doi: 10.7690/bgzdh.2019.10.013

基于随机森林的智能电表故障及寿命预测模型

黄吉涛[1]，樊　博[1]，周媛奉[1]，胡婷婷[1]，梁　飞[1]，曾晓东[2]
（1. 国网宁夏电力有限公司电力科学研究院，银川 750011；
2. 四川大学电气信息学院，成都 610065）

摘要：为解决智能电表累积采集信息量大、故障信息种类繁多和突发性强的问题，构建一种基于随机森林的电表故障及寿命预估模型。依靠大数据分析理论，通过对海量电表的累积数据进行挖掘分析，建立智能电表的故障预测及寿命预测模型来对故障和寿命进行预测，并同其他模型进行比较。实验结果表明：该预测模型是有效的和准确的，可为数据挖掘在智能电表管控研究提供参考。

关键词：智能电表；数据挖掘；随机森林；故障预估
中图分类号：TP391　　**文献标志码：**A

Fault and Life Prediction Model for Smart Electric Meter Based on Random Forest

Huang Jitao[1], Fan Bo[1], Zhou Yuanfeng[1], Hu Tingting[1], Liang Fei[1], Zeng Xiaodong[2]
(1. *Electric Power Research Institute of State Grid Ningxia Power Co., Ltd., Yinchuan* 750011, *China*;
2. *College of Electrical Engineering & Information Technology, Sichuan University, Chengdu* 610065, *China*)

Abstract: In order to deal with the smart electric meters' mass information, varied fault types and sudden faults, the fault and life prediction model for smart electric meter based on random forest is established. Based on the theory of large data analysis, this paper excavates the accumulative data of large numbers of smart electric meters and forecast its fault and life, compare with other models. The experimental results demonstrate the effectiveness and correctness of the proposed model. It can provide reference for data mining in the research of smart electric meter.

Keywords: smart electric meter; data mining; random forest; fault prediction

0　引言

近年来，各大电力公司已逐步建成省级计量自动化系统，截至 2014 年，已累积安装智能电表 2.2 亿只[1]。随着智能电表在我国的普及，所记载的数据量也呈爆发式增长。如何利用大数据的优势，为智能电表的故障管控系统寻求一种数据支撑的分析方法，是目前智能电网炙手可热的研究问题[2]。

数据挖掘技术的主要目的就是从大数据集中探索隐藏的各个变量之间的关系。数据挖掘技术涉及到统计学习、人工智能和机器学习[3]3 个方面。此外，数据挖掘技术还用于对研究对象进行分析和预测[4]，用于分析的数据挖掘技术大多用到了聚类算法和关联规则，而用于预测的数据挖掘技术主要用到了分类和回归的算法，其中包括决策树、人工神经网络、遗传算法、K 近邻、朴素贝叶斯等[5]。一般来说，对大数据集进行数据挖掘的过程包括 7 个步骤。这些步骤可以被定义为“数据清理”“数据集成”数据选择”“数据转换”“数据挖掘”“模式评估”和“分析报告”[6]。

由于电表的数据多为非结构化，很难对观测得到的数据进行直接分析，而电表的属性信息又较为繁多，很难判断是否某一属性对电表的故障及寿命起着决定性影响；因此，有学者提出利用电表的海量数据进行关联分析，通过数据挖掘技术找出传统统计方法难以探究的电表大数据与性能状态间的隐性关系。文献[7]采用了两层前馈神经网络对电表数据进行挖掘分析，以此评估 63 项检测数据与 2 种故障之间的关联；文献[8]分别采用了朴素贝叶斯模型、决策树模型和 softmax 神经网络模型对电表数据进行分析，针对已使用电表和未使用电表构建 2 类预测模型，实验结果证明了电表信息数据和电表故障的关联性。

笔者提出了基于随机森林（random forest，RF）的智能电表数据管理模型，通过对智能电表的基本属性信息进行挖掘分析，以电表的故障类型与使用寿命作为模型的输出标签进行预测，在电表投放使用前可以获取每只电表的预估寿命和可能发生的故障类型，为电表的监测和轮换周期提供分析依据，并验证了笔者所提模型的准确性。

收稿日期：2019-05-14；**修回日期：**2019-06-17
基金项目：国家电网公司科技项目（宁电发展[2018]54 号）
作者简介：黄吉涛（1982—），男，宁夏人，硕士，高级工程师，从事管理信息系统研究。E-mail: 250682629@qq.com。

1 基于随机森林的数据管理模型

1.1 随机森林算法介绍

随机森林是决策树的延伸方法之一[9]，是一种由多颗决策树构成的集成学习算法，由于决策树易发生过拟合的现象[10]，为了改善该缺点，随机森林的预测结果由多颗决策树各自独立的投票决定，决策树的组合使得数据集的并行训练成为可能。在数据集规模庞大、性质复杂的情况下，单一的决策树不足以获取同步相量数据中的数据信息[11]，并且单一的树需要更多的时间来分类整个数据集，采用多个决策树并行工作，对数据集进行分类的速度和分类精度都是非常高效的。随机森林的分类原理如图 1 所示。

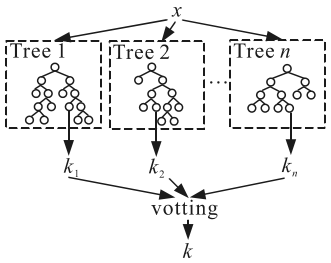

图 1 随机森林分类原理

随机森林并不是使用所有的变量来分割树节点，而是在每个节点处选择变量的随机子集来获得节点的最佳分割。这样随机化的主要目的是去除相关的决策树，使得所有树的集合具有较低的方差。构造随机森林的方法一般包括以下步骤[12]：

1) 在原始数据中提取出 *n*-tree 个样本子集。

2) 利用每个样本子集生成决策树，在树的每个节点，随机选择变量 *M* 进行分裂。继续增长树，使得每个终端节点的节点数不小于节点的大小。

3) 采用投票机制统计 *n*-tree 个决策树的结果进行分类。

随机森林采用多颗决策树并行工作的模式，对数据进行随机有放回的抽样，其预测能力相对优于单分类的模型，适合应用于大数据集，其分类模型也是被普遍认为具有高精度运算结果的模型之一。

1.2 电表故障与寿命预测模型

笔者所用的智能电表数据由宁夏计量中心提供，电表可供分析的基本信息共 22 条，其中包括：“生产厂家”“到货批次”“计度器方式”“接入方式”“脉冲常数”“失压/失流判断”等各类电表的属性数据。将这些数据转化为对应的数字标签，分别作为预测模型的输入特征量，通过建立 2 个模型来分别对故障和寿命进行预测。笔者所用到的电表数量共 15 万余只，采用其中 80%的数据作为训练数据，20%作为预测数据对模型进行测试。预测的思路如图 2 所示。

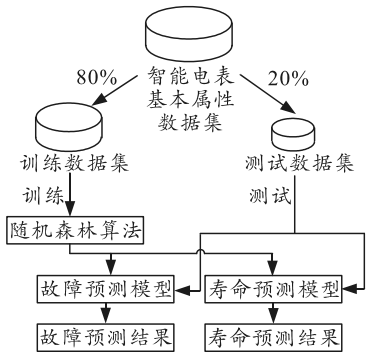

图 2 智能电表故障与寿命的预测

如图 3 所示，在构建故障预测模型时，通过对 15 万只电表的拆除原因进行统计，得到最易发生的前 8 类故障。

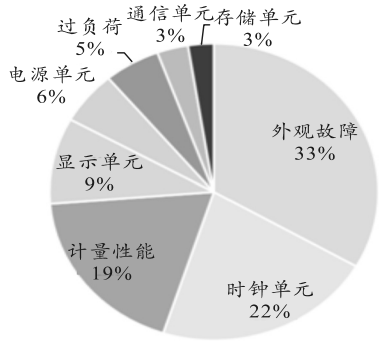

图 3 前 8 类频发故障

由于提取的电表基本信息条目较少，预测模型的准确度会与故障类别数成反比[13]；因此，笔者只考虑这 22 项电表的基本属性与前 3 种频发故障的关联性。同时，通过统计电表的安装时间与拆除时间，可以获得故障电表的使用寿命，在进行电表的故障预测时，将使用寿命也作为模型的输入特征之一，与 22 项基本属性信息合并，转化为由 23 个数字构成的输入特征量。

2 实验结果分析

为验证笔者所提预测模型的有效性，对计量中心提供的电表数据进行实验验证。

2.1 电表故障预测结果

在电表的故障预测模型中，输入含使用时间在内的 23 个特征向量，预测电表最易发生的前 3 类故障：外观故障、时钟单元、计量性能，实验的预测准确度与所用预测时间如表 1 所示。其中训练预测

模型所用数据为 44 590 只电表数据，测试所用数据为 11 147 只电表数据。

表 1　3 类故障预测结果

n-tree	10	30	40	50	80
预测准确度/%	75.1	75.8	76.0	76.5	76.4
预测时间/s	1.8	4	5	7	10

由表可知：利用笔者构造的随机森林预测模型，通过 23 项智能电表的基本信息预测电表可能会发生的 3 种故障的准确率为 76%左右。在文献[8]中，3 种模型对 3 类故障的最优预测值由决策树模型提供，为 69.1%。与笔者预测准确度相比较，可验证所提出模型的正确性和有效性。

为与文献[7]比较，笔者将故障预测的种类减少，只预测电表最易发生的前 2 类故障：外观故障、时钟单元，实验的预测准确度与所用预测时间如表 2 所示。其中训练预测模型所用数据为 33 251 只电表数据，测试所用数据为 8 312 只电表数据。

表 2　2 类故障预测结果

n-tree	10	30	40	50	80
预测准确度/%	85.8	86.1	86.6	86.5	86.3
预测时间/s	1.3	3	3.8	4.8	7

由表可知：基于随机森林的电表故障预测模型，利用 23 项电表的基本信息预测 2 类高发故障时，预测的准确度为 86%左右。相比于文献[7]中的最优准确度 75%，笔者所提出的模型具有更高的预测精度。

同时，根据表 1、表 2 的结果可知：改变模型唯一参数 *n*-tree 的值，预测准确度不会有太大的波动，但随着 *n*-tree 的增加，预测所用时间逐渐增长。如何权衡参数的选择，可在实际操作中根据不同需求进行更改。

2.2　电表使用寿命预测结果

在对电表的使用寿命进行预测时，笔者首先统计了 15 万余只故障拆除电表的使用时间，统计结果如图 4 所示。

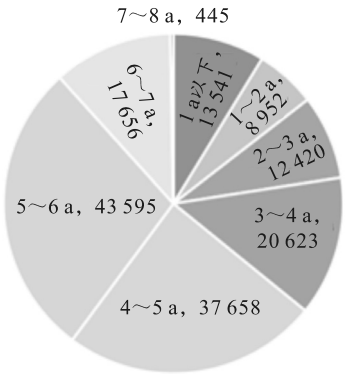

图 4　故障拆除电表使用年限

通过统计现有的故障拆除表的使用年限，将使用寿命划分为 4 个时间段：使用时间在 2 a 以内、使用时间为 2～4 a、使用时间为 4～6 a、使用时间在 6 a 以上。输入为 22 项电表的基本信息，预测准确度与所用预测时间如表 3 所示。其中，训练预测模型所用数据为 113 062 只电表数据，测试所用数据为 28 265 只电表数据。

表 3　4 类寿命年限预测结果

n-tree	10	30	40	50	80
预测准确度/%	74.9	75.3	75.6	75.4	75.5
预测时间/s	5.1	9.1	11.6	14	22

由表可知，将智能电表的寿命划分为 4 个时间段进行预测的准确度在 75%左右。同时，模型会给出每只电表在每个寿命段内的得分，例如随机抽取一只电表的寿命预测结果，在第 1 类的得分为 0.011 3，在第 2 类的得分为 0.286 0，在第 3 类的得分为 0.700 2，在第 4 类的得分为 0.002 6，那么该电表的寿命在 4～6 a 的概率为 70%，其寿命在 2～4 a 内的概率为 28.6%，在 2 a 以下的概率为 1.13%，在 6 a 以上的概率为 0.26%，可以判断该电表的寿命最有可能在 4～6 a。

为了提高寿命预测模型的预测准确度，将使用寿命划分为 3 个时间段：使用时间在 3 a 以内、使用时间为 3～6 a、使用时间在 6 a 以上。同样输入 22 项电表的基本信息，预测准确度与所用预测时间如表 4 所示。其中训练数据与测试数据和表 3 实验数据相同。

表 4　3 类年限寿命预测结果

n-tree	10	30	40	50	80
预测准确度/%	89.4	89.5	89.6	89.6	89.7
预测时间/s	4.6	8.7	11.4	13.8	21

由表可知，将智能电表的寿命划分为 3 个时间段进行预测的准确度在 89%左右。相比于划分为 4 个时间段，预测准确度有了明显提升。因此，模型的预测年限可根据具体需求划分为不同时间段，以更好地为电表的检测和轮换周期提供参考依据。

3　结论

笔者基于随机森林算法，构造了智能电表的故障预测及寿命预测 2 种模型，通过宁夏计量中心提供的 15 万余只电表数据，进行了实验测试，可以得到以下结论：

1) 笔者构造的随机森林模型可以有效地对电表故障及寿命进行预估，并通过与其他算法进行比较，验证了该模型具有较高的预测准确度。

2) 通过实验结果分析，模型的预测准确度与预测模型输出的类别数成反比，在实际操作中，可根据需求调节预测的种类。

3) 改变模型唯一参数 *n*-tree 的值，预测准确度不会大幅变动，但预测时间会随之改变；因此，*n*-tree 的值设置较小为宜。

参考文献：

[1] 徐大青, 栾文鹏, 王鹏, 等. 智能电表数据分析方法及应用[J]. 供用电, 2015, 32(8): 25-30.

[2] 栾文鹏, 余贻鑫, 王兵. AMI 数据分析方法[J]. 中国电机工程学报, 2015, 35(1): 29-36.

[3] KAYRI M, KAYRI I, GENCOGLU M T. The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data[C]. International Conference on Engineering of Modern Electric Systems. IEEE, 2017: 1-4.

[4] 徐辉增. 关联规则数据挖掘方法的研究[J]. 科学技术与工程, 2012, 12(1): 60-63.

[5] AKI A, REDDY D K M, REDDY Y K, et al. Analyzing the real time electricity data using data mining techniques[C]//Smart Technologies For Smart Nation (SmartTechCon), 2017 International Conference On. IEEE, 2017: 545-549.

[6] CAO M, ZHANG X, LI B, et al. Prediction with random forest involving sampling and feature selection strategies [C]//2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2018.

[7] 祝宇楠, 徐晴, 刘建, 等. 数据挖掘在智能电能表故障分析中的应用[J]. 电力工程技术, 2016, 35(5): 19-23.

[8] 贺宁. 智能电表故障大数据分析探究[J]. 中小企业管理与科技, 2016(7): 142-145.

[9] 邹媛. 基于决策树的数据挖掘算法的应用与研究[J]. 科学技术与工程, 2010, 10(18): 4510-4515.

[10] VIJAYAKUMAR V, CASE M, SHIRINPOUR S, et al. Quantifying and Characterizing Tonic Thermal Pain Across Subjects From EEG Data Using Random Forest Models[J]. IEEE Transactions on Biomedical Engineering, 2017, 64(12): 2988-2996.

[11] 黄爱辉. 决策树 C4.5 算法的改进及应用[J]. 科学技术与工程, 2009, 9(1): 34-36.

[12] BABOO, SANTHOSH S, IYYAPPARAJ E. A classification and analysis of pulmonary nodules in CT images using random forest[Z]. 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018.

[13] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工), 2014, 44(1): 137-141.

****

（上接第 52 页）

[4] 杨纯录. 软件测评师教程[M]. 北京: 清华大学出版社, 2005: 14-28.

[5] 乔勇诚. 探讨软件测试“误区”[J]. 通信技术, 2011, 44(8): 149-151.

[6] CEM K, JACK F. 计算机软件测试[M]. 北京: 机械工业出版社, 2004: 6-7.

[7] GLENFORD J. Myers Tom Badgett Corey Sandler. 软件测试的艺术[M]. 北京: 机械工业出版社, 2012: 34-66.

[8] 周晓波. 构件回归测试方法研究与实现[D]. 昆明: 昆明理工大学, 2012.

[9] 王小丽, 段永颢. 软件回归测试用例选取方法研究[J]. 空间控制技术与应用, 2010, 36(3): 47-50.