DOI:10.19651/j.cnki.emt.1903364

基于随机森林的变压器油中溶解气体浓度预测*

徐肖伟[1] 李鹤健[2] 于 虹[1] 刘可真[2] 赵勇军[3] 盛戈皞[4]
(1.云南电网有限责任公司电力科学研究院 昆明 650217; 2.昆明理工大学 电力工程学院 昆明 650504;
3.云南电力技术有限责任公司 昆明 650000; 4.上海交通大学 电气工程系 上海 200240)

摘 要: 对油中溶解气体浓度进行分析及发展趋势预测,可以为变压器的状态评估提供重要的依据。传统的离线DGA方法因易导致延迟判断变压器的运行状态,造成一定的经济损失,现已不适用于油中溶解气体浓度分析及预测。因此,提出一种基于随机森林的变压器油中溶解气体浓度预测模型,以更准确地分析与预测油中溶解气体浓度。该模型以7种气体浓度构成特征向量空间,作为可视输入,并以目标气体浓度作为输出。试验结果表明,相较于传统的机器学习方法(BPNN、RBF和SVM),随机森林模型能更准确地预测油中溶解气体浓度,且需要调整参数少、训练效率高。通过算例分析,验证了该方法的有效性。

关键词: 变压器;油中溶解气体;随机森林;预测
中图分类号: TM411;TP181 **文献标识码:** A **国家标准学科分类代码:** 470.4037

Concentration prediction of dissolved gases in transformer oil based on random forest

Xu Xiaowei[1] Li Hejian[2] Yu Hong[1] Liu Kezhen[2] Zhao Yongjun[3] Sheng Gehao[4]
(1. Electric Power Research Institute, Yunnan Power Grid Co., Ltd., Kunming 650217, China;
2. Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming 650504, China;
3. Yunnan Electric Power Technology Co., Ltd., Kunming 650000, China;
4. Department of Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: The analysis and prediction of the trend the dissolved gas will have in oil can provide crucial basis for estimating the running status of the transformer. The traditional off-line DGA excruciatingly leads to deferred estimation of the running status of the transformer and hence the financial loss will not be prevented. At present, the traditional off-line DGA does not apply in the interest. Therefore, a random forest(RF) model is first suggested for more accurately analyzing and predicting the trend of the dissolved gas in oil. Seven types of gas build the feature vector space as the input of the model, and hence the output is the objective dissolved gas in oil. Experiments show that the RF model is a more accurate approach to predict the dissolved gas by comparison with the traditional machine learning methods (BPNN, RBF and SVM), the RF model is highly efficient in training and need to adjust less parameters. As a result, case analysis verifies effectiveness of the proposed model.

Keywords: transformer, dissolved gas in oil, random forest, prediction

0 引 言

通过油中溶解气体分析[1](dissolved gas analysis, DGA)可以辨识变压器内外部潜伏性的故障及其发展态势,是电力行业公认的一种诊断变压器故障的可行方法[2-3]。因此,对油中溶解气体浓度进行预测,可为电力变压器的运行状况提供依据[4],具有重要现实意义。

传统的离线DGA方法具有周期长、依靠人员经验、误差大、操作复杂等特点,不能实时准确地分析油中溶解气体的浓度,也不能快速精准地预测气体溶解浓度的发展趋势,导致延迟判断变压器的运行状态,容易产生一定的经济损失[5]。随着20世纪80年代人工智能领域机器学习方法的兴起,机器学习方法以其优异的分类或预测性能[6],被引入到变压器状态控制领域,用于判别与预测变压器的运行状

收稿日期:2019-07-20
*基金项目:国家自然科学基金资助项目(51477100)、云南电网有限责任公司科技项目(YNKJXM20180736)资助

态[7]。相较于传统的离线 DGA 方法,机器学习方具有实时性、误差较小、操作简单等优点[8],对保证变压器运行状态的稳定性有着重要的意义,且在使用过程中不需要人员的介入,因此也是实现变压器智能化控制的有效途径。目前,在变压器 DGA 预测以及故障诊断中应用的典型机器学习方法包括人工神经网络(artificial neural networks,ANN)、支持向量机(support machine learning,SVM)等[9-11]。一般而言,机器学习模型的泛化性能越好,发展趋势的预测能力越强。机器学习方法的泛化性能往往跟其所应用的对象有关,无法脱离具体应用对象来比较哪一种机器学习方法更有效。因此,随着近年来众多新的机器学习方法的提出,尝试将这些新方法应用到 DGA 预测中是颇具研究价值的。

随机森林(random forest,RF)是一种结合集成学习(ensemble learning,EL)与决策树的新型机器学习方法[12,13],由 Leo Breiman 于 2001 年提出的。文献研究表明,RF 在预测上具有优异的性能,具体表现在:需调整参数少、训练效率高、容错性好、不易过拟合等[14-15]。鉴于此,本文提出一种基于随机森林的油中溶解气体浓度预测模型,利用该模型预测多种气体的溶解浓度,并在此基础上进行对比实验分析,以验证该模型的有效性与实用性。

1 油中气体浓度预测

变压器在投运过程中,因绝缘油和固体绝缘老化裂解等分解出极少量气体,主要为氢气(H_2)、甲烷(CH_4)、乙烷(C_2H_6)、乙烯(C_2H_4)、乙炔(C_2H_2)、一氧化碳(CO)、二氧化碳(CO_2)。当变压器的运行状态处于异常时,油中溶解气体的浓度会相对应地迅速增加[16]。例如变压器油过热时,CH_4 和 C_2H_4 的溶解浓度迅速增加。因此,在评估变压器的运行状态时,应该充分考虑油中溶解气体的浓度,并以此作为重要依据来预测变压器的运行状态。

假设变压器油中各类溶解气体浓度可以用一个列向量来表示,即 $X_t=[x_t^1, x_t^2, \cdots, x_t^n]^{\mathrm{T}}$,其中 x_t^n 为第 n 种气体在 t 时刻的溶解浓度,且在文中 n 为溶解气体种类数目,因此油中气体溶解浓度预测问题可以被视为多变量、较复杂的预测问题。进一步地,预测任务可以描述为 $\boldsymbol{X}_{t+1}=f(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_t, \lambda)$,其中 λ 为模型 f 中的参数向量,预测值为 $\hat{\boldsymbol{X}}_{t+1}=f(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_t, \hat{\lambda})$,则损失函数可以表示如下:

$$J(\lambda)=\frac{1}{m}\sum_{t=1}^{m}(x_t-\hat{x}_t)^2 \tag{1}$$

其中,损失函数 $J(\lambda)$ 代表了模型的预测误差,也是评估模型有效性的重要指标。当预测模型的预测值越逼近实际值时,损失函数 $J(\lambda)$ 也越小,预测效果也越好。此外在本研究中,n 取值为 7,表示本模型预测涉及 7 种气体。

2 随机森林回归模型

2.1 随机森林回归原理

随机森林是一种 EL 类算法,由多棵决策树组成的,其本质是决策树算法的组合与改进[17]。如图 1 所示,随机森林由多棵 CART(classification and regression tree)构成,可用 CART 的集合来表示,即:$\{h(X, \Theta_k)\mid k=1, 2, \cdots, N\}$,$X$ 表示输入向量,Θ_k 表示生成 k 棵子树,其集合中生长的子树都是基于 Bootstrap 方法抽取的独立样本[18],且子树都具有相同的分布,最后统计得出最终预测结果。

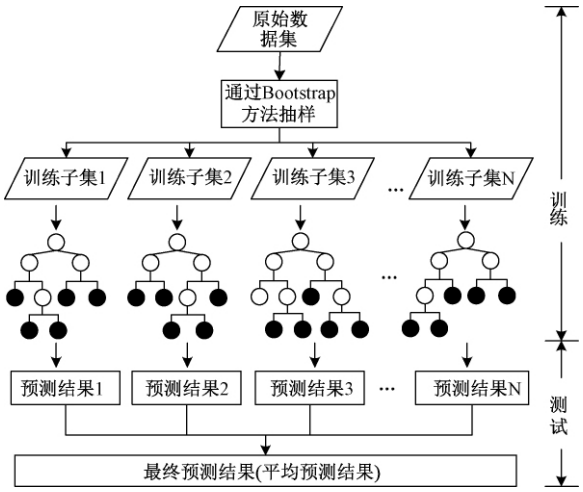

图 1 RF 回归原理

在 RF 回归算法中,最终预测结果是在子树都有预测值的基础上,按照所有子树预测的平均值来作为最终预测的结果。RF 模型性能跟两个因素有关,一是子树的整体性能,二是子树间的差异度。子树的整体性能可以从两方面保证,一是单个子树性能足够好,二是子树棵数恰当。子树间差异度决定了 RF 模型对模式空间的覆盖能力,差异度越大,回归预测效果越好。RF 模型从两方面获取子树间的差异:1)数据上,利用了 Bootstrap 抽样方法,有放回地从原始训练集中抽样产生 N 个独立同分布的训练数据集;2)结构上,在生成子树时,从特征集中随机抽取出一个子集对子树的结点进行分裂。因此,分裂随机特征子集容量成为 RF 模型在使用中必须确定的关键参数。

本文采用了 Forest-RI 形式,若训练集有 M 维,随机选择 $F(F\leqslant M)$ 个特征向量进行,如果 F 取得足够小,随之基决策树间的相关性趋于减弱;同时,基决策树集成的效果随着 F 的增大而提高。综合考虑,通常需要按照经验公式(2)确定 F 值。

$$F=1+\log_2 d \tag{2}$$

式中:d 为原始输入特征数。

2.2 随机森林回归模型的构建过程

本文所提 RF 预测模型如图 2 所示,随机森林回归模型构建步骤可归纳如下:

1)采用 Bootstrap 抽样方法,从训练数据集中抽取 N 个训练样本子集,形成训练集 $S_i\{i=1, 2, \cdots, N\}$。

2)针对上述每一训练集,生成对应的子树 $CART_1$、

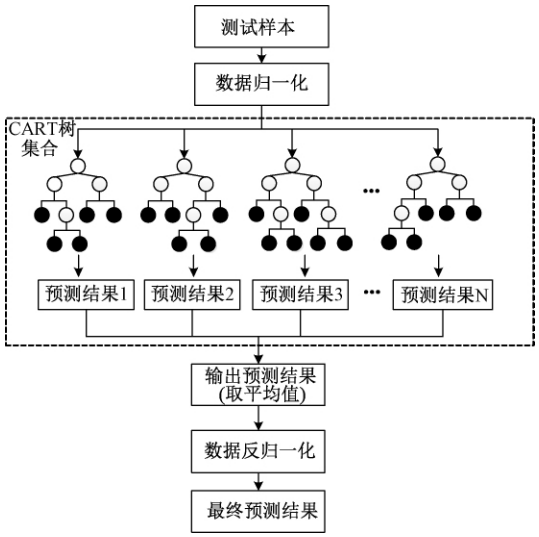

图 2　RF 预测模型

$CART_2$、…、$CART_N$。具体可分为下述两小步：

(1)若训练集有 M 维，则从 M 个属性中随机抽 F 个特征向量作为当前节点的分裂属性集。

(2)以这 F 个属性作为特征向量，对该节点进行分裂，子树完整生长而不剪枝。

3)利用测试数据集测试上步模型的性能，得到子树输出的预测值 $CART_1(Test)$、$CART_2(Test)$、…、$CART_N(Test)$。

4)采取以平均值的方式，统计 N 棵决策树输出的预测值，并将所有子树输出的平均预测值反归一化作为最终预测值。

3　算例分析

为了验证模型的性能，进行实验验证。实验计算机硬件配置为：CPU2.50 GHz，内存 4.00 GB；软件环境为 MATLAB2016a；操作系统为 Windows10，采用 MATLAB 自带随机森林工具箱创建随机森林，其它通过编程实现。

3.1　数据集划分

文中以某 220 kV 变压器油色谱在线监测装置 2018 年 7 月 11 日到 2018 年 12 月 29 日的油色谱数据为例，其中监测周期为 1 天，共 171 组数据。表 1 给出了 10 个样本数据示例，在本文研究中，所有溶解气体浓度的单位均为 μL/L。

表 1　样本示例　　μL/L

日期	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
2018-07-11	18.80	6.29	2.41	11.18	4.62	139.74	1 398.94
2018-07-12	17.46	5.83	2.28	10.55	4.24	129.99	1 271.93
2018-07-13	18.30	6.03	2.27	10.94	4.43	135.35	1 335.57
2018-07-14	17.17	5.42	2.04	9.89	3.98	121.84	1 144.18
2018-07-15	15.77	5.62	1.98	9.67	4.19	129.17	1 207.66
2018-07-16	16.21	5.03	1.87	9.36	3.65	114.24	1 067.34
2018-07-17	17.59	5.53	2.08	10.10	4.05	128.12	1 180.42
2018-07-18	17.52	5.50	2.08	10.12	4.04	128.11	1 180.42
2018-07-19	16.03	4.85	1.82	8.95	3.46	111.54	1 039.99
2018-07-20	16.45	4.86	1.74	9.02	3.56	116.23	1 054.89

如表 2 所示，在本文随机森林回归模型应用及对比试验中，将全部样本数据中 137 组作为训练集，其余 34 组作为测试集，并在此基础上展开对比试验，以验证本文回归模型的有效性。本文试验结果的评价标准均采用平均相对百分误差和最大相对百分误差两个指标，其表达如下：

$$\delta = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{\hat{x}_t - x_t}{x_t}\right| \times 100\% \tag{3}$$

$$\max\delta = \max\left|\frac{\hat{x}_t - x_t}{x_t}\right| \times 100\% \tag{4}$$

3.2　RF 回归模型预测结果

本节分析以油中溶解乙炔(C_2H_2)浓度的预测为例。在本次试验过程中，RF 回归模型的两大关键参数取值分别是：子树棵树(ntree)取默认值为 500，分裂特征子集容量(mtry)根据 1.1 小节所示经验公式(2)取值为 4。在 3.1 节所描述试验数据集的基础上，本节采用 RF 回归模型预测油中溶解乙炔气体浓度，试验研究的测试结果表示如表 3 所示。

表 2　训练集与测试集容量

模型	训练样本数	测试样本数
RF	137	34
BP	137	34
RBF	137	34
SVM	137	34

表 3 详细给出了 34 个测试样本的预测结果，并进一步依据 3.1 节所示两个指标计算得到平均相对百分误差和最大相对百分误差。图 3 直观地显示了 34 个测试样本预测结果与真实值的拟合程度。试验结果表明，RF 回归模

表 3 RF 回归模型预测 C_2H_2 浓度的结果

编号	真实值/(μL/L)	RF 预测值/(μL/L)	相对误差/%	编号	真实值/(μL/L)	RF 预测值/(μL/L)	相对误差/%
1	1.937 8	1.860 9	3.97	18	2.009 4	1.931 6	3.87
2	2.314 6	2.363 6	2.12	19	2.675 7	2.590 8	3.17
3	3.459 3	3.610 3	4.36	20	2.126 4	2.200 0	3.46
4	2.122 8	2.054 0	3.24	21	3.178 6	3.333 2	4.86
5	3.707 8	3.606 4	2.73	22	2.785 5	2.914 1	4.62
6	1.894 3	1.825 4	3.64	23	3.902 5	3.760 0	3.65
7	2.559 6	2.445 5	4.46	24	2.105 6	2.209 8	4.95
8	2.287 4	2.187 8	4.36	25	2.861 0	2.750 8	3.88
9	2.091 6	1.993 8	4.67	26	4.106 0	3.960 6	3.54
10	4.354 0	4.147 9	4.73	27	3.639 8	3.787 7	4.06
11	3.049 5	3.126 9	2.54	28	3.422 4	3.515 2	2.71
12	3.138 7	3.288 4	4.77	29	2.740 8	2.841 7	3.68
13	2.807 9	2.685 5	4.36	30	2.772 8	2.634 4	4.99
14	4.368 6	4.290 9	1.78	31	4.617 8	4.591 9	0.56
15	4.457 8	4.393 0	1.46	32	1.878 4	1.792 1	4.59
16	1.918 7	1.826 6	4.80	33	2.450 4	2.331 4	4.85
17	4.091 7	4.216 3	3.05	34	3.385 8	3.449 5	1.88
平均相对误差 3.66%				最大相对误差 4.99%			

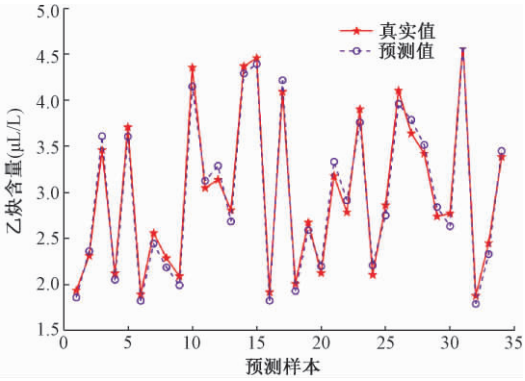

图 3 RF 回归模型预测 C_2H_2 浓度的结果

型的平均测试相对误差为 3.66%,最大测试相对误差为 4.99%。上述结果表明,针对油中溶解乙炔浓度预测,RF 回归模型具有优异且稳定的性能。

3.3 与其他方法对比

在本领域中,最常用的机器学习方法有 BPNN、RBF、SVM 等传统机器学习方法。本节将 RF 回归模型与上述 3 种方法进行比较研究,以对比验证 RF 回归模型的有效性。BPNN 模型、RBF 模型和 SVM 模型参数设置如表 4 所示。

从表 5 中,可以看出,在相同数据集基础上,与常用的传统机器学习方法相较而言,无论是从平均相对误差方面,还是最大相对误差方面,RF 回归模型预测油中溶解乙炔浓度的性能都是最优的。综上所述,可以认为 RF 回归模型预测油中溶解气体浓度比 BPNN、RBF、SVM 模型都更有效。此外,RF 回归模型具有较少调参、容错性好、不易过拟合等优点,在小样本的情况下,RF 回归模型仍有不俗的表现。

表 4 BPNN 模型、RBF 模型与 SVM 模型参数

BPANN					
隐层单元数	激活函数	学习率	训练方法	目标值	训练次数
16	tansig	0.01	traingd	0.0001	1 000
RBF		SVM			
平滑因子	核函数	高斯核函数参数 σ		惩罚因子 C	
30	RBF	8.4193		1	

表 5 不同模型下 C_2H_4 预测结果 %

模型	平均相对误差	最大相对误差
BPNN	5.47	13.04
RBF	4.78	15.79
SVM	4.59	7.18
RF	3.66	4.99

3.4 其余油中溶解气体浓度预测结果

如前所述,3.2 节与 3.3 节均是以乙炔气体溶解浓度为例,验证了文中所提模型的有效性。同理,在相同的数据集和模型的基础上,对其他几种溶解气体浓度进行预测,结果如表 6 所示。

表 6　其他气体浓度预测结果　　%

气体	平均相对误差				最大相对误差			
	RF	BPNN	RBF	SVM	RF	BPNN	RBF	SVM
H_2	3.78	5.84	4.92	4.81	5.56	13.13	15.44	8.72
CH_4	2.21	5.56	4.83	4.52	3.89	14.19	15.32	6.89
C_2H_6	2.43	5.62	5.01	4.76	4.01	13.88	15.68	7.25
C_2H_4	3.85	5.92	5.56	5.32	5.47	14.37	16.16	8.41
CO	4.27	7.56	6.17	5.89	6.37	18.25	18.32	8.92
CO_2	4.68	8.21	6.64	6.14	6.69	18.56	18.57	8.89

由表 6 可知，RF 回归模型的预测平均相对误差和最大相对误差均低于对比研究的 3 种传统机器学习方法(BPNN、RBF、SVM)，具有较高的预测稳定性和可靠性。

4　结　　论

本文在 RF 回归理论的基础上，提出基于 RF 的 DGA 预测模型。以某 220 kV 变压器油中气体溶解度数据为例，对文中所提的 RF 回归模型进行了训练与测试，并与经典的 BPNN、RBF、SVM 机器学习方法进行对比试验。试验结果表明，在相同数据集的基础上，相较于经典的 BPNN、RBF、SVM 机器学习方法，RF 预测准确性明显更好，且更具稳定性。此外，RF 回归模型仅涉及两大关键参数调整，对使用者的经验水平要求较低，从而降低了算法使用难度。因此，可以认为 RF 回归模型是预测油中溶解气体浓度的有效工具。然而文中油中溶解气体浓度样本数据有限，无法得知大数据平台下 RF 回归模型的预测精度，下一步将研究大数据下 RF 回归模型的改进与应用。

参考文献

[1] 彭刚，周舟，唐松平，等. 基于时序分析及变量修正的变压器故障预测[J]. 电子测量技术，2018，41(12)：96-99.

[2] 代杰杰，宋辉，杨祎，等. 基于油中气体分析的变压器故障诊断 ReLU-DBN 方法[J]. 电网技术，2018，42(2)：658-664.

[3] 公茂法，柳岩妮，王来河，等. 基于混沌优化粒子群 BP 神经网络的电力变压器故障诊断[J]. 电测与仪表，2016，53(15)：13-16,32.

[4] 江秀臣，盛戈皞. 电力设备状态大数据分析的研究和应用[J]. 高电压技术，2018，44(4)：1041-1050.

[5] 张施令，姚强. 基于 WNN-GNN-SVM 组合算法的变压器油色谱时间序列预测模型[J]. 电力自动化设备，2018，38(9)：155-161.

[6] 董浩，李明星，张淑清，等. 基于核主成分分析和极限学习机的短期电力负荷预测[J]. 电子测量与仪器学报，2018，32(1)：188-193.

[7] 代杰杰，宋辉，盛戈皞，等. 采用 LSTM 网络的电力变压器运行状态预测方法研究[J]. 高电压技术，2018，44(4)：1099-1106.

[8] 李璐，吴其洲，刘桢杞，等. 基于随机森林的铝铸件内部缺陷类型识别研究[J]. 国外电子测量技术，2018，37(1)：64-68.

[9] 张奎，王建南，王肖峰. 基于神经网络的变压器故障诊断[J]. 电子测量技术，2017，40(12)：98-101.

[10] 唐勇波，丰娟. KTA-SVM 的变压器油中溶解气体浓度预测[J]. 控制工程，2017，24(11)：77-81.

[11] 司马莉萍，舒乃秋，左婧，等. 基于灰关联和模糊支持向量机的变压器油中溶解气体浓度的预测[J]. 电力系统保护与控制，2012，40(19)：41-46.

[12] 熊鹏文，林虹，宋爱国，等. 基于随机森林回归的手臂末端力的软测量方法[J]. 仪器仪表学报，2017，38(10)：2400-2406.

[13] CUTLER A，CUTLER D R，STEVENS J R. Random forests[J]. Machine Learning，2011，45(1)：157-176.

[14] 黄晗，孙堃，刘达. 基于随机森林的电力系统小时负荷预测研究[J]. 智慧电力，2018，46(5)：8-14.

[15] 黄青平，李玉娇，刘松，等. 基于模糊聚类与随机森林的短期负荷预测[J]. 电测与仪表，2017，54(23)：41-46.

[16] 杨廷方，张航，黄立滨，等. 基于改进型主成分分析的电力变压器潜伏性故障诊断[J]. 电力自动化设备，2015，35(6)：149-153,165.

[17] 王奕森，夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术，2018，12(1)：49-55.

[18] 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)：32-38.

作者简介

徐肖伟，硕士，高级工程师，主要从事高电压试验技术、变压器绝缘方面的试验研究工作。

E-mail：40858480@qq.com

刘可真(通信作者)，教授，博士，从事电力数据挖掘和特高压直流输电保护与控制研究。

E-mail：liukzh@foxmail.com