Operations Research and Fuzziology 运筹与模糊学, 2021, 11(4), 462-469
Published Online November 2021 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2021.114052

Hans汉斯

基于随机森林的变压器油温预测方法

顾颖歆

国网江苏省电力有限公司，营销服务中心，江苏 南京

收稿日期：2021年10月12日；录用日期：2021年11月11日；发布日期：2021年11月19日

摘 要

变压器的油温可以有效反映电力变压器的工况，但要预测特定用户区域的未来需求是困难的，因为它随工作日、假日、季节、天气、温度等的不同因素变化而变化。现有预测方法不能适用于长期真实世界数据的高精度长期预测，管理人员不得不根据经验值做出决策，而经验值的阈值通常远高于实际需求而导致浪费，且任何错误的预测都可能产生严重的后果，因此需要一种有效的方法来预测未来的用电量。随机森林(Random Forest，简称RF)是Bagging的一个扩展变体，其原理简单、容易实现、计算开销小，但又具有强大的性能，代表目前最先进的集成学习技术水平的方法。本文通过收集到的变压器数据集，利用随机森林回归的预测方法，对变压器的油温变化进行预测。

关键词

变压器，机器学习，结果预测，随机森林回归

Prediction Method of Transformer Oil Temperature Based on Random Forest Method

Yingxin Gu

State Grid Jiangsu Electric Power Co., Ltd., Marketing Service Center, Nanjing Jiangsu

Received: Oct. 12th, 2021; accepted: Nov. 11th, 2021; published: Nov. 19th, 2021

Abstract

The oil temperature of a transformer can effectively reflect the working condition of a power transformer, but it is difficult to predict the future demand of a specific user's area because it varies with different factors such as weekdays, holidays, seasons, weather, and temperature. The existing forecasting methods cannot be applied to high-precision and long-term forecasting of long-term real-world data, and managers have to make decisions based on empirical values, which are usually much higher

文章引用: 顾颖歆. 基于随机森林的变压器油温预测方法[J]. 运筹与模糊学, 2021, 11(4): 462-469.
DOI: 10.12677/orf.2021.114052

than the actual demand and lead to waste, and any wrong prediction may result in serious consequences, so an effective method is needed to forecast future electricity consumption. Random Forest (RF) is an extended variant of Bagging with simple principles, easy implementation, low computational overhead, yet powerful performance, and is a method representing the state of the art in integrated learning. In this paper, we use the prediction method of Random Forest regression to predict the oil temperature variation of transformers from the collected transformer data set.

Keywords

Transformer, Machine Learning, Prediction, Random Forest Regression

1. 引言

近年来随着社会发展速度的日益加快，对重要大容量输电、变电设备的安全要求更高，电力系统大型变压器的绝缘能力逐渐丧失是大部分变压器寿命终结的主要原因，而变压器运行时的绕组温度、油温是影响变压器绝缘能力的重要因素，所以变压器相关研究的重点一直集中在变压器温度的绕组温度在线监测上，但要预测特定用户区域的未来需求是困难的，因为它随工作日、假日、季节、天气、温度等不同因素变化而变化。现有的预测方法不能适用于长期、真实数据的高精度长期预测，并且任何的错误预测都可能产生严重后果[1] [2]。由于当前没有一种有效的方法来预测未来用电量，管理人员不得不根据经验值做出决策，而经验值的阈值通常远高于实际需求，采用保守的策略额必然会导致不必要的电力浪费和设备折旧浪费。

因为变压器的油温可以有效反映电力变压器的工况，所以通过对变压器的油温进行在线检测，并根据检测数据预测变压器的油温的同时设法避免不必要的浪费，这是维护变压器安全运行的有效策略之一。

2. 项目概述

2.1. 研究背景与意义

变压器作为日常生活中常见的线路传输关键组件，其相关技术复杂，并且价格是非常昂贵的。作为电网电力传输的关键设备之一，变压器运行状况对电网的安全有较大的影响。变压器内部结构的温度决定着变压器负载能力及其内部绝缘系统的性能，进而影响着变压器的寿命。从上个世纪开始，变压器就开始使用油作为绝缘和冷却的媒介，这也标志着油浸式变压器的诞生。由于油具有防水防潮的效果，可以延缓机械的老化，这也是油浸式变压器的优点。因此现在油浸式变压器依旧占据庞大的份额，而油浸式变压器的油温也是油浸式变压器的研究热点。

目前变压器油温的温度预测相关研究很多，有通过对变压器传热过程构建热路模型，进行油温预测的；基于变压器修正热路模型，推导油浸式变压器顶层油温的温度；通过对变压器绕组发热机理和变压器油粘滞度分析，构造修正参数进行预测的；还有利用 BP 神经网络进行油浸式变压器油温的预测工作等等。因此对油浸式变压器油温的异常预测是一个非常有研究价值和理论意义的研究领域[3] [4]。基于大数据挖掘技术的油温异常预警是当前最具有发展潜力的一项技术。本文将大数据及人工智能技术于与电力行业进行结合，提出一种基于随机森林方法的变压器油温预测方法。

2.2. 随机森林介绍

在机器学习中，随机森林是一个包含多个决策树的分类器。随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习(Ensemble Learning)方法。随机森林的主要思想是集成思想，其构建主要包括有两个方面：数据的随机性选取，以及待选特征的随机选取[5] [6]。

首先，从原始的数据集中采取有放回的抽样，构造子数据集，子数据集的数据量是和原始数据集相同的。不同子数据集的元素可以重复，同一个子数据集中的元素也可以重复。第二，利用子数据集来构建子决策树，将这个数据放到每个子决策树中，每个子决策树输出一个结果。最后，如果有了新的数据需要通过随机森林得到分类结果，就可以通过对子决策树的判断结果的投票，得到随机森林的输出结果了。

与数据集的随机选取类似，随机森林中的子树的每一个分裂过程并未用到所有的待选特征，而是从所有的待选特征中随机选取一定的特征，之后再在随机选取的特征中选取最优的特征。这样能够使得随机森林中的决策树都能够彼此不同，提升系统的多样性，从而提升分类性能。

根据上述分析，可得在不同传输功率下，变换器回流功率最小时内移相比 D_1 的取值。

2.3. 电力变压器数据

下面对本研究分析所使用的变压器数据集进行介绍：1) ETT-small：含有 2 个电力变压器(来自 2 个站点)的数据，包括负载、油温；2) ETT-large：含有 39 个电力变压器(来自 39 个站点)的数据，包括负载、油温；3) ETT-full：含有 69 个电力变压器(来自 39 个站点)的数据，包括负载、油温、位置、气候、需求。

通过收集 2 年的数据，用来预测电力变压器的油温并研究电力变压器极限负载能力。本文将使用 ETT-small 的数据进行随机森林模型的建设，以下将对这一数据集进行介绍。该数据集提供了两年的数据，每个数据点每分钟记录一次(用 m 标记)，它们分别来自中国同一个省的两个不同地区，分别名为 ETT-small-m1 和 ETT-small-m2。每个数据集包含 2 年 365 天 24 小时 60 分钟 = 1,051,200 数据点。此外，我们还提供一个小时级别粒度的数据集变体使用(用 h*标记)，即 ETT-small-h1 和 ETT-small-h2。每个数据点均包含 8 维特征，包括数据点的记录日期、预测值“油温”以及 6 个不同类型的外部负载值。

具体来说，数据集中包含短周期模式，长周期模式，长期趋势和大量不规则模式。为了更好地表示数据中长期和短期重复模式的存在，我们在图 1 中绘制了 ETT-small-h1 数据集中 OT (油温)的自相关图，使用蓝色曲线表示，它保持了一些短期的局部连续性，而其他的变量(各类负载)则显示出了短期的日模式(每 24 小时)和长期的周模式(每 7 天)。

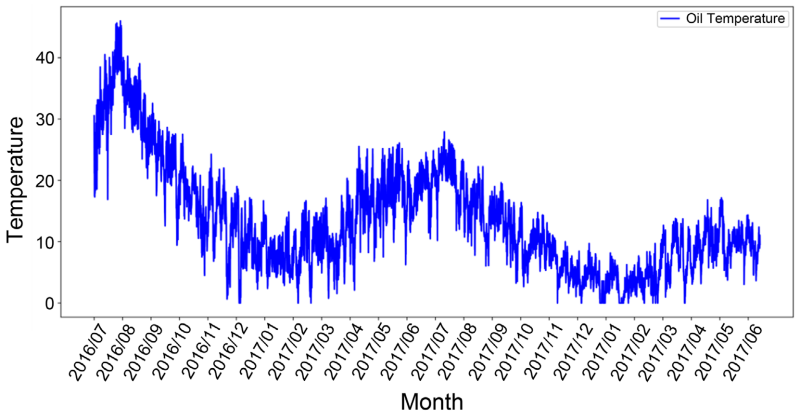

Figure 1. Oil temperature changes monitored over a two-year period
图 1. 两年期间内监测油温变化图

数据集是使用.csv 形式进行存储的，在表 1 中给出了一个数据的样例。其中第一行(8 列)是数据头，包括了“HUFL”，“HULL”，“MUFL”，“MULL”，“LUFL”，“LULL”和“OT”，每一列的详细意义展示在表 1 中。

Table 1. Data set label interpretation
表 1. 数据集标签解释

字段	Date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
介绍	记录的时间	高有效负载	高无效负载	中有效负载	中无效负载	低有效负载	低无效负载	油温(目标)

3. 随机森林方法建立的步骤

3.1. 数据预处理

油温预测的任务目标是使用收集到的各时间段负载信息，预测某一时间段的油温，属于回归任务。在进行机器学习建模时，首先要对收集到的数据进行预处理。使用 pandas 的 read_csv 查看一下数据集。

这里使用.describe 来查看数据集的信息，如每一列的数据量、平均值、最小值、最大值等；也可以使用.head(5)查看数据集的前五行信息了解标签分布；使用.shape 来查看数据集规模；使用.info()来查看数据集的缺失情况[7]。

通过查看数据集可以看出，本数据集一共有 17,420 条记录，每个样本有 8 个特征，包含一个时间信息和七个特征。由数据集的信息可以看出，不存在数据缺失的情况(即各个列值均为 17,420)，所以不需要进行数据填补，如果数据有缺失，则需要决定用什么样的数值进行数据填补。对于数据集中的时间数据，我们可以使用 datetime 的包进行转换，目的就是有些工具包在绘图或者计算的过程中，需要标准的时间格式。

本数据集中所有的数据类型都是浮点型，如果数据集中有些不是数值特征而是字符串，计算机可能无法识别，那就需要进行转换。图 2 展现了一种常用的数据转换方法。为了直观的观察数据，也可以画出各个数据随时间的变化图。

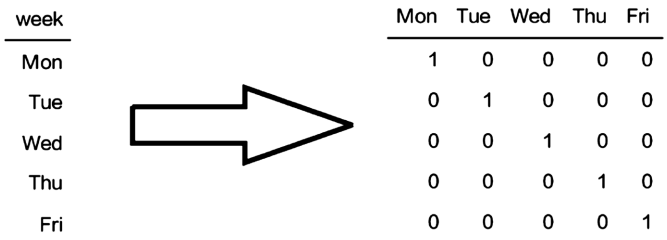

Figure 2. Character encoding
图 2. 特征编码

仅凭数据的变化图无法确定数据的特征重要性差别，还需要进行后续的编程计算。

经过上述处理后的数据，将标签和数据格式进行转换，准备分类训练集与测试集。Sklearn 中有专门进行训练集和测试集划分的函数 train_test_split (features, labels, test_size =, random_state =)，在函数中 features 表示处理完的数据集，labels 为标签，test_size 为测试集所占总数据的比例(若是整数则为样本的数量)，random_state 若为同一数字(非 0)则所划分的方式固定，填 0 或者不填则每次随机划分。

3.2. 随机森林回归模型

为了验证回流功率优化策略的有效性，完成了数据集的划分就可以开始建设模型，首先导入工具包，

使用 sklearn 中的 Random Forest Regressor 建立随机森林回归模型。

下面对随机森林回归器的主要参数进行介绍：

n_estimators：指定随机森林中的分类器的个数，默认为 10。一般来说 n_estimators 太小容易欠拟合，太大计算量大，故需要参数调优选择一个适中的数值；

oob_score：是否采用袋外误差来评估模型，默认为 False；

criterion：及 CART 树划分对特征的评价标准，默认为基尼指数，还可以选择信息增益；

max_features：建立决策树时选择的最大特征数目(从原始特征中选取多少特征进行建立决策树)，默认为 auto，意味着考虑 sqrt (n_features)个特征；还可以为整数，即直接指定数目；浮点数，即指定百分比；sqrt 与 auto 相同；log2 即指定 log2 (n_features)；如果是 None，则为最大特征数 n_features；

max_depth：决策树的最大深度，默认是不进行限制的，如果是模型样本量多，特征也多的情况，推荐限制修改这个，常用的可以取值为 10~100 之间；

min_samples_split：限制子树继续划分的条件，如果某节点的样本数目小于此值，则不会再继续划分，默认为 2，样本量非常大的时候，应该增大这个值；

min_samples_leaf：叶子节点的最小样本数目，如果某叶子节点数目小于样本数，则会和兄弟节点一起被剪枝，默认为 1，数据量大的时候可以增大这个值；

min_weight_fraction_leaf：叶子节点最小样本权重，这个值限制了叶子节点所有样本权重和最小值，如果小于最小值，则会和兄弟节点被剪枝。默认为 0，就是不考虑权重。通常来说，若样本中存在较多的缺失值，或者分类树样本的分布类别偏差很大，就会引入样本权重，这时就需要考虑这个值了；

max_leaf_nodes：最大叶子节点数，通过限制最大叶子节点数目来防止过拟合，默认为 None，即不进行限制，如果特征分成很多可以加以限制；

min_impurity_split：节点划分最小不纯度，这个值限制了决策树的增长，如果某节点的不纯度小于这个阈值，则该节点不在生成子节点，即为叶子节点，一般不推荐改动，默认值为 1e−7；

min_impurity_decrease：若一个节点被分割，如果这个分割导致大于或等于该值。默认为 0；

bootstrap：构建树时是否使用 bootstrap 采样，默认为 True；

n_jobs：设置程序的并行作业数量，默认为 1，如果为−1，则作业数目为核心数；

random_state：随机数的设置；

verbose：控制构建树过程中的详细程度。

大部分的参数具有默认的格式，根据预测要求的不同，选择不同的参数对数据进行预测，不同的参数选择会得到不同的预测结果，我们首先选择分类器个数为 1000，随机数为 42，其他参数暂取默认值，建立随机森林回归模型。

对于回归任务，评估的主要方法如表 2，本文使用 MAPE 指标进行评估，MAPE 是平均绝对百分误差，范围[0, +∞)，MAPE 为 0%表示完美模型，MAPE 大于 100%则表示劣质模型。MAPE 的值越小，说明预测模型拥有更好的精确度。

Table 2. Main evaluation methods of regression model
表 2. 回归模型主要评估方法

指标	描述	metrics 方法
MAE	平均绝对误差	from sklearn.metrics import mean_absolute_error
MSE	平均方差	from sklearn.metrics import mean_squared_error
R-Squared	R 平方值	from sklearn.metrics import r2_score

建立的随机森林模型平均绝对误差为 27.4512534252。

3.3. 特征重要性

在数据分析和特征提取的过程中，出发点都是尽可能多地选择有价值的特征，因为初始阶段能得到的信息越多，建模时可以利用的信息也越多。在机器学习项目中，常常会有在建模之后，又想到一些可以利用的数据特征，再回过头来进行数据的预处理和体征提取，然后重新进行建模分析的情况发生。

反复提取特征后，最常做的就是进行实验对比，但是如果数据量非常大，进行一次特征提取花费的时间就相对较多，所以在开始阶段需要尽可能地完善预处理与特征提取工作，或者多制定几套方案进行对比分析。

在 sklearn 中有用于计算特征重要性的函数(.feature_importance_)，可以用来调用各项特征的重要性，对特征重要性进行分析排序，是随机森林机器学习的必经步骤。计算后的特征重要性可以通过打印或可视化方式展示出来，如图 3 所示。

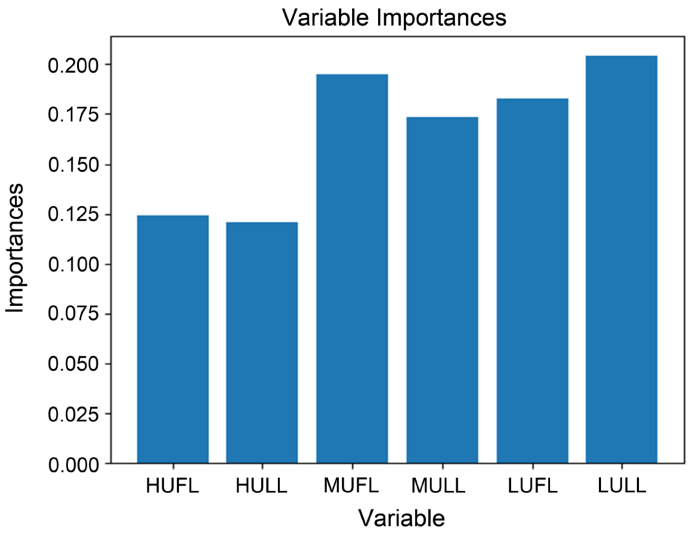

Figure 3. Feature importance data and visual mapping
图 3. 特征重要性数据及可视化绘图

根据不同的特征重要性形式，可以考虑对建模方式进行改进。例如某些特征重要性很低，那就考虑在建模时将其忽略，减少数据的冗杂程度，改进程序运行速率。或者直接取重要性较高的前几位特征进行建模，但不能只凭特征重要性就否定部分特征数据，一切还要通过实验进行判断。

3.4. 预测结果分析

以上步骤已经完成了随机森林回归预测模型的搭建，已经可以对数据进行预测了，通过 matplotlib 绘图，展示在测试集上进行预测的预测值与实际值的差异，如图 4 所示。

可以看出，使用 17420 条样本数据所建立的模型随机森林预测得分只有 0.47 左右，预测的效果也不是很理想。接下来使用数据量更多的数据集，按照同样的步骤进行随机森林回归模型的搭建，特征重要性如图 5。

预测效果如图 6 所示，可以看到，模型得分提高了很多，达到 0.767 左右，预测的效果变得更加优秀，同时平均绝对百分误差从原来的 27.4 降低到 20.2 左右，模型预测效果得到很大改善。可见，随着数据样本量的增大，随机森林模型的预测也会更加准确。

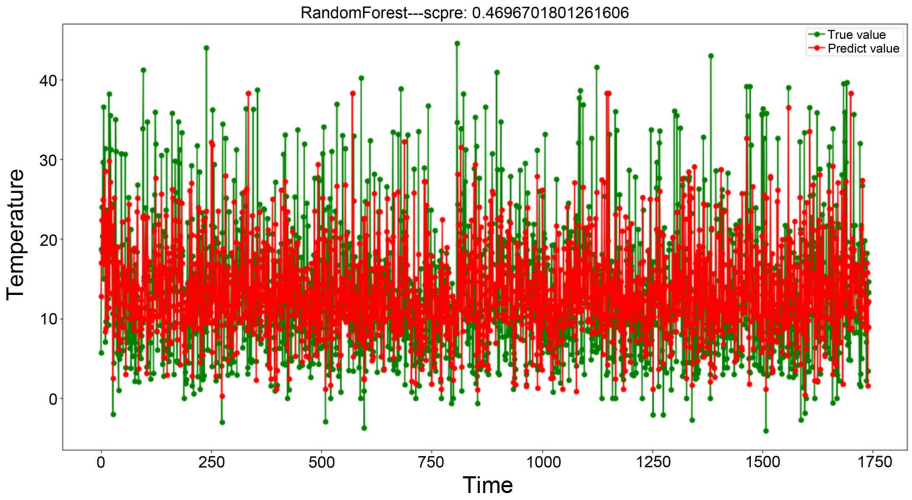

Figure 4. Predicted value versus actual value
图 4. 预测值与实际值

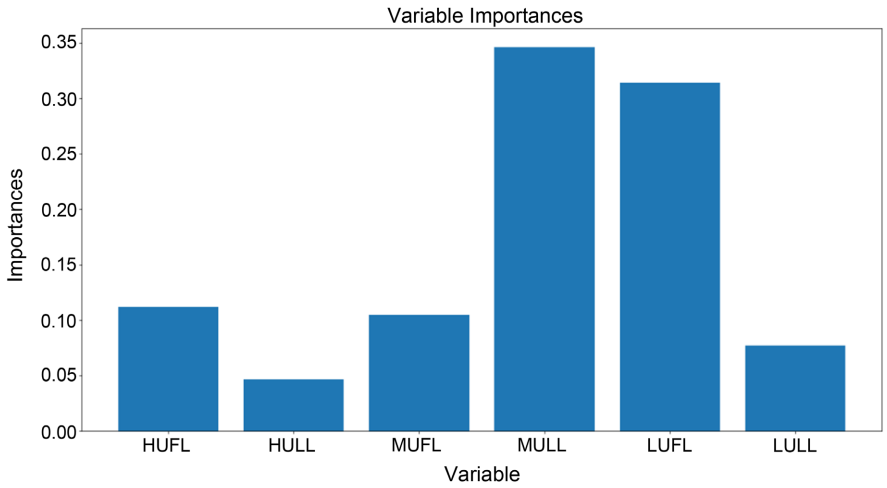

Figure 5. More sample data feature importance
图 5. 高样本数据特征重要性

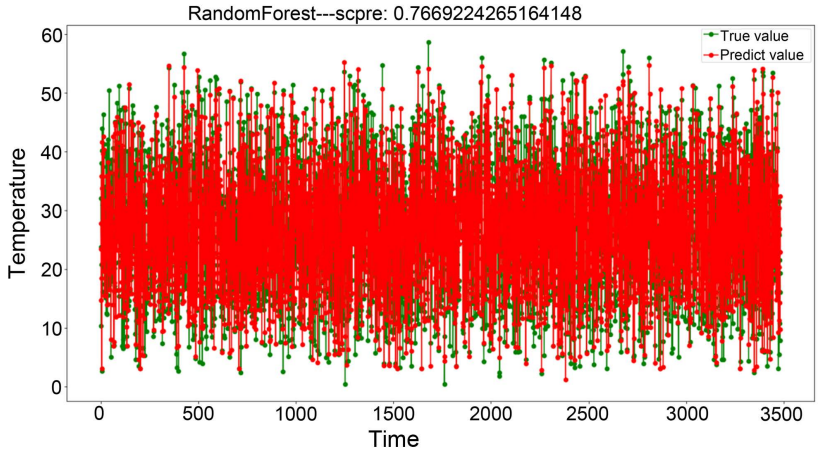

Figure 6. Prediction results of more sample random forest regression modeling
图 6. 高样本随机森林回归建模预测结果

4. 结论

通过建立随机森林模型，我们可以对变压器油温变化进行近似的预测，但其缺陷也很明显。首先，随机森林的调参是一个反复的过程，机器学习建模任务在实验结果确定之后，经常需要再回过头来反复对比不同的参数、不同的预处理方案，最终得到的模型也不一定就是最佳结果，所以预测模型的结果准确度需要不断改进。其次，本文仅提出了一种随机森林的建模思想，缺乏相关方法的比较对比，内容较为单一，难以形成高精度的长期预测。变压器油温的高精度长期预测，对维护变压器安全运行具有重要意义，并且任何错误预测都可能产生严重后果，所以对油浸式变压器的油温预测研究应该参考多种方法，并总结出行之有效的综合监测预测方案。

参考文献

[1] 牟龙华, 石林, 许旭锋, 等. 智能换流变压器在线监测系统的设计与建模[J]. 电力系统及其自动化学报, 2013, 25(1): 23-28.

[2] 程彤, 颜伟, 文雨, 等. 基于变压器参数修正的变电站状态估计[J]. 重庆大学学报(自然科学版), 2006, 29(3): 32-35.

[3] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[4] 谢文睿, 秦州. 机器学习公示详解[M]. 北京: 人民邮电出版社, 2021.

[5] 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(工学版), 2014(1): 142-146.

[6] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报(昆虫知识), 2013, 50(4): 1190-1197.

[7] Chen, D.Y. Python 数据分析活用 Pandas 库[M]. 武传海, 译. 北京: 人民邮电出版社, 2020.