doi:10.3969/j.issn.1008-5327.2023.04.019

基于随机森林算法的智能电表故障诊断及寿命预测模型设计

车　玲[1]，黄勇华[1]，姜林林[1]，车恩羽[2]
（1. 南通职业大学　电子信息工程学院，江苏　南通　226007；2. 西南大学　计算机与信息科学学院，重庆　400715）

摘　要：为及时发现并处理智能电表故障，延长其使用寿命，依据某地级市用电大数据进行数据挖掘和分析，基于随机森林（Random Forest，RF）算法建立智能电表故障诊断及寿命预测模型，并与其他模型进行实验比较。结果表明，构建的预测模型能实现智能电表的故障诊断与使用寿命预测，且有效性和准确性优于其他模型，具有工程应用价值。

关键词：智能电表；随机森林算法；故障诊断；寿命预测
中图分类号：TM933.4　　**文献标志码：**A　　**文章编号：**1008-5327(2023)04-0086-05

Fault Diagnosis and Life Prediction Model Design for Smart Electricity Meter Based on Random Forest Algorithn

CHE Ling[1], HUANG Yong-hua[1], JIANG Lin-lin[1], CHE En-yu[2]
(1. School of Electronic Engineering and Information Technology, Nantong Vocational University, Nantong 226007, China;
2. School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: To identify the fault occurring in smart electricity meters and deal with it timely, data mining and analysis are carried out according to the big data of electricity consumption of the prefecture-level city, and a fault diagnosis and life prediction model is established using Random Forest (RF) algorithm. Comparative experiments are conducted with other models. The results show that the developed prediction model can be used to diagnose the fault and predict the service life of smart electricity meters, and its effectiveness and accuracy are better than other models. It is of application value in engineering.

Keywords: smart electricity meter; random forest algorithm; fault diagnosis; life prediction

智能电表是一种利用数字技术、网络技术实现多种费率双向计量、多种数据双向通信、用户端控制、防窃电等智能化功能的新型数字电度表[1]。智能电表是智能电网（尤其是智能配电网）数据采集的主要设备之一，对于提高电力系统的运行效率、优化电力资源配置、实现用电侧管理等具有重要意义。由于智能电表的元器件构成复杂且类型多样，运行过程中难免产生损坏和各种故障[2]，如外观故障、时钟单元故障、计量性能故障等。这些故障会影响电表的计量准确性、通信可靠性、使用安全性等，给电力系统带来安全风险和经济损失。因此，及时发现并处理智能电表故障，延长其使用寿命，是保证智能电网正常运行的一项重要任务。本文拟针对智能电表存在数据采集量大、故障数

收稿日期：2023-06-06
基金项目：2022 年南通市职业技术教育学会教育研究课题（NTZJXH007）；2019 年南通职业大学校级项目（19KZ11）
作者简介：车玲（1974— ），女，吉林桦甸人，讲师，硕士，主要研究方向为传感器应用技术和电气控制技术。

据种类繁多等问题，构建智能电表故障诊断及寿命预测模型，以期实现智能电表的故障预测和及时处理，确保智能电表的安全可靠运行。

1　随机森林(RF)算法

目前，对于智能电表的故障诊断和寿命预测，主要采用基于规则或机理的方法[3]。这些方法需要依赖专家知识或者物理模型，往往缺乏通用性和适应性，无法有效处理复杂的非线性关系和多因素耦合问题。而且，这些方法往往只能在故障发生后进行诊断，无法提前预测故障发生的可能性和时间。为克服上述方法的局限性，提出一种基于随机森林(Random Forest，RF)算法的智能电表故障诊断及寿命预测模型。随机森林(RF)是一种集成学习方法，可通过构建多个决策树并进行投票或取平均值来提高预测性能[4]。

1.1　随机森林(RF)算法框架结构

随机森林(RF)算法是集成算法的一个子集，利用随机方法构建具有多棵决策树的森林，并根据决策树的投票选择决定最终分类结果。随机森林算法采用有放回的采样，即每棵树从训练集中选取固定数量的样本，选取后再放回到原始训练集中。图 1 为随机森林算法建立的决策树框架结构。

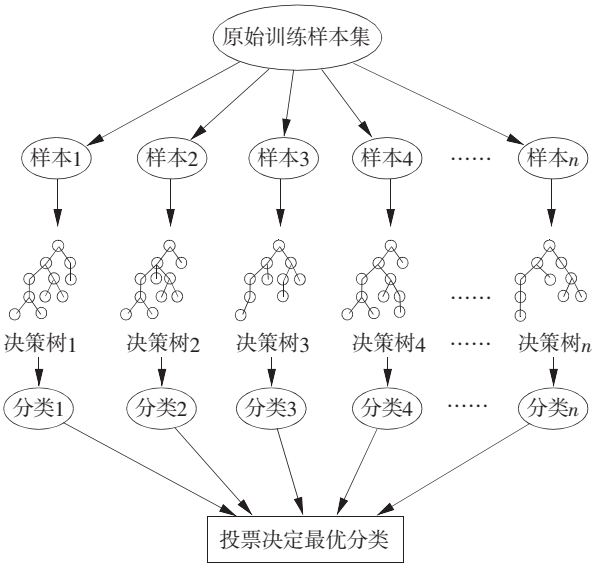

图 1　随机森林算法框架结构

随机森林(RF)算法的构建步骤如下：

1）从原始训练样本集随机抽取 n 个相互独立的训练样本，作为每棵决策树的根节点样本。

2）使用生成的 n 个测试样本，构建 n 棵决策树，并从 M 个特征属性中随机抽取 K 个特征属性，从中选择一个最合适的特征属性作为分裂节点。构建的决策树不进行剪枝，保证其完整生长。

3）建立随机森林后，利用测试样本进入每一棵决策树，进行类型输出和回归输出，并以投票方式输出最终类别。

随机森林(RF)算法具有以下优点：

1）可处理高维度、非线性、非平衡数据；

2）可同时进行分类和回归分析；

3）可评估各特征的重要性；

4）可抵抗噪声和过拟合。

1.2　智能电表故障与寿命预测模型设计

根据大数据分析理论，对某智能电表的海量累积数据进行挖掘分析，并从中提取与故障和寿命相关的特征变量，提出一种基于随机森林(RF)算法的智能电表故障及寿命预测模型。首先，收集和整理电表的特征数据，如用电功率、用电质量、用电计费等特征，以及电表的故障标签，如正常、异常、损坏等类别；其次，对数据进行预处理，如处理缺失值、异常值、噪音等问题，以及进行特征选择、特征编码、特征归一化等操作，使数据符合随机森林算法的输入要求；再次，将数据集中 85 %的数据作为训练样本，训练随机森林分类器和回归器；最后利用数据集中 15 %的数据作为测试数据，评估分类器和回归器性能。预测流程如图2 所示。

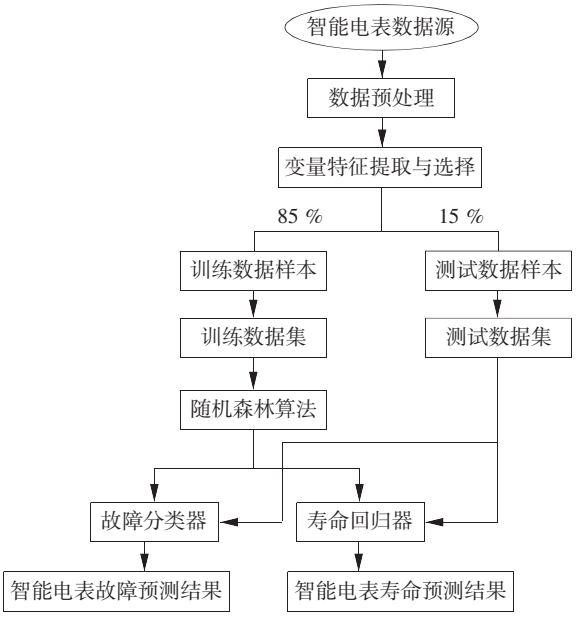

图 2　智能电表故障及寿命预测流程

对某地级市供电公司提供的真实数据进行实验验证，并与支持向量机回归、线性回归等几种常用回归模型的预测结果进行横向比较，验证模型预测的准确度和信用度。

2　预测模型构建

2.1　数据来源与处理

使用某地级市供电公司提供的真实数据进行模型构建与验证，所研究的智能电表用户包括工业和大型商业用户，约 20 万。每个用户对应一个唯一编号，有相应的计费信息和转账信息，还有一个或多个智能电表编号，并有相应的实时功率数据和质量字节数据。

实时功率数据是指每 15 分钟记录一次用户用电功率（kW），每天共 96 条记录。计费数据是指每月记录一次用户用电量值（kW·h），每年共 12 条记录。转账信息是指每次用户缴纳电费时记录其使用的银行账户信息。质量字节数据是指每 15 分钟记录一次用户用电质量信息（8 位二进制数），每天共 96 条记录。每个二进制位代表一个警报类型。表 1 显示用户用电质量信息及其含义。

表 1　用户用电质量信息含义

位/bit	报警类型	具体含义
0	OV	过电压（OV 为 1）
1	OC	过电流（OC 为 1）
2	CT	时钟超差（CT 为 1）
3	VM	测量时每 15 分钟记录（VM 为 1）
4	PFO	功率因数超限（PFO 为 1）
5	INT	测量时发生非法入侵（INT 为 1）
6	OW	溢出（OW 为 1）
7	IV	测量无效（IV 为 1）

选取 2017 年 1 月至 2019 年 12 月共 36 个月内发生过至少一次故障检修事件的用户作为研究对象。如图 3 所示，根据检修事件记录，智能电表故障类型中时钟单元故障、计量性能故障和外观故障占比较大。因此，将发生过这三类故障之一的用户标记为异常用户。

为保证数据完整性和有效性，在进行模型构建前，对原始数据进行预处理。

1）删除缺失值超过 10 %或异常值超过 5 %的用户数据；

2）删除用电功率为 0 或质量字节值全为 0

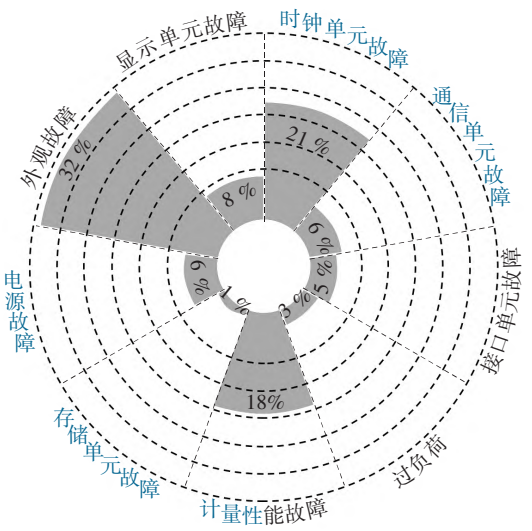

图 3　智能电表故障类型及占比

的无效记录；

3）对于连续变量（如用电功率），采用均值填充法补全缺失值；

4）对于离散变量（如质量字节），采用众数填充法补全缺失值；

5）对于异常值（如用电功率值超过正常范围），采用中位数替换法处理；

6）对于分类变量（如银行账户信息），采用独热编码法转换为数值变量；

7）对于数值变量（如用电功率），采用标准化法转换为标准正态分布。

预处理后的数据，可进行有效的特征提取，减少无效数据特征占比，提高数据应用的准确度。

2.2　特征提取与选择

建立智能电表的故障预测及寿命预测模型，需要从原始数据中提取出与故障和寿命相关的特征变量，包括用电功率、用电质量、用电计费、用户编号、智能电表编号、转账信息等特征。为减少特征维度，提高模型效率，采用随机森林（RF）算法对所有特征进行重要性评估。其主要思想是，观察各特征在随机森林树中所做贡献，取平均值，再比较各特征的贡献度。

根据随机森林算法，依据下列原则判断特征的重要性。

1）特征与目标变量的相关性越高，特征的重要性越高。例如，如果目标是预测电表的寿命，那么用电量、用电功率等特征比用电质量更重要。

2）特征在随机森林中每棵树上所做贡献越

大，特征的重要性越高。

3）特征的取值范围和变化程度越大，特征的重要性越高。例如，如果一个特征的取值范围很小，或者取值分布很不均匀，那么这个特征可能比其他特征更易被忽略或更易受噪声影响。

依据以上原则和特征重要性，从用电功率、用电质量、用电计费等三方面进行数据特征提取，共提取特征值 $k = 14$ 个，具体如下：

1）用电功率特征：用电功率反映了用户的用电行为和负荷变化，与电表的损耗和老化有关。从用电功率数据中提取平均功率、最大功率、功率标准差等特征，如表 2 所示。

表 2　提取的用电功率特征

序号	特征	具体含义
1	平均功率	每天或每月的平均用电功率
2	最大功率	每天或每月的最大用电功率
3	最小功率	每天或每月的最小用电功率
4	功率标准差	每天或每月的用电功率的标准差，反映用电波动程度
5	功率峰谷差	每天或每月的最大功率与最小功率之差，反映用电负荷变化幅度
6	功率峰谷比	每天或每月的最大功率与最小功率之比，反映用电负荷变化比例

2）用电质量特征：用电质量反映了用户的用电环境和电网状态，与电表的稳定性和可靠性有关。从质量字节数据中提取质量字节频率、字节比例等特征，如表 3 所示。

表 3　提取的用电质量特征

序号	特征	具体含义
1	质量字节频数	每天或每月出现某一种质量字节值（8 位二进制数）的频次，反映某一种警报类型的发生频率
2	质量字节比例	每天或每月出现某一种质量字节值（8 位二进制数）的频次占总频次的比例，反映某一种警报类型的发生概率
3	质量字节熵	每天或每月所有质量字节值（8 位二进制数）出现次数的熵值，反映用电质量的不确定性和混乱程度

3）用电计费特征：用电计费反映了用户的用电规模和消费水平，与电表的使用强度和寿命有关。从计费数据中提取平均用电量、最大用电量等特征，如表 4 所示。

表 4　提取的用电计费特征

序号	特征	具体含义
1	平均用电量	每年或每季度的平均用电量
2	最大用电量	每年或每季度的最大用电量
3	最小用电量	每年或每季度的最小用电量
4	用电量标准差	每年或每季度的用电量的标准差，反映用电波动程度
5	用电量增长率	相邻两个月或两个季度的用电量之差与前一个月或季度的用电量之比，反映用电增长趋势

3　实验结果与分析

3.1　故障预测结果

采用同一训练集和测试集，将 RF 算法与其他常见分类算法，包括决策树（decision tree，DT）、逻辑回归（logistic regression，LR）、朴素贝叶斯（naive Bayes，NB）、K 近邻（K-nearest neighbor，KNN）及支持向量机（support vector machine，SVM）等算法进行比较。

交叉验证法是模型进行训练和验证较为有效的方法之一。利用交叉验证法，将数据集划分为 5 个子集，包括 4 个训练集和 1 个测试集。然后，依次对训练集和测试集进行 5 次轮换训练和测试。最后，将 5 次测试结果进行平均，得到最终评估指标。

采用的评估指标分为故障预测指标和寿命预测指标。故障预测指标包括召回率（recall）、准确率（accuracy）、$F1$（F1-score）、精确率（precision）；寿命预测指标包括均方误差（mean squared error，MSE）、均方根误差（root mean squared error，RMSE）及平均绝对误差（mean absolute error，MAE）。

各算法在故障预测任务中的评估指标比较如表 5 所示。

表 5　RF 算法与其他分类算法故障预测结果比较

算法	准确率/%	精确率/%	召回率/%	$F1$
RF	93	91	92	0.92
SVM	87	85	86	0.86
LR	84	82	83	0.83
NB	81	79	80	0.80
KNN	78	76	77	0.77
DT	74	72	73	0.73

从表 5 可以看出，RF 算法对准确率、精确率、召回率等的预测准确率达 90 %以上，$F1$ 达 0.92，均高于其他分类算法，表明 RF 算法可以有效识

别异常用户，具有较高准确率，且在故障预测方面表现最优。

3.2　寿命预测结果

将 RF 算法与其他常见回归算法，包括支持向量回归（support vector regression，SVR）、线性回归（linear regression，LR）、岭回归（ridge regression，RR）、LASSO 回归（least absolute shrinkage and selection operator，LASSO）和决策树回归（decision tree regression，DTR）等算法进行比较，各算法在寿命预测任务中评估指标比较结果如表6 所示。

表 6　RF 算法与其他分类算法寿命预测结果比较

算法	MSE	RMSE	MAE
RF	0.12	0.35	0.28
SVR	0.18	0.42	0.34
LR	0.21	0.46	0.37
RR	0.22	0.47	0.38
LASSO	0.23	0.48	0.39
DTR	0.27	0.52	0.43

从表 6 可以看出，RF 算法在寿命预测任务中表现最优，其 MSE（均方误差）、RMSE（均方根误差）和 MAE（平均绝对误差）均为最低。表明 RF 算法可有效估计智能电表的剩余寿命，且具有较高精度。

4　结　语

利用 RF 算法分别建立了智能电表的故障诊断及寿命预测模型，通过对大量的智能电表数据进行分析和处理，提取了用电功率特征、用电质量特征和用电计费特征，并利用 RF 算法评估特征的重要性，再进行特征选择，最终得到优化的模型输入特征向量。利用交叉验证法进行模型训练和验证，并与其他常见的分类回归算法进行比较。结果表明，RF 算法在故障预测和寿命预测方面均表现出较高的准确性和稳定性。研究成果可为智能电表的运行监测和维护管理提供有效的技术支持。

参考文献：

[1] 涂家海，文松，李渊，等.基于物联网的智能电表技术应用思考[J].襄阳职业技术学院学报，2022，22(4)：92-97.

[2] 潘磊，杨延，连浩，等.基于智能电表大数据的异常用电检测[J].计算技术与自化，2020，39(2)：177-183.

[3] 史鹏博，李蕊，李铭凯，等.基于决策树和聚类算法的智能电表误差估计与故障检测[J].计量学报，2022，43(8)：1089-1094.

[4] 申佳灵，易婷，聂勤，等.几种电力数据异常检测算法的对比分析[J].智能城市，2023，9(2)：1-4.

责任编辑　谭　华