**机器学习在变压器故障诊断中的应用**

马洪斌 杨飞 王振平 宋俊锋 候宪华

(国网枣庄供电公司 山东枣庄 277000）

**摘要：**气体分析（DGA）技术是目前对油浸变压器进行故障诊断中最方便、有效的手段之一，它往往能较准确、可靠地发现逐步发展的潜伏性故障，防止由此引起重大事故。在油中溶解气体的诊断方面主要采用IEC三比值法，为及时发现变压器故障隐患发挥了重要作用。但其诊断的准确率往往只能达到80%左右，故障诊断的准确率偏低。机器学习技术被广泛用于探索数据的内在规律和对数据实现分类预测，本文通过对多种分类方法进行比较，验证基于随机森林的变压器故障诊断具有较高的准确性。

**关键词：**机器学习、随机森林、变压器、故障诊断

中图分类号：TM411

文献标识码：A

Application of Machine Learning in Transformer

Fault Diagnosis

MA Hongbin, LIU Jian, WANG Zhenping, SONG Junfeng, HOU Xianhua

(State Grid Zaozhuang Power Supply Company, Shandong Zaozhuang 277000)

Abstract: Gas Analysis (DGA) technology is one of the most convenient and effective means for fault diagnosis of oil-immersed transformers, it can often accurately and reliably detect the latent faults which are gradually developing, and prevent the major accidents caused by it. The IEC three-ratio method is mainly used in the diagnosis of dissolved gas in oil, which plays an important role in the detection of transformer faults. But the accuracy of fault diagnosis can only reach about 80% , and the accuracy of fault diagnosis is low. Machine learning technology is widely used to explore the inherent law of data and to classify and predict the data. This paper verifies the transformer fault diagnosis based on random forest with high accuracy by comparing several classification methods.

KEYWORDS: Machine Learning, Random Forest, Transformer, Fault Diagnosis

**0 引言**

变压器作为电力系统的重要组成部分，其运行状态决定了整个系统的安全稳定。一旦变压器发生故障，将带来巨大的经济损失。因此对变压器的故障进行有效地诊断对保证变压器的安全可靠运行，具有非常重要的意义。随着机器学习的兴起，越来越多的研究人员将机器学习应用的变压器故障诊断中，取得较好的效果。文献[1]提出了基于加权K近邻算法的变压器故障诊断，根据待分类样本在特征空间中K个最近邻样本中的多数样本的类别来进行分类，因此具有直观、无需先验统计知识、无师学习等特点。文献[2]提出了基于油中溶解气体的支持向量机变压器故障诊断，有效解决了传统学习方法的“维数灾难”和“过学习”等问题。文献[3]提出了基于知识粗糙度的多变量决策树在变压器故障诊断系统中的应用，有效地简化了决策树．减少诊断信息的冗余性，诊断效率高．结果易于被人理解。

**1 机器学习方法**

**1.1 支持向量机**

支持向量机(Support vector machine，SVM)是在统计学习理论的基础上发展出的一种模式识别方法，是一种全局最优求解算法而不是求得局部极小值，具有很好的泛化能力。SVM具有坚定的理论基础在人脸识别、文本自动分类等众多领域展现获得广泛的应用[4]。

SVM最大的局限在于其核函数的选择，通常采用的核函数有线性核函数、多项式核函数、径向基核函数和二层神经网络，不同的核函数的选择对其模式识别的结果有较大的影响。主要应用SVM进行二分类，在对多分类问题的处理上需要更进一步的研究。

**1.2逻辑回归**

逻辑回归又称为逻辑斯蒂回归分析，因其训练速度较快具有较好的分类效果，广泛应用在分类中。逻辑回归模型是线性回归模型和逻辑回归模型的结合。

逻辑回归主要是寻找预测函数和损失函数，预测函数是输入变量预测输出变量结果概率的一个函数。要评价预测函数预测的结果与实际值之间的差别，需要构造一个损失函数作为衡量标准,损失函数表示了训练样本输入之后预测值与真实值的平均差异。损失函数也为目标函数，通常我们就是通过优化算法来使损失函数最小[5]。

**1.3 K-最近邻分类**

K-最近邻就是给定一个样本，确定其类别号，找到样本在训练集中最近的k个样本，然后找出k个样本中出现次数最多的一类，其标号就是样本的标号。

在最近邻算法中，通常采用的是欧氏距离，对每个训练样本在进行距离计算时没有赋予加权，因此每个样本的差异很难体现。通常通过高斯核函数引出距离公式，通过训练集学习核参数，得到一个与训练集对应的加权矩阵。再对测试集进行分类时通过引入的加权矩阵就可以表现出样本之间的差异，从而提高分类的效率[6]。

**1.4 贝叶斯分类**

贝叶斯理论假设的是事件的结果是不确定，通过事件发生的概率来量化它。事件过去发生的概率已知的话，则可以通过数学方法计算出未来发生的概率。

朴素贝叶斯模型假定所有的条件属性相互独立，条件属性共有的父结点就是类属性的结点。假设就是条件独立假设。假定构造出来的模型结构简单，降低了贝叶斯网络的复杂性。

实际中的数据往往具有复杂性、复杂性、不完整性，这样在构造贝叶斯模型时其条件独立假设就难以满足[7]。

**1.5 随机梯度下降**

随机梯度下降是在支持向量机原问题的形式下给定的。梯度下降进行迭代更新使用的梯度是训练样本计算出来的，每一次的迭代需要遍历整个训练集，计算量大。此外还容易陷入局部极小值，无法计算全局最优解[8]。随机梯度下降不是计算梯度的准确值，是用梯度的无偏差估计来代替梯度。

**1.6 决策树**

决策树学习主要是从一组无规则、无次序的事例中推理出以决策树形式来表示的分类规则。决策树的创建过程中由于训练样本过少或数据集中存在噪声，导致决策树的许多分枝反映出训练样本中异常现象。决策树会对训练样本过度拟合，而决策树的剪枝正是解决过拟合问题。剪纸分为预剪纸和后剪纸，预剪纸提前做一些判断，会导致决策树过早停止构造，但是效率高，适合数据量大的情况，通常应用于数据预测中。决策树结构简单，便于理解，算法描述简单，分类速度快，在数据处理中广泛使用。

通常我们采用的是Quinlan提出的基于信息熵的决策树算法，它是通过非递增学习算法形成的。决策树中一个叶节点代表了一个类别属性值，而非叶节点则代表了非类别属性，树枝代表了这个属性的值。非叶节点是通过属性值中最大信息量的非类别属性相关联。通常采用熵值来表示一个非叶节点的信息量[9]。

**1.7随机森林**

随机森林是利用重抽样方法从原始样本中抽取多个样本，对每个样本进行决策树建模，然后组合多个决策树，形成一个森林。通过形成的森林对未知的样本进行预测，选取投票最多的分类。因此随机森林相比较决策树可以大大提高分类的准确率[10]。

随机森林的实现过程如下：

（1）原始训练集为N，通过Bootstrap法有放回地随机抽取k个新的自助样本集，训练构造出k个决策树模型。

（2）每一个决策树的每个节点的m个变量中，计算不纯度选取具有分类能力的变量作为分类属性点，，其阈值通过分类点确定。

（3）将多个决策树组成随机森林，用随机森林对预测数据进行模式识别，以分类器投票结果决定分类结果。

**2 实例分析**

对于大型电力变压器，目前几乎都是用油来绝缘和散热，电力变压器油与油中的固体有机绝缘材料在运行电压下因电、热、氧化和局部电弧等多种因素作用会逐渐老化、裂解，产生少量的、、和等低分子烃类，以及、和等气体，并多数溶解在油中。油中溶解气体的组分和含量在一定程度上反映出电力变压器绝缘老化或故障的程度，可以作为反映电力设备异常的特征量。通过对运行中的电力变压器定期分析溶解于油中的气体组分、含量和产气速率，能够及早发现电力变压器内部存在的潜伏性故障。

以,,,,作为特征向量进行变压器故障模式识别。综合变压器的故障原因，将变压器的故障类型分为中低温过热、高温过热、低能量放电和高能量放电，分类问题主要是数值型结构，因此将变压器的状态正常、中低温过热、高温过热、低能量放电和高能量放电分别标注为0、1、2、3、4。对采集到的变压器油中气体数据115组，分成训练数据（75组）和测试数据（40组），采用支持向量机、最近邻分类、随机梯度下降、逻辑回归、贝叶斯分类、决策树和随机森林进行模式识别，其结果如下：

表1 不同分类方法的模式识别结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实际故  障标签 | 决策树 | SVM | KNN | 随机梯度下降 | 随机  森林 | 逻辑  回归 | 贝叶斯 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 4 | 4 | 0 | 4 | 3 | 4 | 4 | 4 |
| 3 | 4 | 0 | 4 | 4 | 4 | 4 | 4 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 2 | 0 | 1 | 4 | 0 | 1 | 0 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 1 | 1 | 0 | 2 | 2 | 1 | 1 | 2 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 2 | 2 | 1 | 1 | 2 |
| 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 2 | 2 | 1 | 1 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 3 | 0 | 0 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 0 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 4 | 0 | 4 | 4 | 4 | 4 | 4 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 4 | 4 | 0 | 1 | 3 | 0 | 0 | 0 |
| 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 |
| 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 |
| 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 |
| 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 正确/总数 | **35/40** | **19/40** | **33/40** | **30/40** | **36/40** | **34/40** | **32/40** |

表2 不同分类方法的模式识别结果准确率

|  |  |  |  |
| --- | --- | --- | --- |
| 分类方法 | 标准化处理分类 | 正则化处理分类 | 直接分类 |
| 决策树 | 0.675 | 0.775 | 0.875 |
| 支持向量机 | 0.875 | 0.725 | 0.475 |
| 最近邻 | 0.825 | 0.825 | 0.825 |
| 贝叶斯 | 0.7 | 0.75 | 0.8 |
| 随机梯度下降 | 0.8 | 0.625 | 0.75 |
| 随机森林 | 0.875 | 0.85 | 0.9 |
| 逻辑回归 | 0.775 | 0.7 | 0.85 |

通常在使用训练样本获得模型时需要对数据进行预处理，数据的标准化或正则化。标准化是为了方便数据的下一步处理，而进行的数据缩放等变换，并不是为了方便与其他数据一同处理或比较；正则化而是利用先验知识，在处理过程中引入正则化因子增加引导约束的作用，在逻辑回归中使用正则化，可有效降低过拟合的现象。

通过分类结果可以看出数据的处理对最近邻分类没有影响，分类结果均为82.5%，对数据标准化、正则化可以提高支持向量机分类的准确率。随机森林的分类准确率比决策树分类的准确率要高，不对数据进行处理时，故障类型分类的准确率最高为90%。数据的形态对随机森林的结果影响不大，其分类的准确率均为7种分类方法的最高。

**3 结论**

分别采用支持向量机、最近邻分类、逻辑回归、随机梯度下降、贝叶斯分类、决策树和随机森林对变压器的五种状态进行模式识别。可以看出在故障类型的诊断上均具有较高的准确率，随机森林作为决策树的组合分类模型，其故障诊断的准确率最高，可以广泛地应用在实际变压器故障诊断中。

**4 参考文献**

[1]刘君,游家训,梁薇等.基于加权K近邻算法的变压器故障诊断[J].电力系统及其自动化,2010,32(5):59-62.

[2]祖文超.基于油中溶解气体的支持向量机变压器故障诊断[D].华北电力大学,2013.

[3]黎静华,栗然.基于知识粗糙度的多变量决策树在变压器故障诊断系统中的应用[J].电力自动化设备,2005,25(10):40-43.

[4]肖燕彩,张清.基于模糊支持向量机的变压器故障诊断[J].北京交通大学学报,2012,36(1):117-121.

[5]安波.基于逻辑回归模型的垃圾邮件过滤系统的研究[D].哈尔滨工程大学,2009.

[6]朱鹏飞,胡清华.基于核距离学习的K近邻分类[J].计算机科学,2009,36(8):59-61.

[7]董立岩,刘光远，范淼淼等.混合式朴素贝叶斯分类模型[J].吉林大学学报(信息科学版),2007，25(1):57-61.

[8]陶秉墨,鲁淑霞.基于自适应随机梯度下降方法的非平衡数据分类[J].计算机科学,2018,45(6):487-492.

[9]赵锦阳,卢会国,蒋娟萍等.基于改进决策树的故障诊断方法研究[J].成都信息工程大学学报,2018,33(6):624-631.

[10]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.

作者简介：

姓名：马洪斌；出生年份：1991-7-6；性别：男；主要从事的工作：变压器检修、维护，故障诊断及负荷预测；作者Email：[457395074@qq.com](mailto:457395074@qq.com)。