第42卷，第11期
2022年11月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 11, pp3395-3402
November, 2022

基于 IFSR 异常样本剔除的落叶松木材密度近红外优化模型的研究

张哲宇，李耀翔*，王志远，李春旭

东北林业大学工程技术学院，黑龙江 哈尔滨 150040

摘 要 木材密度可以反映木材的干缩性、抗压抗拉强度等多种物理性质，是重要的木材物理特性。采用近红外光谱技术能够实现木材密度的快速预测，可克服传统检测方法耗费人力、物力、时间的弊端，但建模结果往往受异常样本的影响。为准确识别并剔除样本集中的异常样本，提出一种孤立森林结合学生化残差方法(IFSR)，在利用孤立森林集成特征的优点基础上考虑样本对模型的影响度，可同时检测异常样本与强影响样本。该研究对 181 个落叶松木材样本的近红外光谱及其在常温下的气干密度进行了测定。通过对比多种方法预处理和特征选择方法，确定采用标准正态变量变化(SNV)＋去趋势处理(DT)＋均值中心化(MC)＋标准化(Auto)方法进行预处理，采用竞争性自适应重加权算法(CARS)进行特征波段选择，消除噪声及无关信息对算法的影响，简化数据集，提高算法剔除异常样本的准确性。为验证 IFSR 方法剔除异常样本的能力，将其与蒙特卡洛交互验证(MCCV)、马氏距离(MD)等其他六种异常检测方法对比分析，建立偏最小二乘(PLS)模型对其进行异常检测性能评价。同时在上述基础上采用粒子群寻优-支持向量机回归(PSO-SVR)，BP 神经网络(BPNN)与 PLS 分别建立落叶松木材密度近红外预测模型。结果表明，IFSR 结合 PSO-SVR 方法得到的优化模型预测能力最强，IFSR 可有效剔除奇异样本，提高模型精度。

关键词 近红外；木材密度；异常值检测；孤立森林算法；支持向量机回归

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)11-3395-08

引 言

木材密度可以很好地表征木材的干缩性、抗压抗拉强度等物理性质，同时还是确定加工价值与工艺需求的重要因素，是提高木材利用率中应重点研究的木材材性之一[1-2]，准确、实时地估算木材密度对木材材性预测及合理造材具有重要意义。近红外光谱技术是一种快速、无损的检测技术[3-4]，在近红外定量预测过程中，所采集的样本数据可能存在因人为因素或仪器因素出现的奇异样本或偏离整体的强影响样本，这些异常样本会带偏模型预测的方向，使模型预测结果变得不可靠[5]。因此有必要在近红外建模过程中剔除上述异常现象，以提高模型的精度。目前常用的剔除光谱异常样本的方法包括：蒙特卡洛交互验证[6]，马氏距离[7]，杠杆值检验[8]及光谱残差检验[9]等。近年来新方法也层出不穷，尹宝全等提出一种联合光谱数据 X 与组分信息 Y 的 ODXY 异常样本剔除算法，通过对羊肉近红外样本的异常剔除，证明该算法能很好地提高模型的泛化能力[10]。Brownfield 等将排序差异和算法(sum of ranking differences，SRD)与 Procrustes 分析相结合，通过同时跨窗口评估光谱与组分的异常值，调整参数值来提高异常检测的效率和准确率[11]。以上方法虽然可以有效识别异常样本，但大多受经验阈值或建模偏差的影响，容易在建模前的剔除过程中出现误判，且对复杂样品的异常样本剔除能力相对较差，从而降低了模型的泛化能力及准确性。

孤立森林算法(isolation forest，iForest)在统计学领域被广泛应用于识别高维复杂数据的异常值，其认为特征空间中异常样本是孤立的，可以选取子样本，使用随机超平面构建孤立树 iTree，递归地连续分割数据集，其中正常样本需要分割到孤立树较深层的叶子节点，需要较长的分割路径，而异常样本靠近孤立树根节点，只需较短的分割路径就能孤立出来。孤立森林算法不假设样本与背景空间的概率分布，是一种采用特征集成方法的无监督异常检测方法。目前孤立森林算法已开始应用于高光谱影像的异常识别[12-13]等方面。

收稿日期：2021-09-23，**修订日期**：2022-01-19

基金项目：国家重点研发计划项目(2017YFC0504103)，双一流专项-创新人才培养项目(000/41113102)，黑龙江省应用技术研究与开发计划项目(GA19C006，GA21C030)资助

作者简介：张哲宇，1997 年生，东北林业大学工程技术学院博士研究生 e-mail：1453029789@qq.com

*通讯作者 e-mail：yaoxiangli@nefu.edu.cn

近红外光谱数据的高维性及复杂性，在一定程度上限制了其建模精度及普及性，将孤立森林算法应用于近红外光谱数据将会大大提高数据分析的有效性，但是在实际应用中也遇到两个主要问题，其一近红外光谱谱峰重叠严重，且各个波段共线性强，采用孤立森林划分时不易有效地区分无效波段与特征波段；其二光谱数据全谱波段信息量较大，可能会出现建完孤立树后遗漏有效特征。针对以上两点，在使用孤立森林算法检测近红外光谱异常值前应对光谱数据进行预处理及特征波段选择，以减少噪声等对背景的干扰，同时简化光谱数据，增强光谱特征对比度。

为利用 iForest 高效检测异常目标的优点，同时克服将其直接应用于近红外光谱分析的困难，本研究提出一种孤立森林结合学生化残差方法(isolation forest-studentized residual，IFSR)。首先通过对光谱数据预处理降低噪声、基线漂移等的影响，提高光谱分辨率。再通过选择特征波段，简化光谱数据，突出强相关波长，降低特征峰重叠给 iForest 带来的不确定性，利用 iForest 计算的异常得分，代入计算学生化残差，考虑每个样本对模型的影响程度，若异常得分过大或残差值过大，则可认定该样本为异常样本。

以落叶松木材密度为研究对象，分别采用多种预处理方法与特征波长选择方法对光谱数据进行处理，对比 IFSR 方法与不同异常样本剔除方法处理样本集后的建模效果，验证 IFSR 的异常识别能力。再基于常用的近红外定量分析建模方法：偏最小二乘交叉验证[14](partial least squares，PLS)、BP 神经网络[15](back propagation neural network，BPNN)以及支持向量机回归[16](support vector regression，SVR)建立预测模型，通过对比来得到最优的近红外模型。

1 实验部分

孤立森林结合学生化残差的异常样本剔除算法(IFSR)，其运行共分三步。其原理如图 1 所示。

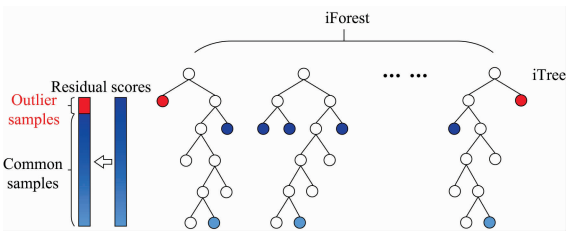

图 1 IFSR 算法原理图

Fig. 1 IFSR algorithm schematic diagram

第一步是对光谱数据进行预处理和特征波段选择，即先简化光谱数据，去除噪音及无关特征对异常样本识别及建模的干扰，提高 IFSR 算法的准确性和对异常样本的敏感性。第二步是训练，即在样本集吸光度矩阵 x 中随机选取一个特征，并在 x 的范围内构建 iTree 进行二叉划分，构建一棵 iTree 时，从 n 个样本中均匀抽样 Ψ 个样本，作为这棵树的训练样本，将大于和小于该值的样本归于左右叶子节点，继续在左右叶子节点重复上述过程，直到达到终止条件：数据不可再分且达到树的最大高度 $l=\log_2(\Psi)$。第三步是预测，即记录测试样本从 iTree 的根节点到外部叶子节点所走过的边数，记为路径长度 $h(x)$。为标准化样本集吸光度矩阵 x 的路径长度 $h(x)$，需要计算树的平均路径长度 $c(n)$

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{1}$$

式(1)中，$H(i)$为调和数($i=1, 2, \cdots, n-1$)，该值可以被估计为 $H(i)+0.577\ 215\ 664\ 9$[17]。最后将 $h(x)$代入，计算样本集 x 的异常得分 S

$$S = 2^{-\frac{E[h(x)]}{c(n)}} \tag{2}$$

式(2)中，$E[h(x)]$为样本 x 在孤立森林的路径长度的期望。当 S 接近 1 时样本被识别为异常样本，当 S 接近 0 时，样本被识别为正常样本，当 S 在 0.5 附近时，无法明确区分样本是否异常。此时利用学生化残差将异常得分考虑在内，计算校正集均方根误差，见式(3)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}} \tag{3}$$

则学生化残差 R_i 为式(4)

$$R_i = \frac{y_i - \hat{y}_i}{\sqrt{\mathrm{RMSE}(1-S_i)}} \tag{4}$$

很明显 R_i 在考虑特征空间异常样本的同时考虑了每个样本对模型的影响度，可以更好地检测异常样本。

1.1 仪器与样本

实验采用美国 ASD 公司制造的 LabSpec® Pro FR/A114260 便携式物质成分分析光谱仪测量近红外光谱。该仪器可选择的光谱范围为：350～2 500，1 000～2 500，1 000～1 800，1 800～2 500，350～1 800 和 350～1 050 nm。光谱分辨率为：3 nm@700 nm，10 nm@1 400 nm，10 nm@2 100 nm。光谱采样间隔为：1.4 nm@350～1 050 nm，2 nm@1 000～2 500 nm。本工作采用的所有算法均在 MATLAB R2017a 软件上操作。

所用样本采自黑龙江省方正县高楞镇星火林场(N45°43′5.73″，E129°13′34.37″)，在落叶松天然次生林区，分别在向阳与背阴面共设立 4 块样地，每块样地大小为 20 m×20 m，在每块样地中选取 3 棵标准木；各标准木经伐倒后，用便携油锯在标准木胸径(胸高 1.3 m 处)附近自下而上连续锯截多个木圆盘，带回实验室经手工剥皮后，在木圆盘上过树芯截取木条，共得到 181 个 2 cm×2 cm×4 cm 的落叶松木材样本，并对每个样本标号记录。在通风干燥的室温(20 ℃)环境中将样本放置 4 周，测得样本的平衡含水率约为 10%，参照《木材密度测定方法(GB/T 1933—2009)》测定木材气干密度。

1.2 光谱数据采集与样本划分

用 80 目的砂纸打磨木材样本各个面各 5 次，使其表面粗糙度参数 Ra 接近 12.5 μm。在样本横切面的两个不同位置用光纤探头扫描各 1 次，每次扫描时间约为 1.5 s，设定扫描期间对样本连续扫描 30 次。取两次测量的平均值为原始光谱数据。得到原始光谱的吸光度如图 2 所示。由图可知在 1 440，1 894 和 2 395 nm 附近处存在明显的吸收峰，且此三

处波段对应的吸收峰在水分子 H—O 键的二倍频吸收带附近，但 1 840～2 500 nm 的光谱存在较大的噪声，因此需要对光谱数据进行降噪及预处理。

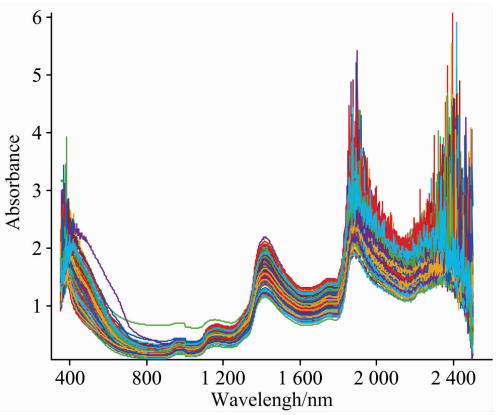

图 2　落叶松木材样本近红外原始光谱图

Fig. 2　Original near-infrared spectra of larch wood samples

采用光谱-理化值共生距离法 SPXY(sample set partitioning based on joint X-Y distance)方法[18]划分样本集，选取样本集的 1/3 作为预测集，2/3 作为校正集，共得到校正集样本 121 个，预测集样本 60 个。SPXY 方法考虑了样本集光谱所含理化值的权重以及与木材密度的关系，以光谱和木材密度为基本参数通过计算样本空间距离来进行样本集划分，所划分的数据集更具有代表性。如表 1 所示，校正集与预测集的木材气干密度均值与标准差均相差不大，样本分散较为均匀。

表 1　校正集与预测集样本统计分析(g·cm^{-3})

Table 1　Statistical analysis of correction set and prediction set results (g·cm^{-3})

样本集	样本数/个	均值	最大值	最小值	标准差
校正集	121	0.521 3	0.743 5	0.411 0	0.066 2
预测集	60	0.525 8	0.620 6	0.413 0	0.066 3

2　结果与讨论

2.1　光谱预处理方法对比与分析

由于近红外光谱在扫描过程中主要靠漫反射来获取物质信息，其光谱通常不仅只包含实验所需要的信息，还包括了很多诸如各种噪音、散射光、以及来自样本本身内部的杂质信息等等，这些噪音及无关信息会干扰预测模型的精度，同时增加识别异常样本的难度。直接分析含有较多干扰信息的光谱数据容易出现误判，可能错误地删除非异常样本，因此本研究在进行异常样本的识别及剔除之前采用光谱预处理方法排除上述杂质信息，从而提高识别的准确度及后续建模的质量和精度。

采用多元散射校正(multiplicative scatter correction，MSC)，标准正态变量变换(standard normal variate transformation，SNV)，去趋势(detrending，DT)，移动平均平滑(moving average smoothing，MAS)，Savitzky-Golay 卷积平滑(Savitzky-Golay smoothing，SGS)，与均值中心化(mean centering，MC)，标准化(autoscaling，Auto)相结合对原始光谱进行预处理，结果如表 2 所示。由表 2 可以看出经预处理后的光谱数据的预测集决定系数 R^2 及均方根误差(root mean squared error of prediction，RMSEP)相较原始光谱均有较大改善，均可以很好地校正光谱的基线漂移及去噪，其中采用 SNV＋DT＋MC＋Auto 联合光谱预处理方法，主因子个数为 5，预测集 R^2 为 0.721 1，RMSEP 相较原始光谱从0.042 2 降为 0.034 7，是不同预处理方法预测结果中的最小值。综合考虑，确定采用 SNV＋DT＋MC＋Auto 作为异常样本剔除及建模前的预处理方法。

表 2　基于不同预处理方法的落叶松木材密度预测结果

Table 2　Prediction results of larch wood density based on different pretreatment methods

预处理方法	主因子数	校正集		预测集	
		R^2	RMSEC	R^2	RMSEP
none	8	0.476 6	0.048 0	0.587 1	0.042 2
MSC	7	0.476 2	0.048 1	0.591 1	0.042 1
SNV	7	0.480 7	0.043 2	0.652 7	0.038 8
SNV＋DT	6	0.495 5	0.044 1	0.711 0	0.037 4
MAS	8	0.554 9	0.040 0	0.701 6	0.039 5
SGS	8	0.548 5	0.042 3	0.699 4	0.040 1
MSC＋MC＋Auto	5	0.525 7	0.048 8	0.689 2	0.034 9
SNV＋DT＋MC＋Auto	5	0.534 9	0.045 3	0.721 1	0.034 7
MAS＋MC＋Auto	6	0.444 1	0.049 9	0.588 1	0.042 1
SGS＋MC＋Auto	6	0.423 2	0.050 1	0.587 7	0.042 2

2.2　光谱特征波段选择与分析

对预处理后的光谱数据，尽管对其进行了降噪，基本消除了基线漂移等对光谱数据的影响，但光谱中还存在大量冗余信息，其共线性对异常样本识别以及后续建模仍有较大影响，因此需要进一步分析光谱数据，提取特征信息。工作中采用竞争性自适应重加权算法(competitive adaptive reweighted sampling method，CARS)提取特征波段。采用 CARS 方法时，设定蒙特卡洛采样次数为 50，以 10 折交叉验证构建最大潜变量因子数为 15 的偏最小二乘模型。

模型整体预测效果相比无特征选择提升明显，其中 R^2 从 0.437 7 提高到 0.894 2，RMSEP 从 0.045 2 降低为 0.019 6。CARS 方法的特征波段选择结果如图 3 所示，图 3(a)表示选取的特征变量数随波长变量子集数的增加的变化趋势图，整体呈逐渐减小趋势，且减小速度逐渐变缓。对比图 3(b)中的交叉验证均方根误差(root mean squared error of cross validation，RMSECV)结果，曲线呈先减小后增加的趋势，随着无关信息的剔除 RMSECV 逐渐减小，模型效果渐优，但当部分有用信息被剔除时，RMSECV 则趋于增加，模型出现过拟合现象，可以确定在波长变量子集数为 40 时得到最优的特征波段集。图 3(c)表示各波段的稳定度随波长变量子集的变化轨迹，其中星号线表示最小 RMSECV 对应的

子集数，稳定度为选择特征波段的主要依据，当有用信息的稳定度变为 0 则 RMSECV 对应也会增加。

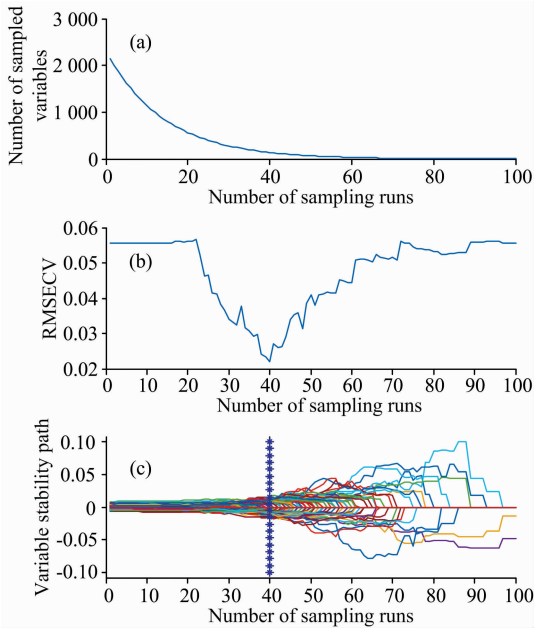

图 3 CARS 波段选择变化趋势图

(a)：选取变量数；(b)：RMSECV；(c)：变量稳定度轨迹

Fig. 3 CARS band selection trend chart

(a)：Number of sampled variables；(b) RMSECV；(c) Variable stability path

2.3 光谱异常值识别方法对比与分析

采用上述波段选择后的光谱数据样本集，剔除其中可能存在的对模型产生强影响的极端样本或异常样本。奇异样本可能存在于光谱数据或组分指标的真值，其可能是由于测量时的人为误差或仪器误差造成的，剔除这些奇异样本是确保预测精度和模型准确性的必要步骤。为验证 IFSR 算法对经预处理及特征选择后的光谱数据的异常筛选能力，分别应用蒙特卡洛交互验证(Monte Carlo cross validation，MCCV)、马氏距离(Mahalanobis distance，MD)、高杠杆值检验(high leverage，HL)、杠杆值与学生化残差 t 检验(high leverage-studentized residual，HLSR)、光谱残差检验(spectral residual，SR)以及基于 XY 变量联合的 ODXY 算法共六种算法与孤立森林算作对比，对上述数据进行异常样本剔除，并在异常样本剔除后建立偏最小二乘交叉验证模型，根据模型的预测能力进行评估。

对于 MCCV 方法，设定蒙特卡洛循环次数为 1 000 次，假设各样本的预测残差均满足正态分布，引入设定参数 q，分别计算预测残差均值 $m(i)$与预测残差标准差 $s(i)$的阈值 T_m和 T_s，超出阈值的样本即为异常样本。设定参数 q 根据 3σ 准则设为 3，T_m，T_s的计算公式如式(5)和式(6)

$$T_m = \frac{\sum_{i=1}^{n} m(i)}{n} + q\sqrt{\frac{\sum_{i=1}^{n}(m(i)-\bar{m})^2}{n}} \tag{5}$$

$$T_s = \frac{\sum_{i=1}^{n} s(i)}{n} + q\sqrt{\frac{\sum_{i=1}^{n}(s(i)-\bar{s})^2}{n}} \tag{6}$$

对于 IFSR 方法，设定 iTree 数量为 100，iTree 训练子样本容量为 256，iTree 的最大特征容量为经特征波长选择方法简化后的光谱波长数，最大迭代次数为 50 次，学生化残差检验的 t 值查阅 t 分布临界值表确定，本研究中 t 临界值为 2.601，其余参数均为默认值。基于 IFSR 方法剔除结果如图 4 所示，共剔除了样本编号为 33，39，107，146，150，172，175 的 7 个异常样本。

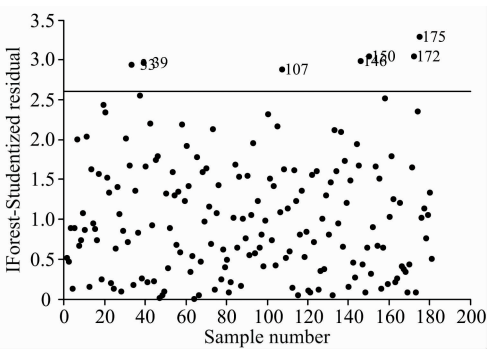

图 4 基于 IFSR 方法的异常样本剔除结果

Fig. 4 Results of abnormal sample elimination based on IFSR method

对其他异常样本剔除方法的阈值采用逐一放回法进行确定，规定每种方法先预选出 20 个异常值，再按照次序从最后一个剔除的样本开始放回，若模型性能没有变差则保留放回，否则剔除，得到最佳模型性能的异常样本数对应的阈值即为最佳阈值。基于六种异常样本剔除方法的样本剔除结果分别如图 5(a—f)所示。

为进一步确认 IFSR 方法的异常样本识别能力，将上述几种异常样本剔除方法剔除后的样本集重新用 SPXY 方法按照 3∶1 划分校正集与预测集，分别建立偏最小二乘交叉验证模型，并对各模型对比评价，得到未经异常样本剔除(Full)与经 IFSR 方法及六种对照方法所建交叉模型的预测结果如表 3 所示。

从表 3 可以看出经异常剔除后的结果均比未剔除更优，且 IFSR 方法的剔除效果均为最优。对比剔除样本编号，IFSR 相较 MD、HL 方法在剔除空间距离异常点的同时还剔除了残差过大所造成的强影响点；而 IFSR 相较传统 SR 方法多考虑了空间距离的影响因素；MCCV 方法虽然结合了 X 与 Y 变量间的关系进行分析，但也会由于异常导致模型过拟合造成误判；ODXY 方法也联合 XY 两变量进行分析，但其以平均光谱为参考进行关联分析也会由于异常样本导致偏差从而造成误判。IFSR 方法虽然并未结合 XY 两变量综合考虑，但其以二叉树对光谱数据进行切割，对不同空间维度均可进行切割，且该方法具有随机性并结合了集成学习的优点，能在循环中快速找到异常样本，同时切割过程无需建模，大大提高了搜索速度。

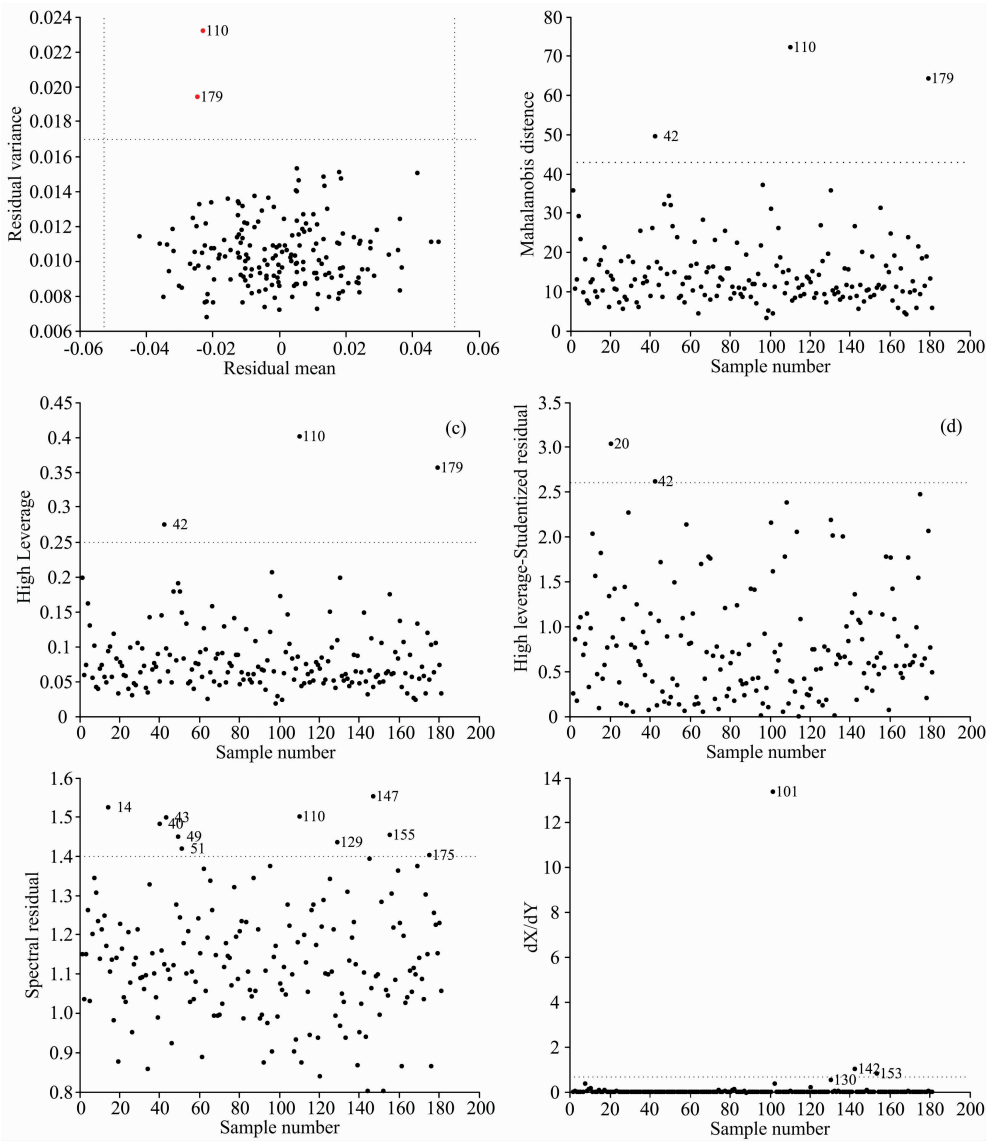

图 5　基于六种异常样本剔除方法的样本剔除结果

(a)：MCCV；(b)：MD；(c)：HL；(d)：HLSR；(e)：SR；(f)：ODXY

Fig. 5　Sample removal results based on six abnormal sample removal methods

(a)：MCCV；(b)：MD；(c)：HL；(d)：HLSR；(e)：SR；(f)：ODXY

表 3　基于不同异常值剔除方法的落叶松木材密度建模及预测结果

Table 3　Modeling and prediction results of larch wood density based on different outliers elimination methods

模型	剔除样本数	主因子数	校正集		预测集	
			R^2	RMSECV	R^2	RMSEP
Full-PLS	0	12	0.887 5	0.023 1	0.894 2	0.019 6
IFSR-PLS	7	14	0.869 0	0.024 3	0.921 5	0.016 4
MCCV-PLS	2	11	0.882 7	0.023 7	0.894 3	0.019 6
MD-PLS	3	11	0.871 0	0.024 8	0.895 2	0.019 6
HL-PLS	3	11	0.871 0	0.024 8	0.895 2	0.019 6
HLSR-PLS	13	12	0.924 2	0.019 4	0.904 7	0.018 7
SR-PLS	9	12	0.895 3	0.022 5	0.849 8	0.023 0
ODXY-PLS	4	12	0.875 5	0.024 7	0.908 1	0.018 3

2.4　建模方法对比与分析

为得到最优的落叶松木材密度近红外预测模型，对经 IFSR 剔除异常样本后重新划分好的样本集分别采用 PSO-SVR，PLS 和 BPNN 三种建模方法建模并确定最优方法。所采用的 PSO-SVR 算法基于 LIBSVM 工具箱，确定核函数为径向基核函数，惩罚因子 c 与核参数 g 通过粒子群算法(particle swarm optimization，PSO)确定，PSO 算法中设定种群规模大小为 200，个体学习因子 $c_1=1.5$，社会学习因子 $c_2=1.7$，最大迭代次数为 200，交叉验证折数为 10 折，主成分因子数为上述交叉验证偏最小二乘测试 IFSR 方法时所确定的主因子数。其中 PSO 参数寻优适应度曲线及校正集与预测集的拟合曲线可视化结果如图 6 所示。由图 6(a)可知，在惩罚因子 $c=30.029\ 1$，核参数 $g=0.01$ 时的预测效果最优，此

时预测集 R^2 为 0.932 1，RMSEP 为 0.015 4。PSO-SVR 模型的预测效果很好。

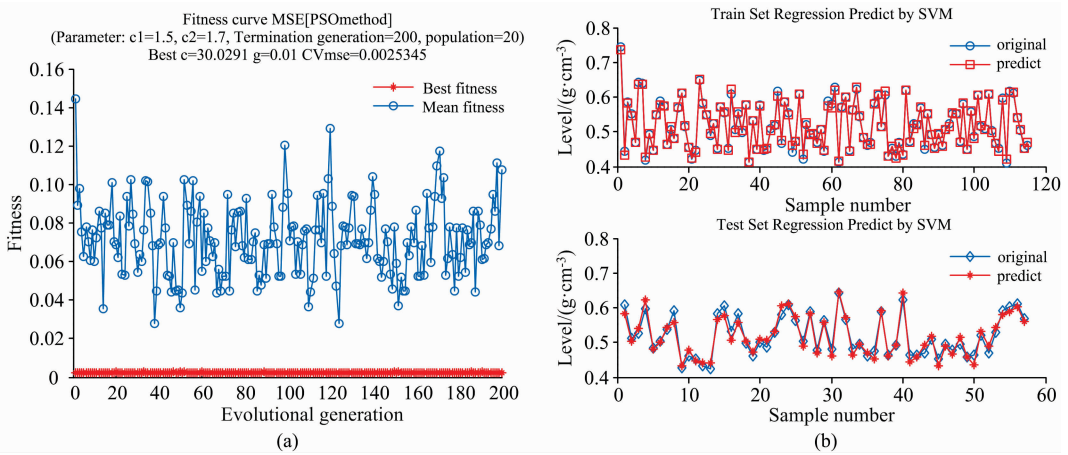

图 6 PSO-SVR 预测结果

(a)：PSO 参数寻优适应度曲线；(b)：校正集与预测集的拟合曲线

Fig. 6 PSO-SVR prediction results

(a): PSO parameter optimization fitness curve; (b): Fitting curve of correction set and prediction set

采用的 BPNN 算法基于 MATLAB 神经网络工具箱，经过多次调试确定 BPNN 的训练参数：学习速率为 0.01，训练要求精度为 0.000 1，最大训练次数为 2 000 次。BPNN 预测集拟合曲线如图 7 所示。从图中 7 可知，预测集的 R^2 为 0.913 1，RMSEP 为 0.017 7，BPNN 也可以很好地预测木材密度。

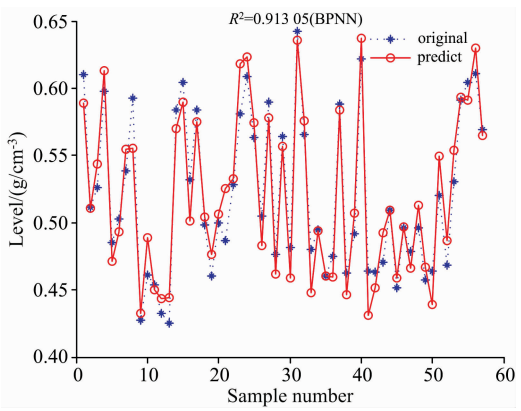

图 7 BPNN 预测集拟合曲线

Fig. 7 BPNN prediction set fitting curve

表 4 基于 PSO-SVR，BPNN，PLS 方法的落叶松木材密度建模及预测结果

Table 4 Modeling and prediction results of larch wood density based on PSO-SVR, BPNN and PLS methods

模型	校正集		预测集	
	R^2	RMSECV	R^2	RMSEP
PSO-SVR	0.993 3	0.005 5	0.932 1	0.015 4
PLS	0.869 0	0.024 3	0.921 5	0.016 4
BPNN	0.964 5	0.012 7	0.913 1	0.017 7

分别对 IFSR 剔除异常样本后的落叶松木材密度近红外样本集进行 PSO-SVR，BPNN 及 PLS 建模，得到的预测结果如表 4 所示，从表 4 可以看出，PSO-SVR 与 PLS 方法的建模效果优于 BPNN，而 PSO-SVR 方法的建模效果最优，R^2 为 0.932 1，RMSEP 为 0.015 4，由此证明对于小样本数据，支持向量回归要优于神经网络，且 SVR 考虑了非线性因素，所建模型的预测能力优于线性的 PLS。PSO-SVR 所建落叶松木材密度近红外模型的预测结果如图 8 所示。

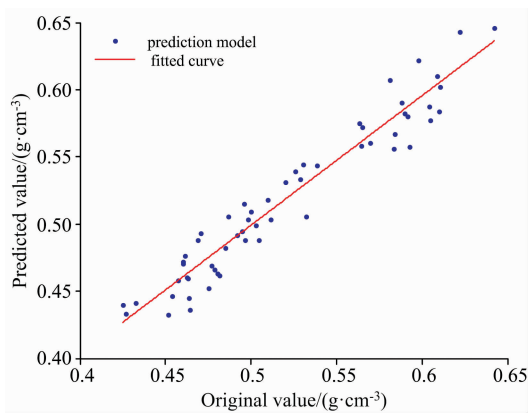

图 8 PSO-SVR 模型预测结果

Fig. 8 Prediction results of PSO-SVR model

通过对预测结果与真实值进行残差分析(图 9)，残差值均匀分布在横轴两端，证明预测值是等方差分布的，且在±0.04范围内预测值具有很强的解释性，进而证明预测模型具有较强的可靠性。预测结果表明，基于 IFSR 的异常样本剔除方法能在建模前准确地识别样本集中的异常样本，尤其针对高维且多变量的数据集具有明显效果，由于结合了学生化残差检验，只需查表即可确定阈值，避免了根据经验或多次实验确定阈值的复杂过程。相对传统异常样本剔除方法更加准确简便。

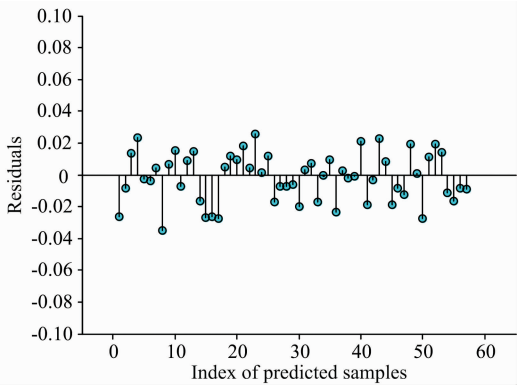

图 9　预测结果残差分析

Fig. 9　Residual analysis of calibration prediction results

3　结　论

基于统计学方法孤立森林算法，提出了一种孤立森林结合学生化残差方法(IFSR)，并通过预处理及特征变量选择消除数据中的噪音及无效信息，使其可用于识别并剔除高维、高共线性的近红外光谱数据中的异常样本值，且在识别过程中只需查阅 t 分布临界值表即可确定阈值，避免了设定阈值的问题。

为验证 IFSR 的可靠性，将该算法用于剔除落叶松木材密度样本中的异常值，并建立近红外预测模型，经与多种传统异常样本剔除方法对比，证明用 IFSR 可以有效剔除近红外光谱中的异常样本值，所得预测模型稳健性好，预测精度高。但 IFSR 方法也有一定的局限性，如未将光谱数据 X 与真实值 Y 联合考虑来分析样本中的异常值，下一步可从此方向，结合阔叶材、竹材等多种样本进一步优化算法。

References

[1] HE Xiao-yu, WANG Yan-wei, HUANG Rong-feng, et al(何啸宇, 王艳伟, 黄荣凤, 等). Forest Engineering(森林工程), 2020, 36(6): 72.

[2] Balasso Michelle, Hunt Mark, Jacobs Andrew, et al. Forest Ecology and Management, 2021, 491: 118992.

[3] GAO Ming-yu, NI Hai-ming, ZHANG Bo-yang, et al(高明宇, 倪海明, 张博洋, 等). Forest Engineering(森林工程), 2021, 37(4): 66.

[4] JIANG Xin-bo, SONG Jing, XIA Peng(姜新波, 宋　靖, 夏　鹏). Forest Engineering(森林工程), 2022, 38(1): 34.

[5] Cappozzo A, Duponchel L, Greselin F, et al. Analytica Chimica Acta, 2021, 1153(3): 338245.

[6] Wang B, He J, Zhang S, et al. Journal of Food Process Engineering, 2021, 44: 10.

[7] SHI Lu-zhen, ZHANG Jing-chuan, WANG Yan-qun, et al(石鲁珍, 张景川, 王彦群, 等). Journal of Chinese Agricultural Mechanization(中国农机化学报), 2016, 37(6): 99.

[8] Silalahi D D, Midi H, Arasan J, et al. Symmetry, 2021, 13: 4.

[9] WANG Lin, MA Xue-jie, MENG Dan-rui, et al(王　林, 马雪洁, 孟丹蕊, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(9): 2774.

[10] YIN Bao-quan, SHI Yin-xue, SUN Rui-zhi(尹宝全, 史银雪, 孙瑞志). Journal of University of Science and Technology of China(中国科学技术大学学报), 2016, 46(3): 208.

[11] Brownfield B, Kalivas J H. Analytical Chemistry, 2017, 89(9): 5087.

[12] HUANG Yuan-cheng, XUE Yuan-yuan, LI Peng-fei(黄远程, 薛园园, 李朋飞). Acta Geodaetica et Cartographica Sinica(测绘学报), 2021, 50(3): 416.

[13] LI Xin-peng, GAO Xin, YAN Bo, et al(李新鹏, 高　欣, 阎　博, 等). Power System Technology(电网技术), 2019, 43(4): 1447.

[14] Amirvaresi A, Nikounezhad N, Amirahmadi M, et al. Food Chemistry, 2021, 344: 128647.

[15] LIU Xiu-ying, YU Jun-ru, WANG Shi-hua(刘秀英, 余俊茹, 王世华). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2020, 36(22): 308.

[16] Leng Tuo, Li Feng, Chen Yi, et al. Meat Science, 2021, 180: 108559.

[17] Li S T, Zhang K Z, Duan P H, et al. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(1): 319.

[18] Sun Y, Yuan M, Liu X, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 10: 258.

NIR Model Optimization Study of Larch Wood Density Based on IFSR Abnormal Sample Elimination

ZHANG Zhe-yu, LI Yao-xiang*, WANG Zhi-yuan, LI Chun-xu
College of Engineering and Technology, Northeast Forestry University, Harbin 150040, China

Abstract Wood density is an important physical property of wood which can reflect a variety of physical properties such as wood shrinkage, compressive and tensile strength. Using near-infrared spectroscopy technology can rapidly predict wood density, which can overcome the disadvantages of traditional detection methods that consume workforce, material resources and time. However, the modeling results are often affected by abnormal samples. In order to accurately identify and eliminate abnormal samples in the sample set, an isolation forest combined with the studentized residual method (IFSR) was proposed. Based on the advantages of integrated features of isolated forests, the influence of samples on the model is considered, and abnormal samples and strong influence samples can be detected simultaneously. This study measured the near-infrared spectra of 181 larch wood samples and their air-dry density at room temperature. By comparing a variety of preprocessing and feature selection methods, the preprocessing method was determined to adopt the standard normal variable change (SNV) + detrending processing (DT) + mean centralization (MC) + standardization (Auto) method and the feature wavelength selection was determined to adopt competitive adaptive reweighted sampling (CARS) method. Eliminated the influence of noise and irrelevant information on the algorithm, simplified the dataset, and improved the algorithm's accuracy in removing abnormal samples. In order to verify the ability of the IFSR method to eliminate abnormal samples, it was compared with the other six anomaly detection methods such as Monte Carlo Interactive Verification (MCCV), Mahalanobis Distance (MD), etc. The partial least squares (PLS) model was established to evaluate its anomaly detection performance. At the same time, the particle swarm optimization-support vector machine regression (PSO-SVR), BP neural network (BPNN) and PLS were used to establish the near-infrared prediction model of larch wood density. The results show that the optimized model obtained by IFSR combined with PSO-SVR has the strongest predictive ability, and IFSR can effectively eliminate singular samples and improve the model's accuracy.

Keywords Near-infrared spectrum; Wood density; Outlier detection; Isolation forest algorithm; Support vector regression

(Received Sep. 23, 2021; accepted Jan. 19, 2022)

* Corresponding author