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，pp3482-3488
November，2022

改进的随机蛙跳算法对农机润滑油污染浓度的近红外光谱检测研究

韩嘉庆[1]，周桂霞[1*]，胡　军[1*]，程介虹[2]，陈争光[2]，赵胜雪[1]，刘奕伶[1]

1. 黑龙江八一农垦大学工程学院，黑龙江 大庆　163319
2. 黑龙江八一农垦大学电气与信息学院，黑龙江 大庆　163319

摘　要　润滑油是农业机械正常作业的必要物资，农业机械发动机工作的动力性、安全性、经济性以及寿命与润滑油状况有着紧密联系。污染浓度作为油液的综合评价指标，常规的实验室检测耗时长、成本高，所以开发高效的润滑油污染浓度检测技术具有重要意义。提出了一种基于近红外光谱技术的农机润滑油污染浓度的检测方法，同时针对随机蛙跳(RF)特征波长选择算法中迭代次数大，结果再现性低等缺点，提出了一种迭代保留信息变量的随机蛙跳(IRIV-RF)特征波长选择算法。该算法一方面利用迭代保留信息变量(IRIV)算法提取出强信息变量和弱信息变量，将其作为RF算法中的初始变量集，消除初始变量集的随机性对结果再现性的影响。另一方面通过对变量按被选概率值由大到小正向排序后，从首个波长开始依次增加一个波长建立偏最小二乘回归(PLSR)模型，选择交叉验证均方根误差(RMSECV)值最小时的变量子集为特征波长，消除RF算法所提取的特征波长数量的不确定性。利用近红外光谱仪采集自行配制的101份不同污染浓度的农机润滑油原始光谱数据，选用三种不同的预处理方法分别对原始光谱进行处理，确定最佳的预处理方法为变量标准化(SNV)。在此基础上通过RF，IRIV和IRIV-RF三种算法分别对全谱进行特征波长选择，并建立PLSR模型。通过对全谱-PLSR，RF-PLSR，IRIV-PLSR以及IRIV-RF-PLSR模型的预测精度进行比较，结果表明，经过IRIV-RF算法提取特征波长后所建立的PLSR模型预测精度最高，预测相关系数(R_p)为0.965 7，预测均方根误差(RMSEP)为9.0584，显著提升了预测精度与运行效率，降低模型复杂程度。IRIV-RF是一种有效的特征波长选择算法，研究证明了近红外光谱联合改进的IRIV-RF算法检测农机润滑油污染浓度的可行性，为鉴定润滑油品质提供了一种新的思路。

关键词　特征波长选择；随机蛙跳；迭代信息保留变量；农机润滑油；污染浓度；近红外光谱

中图分类号：O433.4　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2022)11-3482-07

引　言

润滑油农业机械化的必要条件。润滑油作为农机发动机良好工作的润滑剂，不仅起到润滑和抗磨损的功用，也具备冷却、清洁、防锈、防腐和抗氧化的功能。农机作业环境恶劣且工况多变，长期高负荷的作业导致润滑油变质速度加快，使得部分农机润滑油在保质期内已经提前变质却未得到及时更换；同时多数进口农业机械要求使用专用润滑油且需要严格按照规定时长进行换油，但有时农机作业量少且工作负荷轻，存在润滑油还未变质就被强制换油，造成了能源的浪费，成本的提高。污染浓度作为评价油液质量的综合指标，不仅可以体现出润滑油中机械杂质的多少，也能体现出油液的使用状态。常规的的实验室检测耗时长，检测成本高，因此，开发出一种高效的润滑油污染浓度检测方法具有重要意义。

近红外光谱技术具有适应性广、无损无污染、分析速度快、精密度高等优点，是油品质量检测极其重要的分析手段之一。Balabin等[1]利用近红外光谱技术结合概率神经网络(probabilistic neural network，PNN)、K最近邻(K-nearest neighbor method，KNN)等七种方法来鉴别润滑油种类；Alves等[2]利用近红外光谱技术结合支持向量机(support vector machine，SVM)对石蜡润滑油中环烷烃和植物油的含量进行测定；刘晨阳等[3]利用近红外光谱，提出一种量子遗

收稿日期：2021-10-08，**修订日期**：2022-03-20

基金项目：国家大豆产业技术体系岗位专家项目(CARS-04-01A)，国家重点研发计划项目(2017YFC1601905-04)，黑龙江省重点研发计划项目(GA21B003)，黑龙江八一农垦大学三横三纵支持项目(DJH201808)资助

作者简介：韩嘉庆，1997年生，黑龙江八一农垦大学工程学院硕士研究生　　e-mail：759326496@qq.com

*通讯作者　　e-mail：357652493@qq.com；gcxykj@126.com

传(quantum genetic algorithm，QGA)与 BP 神经网络(back propagation neural network，BPNN)结合的算法对润滑油粘度进行定量分析的方法；陈彬等[4]利用近红外光谱技术结合无信息变量消除法(uninformative variable elimination，UVE)，提出了一种润滑油中含水率的检测方法；张瑜等[5]利用近红外光谱结合最小二乘支持向量机(least squares support vector machine，LSSVM)对润滑油的酸值进行检测。通过上述可以发现，较多国内外学者均应用近红外光谱对润滑油含水率、粘度、酸值以及润滑油品牌的鉴别进行了相关研究，而对于润滑油污染浓度的研究较少。

在近红外光谱分析中，特征波长的提取[6]一直是热门话题。近红外光谱特征波长选择方法众多，其中随机蛙跳(random frog，RF)[7]是目前较为高效且新式的特征波长选择算法之一。其原理是由于各变量被选概率值的大小不同，所以通过多轮迭代，将被选概率值较高的部分变量作为所选特征波长。通过对 RF 算法原理的分析，发现 RF 算法在数据降维方面具有一定优势，但也存在明显的缺陷。RF 算法中初始变量集的选择是随机的，存在较大的不确定性，无法保证初始信息的合理性与有效性，可能会选择到部分无信息变量或干扰变量，导致所得结果的再现性较低。因此 RF 算法中的迭代次数 N 需要足够大，以确保算法在运行过程中可以遍历所有数据集，但这也导致算法需要处理的数据量过大，运行效率较低。同时，RF 算法中对被选变量概率值的标定无理论依据，易受客观因素影响[8]，导致所选取特征波长的数量存在不确定性。

因此，以农机润滑油为研究对象，利用所获取的近红外光谱数据，对 RF 算法进行改进，提出一种迭代保留信息变量的随机蛙跳(iteratively retains informative variables-random frog，IRIV-RF)特征波长选择算法。分别利用全谱以及 RF 算法、迭代保留信息变量(iteratively retains informative variables，IRIV)[9]算法和改进的 IRIV-RF 算法提取出的特征波长建立偏最小二乘回归(partial least squares regression，PLSR)模型[10]，比较四种模型的预测精度，用来证明本文提出的 IRIV-RF 算法的有效性。

1　实验部分

1.1　试验材料与样本配置

试验材料分别选用未开封使用的约翰迪尔 CK-4/SN 机油和使用远超正常工作小时的同品牌同型号污染润滑油，此型号润滑油适应性强，属于冬夏通用型机油，应用于众多约翰迪尔大马力拖拉机中，具有代表性。为得到不同污染浓度的润滑油的近红外光谱数据，使用混匀仪，将未开封的新油与污染后的旧油按照不同比例进行混合，依次配置了污染度浓度为 0%，1%，2%，3%，4%，…，100%的共 101 份润滑油样本，以此模拟农机发动机工作过程中润滑油污染浓度的逐渐变化过程。本文以农机润滑油污染浓度(0～100)作为因变量进行波长选择及近红外光谱数据建模预测分析。

1.2　光谱采集

使用德国布鲁克品牌的 TANGO 系列近红外光谱仪采集 101 份润滑油样本的近红外光谱。光谱采集软件为 OPUS6.5，波长范围为 1 000～2 500 nm，控制室温在 25 ℃左右，开机预热 30 min 后进行背景测量，在后续光谱采集过程中，每隔 30 min 重新测量背景，避免背景干扰。采集过程中应避免润滑油样本中有气泡产生，因为润滑油样本中存在气泡会使光谱采集过程中产生光的散射等现象，降低光谱数据的准确性，所以部分产生气泡的润滑油样本，均静置 15～20 min，待气泡自行排出后进行光谱采集。为最大限度避免采集过程中所产生的误差，每份润滑油样本均采集三次光谱数据，将其平均值作为各润滑油样本的原始光谱。

1.3　样本集划分与光谱预处理

SPXY(sample set portioning based on joint X-Y distances)算法[11]是由 Galvao 等提出的样本选择方法，原理是分别以光谱变量和理化值为特征参数，计算各样本间的距离，最大程度保证样本分布的代表性，具体实现过程参照文献[12]。该算法可以有效覆盖多维向量空间，避免样本间因差异过小所引起的预测模型过拟合或预测效果差的现象，在分析光谱变量特征与理化值之间的关系的同时进行样本选择，改善模型的预测能力与稳健性。

光谱除含有样本自身的化学信息以外，也存有部分无用信息以及噪声信息。样本自身基质状态的改变、光的散射以及光谱仪工作状态的改变等都会导致光谱出现基线漂移，多元散射等噪声信息。因此，需要对原始光谱进行预处理来消除噪声、校正斜坡背景、压缩数据以及消除其他无用因素对光谱信息的干扰，为所建立模型的鲁棒性和待测样本预测的精度奠定良好的基础[12]。目前在近红外光谱技术中，卷积平滑法(savitzky-golay，SG)、基线矫正(baseline correction，BC)以及变量标准化(standard normal variate，SNV)等是较为常用的预处理方法。

1.4　IRIV-RF 算法

1.4.1　初始变量集的选择

通过对 RF 算法的原理和步骤进行深入研究，该算法体具体实现过程参照文献[8]的描述，可以发现 RF 算法中初始变量集的选择是随机的，存在较大的不确定性，可能会选择到一些无信息变量或干扰变量，导致算法迭代次数较大，结果再现性较低，运算时间过长。而 IRIV 算法是一种将各变量随机组合并判断各变量之间相互作用，基于二进制矩阵重排过滤器提出的特征波长选择算法，通过评估各变量与模型之间的利害关系，将全部变量划分为对建模有益的强信息变量和弱信息变量以及对建模无益甚至产生危害的无信息变量和干扰信息变量，该算法具体实现过程参照文献[10]的描述。本工作通过 IRIV 算法对 RF 算法进行改进，利用 IRIV 算法提取出强信息变量和弱信息变量，将其作为 RF 算法中的初始变量集，确保初始变量集的合理性与有效性，避免初始变量集的随机性，剔除对建模无意义的无信息变量以及对建模有害的干扰信息变量，减少迭代次数，提高结果的再现性。

1.4.2　建模波长的优选

RF 算法中，通常基于各变量被选概率值的大小进行变量选择，一般选择由大到小正向排序的前十个变量作为特征

波长，或者根据客观标定的概率值，一般选择大于或者等于此客观标定的概率值的变量作为符合要求的特征波长[13]，导致经过 RF 算法所提取的特征波长数量的选择存在很大的不确定性。

为避免所选建模波长数量的不确定性，对各变量按其被选概率值进行由大到小正向排序，通过向前选择法对排序后的变量从首个波长开始，依次增加一个波长，建立光谱数据与理化值的 PLSR 模型，计算各模型的交叉验证均方根误差(root mean square error of cross validation，RMSECV)值，以出现最小 RMSECV 值时对应的变量子集为最终所选的特征波长。RMSECV 使用式(1)计算

$$\mathrm{RMSECV}=\sqrt{\frac{\sum_{i=1}^{n}(y_{i,\ \mathrm{actual}}-y_{i,\ \mathrm{predicted}})^2}{n-1}} \tag{1}$$

式(1)中，$y_{i,\ \mathrm{actual}}$为第 i 个农机润滑油样本污染浓度实测值，$y_{i,\ \mathrm{predicted}}$为经过该模型所计算出的第 i 个农机润滑油样本污染浓度预测值，n 为全部样本中校正集的样本数量。

IRIV-RF 算法通过 IRIV 算法得到强信息变量与弱信息变量，将其作为 RF 算法中的初始变量集；通过依次增加一个变量并建立光谱数据与理化值的 PLSR 模型后找到出现最小 RMSECV 值时对应的变量子集。这样 IRIV-RF 算法就可以找到最优预测精度所对应的特征波长，通过两次改进提高了结果再现性与预测精度。

1.5 模型建立与评价方法

目前，在近红外光谱分析中，PLSR[14]是比较高效且常用的建模方法。它将光谱数据压缩成为潜在变量的正交结构，描述参考值与光谱信息之间的最大协方差。PLSR 同时拥有多元线性回归分析、主成分分析以及典型相关分析的分析特性，便于有效消除自变量可能存在多重共线性所导致模型预测失真等缺陷，具有较高的预测稳定性，能更加准确进行信息的识别。

模型评价的目的是为了对所建立的模型预测的精度、可信度以及建模的效果进行验证评价，主要通过校正相关系数(R_c)、校正均方根误差(RMSEC)、预测相关系数(R_p)、预测均方根误差(RMSEP)四个参数来评价建模的效果以及模型的预测能力。其中，若相关系数的数值越大且越趋近于 1，则说明对应特征变量与理化值两者的相关性越强，模型的拟合程度越高；若均方根误差的数值越小，则说明模型的稳定性越高，预测能力越好。

2 结果与讨论

2.1 不同光谱预处理方法的选择

利用 SPXY 算法将 101 个润滑油样本划分为两部分，分别为 70%的校正集和 30%的预测集，即校正集由 70 个样本构成，预测集由 31 个样本构成。选用 SG、BC 以及 SNV 三种不同的预处理方法分别对原始光谱进行处理，其中 SG 去噪时选择移动窗口数为 7，用于拟合的多项式次数为 3。同时对原始光谱数据以及三种不同预处理后所得的光谱数据分别建立 PLSR 模型，根据所建立的 PLSR 模型的建模效果与预测精度确定出最佳的预处理方法。原始光谱的 PLSR 模型以及三种不同预处理后所得的光谱数据建立的 PLSR 模型结果如表 1 所示。

表 1 不同预处理方法的模型结果

Table 1 Performance results of models with different Pretreatment methods

预处理方法	R_c	RMSEC	R_p	RMSEP
原始光谱	0.862 4	16.027 5	0.843 9	15.697 6
SG 平滑	0.853 9	14.798 3	0.846 9	15.239 9
BC 基线矫正	0.859 9	14.908 9	0.850 2	15.435 7
SNV 变量标准化	0.887 8	13.114 5	0.875 3	13.225 7

通过表 1 可以看出，经过 SNV 预处理后所得的光谱数据建立的 PLSR 模型的均方根误差和相关系数与原始光谱、SG 以及 BC 预处理后的光谱数据建立的 PLSR 模型的均方根误差和相关系数比较，可以发现 SNV 预处理建立的 PLSR 模型的均方根误差最小且相关系数更趋近于 1，故选用 SNV 进行光谱预处理。图 1(a)为原始光谱，图 1(b)为 SNV 预处理后得到的光谱，可以发现通过 SNV 预处理后的光谱曲线的波峰波谷的位置特征越发分明，解决了原始光谱中存在基线漂移的问题，所以后续的特征波长选择和建模均应用 SNV 预处理后所得到的光谱数据。

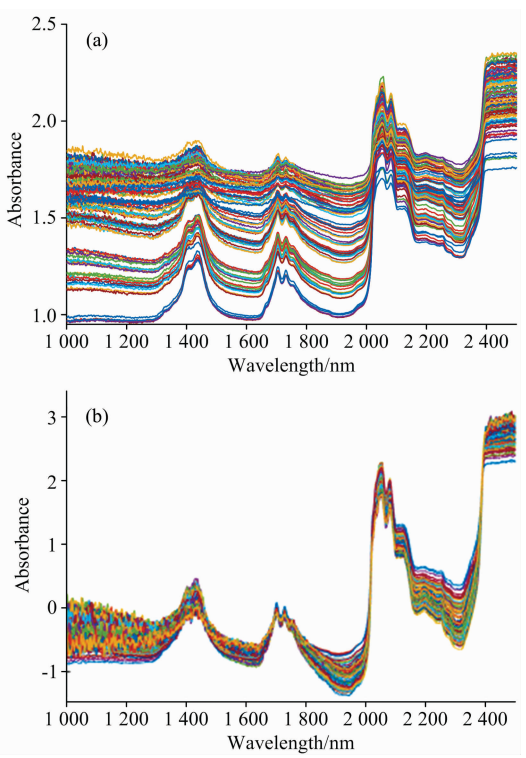

图 1 原始光谱(a)及 SNV 预处理后的光谱(b)

Fig. 1 Original spectra (a) and SNV preprocessed spectra (b)

2.2 特征波长选择

2.2.1 RF 变量选择结果

对 RF 进行初始化参数设置，N 设定为 10 000，Q 设定

为 10，开始运行。全谱经过 RF 算法进行波长选择后，各变量被选择的概率值如图 2(a)所示，经过多次试验，选用不同概率值所对应的变量作为特征波长分别建模，进行对比分析后，发现选择概率值大于等于 0.24 的变量为特征波长建模可以获得最佳的建模效果，此时的特征波长共计 24 个，如图 2(b)所示，分别为 2 435，2 260，2 396，2 290，2 175，2 498，1 953，2 430，2 427，1 400，1 194，1 207，1 770，1 740，1 743，1 409，2 173，1 698，2 487，1 163，1 364，1 049，1 283和 1 840 nm。

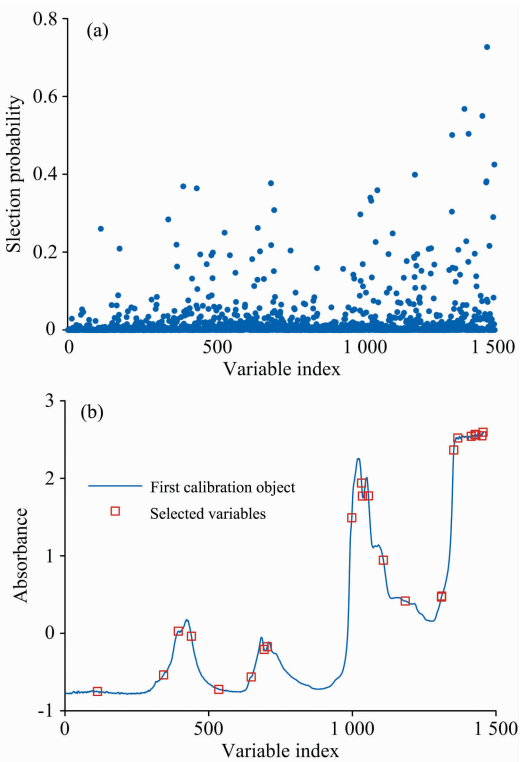

图 2　RF 运行结果

Fig. 2　The results of random frog selection

2.2.2　IRIV 变量选择结果

对 IRIV 进行初始化参数设置，设定最大主成分数为 10，交叉验证次数为 5，开始运行。全谱经过 IRIV 算法波长选择后，选择出强信息变量和弱信息变量共计 29 个，如图 3 所示，分别为 1 013，1 037，1 066，1 078，1 083，1 084，1 171，1 172，1 212，1 229，1 323，1 363，1 364，1 402，1 408，1 410，1 425，1 474，1 486，1 674，1 675，1 991，2 021，2 028，2 030，2 290，2 352，2 427 和 2 430 nm。

2.2.3　IRIV-RF 变量选择结果

在 IRIV 得到的 29 个强信息变量和弱信息变量的基础上进行 RF 波长选择。根据被选概率值对各变量进行由大到小正向排序，从被选概率值最大的波长开始，按照所排顺序依次增加一个波长建立光谱数据与农机润滑油污染浓度的 PLSR 模型。IRIV-RF 变量选择的结果如图 4 所示，其中图 4(a)中红色正方形标记处为最小的 RMSECV 值，为 9.134 2，得到满足条件的特征波长共计 17 个，如图 4(b)所示，分别为：1 408，1 674，2 427，1 083，2 290，1 066，1 212，1 675，2 352，1 991，1 013，1 171，1 363，1 364，2 021，1 172 和 1 323 nm。

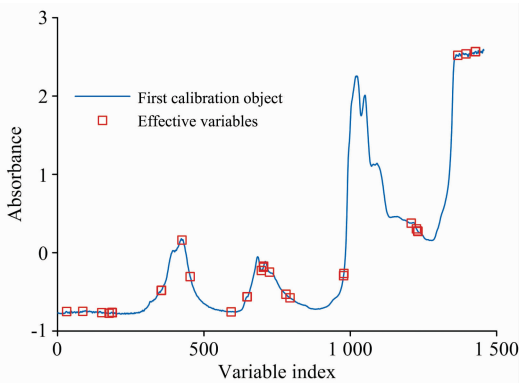

图 3　IRIV 所得强、弱信息变量

Fig. 3　Strongly and weakly informative variables obtained by IRIV

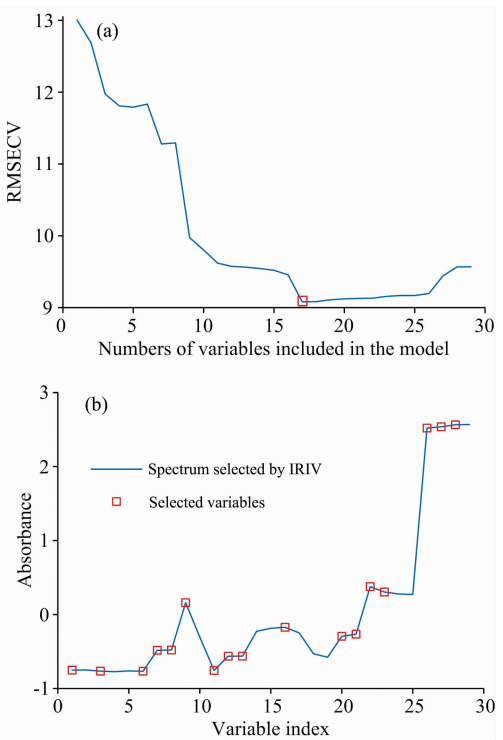

图 4　IRIV-RF 变量选择的结果

Fig. 4　Variable selection results of IRIV-RF

2.3　模型建立与比较

将全谱、RF、IRIV 以及 IRIV-RF 三种算法所选择出的特征波长，分别建立 PLSR 模型，计算得到各模型的校正相关系数、预测相关系数、校正均方根误差以及预测均方根误差的数值，如表 2 所示，通过四个评价参数来比较各模型的预测能力。

通过表 2 可以发现，对全谱提取特征波长后所建立模型的各项参数均优于全谱，证明了对全谱提取特征波长的必要

表 2 不同波长选择方法下模型的结果

Table 2 Performance results of models with different wavelength selection methods

模型	变量数	R_c	RMSEC	R_p	RMSEP
FULL-PLSR	1 500	0.887 8	13.114 5	0.875 3	13.225 7
RF-PLSR	24	0.926 7	11.172 0	0.913 3	11.434 6
IRIV-PLSR	29	0.938 8	10.126 3	0.931 6	10.676 1
IRIV-RF-PLSR	17	0.972 9	8.738 6	0.965 7	9.058 4

性。同时，改进的 IRIV-RF 算法模型的各项评价参数均优于 RF 算法以及 IRIV 算法，其中经过 RF 算法所选特征波长建立的 PLSR 模型的 R_p 为 0.913 3，RMSEP 为 11.434 6，而经过改进后 IRIV-RF 算法所选特征波长建立的 PLSR 模型的 R_p 为 0.965 7，RMSEP 为 9.058 4，提升了预测精度。这是因为在 RF 算法初始变量集的选择过程中可能会选取到部分干扰信息变量或无信息变量，导致 RF 算法的不确定程度极高，结果再现性较低，算法迭代次数过大，运行时间较长。本文提出的 IRIV-RF 算法，首先通过 IRIV 算法对全谱进行特征波长初选，提取强信息变量和弱信息变量，避免光谱中无信息变量和干扰变量对光谱的影响，将其初选结果作为 RF 算法的初始变量集，保证了 RF 算法初始变量集选择的有效性，有益于特征波长的选择，可以提高模型的预测精度，改善 RF 算法收敛速度慢、迭代次数大、运行时间较长等问题。

图 5 为预测集样本的全谱-PLSR，RF-PLSR，IRIV-PLSR 以及 IRIV-RF-PLSR 模型的农机润滑油污染浓度的预测值与化学值的散点图。从图中可以清晰地观测出，经过改进的 IRIV-RF 算法所选特征波长建立的 PLSR 模型的预测效果最优。

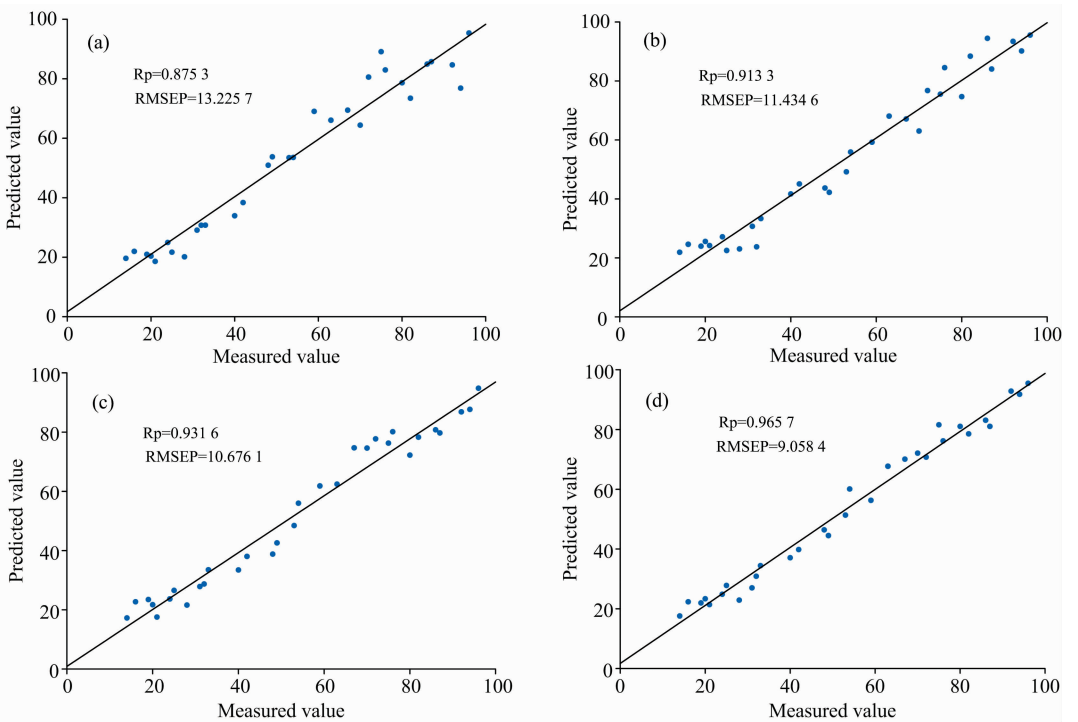

图 5 PLSR 模型下农机润滑油污染浓度的散点图

(a)：全谱-PLSR；(b)：RF-PLSR；(c)：IRIV-PLSR；(d)：IRIV-RF-PLSR

Fig. 5 Scatter plots of PLSR models for contamination concentration of agricultural machinery lubricating oil

(a)：Full spectrum-PLSR；(b)：RF-PLSR；(c)：IRIV-PLSR；(d)：IRIV-RF-PLSR

3 结 论

配置不同污染浓度的农机润滑油，并采集其近红外光谱，结合化学计量学方法对农机润滑油污染浓度进行检测研究。提出了一种 IRIV-RF 特征波长选择算法，该算法的初始变量集为利用 IRIV 算法提取出的强信息变量和弱信息变量，保证了初始变量集中的变量均是对建模有益的有效信息变量，解决 RF 算法结果再现率较低、算法迭代次数过多大以及运行时间较长等问题。将改进的 IRIV-RF 特征波长选择算法应用于农机润滑油污染浓度的光谱数据集中，通过对比全谱-PLSR，RF-PLSR，IRIV-PLSR 以及 IRIV-RF-PLSR 四个模型的预测精度可以发现，IRIV-RF-PLSR 模型的建模效果最佳，预测精度最高，同时建模的复杂程度与算法运行时间均有所降低。证明本文提出的 IRIV-RF 是一种有效的特征波长选择算法，近红外光谱联合改进的 IRIV-RF 算法为农机润滑油污染浓度的检测提供了一种较为快速准确且简便的方法，为鉴定润滑油品质提供了一种新的思路。

References

[1] Balabin R M, Safieva R Z, Lomakina E I. Microchemical Journal, 2011, 98(1): 121.
[2] Alves J, Poppi R J. Analytical Methods, 2013, 5(22): 6457.
[3] LIU Chen-yang, TANG Xing-jia, YU Tao, et al(刘晨阳, 唐兴佳, 于　涛, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(5): 1634.
[4] CHEN Bin, LIU Ge(陈　彬, 刘　阁). Acta Photonica Sinica(光子学报), 2014, 43(2): 230001.
[5] ZHANG Yu, WU Di, HE Yong, et al(张　瑜, 吴　迪, 何　勇, 等). Infrared(红外), 2011, 32(12): 39.
[6] Yun Y H, Li H D, Deng B C, et al. Trends in Analytical Chemistry, 2019, 113(7): 102.
[7] Li H D, Xu Q S, Liang Y Z. Analytica Chimica Acta, 2012, 740: 20.
[8] CHEN Li-dan, ZHAO Yan-ru(陈立旦, 赵艳茹). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2014, 30(8): 168.
[9] Yun Y H, Wang W T, Tan M L, et al. Analytica Chimica Acta, 2014, 807(17): 36.
[10] Zheng X, Li Y, Wei W, et al. Meat Science, 2019, 149(3): 55.
[11] Chen Y, Luo P, Zhao Z Y, et al. Physics Letters A, 2017, 381(40): 3472.
[12] XIE Yue, LI Fei-yue, FAN Xing-jun, et al(谢　越, 李飞跃, 范行军, 等). Chinese Journal of Analytical Chemistry(分析化学), 2018, 46(4): 609.
[13] LONG Yan, LIAN Ya-ru, MA Min-juan, et al(龙　燕, 连雅茹, 马敏娟, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2019, 35(13): 270.
[14] Yu H W, Liu H Z, Wang N, et al. Analytical Methods, 2016, 8(41): 7482.

Near-Infrared Spectroscopy Detection of Pollution Concentration of Agricultural Machinery Lubricating Oil Based on Improved Random Frog Algorithm

HAN Jia-qing[1], ZHOU Gui-xia[1*], HU Jun[1*], CHENG Jie-hong[2], CHEN Zheng-guang[2], ZHAO Sheng-xue[1], LIU Yi-ling[1]
1. College of Engineering, Heilongjiang Bayi Agricultural University, Daqing　163319, China
2. College of Electrical and Information, Heilongjiang Bayi Agricultural University, Daqing　163319, China

Abstract　The use of lubricating oil is necessary for the normal operation of agricultural machinery. The power performance, safety, economy and life of agricultural machinery engines are closely related to the condition of lubricating oil. The pollution concentration is the comprehensive evaluation index of oil, routine laboratory testing takes a long time and costs a lot, so it is of great significance to develop efficient detection technology for lubricating oil pollution concentration. This paper takes agricultural machinery lubricating oil as the research object. A method for detecting pollution concentration of agricultural machinery lubricating oil based on near-infrared spectroscopy is proposed. At the same time, aiming at the shortcomings of the Random Frog (RF) feature wavelength selection algorithm, such as a large number of iterations and low reproducibility, and iteratively retains informative variables-Random Frog (IRIV-RF) feature wavelength selection algorithm is proposed. On the one hand, IRIV-RF uses the iteratively retains informative variables (IRIV) algorithm to filter the strong and weak information variables. It is used as the initial variable subset of RF to eliminate the effect of the randomness of the initial variable set on the reproducibility of the results. On the other hand, IRIV-RF builds a Partial least squares regression (PLSR) model by arranging the variables in descending order of the selected probability values and then adding one wavelength at a time, starting with the first. The variable subset with the minimum Root Mean Square Error of Cross Validation (RMSECV) value is selected as the characteristic wavelength to eliminate the uncertainty of the number of characteristic wavelengths extracted by the RF algorithm. The original spectrum data of 101 samples of agricultural machinery lubricating oil with different pollution concentrations are collected by near-infrared spectrometer. Three different pretreatment methods are used to process the original spectrum, and the optimal pretreatment method is Standard Normal Variate (SNV). On this basis, the characteristic wavelength of the whole spectrum is selected by RF, IRIV and IRIV-RF algorithms, and the PLSR model is established. By comparing the prediction accuracy of full-spectrum PLSR, RF-PLSR, IRIV-PLSR and IRIV-RF-PLSR models, the results show that the prediction

accuracy of the PLSR model based on the IRIV-RF algorithm is the highest, the Correlation Coefficient of Prediction (R_p) is 0.965 7 and the Root Mean Square Error of Prediction (RMSEP) is 9.058 4. It significantly improves the prediction accuracy and operation efficiency, reducing the model's complexity. It is proved that the proposed IRIV-RF algorithm is an effective characteristic wavelength selection algorithm, and the feasibility of near-infrared spectroscopy combined with the improved IRIV-RF algorithm to detect the pollution concentration of agricultural machinery lubricating oil is proved, which provides a new idea for identifying the quality of lubricating oil.

Keywords Feature wavelength selection; Random frog; Iteratively retains informative variables; Agricultural lubricating oil; Pollution concentration; Near-infrared spectroscopy

(Received Oct. 8, 2021; accepted Mar. 20, 2022)

* Corresponding authors