

采用核主成分分析和随机森林算法的变压器油纸绝缘评估方法

张丽平¹, 吴杰康², 陶飞达¹, 曾振达¹, 梁浩浩², 邹志强¹, 黄智鹏¹, 杨夏¹

(1. 广东电网有限责任公司河源供电局, 广东 河源 525000;

2. 广东工业大学自动化学院, 广东 广州 510006)

摘要: 为了实现对电力变压器绝缘状态的智能评估, 提出了一种融合核主成分分析和集成学习理论的电力变压器油纸绝缘评估方法。在特征提取方面, 通过回复电压法(recovery voltage method, RVM)提取特征量, 并对特征量进行核主成分分析(kernel principle component analysis, KPCA)将低维度的样本空间映射到高维度的核空间, 并按照重要性进行排序; 在评估识别方面, 利用集成学习的思想建立分类器群模型, 克服了单分类器的局限性, 并提高了分类器的分类预测能力。通过实例论证, 融合核主成分分析和集成学习的分类模型在变压器油纸绝缘评估中具有很高的准确性。

关键词: 电力变压器; 油纸绝缘; 回复电压法; 分类器群; 集成学习; 核主成分分析

中图分类号: TM855 文献标志码: A 文章编号: 1003-6954(2018)02-0029-05

DOI:10.16527/j.cnki.cn51-1315/tm.2018.02.007

Oil - paper Insulation Evaluation Method of Transformer Based on Kernel Principal Component Analysis and Random Forest Algorithm

Zhang Liping¹, Wu Jiekang², Tao Feida¹, Zeng Zhenda¹,
Liang Haohao², Zou Zhiqiang¹, Huang Zhipeng¹, Yang Xia¹

(1. Heyuan Power Supply Bureau of Guangdong Power Grid Corporation, Heyuan 525000, Guangdong, China;

2. School of Automation, Guangdong University of Technology, Guangzhou 510006, Guangdong, China)

Abstract: In order to realize the intelligent evaluation of power transformer insulation status, an evaluation method for oil - paper insulation of transformer is proposed based on kernel principal component analysis and random forest algorithm. Recovery voltage method (RVM) is to study insulation aging status of power transformer. In order to overcome the limitations of single classifier, random forest (RF) is introduced to get classifier groups. The randomness of training samples and feature selection can avoid the problem of overfitting. The differences are small among classifiers built with features when there are few features available. The original samples are analyzed by using kernel principle component analysis (KPCA) so as to increase the number of features. Then random forests are constructed to get classifier groups in high dimensional kernel space. Finally, the voting results show that, the differences among classifier groups are improved effectively by KPCA. And results of the state diagnosis based on KPCA and random forests have a higher accuracy.

Key words: power transformer; oil - paper insulation; recovery voltage method; classifier groups; ensemble learning; kernel principle component analysis

0 引言

油浸式变压器是电力系统的核心设备之一, 研

基金项目: 国家自然科学基金项目(51567002, 50767001); 广东省公益研究与能力建设专项资金项目(2014A010106026); 广东电网有限责任公司科技项目(031600KK52160004)

究变压器的绝缘老化对电网的安全稳定运行具有十分重要的作用。油浸式变压器的使用寿命取决于其绝缘状态, 复合油纸绝缘状态则是评估变压器绝缘水平的重要指标^[1]。回复电压法(recovery voltage method, RVM)通过提取回复电压极化谱的特征量来评估变压器油纸绝缘状态。利用回复电压法能够

无损诊断绝缘状态,并且准确地反映变压器油纸绝缘内部缓慢的弛豫过程,在实际应用中取得了很好的效果^[2]。

目前,国内外研究者对回复电压法特征量和变压器绝缘状态展开了定性研究,如文献[3]利用实验证明回复电压最大值越大,变压器油纸绝缘老化越严重。文献[4]通过热加速老化论证了绝缘老化程度随着中心时间常数变小而增大。由于大部分研究都是基于单个特征量进行定性分析,并没有充分挖掘多个特征量之间的内在联系,对于多个特征量变化趋势不一致的情况,将无法做出准确地判断。为了获得变压器绝缘状态更准确的信息,有研究者利用人工智能算法建立了综合评判模型。例如,文献[5]在搜集大量测试数据的基础上,利用层次分析法建立了变压器绝缘的标准状态向量表。文献[6]考虑绝缘状态的不确定性,将各指标模糊化,并结合粗糙集理论对变压器进行状态评估。为了克服单一方法的局限性,研究人员开发出了将不同方法组合起来的综合模型,如粗糙集与贝叶斯理论相结合、神经网络与范例推理、聚类分析和模糊理论的结合等等。

上述方法都是基于单一分类器对绝缘状态进行分类,单一分类器存在局部收敛、过拟合等问题。而分类器群则是基于集成学习的思想,通过不同的样本集和特征集训练多个基分类器,最后对各个分类器的输出进行综合分析得出最终评估结果。研究表明^[7],分类器群不仅可以克服单分类器的局限性,还能够提高分类的准确性。不同的属性特征可以从不同的方面进行分类,集成学习将各分类结果加权综合,能有效地降低陷入局部最优的风险。

鉴于以上分析,在回复电压法所得相关特征量较少的情况下,单个的分类器往往不能提取出有效地分类信息。下面融合核主成分分析与集成学习理论,建立变压器油纸绝缘的分类器群模型。首先对所测样本集进行矩阵变换,利用核函数的映射特性,得到高维的特征属性集,有效地提取出分类特征量。然后在核空间中用随机森林法建立集成学习的分类器群,对各基分类器进行加权求和,得到变压器绝缘油纸状态评估结果。

1 回复电压法

1.1 回复电压法基本原理

• 30 •

RVM 是通过回复电压测试仪,测得回复电压曲线的特征量,进而对变压器油纸绝缘状态进行评估,由于其无损特性,在绝缘状态评估领域得到了广泛应用^[8]。基本原理如下:首先用直流高压使绝缘介质两端极化,保持充电状态一段时间 t_c ,使束缚电荷游离到介质表面;然后撤掉直流电压并且短接两极,这时候会发生放电过程,表面电荷释放,内部则会缓慢地去极化,持续放电时间为 t_d ;当 $t_d < t_c$ 时断开短接线,停止放电过程,这时候残余电荷将在绝缘介质两端建立一个电势差,即形成回复电压。利用多次测量的电压值可以绘制回复电压极化谱^[9],对曲线的特征量进行分析即可以对电力变压器的绝缘状态做出判断。图1和图2分别为回复电压测试电路和理论曲线。

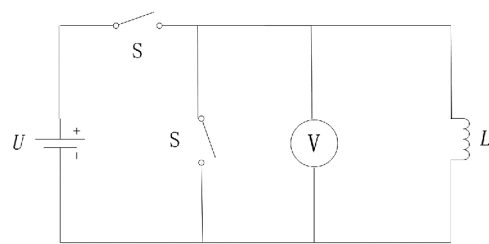


图1 回复电压测试电路

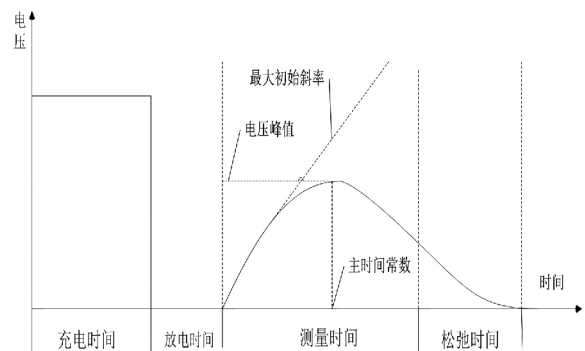


图2 回复电压曲线

1.2 特征量的选取

选取回复电压极化谱峰值电压 U_{rmp} 、主时间常数 t_{cdom} 、最大初始斜率 S_{rmax} 作为电力变压器油纸绝缘状态评估指标^[10]。

极化谱峰值电压 U_{rmp} 取回复电压曲线中的极大值,其值的大小与变压器绝缘状态成反比,其表达式为

$$U_{rmp} = \max(U_{rmax}(t_c)) \quad (1)$$

式中: t_c 为测试时间; U_{rmax} 为回复电压最大值。

主时间常数 t_{cdom} 为极化谱峰值所对应的测量时间,其值越小,变压器绝缘状态越差,其表达式为

$$t_{cdom} = \operatorname{argmax}(U_{\max}(t_c)) \quad (2)$$

最大初始斜率 S_{\max} 反映了绝缘介质极化的速率,其值越大,极化响应速度越快,即绝缘状态越差,其表达式为

$$S_{\max} = \max(\lim_{t \rightarrow 0} \frac{du_t}{dt}) \quad (3)$$

式中 u_t 为回复电压值。

2 基于随机森林的分类器群

随着电力系统的不断发展,准确的电力变压器状态评估对电网的规划运行有重大的指导意义。基于变压器状态特征量的分类技术是当前较为常用的分类技术之一,它主要根据变压器的某些属性,例如油中溶解气体、油中糖醛含量、运行年限等统计信息,借助机器学习的分类方法映射到不同的类状态。目前,使用较为广泛的分类技术主要有朴素贝叶斯 (naive bayes, NB)、支持向量机 (support vector machine, SVM)、决策树 (decision tree, DT)、神经网络 (neural networks, NN) 等。但是这些单分类器存在过拟合、大量调参、局部最优等问题,对于复杂的大数据并不能取得很好的效果,于是使用多个元分类器进行分类,综合分类结果形成最终结果的分类器群的思想应运而生。随机森林作为一种集成算法被提出来,能够有效地克服单分类器的缺陷,且速度快、可扩展,更不需要调整大量参数^[11]。

集成学习的集成效果取决于单个学习器的准确性和各学习器之间的差异性。如何提高这两方面的性能,是集成学习的研究重点。随机森林中为了提高各基学习器的差异性,对样本和属性都进行了随机选择,差异性的扩大进一步提升了随机森林模型的泛华能力^[12]。

基本步骤分为两个部分:随机采样与完全分裂。

1) 随机采样:首先对输入的样本矩阵进行行采样,也就是随机抽取一定量的样本,每一次抽取都采用有放回的方式,保证抽取的样本个数小于原始的样本集,这样得到的样本集可能存在重复的样本。用抽样得到的部分样本来训练一个决策树,每棵树输入样本的随机性极大地降低了出现过拟合的概

率;而对于特征量的选取,采用随机的列采样,随机得到的部分属性用于决策树的分裂。

2) 完全分裂:决策树的分裂属性可以选择信息增益^[13]等策略,并且保证每一个节点都不能继续分裂。

按照上面方法得到的每一棵决策树,并没有用到所有的样本和所有的特征属性,所以单棵树分类性能较低。但是把多个较弱的分类器组合成分类器群后,分类能力就会大大增强,并且只有两个参数需要控制,使得分类过程大大简化。

每个决策树都将得到一个分类结果,对分类结果进行投票,选取票数最多的类别将作为最终的结果,具体流程如图 3 所示。

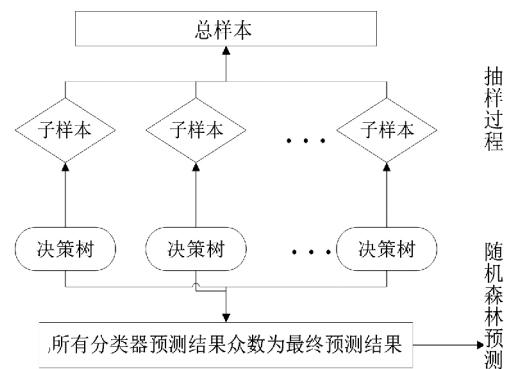


图 3 算法流程

3 核主成分分析

单个分类器之间的差异性是影响分类器群性能的关键因素,而对于特征量偏少的数据集,将会导致各分类器之间差异性不大,往往不能提取出有效的分类信息。而核主成分分析是一种调整数据集维数,并提取主要特征量的有效手段^[14]。通过特定的核函数将输入空间的主成分呈非线性关系映射到新的核特征空间,新的特征空间可以是任意维数,甚至是无穷维。核主成分分析可以增加样本特征量维数,广泛应用于特征提取、回归分析等领域^[15]。

设有一组样本集 $x_k \in R^n (k = 1, \dots, m)$, 通过核函数进行非线性映射,其协方差矩阵表示为

$$C = \frac{1}{n} \sum_{i=1}^n \phi(x_i) \cdot (\phi(x_i))^T \quad (4)$$

式中: ϕ 为非线性映射函数。

由于 $\phi(x)$ 无法直接求出,所以引入核函数^[16]来求解 C 的特征值和特征向量,即

$$n\lambda a = K_a \quad (5)$$

式中: K 为核矩阵 $K(x_i, x_j) = \phi(x_i \cdot x_j)$; a 满足 $v = \sum_{i=1}^n a_i \phi(x_i)$ 。

为了保证 $\sum_{i=1}^n a_i \phi(x_i) = 0$,需要对 K 进行中心化,即

$$\left\{ \begin{aligned} K' &= K - I_n K - K I_n + I_n K I_n \\ (I_n)_{ij} &= \frac{1}{n} \end{aligned} \right. \quad (6)$$

通过求解矩阵方程,可以得到 K' 的特征值和特征向量,利用式(7)可以求得样本矩阵 X 经非线性变换 ϕ 后在 v 上的投影,即

$$V = v \cdot \phi(X) = \sum_{i,j=1}^n a_i K(x_i, x_j) \quad (7)$$

4 评估模型的建立

1) 利用回复电压法测量所需特征量,并建立样本矩阵 X 。

2) z -score 标准化是一种基于均值和标准差的数据处理方法,可有效地减小样本信息的丢失,故采用 z -score 标准化来处理原始数据,其公式为

$$X' = (x - \mu) / \delta \quad (8)$$

式中: X' 为标准化后的样本矩阵; μ 和 δ 分别为 X 的平均数和标准差。

3) 对 X' 进行核主成分分析,核函数用高斯径向基函数,即

$$K = K(x', x_i) = \exp\left(-\frac{\|x' - x_i\|^2}{\sigma^2}\right) \quad (9)$$

式中,核参数 σ 取 0.1。

4) 由于特征值的大小与特征量对样本集影响呈正相关,所以去掉特征值 $\leq 10^{-6}$ 的特征量,剩下的特征量即可看作是核主成分。

5) 将经过核主成分分析的样本集输入随机森林分类器进行训练得到所需评估模型。

5 实例分析

为了获取数据样本,在实验室搭建了变压器油

纸绝缘物理模型,并在 130℃ 下进行加速老化试验。将测得的数据分为 3 类:第 1 类是绝缘纸聚合度大于 500 的,认为其绝缘良好;第 2 类是绝缘纸聚合度在 250~500 范围内的,认为其绝缘较差,需引起注意;第 3 类是绝缘纸聚合度小于 250 的,认为其绝缘劣化严重,需进行检修或退出运行^[17]。

为验证所建立模型的有效性和正确性,将基于核主成分分析的随机森林分类(RFC)模型、BP 神经网络和模糊 C 均值分类器进行比较。BP 网络分类器分为 3 层,其中输入层为 3 节点,隐含层为 9 节点,输出为对应变压器油纸绝缘状态发展阶段 1~3。FCM 选用的聚类系数 $m=2$,分类数 $c=3$ 。

通过实验得到的 330 组样本分为两部分:一部分用于训练分类器;一部分用于分类器的测试。通过对测试结果进行统计分析,分类器群 RFC 模型的准确率达到较高的 92.2%,比单分类器的 BP、FCM 分别高 6.6%、16.7%。比较而言,所建立的 RFC 模型明显优于其他两种典型单分类器。

表1 BP、FCM 和 RFC 模型性能

序号	输入特征量			模型输出			实际状态
	m_1	m_2	m_3	BP	FCM	RFC	
1	0.53	38.5	0.11	1	2	1	1
2	0.29	41.1	0.12	1	1	1	1
3	0.55	35.8	0.15	1	1	2	2
4	0.70	24.3	0.18	2	3	2	2
5	0.83	21.5	0.17	2	3	3	3
6	1.20	21.6	2.10	3	3	3	3

由表 1 列出的部分数据可以看出,第 1 个样本应为第 1 类,但其特征量 m_1 偏大,导致 FCM 将其误判为第 2 类。第 3 个样本由于 m_2 偏大,导致 BP 和 FCM 均误判为第 1 类。第 4 个样本由于 m_2 偏小,导致 FCM 误判为第 3 类。第 5 个样本由于 m_3 偏小,导致 BP 误判为第 2 类。而 RFC 模型得出的结论则与实际状态完全相符。

由于核主成分分析能够把低维的特征量映射成高维,提取变压器状态信息更为充分,而随机选择不同的核特征量来构造不同的基分类器,能够从不同方面学习数据集的状态信息,从而提高了变压器油纸绝缘评估的准确性。

由于实际样本测试中会存在噪声数据,所以对实

验所得样本集添加了5%的随机噪声。对于加噪的特征量进行分类, BP、FCM和RFC模型的准确率分别为73%、69%和87%。3种模型准确率均有所下降,但是RFC模型受到的影响明显要低于BP和FCM,这是由于随机森林理论构造的分类器群能够将单分类器之间的偶然因素抵消,进而缩小分类器群的误差,体现出更好的抗干扰能力和更好的稳定性。

6 结 语

1) 利用回复电压法能够无损地诊断电力变压器油纸绝缘状态,并通过对极化谱特征量的提取分析建立油纸绝缘评估体系。

2) 分类器群能够克服单分类器的局限性,提高分类预测的能力,具有更好的鲁棒性。

3) 采用核主成分分析将低维的特征量非线性地映射到高维的核特征空间,有助于分类器状态信息的提取,并且扩大了单个分类器之间的差异性,提高了分类器群的分类性能。

参考文献

[1] 王有元,袁园,李剑,等. 变压器油纸绝缘可靠性的威布尔混合评估模型[J]. 高电压技术, 2010, 36(4): 842-848.

[2] 郭洪英. 基于回复电压极化谱小波包变换的变压器绝缘老化诊断方法[J]. 电力系统保护与控制, 2016, 44(24): 170-175.

[3] 廖瑞金,孙会刚,袁泉,等. 采用回复电压法分析油纸绝缘老化特征量[J]. 高电压技术, 2011, 37(1): 136-142.

[4] 邹阳,蔡金锭. 油纸绝缘变压器时域极化谱特性实验分析[J]. 电工技术学报, 2015, 30(12): 307-313.

[5] 黄云程,蔡金锭. 融合改进层次分析与灰色关联法评估油纸绝缘状态[J]. 仪器仪表学报, 2015, 36(9): 2083-2090.

[6] 谢松,邹阳,蔡金锭. 基于模糊粗糙集的变压器油纸绝缘状态评估[J]. 仪器仪表技术, 2017, 38(1): 190-197.

[7] 方匡南,吴见彬,朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-37.

[8] 江修波,黄彦婕,张涛. 变压器油纸绝缘水分的回复电压分析法[J]. 高电压技术, 2010, 36(12): 2984-2989.

[9] 周利军,李先浪,王晓剑,等. 基于回复电压曲线的油

纸绝缘状态评估[J]. 高电压技术, 2014, 40(2): 489-496.

[10] Tapan K Saha, Prithwiraj Purkait, Frank Muller. Deriving an Equivalent Circuit of Transformers Insulation for Understanding the Dielectric Response Measurements [J]. IEEE Transactions on Power Delivery, 2005, 20(1): 149-157.

[11] 邱一卉,米红. 基于随机森林和转导推理的特征值提取方法[J]. 厦门大学学报(自然科学版), 2010, 49(3): 333-338.

[12] 赵腾,王林童,张焰,等. 采用互信息与随机森林算法的用户用电关联因素辨识及用电量预测方法[J]. 中国电机工程学报, 2016, 36(3): 604-614.

[13] 贺建章,王海波,季知祥,等. 基于随机森林理论的配电变压器重过载预测[J]. 电网技术, 2017, 41(8): 2593-2597.

[14] 肖慧,刘苏东,黄小燕,等. 基于核主成分的神经网络集合预报建模研究[J]. 计算机仿真, 2010, 27(12): 163-166.

[15] 胡青,孙才新,杜林,等. 核主成分分析与随机森林相结合的变压器故障诊断方法[J]. 高电压技术, 2010, 36(7): 1725-1729.

[16] 王向红,朱昌明,毛汉领,等. 基于核主成分分析及支持向量机的水轮机叶片裂纹源定位[J]. 振动与冲击, 2010, 29(11): 226-229.

[17] 单长星. 基于回复电压法的变压器绝缘测试系统的研制及应用[D]. 成都: 西南交通大学, 2013.

作者简介:

张丽平(1977), 工程师, 研究方向为配电系统运行与控制等;

吴杰康(1965), 教授、博士生导师, 研究方向为电力系统运行与控制等;

陶飞达(1986), 高级工程师, 主要从事配电网管理及新技术的应用研究;

曾振达(1984), 高级工程师, 主要从事配电网管理及新技术的应用研究;

梁浩浩(1988), 硕士研究生, 研究方向为电力系统运行与控制等;

邹志强(1987), 工程师, 主要从事配电网管理及新技术的应用研究;

黄智鹏(1988), 助理工程师, 主要从事配电网管理及新技术的应用研究;

杨夏(1982), 工程师, 主要从事配电网管理及新技术的应用研究。

(收稿日期: 2018-01-06)