基于高维拉曼光谱数据的变压器油纸绝缘老化评估方法研究

陈新岗[1,2]，陈姝婷[1*]，杨定坤[3]，罗 浩[1]，杨 平[1]，崔炜康[1]

1. 重庆理工大学电气与电子工程学院，重庆 400054
2. 重庆市能源互联网工程技术研究中心，重庆 400054
3. 重庆大学输配电装备及系统安全与新技术国家重点实验室，重庆 400054

摘 要 采用激光拉曼光谱技术对变压器油纸绝缘老化状态检测是一种有效的方法。随着样本量的扩充，亟待处理的数据集维度逐渐增大，研究适用于高维拉曼光谱数据的变压器油纸绝缘老化评估方法具有重要的意义。设计与现场变压器内部绝缘结构相似的油纸绝缘环境，进行加速热老化实验并定期采样，获取到10类老化程度依次递增的油样本，采用激光拉曼光谱技术对样本进行检测。选用复合稀疏导数建模法对样本原始拉曼光谱数据预处理，可以一步完成去噪与基线校正；引入差异特征选取方法筛选不同老化程度下光谱中变化显著的特征，计算同一拉曼频移下不同老化程度的特征点数据集方差，选择差异较大的数据序列所对应的拉曼特征变量，设定方差阈值为0.5进行特征选择，每个样本都从1 023个光谱特征点抽取出304个特征点进行后续分析；针对变压器油纸绝缘老化拉曼光谱高维样本数据集，引入多种不同类型的算法对其处理。分别运用K-means聚类算法、Fisher算法与随机森林算法对获取到的样本预处理后的数据建立模型，引入评估准确度、提升度以及Kappa系数对各算法建立的模型判别效果进行评估。结果表明：有监督学习的Fisher算法与随机森林算法效果较好，相对于无监督学习的K-means聚类算法，模型判别能力分别提升了1.166 6和1.95，论证了有监督学习模型在变压器油纸绝缘老化的评估中具有判别优势；从模型判别准确度和Kappa系数来看，强分类器随机森林算法建立的判别模型均高于Fisher判别模型，其准确度提升了10%，且Kappa系数上升了0.111 5，论证了随机森林算法作为由多个单一分类器组成的强分类器，相对单一分类器来说，在变压器油纸绝缘老化的评估中模型的泛化能力较好，且模型较为稳定可靠。通过对三种不同类型的算法对比，确定了在变压器油纸绝缘老化评估中，有监督学习强分类器随机森林算法的判别优势，为变压器油纸绝缘老化的有效评估打下了基础。

关键词 变压器；油纸绝缘；拉曼光谱；高维数据集；老化评估

中图分类号：O657.37 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)05-1463-07

引 言

电力变压器的正常运行关乎电力的安全生产和供电可靠性，是电力系统中非常重要的设备。油浸式变压器一般采用绝缘油和以纤维素为主要成分的绝缘油纸作为其内绝缘结构，运行过程中受到热应力、电场应力、机械应力及环境应力等影响，绝缘能力逐渐降低。能够适时对变压器内部油纸绝缘状态进行准确的评估，对电网的安全、稳定运行具有重要意义[1]。

目前对变压器进行老化检测的方法多为油中溶解气体检测，糠醛检测，聚合度检测等，但是这些检测方法在实际运用中还存在一定的局限性。油中溶解气体的检测步骤较为复杂，需要对气体与油进行分离且不能做到样品的无损检测；糠醛的检测需要用到甲醇萃取，操作过程复杂，对操作人员技术及环境要求较高；聚合度的检测需要停电吊芯，不容易获取相关数据。选择拉曼光谱检测技术对变压器油混合物进行检测，可以不与油样直接接触，检测重复性较好；同时，拉曼光谱法对电磁波抗干扰能力强，降低了变压器油纸绝缘老化评估时对实际检测环境的要求；且拉曼光谱法还可以与

收稿日期：2020-06-05，**修订日期**：2020-09-26

基金项目：国家自然科学基金项目(51977017)，重庆市教委科学技术研究项目(KJ1400917)和重庆理工大学研究生创新项目(YCX20192057)资助

作者简介：陈新岗，1968年生，重庆理工大学电气与电子工程学院教授 e-mail：chenxingang@cqut.edu.cn

*通讯作者 e-mail：490210758@qq.com

光纤传感技术很好的结合起来，能够实现对现场变压器快速准确的评估。

为了将变压器油拉曼检测光谱与能够反映不同老化程度的特征物联系起来，需要探索更适合现有数据特征背景的算法[2]，继而对变压器油纸绝缘老化更精确的评估。为了使样本分布稀疏，重叠性少，易于分类，需要增加实验次数以扩充数据库；随着实验的进行，样本数量逐渐增多，且每条光谱的信息特征点较多，样本数据集维数升高；高维样本数据集的特点一般为：数据规模较大，包含的信息价值密度低，容易引起维数灾难，且对高维样本数据建模的过程中容易出现训练效率低或者时间成本升高等很多问题。因此引入三种不同类型的算法对得到的光谱样本数据进行分析。

本文研究中，通过模拟现场变压器油纸绝缘加速热老化过程，得到不同老化时间的油样本，运用实验室搭建的老化特征物拉曼检测平台，对样本原始拉曼光谱信号进行采集[3-4]；采用复合稀疏导数建模法对原始光谱进行去噪和基线校正处理；引入 Filter 法对差异较大的特征进行选择[5]；基于特征选择后的样本，分别采用 K-means 聚类算法[6]、Fisher 算法[7]和随机森林算法[8]对变压器油纸绝缘老化程度进行判别分析，建立老化状态评估模型，对测试集样本进行归类，诊断样本属于哪一老化天数(老化程度)类别；基于多种评价因素，对比分析各类模型的判别能力。

1 老化评估算法

评估算法有无监督学习和有监督学习之分，无监督学习不考虑已有类别判断，对样本中心进行迭代计算并归类；有监督算法在模型训练时输入已知类别样本信息进行参考，对模型的建立有一定的影响。有监督分类器算法有强分类器和弱分类器之分。强分类器随机抽取训练集的子集，建立多个均具有判别效力的模型，通过投票机制汇总判别结果，进而得出最终分类结果；弱分类器训练数据构建单一判别模型，其分类效率由输入的数据特征决定，输入总体样本中不同的训练集，测试集判别的结果也不相同。本文分别采用代表无监督学习的 K-means 聚类算法、代表有监督学习弱分类器的 Fisher 算法和代表有监督学习强分类器随机森林算法对变压器油纸绝缘老化拉曼光谱分析。

1.1 K-means 聚类算法

K-means 聚类是快速聚类中运用欧氏距离进行样本－聚点计算的一种聚类形式，确定所需要划分的类别数，随机选择相应类别数不相交的初始化聚点，并计算其他各样本到达类聚点的欧氏距离，如式(1)

$$d(x_i, x_j) = [(x_i - x_j)^{\mathrm{T}}(x_i - x_j)]^{\frac{1}{2}} \tag{1}$$

以每个样本最靠近初始聚点原则归类，将样本划分成初始类别后，迭代计算各类别新的聚点并重新归类，直到所有类别聚点不再有变化则迭代结束。

1.2 Fisher 算法

Fisher 算法的原理是通过某些决策函数的计算，将高维数据集样本投影到低维子空间上，使得这些不同类别的数据集样本在低维子空间上的分离性最佳。

设样本训练总体为$\{G_i\}$ ($i \in \{1, 2, \cdots, 10\}$)，$G_i$ 是第 i 类样本的集合。判别函数是构成 Fisher 判别模型的重要部分；Fisher 算法中构造判别函数的原则是不同类别之间距离最大，类别中所有样本距离最小，即要满足式(2)达到最大。

$$\Psi(w) = \frac{\sum_{k=1}^{i}(w^{\mathrm{T}}\mu_k - w^{\mathrm{T}}\bar{\mu})^2}{w^{\mathrm{T}}\left(\sum_{k=1}^{i} v_k\right)w} \tag{2}$$

其中，w^{T} 为投影向量，μ_k 为样本质心，v_k 为协方差矩阵。

Fisher 判别模型建立后，将测试集样本各变量带入判别函数，得到各样本观测值的具体空间位置，计算各样本距离类别组质心位置，距离哪一类别组质心位置最近，就归为此类。

1.3 随机森林算法

随机森林算法是 Breiman 在 2001 年提出的决策树集成分类器，主体思想是将多个单一分类器联系起来，对随机选取的不同特征建立决策树群，之后通过对所有决策树结果进行投票来决定类别归属。该算法在近些年来被广泛运用，在电气研究领域展现了不错的数据处理能力[9]，具有以下优点[10]：能够有效地运用在高维数据集中；能够处理高维数据且不需要降维；内部生成误差为无偏估计；运行效率高；具有较高分类精度且泛化能力强。

采用 bootstrap 重采样技术对训练集样本随机抽取样本子集，每个样本子集通过决策树的生成方式建立独立决策树模型。本文选用 CART(分类回归树)为基础分类器，通过参数调整选择最适合的决策树数量，通过计算 $\log_2^n$(n 为样本中的特征变量个数)，得出指定节点中用于决策树的变量个数。

2 实验部分

2.1 拉曼光谱检测平台设计

设计如图 1 所示的拉曼光谱检测平台，为了避免高温引起过高的暗电流和阅读噪声以提高 CCD 探测器的灵敏度，检测前将其内部工作温度降至零下 10 ℃；为了避免室内光线对样品检测的干扰，整个检测过程在黑暗环境中进行；实验室环境温度为 25 ℃；设置仪器恒定激光功率为 300 mW，数据采集积分时间为 0.3 s，积分次数为 10。

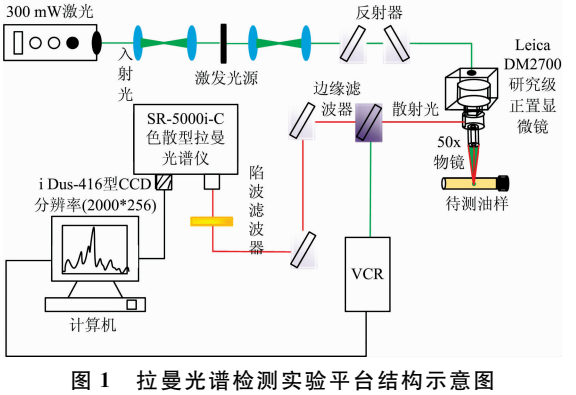

图 1 拉曼光谱检测实验平台结构示意图

Fig. 1 Schematic diagram of Raman spectroscopy detection experimental platform

2.2　变压器油纸绝缘老化实验过程

依据 IEEE 导则进行油纸绝缘加速热老化样本的制备。流程图如图 2。定期取样并获取 10 类老化时间分别为 0，1，3，5，7，9，12，17，21 和 24 d 的 100 个油纸绝缘样本拉曼光谱。

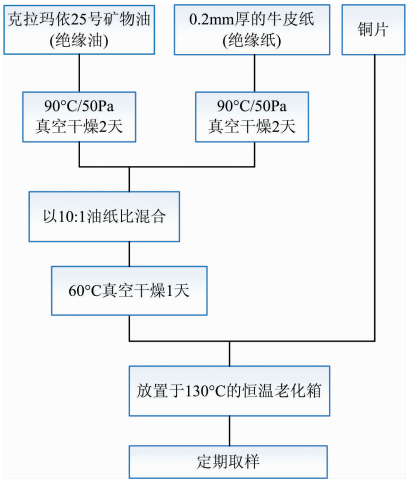

图 2　油纸绝缘加速老化实验流程图

Fig. 2　Flow chart of accelerated aging experiment of oil-paper insulation

3　结果与讨论

从实验中获取到 10 类不同老化天数的变压器油老化拉曼光谱图，图 3 反映了各类光谱图的显著差异。

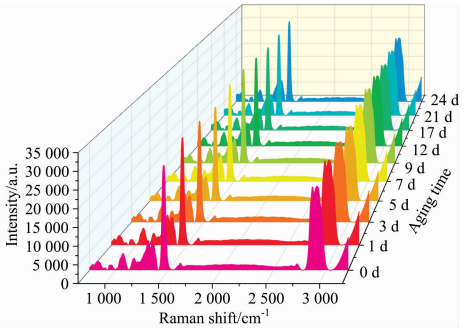

图 3　不同老化天数变压器油拉曼光谱

Fig. 3　Raman spectra of transformer oil for different aging time

3.1　光谱数据预处理

采用复合稀疏导数建模法对原始拉曼光谱数据进行预处理。此方法能够将基线校正和噪声去除两部分合并处理，极大地简化了预处理步骤。

复合稀疏导数建模法的原理是逆行推导光谱信号的分解过程，建立基于构造一个凸优化问题来封装基线和光谱峰的非参数模型，代数推导过程见文献[11]。将原始拉曼光谱信号 y 建模为三个部分，如式(3)

$$y = x + f + w \tag{3}$$

其中，x 为稀疏状峰值信号，f 为低通基线，w 为平稳白色高斯噪声。

通过二次数据保真度项 $\| y-Ly-H\hat{x} \|_2^2$ 提出凸函数优化问题，如式(4)

$$\hat{x} = \mathrm{argmin}_x \left\{ F(X) = \frac{1}{2} \| y - Ly - H\hat{x} \|_2^2 + \sum_i^M \lambda_i \sum_{n=0}^{N_{i-1}} \phi([D_i x]_n) \right\} \tag{4}$$

其中，L，H 分别为低通和高通滤波器，由带状卷积矩阵构成，且满足 $H = I - L$，利用非对称惩罚函数 $\sum_i^M \lambda_i \sum_{n=0}^{N_{i-1}} \phi([D_i x]_n)$ 模拟光谱峰的正则性，$D_i x$ 为谱峰数据 x 的 i 阶差分运算；N_i 为 $D_i x$ 的长度；λ_i 为正则化系数；借鉴 Majorization-Minimization 迭代优化算法求解思路，迭代计算出谱峰信号估计值 $\hat{x}$，提升收敛速度。得出近似于有效峰值信号后，对原始光谱信号减去有效峰值信号部分进行低通滤波处理，过滤掉相对对称的信号，而不引起峰值位置的偏移，得到基线估计值 f；由原始光谱信号减去有效光谱信号 x 和基线估计值 f，得出噪声信号 w。原始拉曼光谱图预处理后的信号分解结果如图 4。经去噪和基线校正后的峰值信号曲线较为平滑，能够更清晰观察到峰值的波动情况。

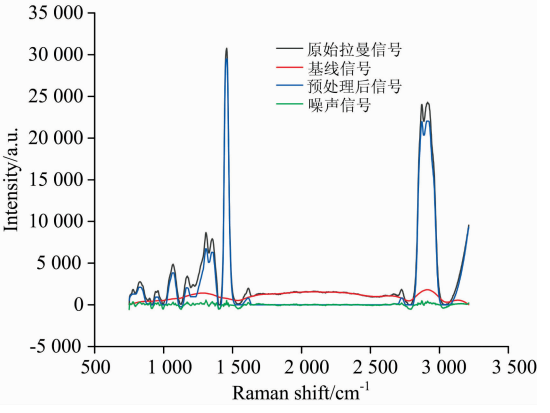

图 4　复合稀疏导数建模预处理

Fig. 4　Preprocessing with compound sparse derivative modeling

3.2　光谱差异特征信息点选择

针对拉曼光谱，前期一般使用特征提取的方法，例如主成分分析，小波包能量熵等方法，这些方法虽然可以快速提取样本中大部分有效信息，但是在整合信息的同时，提取出来的重要成分不能与每个光谱信息点的物理意义联系起来，也不方便对油中溶解物质的老化机理进行后续研究。

如图 5，根据不同老化程度光谱图对比，可以看到其中有很多差异谱段，还有一些谱段处于高度重合状态，若将光谱图全部导入判别模型，可能会因为无效信息过量造成干扰现象。遂采用差异特征选择[12]的方法进行处理，同时保留了光谱信息点包含的物理含义。研究发现，油中溶解的某些化学物质有其对应的拉曼频移特征点[13-14]，对于差异特征点的抽取，有利于进一步探究随着变压器油纸绝缘不断老化，同一拉曼频移下对应的光谱差异特征点强度变化与油中溶解物

质的老化机理关系。

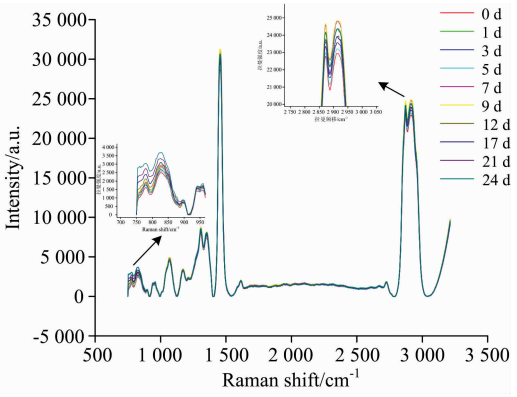

图 5 不同老化程度油样拉曼光谱对比

Fig. 5 Comparison of Raman spectra of oil under different aging degrees

进行特征工程中的特征选择步骤时，较为直接的是 Filter 法，选用方差计算同一拉曼频移下不同老化程度信息点差异，N 为样本数，如式(5)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_i)}{N} \tag{5}$$

方差可以表示一组数据的显著差异，能够较好的将差异特征点提取出来。

如图 6，通过对所有拉曼频移下的数据集方差按从小到大依次排列作图，可以观察到，当方差值小于 0.5 时，基本处于平稳状态，表示不同老化程度下，光谱图部分曲线基本无太大的变化，由于数据预处理出现的微小差异可以忽略不计，当方差大于 0.5 时，曲线走势渐陡，数据集差异显著增大，表示这部分数据集在老化过程中存在某些物质量的变化，能够与不同老化程度下绝缘油中物质的老化机理联系起来，具有可研究性。遂设定方差阈值为 0.5，对每个样本抽取出的 304 个特征信息点进行后续分析。

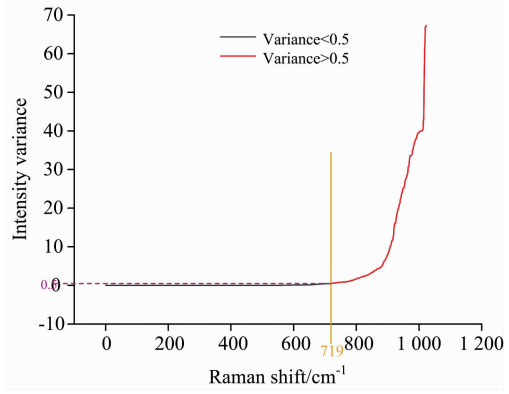

图 6 同一拉曼频移下数据集方差筛选

Fig. 6 Data set variance screening under the same Raman shift

3.3 样本拉曼光谱诊断分类器应用

将预处理后的的 100 组样本数据按 7∶3 比例随机进行训练集和测试集分配，且运用不同算法建模的数据集相同，测试时的数据集也相同。

3.3.1 K-means 聚类算法应用

根据 K-means 算法原理对预处理后的训练集及测试集一起进行聚类处理，预计分为 10 类，选择最大迭代次数为 20 进行计算，如图 7，迭代次数为 9 时，聚点变动趋于稳定。

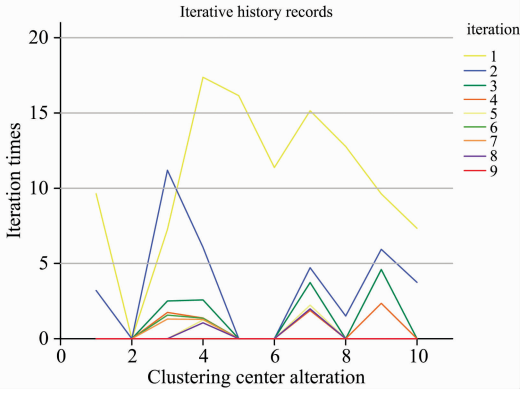

图 7 聚类中心随迭代次数增加的变动

Fig. 7 Changes of cluster centers with increasing number of iterations

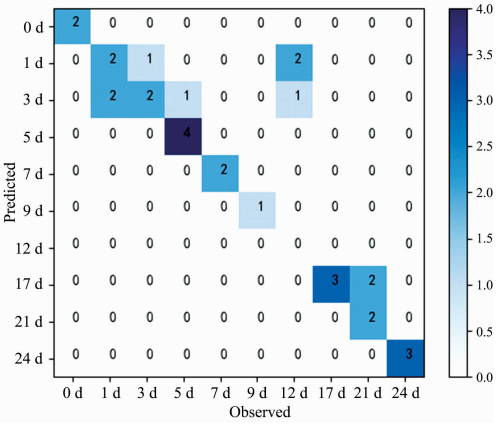

图 8 K-means 聚类算法判别分析结果

Fig. 8 Results of the discriminant analysis with K-means clustering algorithm

通过训练集样本已知类别对划分的类别进行重新标记，统计测试集样本预测类别，并与其实际类别进行比较，如图 8。K-means 聚类算法判别结果显示：一共错判了 9 个测试样本，其中，属于 1，3，5 和 21 d 的部分测试样本错判到相邻类别，表明相邻类别的特征易于混淆，采用无监督聚类的方法对于相邻类别的判别效果并不显著；实际类别为 12 d 的测试样本全都判别到 1 和 3 d 类别，说明此算法在聚类时对于 12 d 的关键特征信息不敏感，导致对这一类别的全部错判。

3.3.2 Fisher 算法应用

根据 Fisher 算法原理对训练集样本总体计算，根据表 1，威尔克 Lambda 表示组内平方和与总平方和的比例，值越小表示组间差异越大，可以看到前三个判别函数的威尔克 Lambda 检验显著性均小于 0.05，即表示用这三个判别函数

建立的模型是有效的。

表 1　判别函数有效性检验

Table 1 Effectiveness test of discriminant function

威尔克 Lambda	卡方	自由度	显著性	威尔克 Lambda
1 直至 9	614.545	135	0.000	0.000
2 直至 9	361.644	112	0.000	0.002
3 直至 9	221.143	91	0.000	0.020
4 直至 9	91.852	72	0.057	0.197
5 直至 9	43.993	55	0.856	0.459
6 直至 9	27.632	40	0.931	0.613
7 直至 9	15.666	27	0.959	0.758
8 直至 9	6.696	16	0.979	0.888
9	1.787	7	0.971	0.969

反映判别函数所能解释的方差变异程度的特征值贡献率恰恰印证了这一结果，如图 9，前三个判别函数累积贡献率显著提升，能够解释的变量占比达到了 98.0%，最大程度的对数据集变量进行了处理，之后的判别函数能解释变量的能力逐渐减弱，累积贡献率曲线趋于平缓，起伏较小。

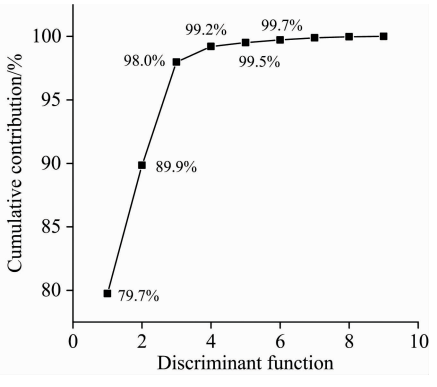

图 9　判别函数累积贡献率

Fig. 9　Cumulative contribution of discriminant function

通过对 70 个样本总体的训练，建立了 3 个主要判别函数。将 30 个测试样本带入判别函数，得到各测试样本的空间坐标；对各测试样本与类质心的距离进行计算，判别细节如图 10，从图 10 可以看出，单一的判别函数不能完全的将不同类别划分出来，需要多种判别函数相结合，才能使不同类别尽可能分开，从而较容易判别未知样本。

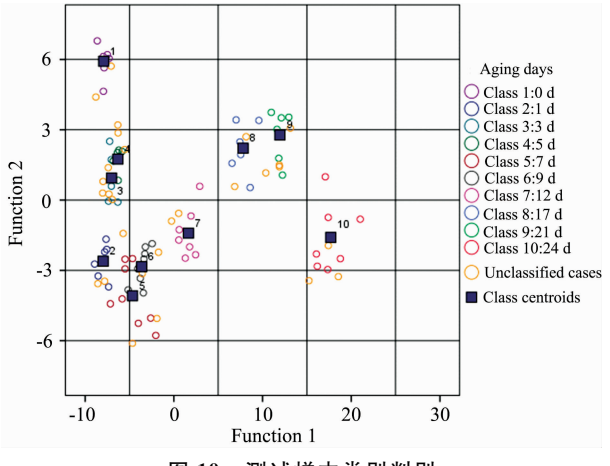

图 10　测试样本类别判别

Fig. 10　Discriminant analysis of test samples

图 11 为 Fisher 模型对测试样本的判别结果，在 1，3，5，12 和 21 d 类分别错判 1 个，由于部分类别组质心分布过于密集，导致相近类别误判情况略为明显。

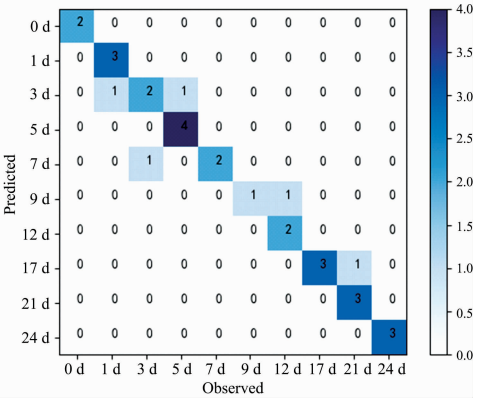

图 11　Fisher 算法判别分析结果

Fig. 11　Results of the discriminant analysis with Fisher algorithm

3.3.3　随机森林算法应用

一般来说，构建随机森林模型时生成的树越多，容错率就越高；但在实际运用中，会选择模型错误率降低至趋近于平稳时所需 CART 的最少棵数，以减少运算量，提高预测速度。

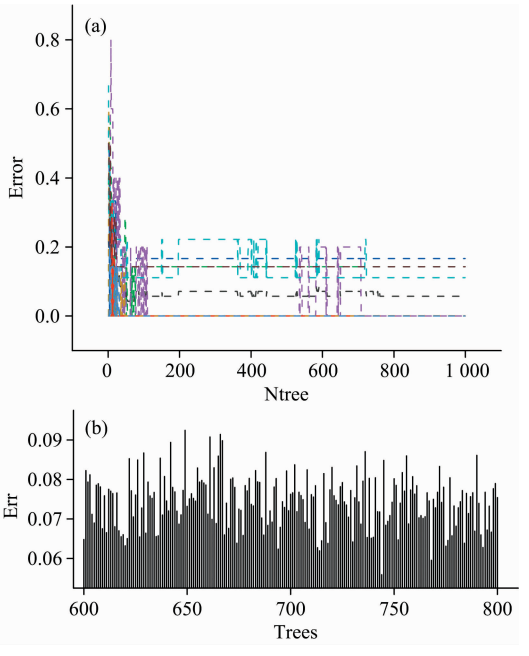

图 12　决策树数量与错误率关系图

Fig. 12　Diagram of the number of decision trees and the error rate

如图 12(a)，设置决策树数量为 1 000，画出错误率和决策树数量的关系图，从图上看出，当决策树数量在 600～800 左右时，错误率稳定。生成 600～800，步长为 1 的数列，迭代计算出错误率最小时需要的棵树，如图 12(b)，当决策树棵树为 744 时，模型最优。

计算指定节点中用于决策树的变量个数，调整好参数后进行随机森林建模，通过计算，OBB 袋外估计错误率为 5.71%，确定了建立的分类模型是较为可靠且稳定的，这与随机森林模型内部计算泛化误差的无偏估计结果一致。将测试集输入已建好的模型中，得到如图 13 的判别结果。

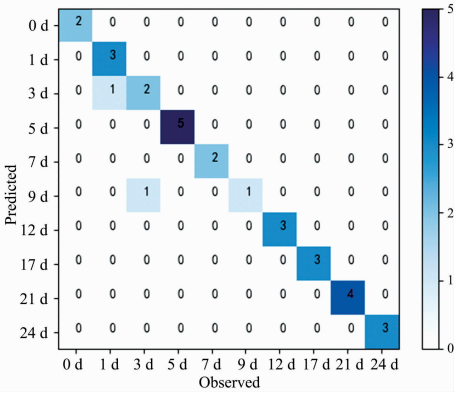

图 13 随机森林判别分析结果

Fig. 13 Results of the discriminant analysis with Random Forest algorithm

运用随机森林模型对测试样本进行类别评估，在 1 和 3 d 类分别错判 1 个。除了出现极少数的相近类别判断错误的问题，实际老化时间为 3 d 类的一个测试样本被判别到 9 d 类，类别错判差异较大。

3.3.4 三种判别方法结果对比

引入多种评价因子[15]对模型及其预测结果进行效果对比。如表 2。

表 2 评估方法效果对比

Table 2 Comparison of evaluation methods

算法类别	评估准确度/%	提升度	Kappa 指数
K-means 聚类算法	70	7.466 7	0.662 9
Fisher 判别算法	83.33	8.633 3	0.813 6
随机森林算法	93.33	9.416 7	0.925 1

计算可知，K-means 聚类算法判别准确率为 70%，Fisher判别算法判别准确率为 83.33%，随机森林算法判别准确率为 93.33%，表明了随机森林算法在变压器油纸绝缘老化拉曼光谱高维数据集处理上的可靠性与有效性；模型提升度是比较模型之间预测能力的提升状况指数，就三种模型的提升度来看，以无监督 K-means 聚类模型为基准，有监督的 Fisher 判别模型和随机森林模型分别提升了 1.166 6 和 1.95，表明了加入已知样本的类别信息，可能会影响模型的分辨能力，使得模型能更好地判别未知样本；Kappa 指数是评价分类结果一致性和信度的重要指标，从三种模型的 Kappa 指数来看，样本判别的实际一致率和随机一致率差别并不显著，但总体来说，随机森林的 Kappa 指数要高于另外两种模型，表明了强分类器在样本增多，数据维度增大后具有良好的表现，随机森林算法经决策树数量和分离节点参数调整后具有较强的分类能力。由于训练集和测试集为程序随机分配，该评价结果也具有一定的普适性。

4 结 论

在实验中进行变压器油加速热老化实验，获取到 10 类不同老化天数的 100 个油老化样本。运用拉曼光谱检测方法对实验室制备不同老化程度油样本进行检测。

(1)选用复合稀疏导数建模法对拉曼光谱进行预处理，能够一步完成去噪和基线校正，且预处理效果较好，与原始光谱曲线相比更平滑。

(2)选取 Filter 法对光谱图中同一拉曼频移差异较大光谱特征点进行特征选择，并设定阈值将差异较大的特征信息点抽取出来，相比于特征提取的方法来说较为直接，且输入数量较少的特征信息点有利于之后建模训练效率的提升。

(3)将样本总体按 7∶3 比例分配训练集与测试集，分别建立变压器油纸绝缘拉曼光谱的 K-means 聚类模型、Fisher 判别模型与随机森林分类模型，通过多种评价因素来验证各模型在高维数据集的分类效率。结果表明，随机森林模型能更准确的评判实验样本的老化程度，判别正确率达到了 93.3%；相比 K-means 聚类算法和 Fisher 算法的判别正确率来看，上升了 23.33%和 10%；有效解决了无监督算法过于依赖数据集的构成和单一分类器在建模时学习的局限性问题，体现了油样本增多后，有监督学习相对于无监督学习，强分类器相对于弱分类器，在变压器油纸绝缘老化评估上的判别优势，为变压器油纸绝缘老化的评估打下了基础。

References

[1] ZOU Jing-xin, CHEN Wei-gen, WAN Fu, et al(邹经鑫，陈伟根，万 福，等). Transactions of China Electrotechnical Society(电工技术学报), 2018, 33(5): 1133.

[2] ZENG Zi-lin, ZHANG Hong-jun, ZHANG Rui, et al(曾子林，张宏军，张 睿，等). Control and Decision(控制与决策), 2014, 29(6): 961.

[3] CHEN Xin-gang, YANG Ding-kun, TAN Hao, et al(陈新岗，杨定坤，谭 昊，等). High Voltage Engineering(高电压技术), 2017, 43(7): 2256.

[4] CHEN Xin-gang, LI Chang-xin, FENG Yu-xuan, et al(陈新岗，李昌鑫，冯煜轩，等). High Voltage Apparatus(高压电器), 2018, 54(9): 117.

[5] ZHANG Tao-tao, HU Ya-nan, LI Yang, et al(张陶陶, 胡亚南, 李　杨, 等). Statistics and Decision(统计与决策), 2017, (4): 18.
[6] LIU Qing-zhen, ZHANG Xiao-yan, CAI Jin-ding(刘庆珍, 张晓燕, 蔡金锭). Power System Protection and Control(电力系统保护与控制), 2019, 47(8): 62.
[7] FAN Zhou, CHEN Wei-gen, WAN Fu, et al(范　舟, 陈伟根, 万　福, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(10): 3117.
[8] FANG Kuang-nan, WU Jian-bin, ZHU Jian-ping, et al(方匡南, 吴见彬, 朱建平, 等). Statistics & Information Forum(统计与信息论坛), 2011, 26(3): 32.
[9] LI Guo, JIANG Xiao-dong(李　国, 江晓东). Proceedings of the CSU-EPSA(电力系统及其自动化学报), 2018, 30(11): 70.
[10] Banfield R E, Hall L O, Bowyer K W, et al. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 173.
[11] Ning Xiaoran, Ivan Selesnick, Laurent Duval. Chemometrics and Intelligent Laboratory Systems, Elsevier, 2014, 139: 156.
[12] Rodriguez-Galiano V F, Luque-Espinar J A, Chica-Olmo M, et al. The Science of the Total Environment, 2018, 624: 661.
[13] CHEN Wei-gen, WAN Fu, GU Zhao-liang, et al(陈伟根, 万　福, 顾朝亮, 等). Transactions of China Electrotechnical Society(电工技术学报), 2016, 31(2): 236.
[14] CHEN Wei-gen, ZHAO Li-zhi, PENG Shang-yi, et al(陈伟根, 赵立志, 彭尚怡, 等). Proceedings of the CSEE(中国电机工程学报), 2014, 34(15): 2485.
[15] Asha Kiranmai S, Jaya Laxmi A. Protection and Control of Modern Power Systems, 2018, 3: 29.

Study on the Evaluation Method of Oil-Paper Insulation Aging in Transformer Based on High Dimensional Raman Spectral Data

CHEN Xin-gang[1, 2], CHEN Shu-ting[1*], YANG Ding-kun[3], LUO Hao[1], YANG Ping[1], CUI Wei-kang[1]
1. Chongqing University of Technology, Chongqing　400054, China
2. Chongqing Energy Internet Engineering Technology Research Center, Chongqing　400054, China
3. State Key Laboratory of Power Transmission Equipment & System Security and New Technology, Chongqing　400054, China

Abstract　Laser Raman spectroscopy is an effective method for detecting the aging state of transformer oil-paper insulation. With the expansion of sample quantity and the gradual increase of data set dimension, it is of great significance to study the evaluation method of oil-paper insulation aging in transformer suitable for high-dimensional Raman spectral data. An oil-paper insulation environment similar to the internal insulation structure of the field transformer was designed, and the accelerated thermal aging experiment was carried out and regularly sampled to obtain ten types of oil samples with increasing aging degrees, then these samples were detected using laser Raman spectroscopy. The compound sparse derivative modeling method was used to preprocess the original Raman spectral data, which can complete the noise elimination and baseline correction in one step. The differential feature selection method was introduced to screen the spectral features with significant changes under different aging degrees, and the variance of the feature point data set with different aging degrees was calculated under the same Raman shift. Furthermore, the Raman feature variable corresponding to the data sequence with a large difference was selected, and the variance threshold was set to 0.5 for feature selection, each sample selected 304 from 1 023 spectral feature points for subsequent analysis. In this paper, many different types of algorithms were introduced to process the high-dimensional sample data set of transformer oil-paper insulation aging Raman spectra. For instance, the K-means clustering algorithm, the Fisher algorithm and Random Forest algorithm were used to establish a model with the preprocessed data of the obtained samples. The evaluation accuracy, lifting degree and Kappa coefficient were introduced to evaluate the discriminant effect of each mathematical model. The results show that supervised learning Fisher algorithm and Random Forest algorithm have a better effect and discriminatory advantage compared with the unsupervised learning k-means clustering algorithm because the discrimination ability of the model is improved by 1.166 6 and 1.95, respectively; Judging from the discrimination accuracy and Kappa coefficient, the discriminant model established by the strong classifier Random Forest algorithm is better than the Fisher discriminant model, for its accuracy is improved by 10%, and the Kappa coefficient is increased by 0.111 5. Compared with a single classifier, a strong classifier composed of multiple single classifiers has better generalization evaluating of transformer oil-paper insulation aging, and the model is more stable and reliable. By comparing three different types of algorithms, the discrimination advantages of the supervised learning strong classifier Random Forest algorithm in evaluating transformer oil-paper insulation aging are determined, which lays the foundation for the effective evaluation of transformer oil-paper insulation aging.

Keywords　Transformers; Oil-paper insulation; Raman spectroscopy; High dimensional data set; Aging assessment

* Corresponding author
(Received Jun. 5, 2020; accepted Sep. 26, 2020)