函数是$L(y,F_{t-1}(x))$, 那么本轮的目标是找到一个弱学习器$h_t(x)$, 使得本轮的损失函数最小, 本轮的损失函数为:

$$h_t(x)=\arg\min_{h\in H}\sum L(y,F_{t-1}(x)+h(x)) \tag{1}$$

然后计算损失函数的负梯度, 来拟合本轮损失的近似值. 损失函数的负梯度表示为:

$$r_{ti}=-\frac{\partial L(y,F_{t-1}(x_i))}{\partial F_{t-1}(x_i)} \tag{2}$$

通常使用平方差来近似拟合$h_t(x)$:

$$h_t(x)=\arg\min_{h\in H}\sum (r_{ti}-h(x))^2 \tag{3}$$

从而本轮的强学习器如下:

$$F_t(x)=h_t(x)+F_{t-1}(x) \tag{4}$$

LightGBM 在 GBDT 算法的基础上进行了改进. 改进的地方主要包含两个方面: Gradient-based One-Side Sampling (GOSS) 和 Exclusive Feature Bundling (EFB).

基于梯度的单面采样 (GOSS). 根据信息增益的定义, 具有较大梯度的那些实例对信息增益会做出更大的贡献, 梯度小的样本进行进一步的学习对改善结果精度帮助并不大. 因此, 为保持信息增益估计的准确对数据实例进行采样时, 应该更好的保留那些具有较大贡献的实例, 删除一部分具有小梯度的样本.

互补特征压缩 (EFB), 是一种可以减少高维数据的特征数目并且使损失最小的一种算法. 这里不是使用所有特征来获得最佳分割点, 而是将某些特征合并到一起降低特征维度来使寻找最优分割点的消耗减少. LightGBM 关于互斥特征的合并用到了直方图 (Histogram) 算法[9]. EFB 通过捆绑合并相互独立的特征, 来减少特征的数量. 这样既降低了内存占用, 又降低了时间复杂度.

LightGBM 拥有更快的训练效率、更高的准确率、更低的内存使用, 并且可支持并行化学习, 故本文在基于 PSO_RF 的双向特征选择的过程中, 使用 LightGBM 根据混淆矩阵计算当前特征子集的精确度, 保证选出最优特征子集. 将经过特征选择方法处理后的数字 X 线摄影机故障数据输入 LightGBM 中训练模型, 通过交叉验证的方法来验证本文所采用方法的准确性, 进而说明本文方法对于 X 线摄影机的故障检测具有积极的推动作用.

2.2 基于 PSO_RF 的双向特征选择算法

特征重要性指特征对于样本预测结果的影响, 好的特征能够提高模型预测的准确率及性能.

随机森林 (Random Forest, RF), 是一种基于决策树的算法. 随机森林在建立决策树的过程中做了一些改进[10,11]. 在整个随机森林算法的过程中, 有两个随机过程: 一是使用 Bootstrap 有放回的随机抽取样本来生成决策树; 二是在划分决策树左右子树时, 随机选择节点的部分特征, 然后在随机选取的特征中选取最优特征划分左右子树. 这两个随机过程一定程度上避免了过拟合的出现.

随机森林依据袋外数据误差率准则去度量特征准确性的, 主要思想是: 对一个相关的特征加入噪声后, 那么用仅对此特征进行变化特征之后的数据进行模型训练, 模型预测的准确率将降低; 反之, 如果某个特征是不重要的, 那么重新训练后模型预测准确率变化不大.

粒子群算法 (Particle Swarm Optimization, PSO) 已经被广泛应用于各种优化问题上. 种群中每个粒子代表一个可行解, 但不一定是最优解, 粒子在迭代过程中, 通过学习历史经验来调整自身的速度和位置矢量, 最终求得最优解.

本文采用基于 PSO_RF 的双向特征选择方法. 首先采用随机森林度量特征重要性并将结果进行降序排序, 其中引入粒子群算法对随机森林进行参数寻优. 然后, 从特征的全集开始搜索, 每次从当前特征子集中删除重要性最低的特征, 组成新的特征子集; 然后进行前向选择部分, 使用 LightGBM 根据混淆矩阵对当前特征子集的精确度进行计算, 如果精确度下降, 则回收刚删除的特征, 如此循环直至结束. 这样, 在特征重要性的基础上, 加上当前特征子集分类精度作为评价指标, 能够降低特征波动性, 保证选出的最优特征子集冗余少且不损失分类精度.

算法的具体过程如算法 1(m 为原始数据中特征总数, U 表示数据集中全部特征).

算法 1. 基于 PSO_RF 的双向特征选择算法

输入: 数据集 D, 特征集$C=\{f_i|i=1,2,\cdots,m\}$
输出: 最优特征子集C

1) 使用粒子群优化的随机森林分别计算特征f_i的特征重要性程度I_i;
2) 根据第一步得到的I_i对特征进行降序排序;
3) 使用 LightGBM 进行评估. 对于步骤 2) 中排序后的特征子集进行后向选择, 子集搜索策略遍历特征空间:

while $C\neq\emptyset$
计算a_i; //计算当前特征子集的精确度
$C=C-f_i$; //删除特征f_i
计算a_{tmp}; //计算删除特征后的精确度
if $a_i>a_{tmp}$: //如果新的精确率小于旧的
$C=C+f_i$; //回收之前删除的特征
end if;
end while;

3 实验分析

对于设备故障检测的研究过程主要包括以下几个步骤: 数据采集、数据预处理、特征选择、模型预测、实验结果分析. 设备故障检测模型流程图如图 1 所示.

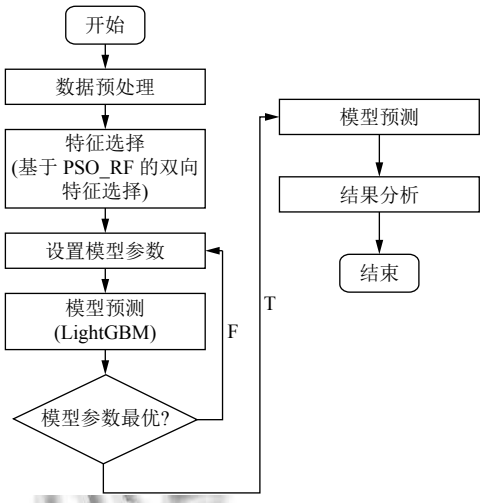

图 1 模型流程图

3.1 数据采集

数据集包括了 1200 条 X 线摄影机的使用信息, 主要包括仪器编号、室内温度、湿度、电压、电压频率、电流等 20 个特征维度. 本实验所用真实数据来自于实验室的自主研发项目“大型仪器共享平台”, 此项目主要是解决高校仪器利用率不高等问题, 目前已服务于多所高校.

3.2 数据预处理

由于原始数据数据不完整, 存在缺失值、重复等问题, 并且数据处理是否得当对训练和预测结果影响非常大. 所以本文对原始数据进行了数据清理、数据变换等操作, 丢弃重复数据, 对缺失值进行删除补齐等操作, 最后对数据进行 MIN-MAX 归一化处理.

3.3 特征选择

好的特征能够提升模型的性能, 降低复杂度, 提高模型的泛化能力, 对模型的正确性和有效性都会有很大的影响.

本文采用基于 PSO_RF 的双向特征选择方法进行特征选择, 对原始数据集中的设备故障有关信息进行特征选择, 为了快速找到模型最优参数, 本文采用了粒子群算法以加快寻找模型的最优参数, 然后通过交叉验证的方法来提高模型的稳定性. 最终选取的 12 个特征被选入最优特征子集, 作为模型的输入变量. 其中包括室内温度、室内湿度、电压、电压频率、电流、地阻、球管温度等.

3.4 模型预测

对于医用设备故障检测, 根据 X 线摄影机故障信息分为电源故障、接线故障、干扰故障、球管故障及其他故障. 本文使用 LightGBM 来建立模型, 通过对原始数据的预处理, 然后通过基于 PSO_RF 的双向特征选择算法进行特征选择, 然后与 CFS 算法进行比较, 验证本文所用特征选择方法的有效性. 将处理过后的 12 个特征变量作为输入变量输入 LightGBM 进行分类预测, 本次实验采用网格搜索算法进行参数寻优, 然后通过交叉验证的方法来计算故障检测模型的分类精度. 本文通过精确率、召回率、$F1$ 值以及运行时间方面, 通过与 GBDT、随机森林模型进行对比分析, 可以得出 LightGBM 在处理该问题上更有优势.

3.5 评价标准

为了评价不同模型之间的性能, 需要统一的评价标准. 本文采用混淆矩阵对模型的优劣进行评价. 混淆矩阵见表 1.

表 1 混淆矩阵

	Positive	Negative
True	*TP*	*FN*
False	*FP*	*TN*

本文采用精确率、召回率及 $F1$ 值作为评价标准. 评价指标的定义如下:

$$precision = \frac{TP}{TP+FP} \tag{5}$$

$$recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1 = \frac{2*precision*recall}{precision+recall} \tag{7}$$

其中, *TP* 表示被模型预测为正的正样本, *FP* 表示被模型预测为正的负样本, *FN* 表示被模型预测为负的正样本, *TN* 表示被模型预测为负的负样本. 另外, 如果样本数据量大, 模型复杂, 计算速度也是评价一个模型优劣的指标. 实验中采用 Python 的 time.clock() 来计算模型的运行时间.

3.6 实验结果分析

经过数据预处理, 然后使用基于 PSO_RF 的双向特征选择算法进行特征选择, 并与 CFS 进行对比, 表 2 是两者的结果对比, 其中, X_{num} 是特征数量.

表 2 不同特征选择算法构建故障检测模型结果对比

算法	X_{num}	精确率	$F1$ 值
CFS	13	0.8626	0.8592
基于 PSO_RF 的双向特征选择	12	0.8973	0.9026

从结果可以看出, 本文采用的基于 PSO_RF 的双向特征选择算法分类精度与 $F1$ 值均优于 CFS, 尤其是 $F1$ 值达到了 90.26%, 比其 CFS 提高了 4.34%, 说明本文采用的特征选择方法是有效的, 并取得了不错的分类效果.

通过上述实验, 最后选出了最优特征子集, 然后用 LightGBM 算法建立故障检测模型, 并采用 10 折交叉验证对数据进行分组训练并测试. 对每组训练集与测试集, 分别采用 LightGBM、GBDT 和随机森林进行训练测试, 并取其平均值作为最终结果, 结果如表 3 所示.

表 3 不同模型对故障诊断结果对比

模型名称	精确率	召回率	$F1$ 值	运行时间 (s)
随机森林	0.8628	0.8958	0.8790	0.3260
GBDT	0.8558	0.9447	0.8980	1.0042
LightGBM	0.9027	0.9261	0.9143	0.2691

从表 3 可以看出, 在精确率方面, LightGBM 的结果高于其他 3 种模型, 精确率达到 90.27%; 在召回率方面, LightGBM 略低于 GBDT, 好于随机森林; 在 $F1$ 值方面, LightGBM 是 3 种模型中结果最好的; 在运行时间方面, LightGBM 的运行时间优于其他两种模型, 提升了计算速度.

综合以上, 本文采用的设备故障检测方法通过基于 PSO_RF 的双向特征选择算法筛选出最优特征子集, 然后在其基础上使用 LightGBM 对设备故障进行分类检测, 在结果上达到了更高的准确率, 并且提高了计算速度, 运行时间保持在 0.3 s 以内, 与文中其他的模型相比具有较大的性能提升.

4 结语

本文以设备故障检测为应用背景, 然后根据设备的数据信息, 提出一种基于 PSO_RF 的双向特征选择和 LightGBM 的故障检测方法并将其应用于 X 线摄影机故障检测中, 取得了不错的效果. 本文中, 对原始数据进行数据清理等预处理后, 采用基于 PSO_RF 的双向特征选择方法进行特征选择, 删除无关特征, 然后利用 LightGBM 训练速度快、精度高等特点, 构建了故障检测模型. 实验结果表明, 该模型在精确率、召回率、*F*1 值以及模型训练时间上有着很好的表现, 可以为设备故障检测提供依据, 对设备的维护维修、提高设备利用率具有积极推动作用.

参考文献

1 樊宁, 高凤岐. 基于贝叶斯网络的电子设备故障诊断技术研究. 仪表技术, 2010, (9): 51–53. [doi: 10.19432/j.cnki.issn1006-2394.2010.09.019]

2 刘柱, 朱永利, 郝宁. 基于贝叶斯网络分类器的电容型电流互感器绝缘故障诊断. 电气时代, 2010, (9): 92–94.

3 孙琛, 田晓声. 基于 XGBOOST 算法的变压器故障诊断. 佳木斯大学学报 (自然科学版), 2019, 37(3): 378–380.

4 吉敏. 基于 PCA-SVM 的轴承故障诊断研究. 电子设计工程, 2019, 27(17): 14–18. [doi: 10.14022/j.cnki.dzsjgc.2019.17.004]

5 黄帅, 吴娟, 李琳琳, 等. 粒子群优化的 SVM 提升钢丝绳故障诊断. 机械科学与技术, 2020, 39(2): 282–287. [doi: 10.13433/j.cnki.1003-8728.20190121]

6 苗玉. 改进粒子群算法优化 SVM 的故障诊断方法研究. 机械工程与自动化, 2019, (6): 14–15, 18. [doi: 10.3969/j.issn.1672-6413.2019.06.005]

7 Arpaia P, Cesaro U, Chadli M, *et al*. Fault detection on fluid machinery using hidden Markov models. Measurement, 2020, 15: 107126. [doi: 10.1016/j.measurement.2019.107126]

8 王思宇, 陈建平. 基于 LightGBM 算法的信用风险评估模型研究. 软件导刊, 2019, 18(10): 19–22.

9 Pu QM, Li YH, Zhang H, *et al*. Screen efficiency comparisons of decision tree and neural network algorithms in machine learning assisted drug design. Science China Chemistry, 2019, 62(4): 506–514. [doi: 10.1007/s11426-018-9412-6]

10 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法. 吉林大学学报 (工学版), 2014, 44(1): 137–141. [doi: 10.13229/j.cnki.jdxbgxb201401024]

11 Arora N, Kaur PD. A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 2020, 86: 105936. [doi: 10.1016/j.asoc.2019.105936]