第40卷，第2期
2020年2月

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

Vol. 40, No. 2, pp562-566
February, 2020

FT-NIR 光谱半定性判别方法应用于土壤总氮的波段优选

辜　洁[1]，陈华舟[1,2*]，陈伟豪[1]，莫丽娜[1]，温江北[2]

1. 桂林理工大学理学院，广西 桂林　541004
2. 广东星创众谱仪器有限公司，广东 广州　510663

摘　要　总氮是衡量土壤肥力的重要成分指标。传统的检测土壤总氮含量的化学方法操作复杂且费时费力，采用傅里叶近红外(FT-NIR)对土壤总氮的含量实现直接快速定量分析；然而，利用偏最小二乘(PLS)等线性分析方法定量预测土壤样本的总氮含量，定标预测模型有可能被理想化，不利于在线检测的实际应用。考虑给定量分析模型添加容错机制，将FT-NIR定量分析转化为半定性判别分析，以加强光谱模型的应用能力，提出区间间隔搜索主成分分析逻辑回归(iPCA-LR)方法，结合PLS的先验定量预测值，通过设定 $r=0.05$，0.10，0.15三个不同的容错阈值范围，给样本赋予先验判别标记，将定量分析模式转换为半定性判别模式，建立土壤总氮的FT-NIR半定性判别模型，同时，对比讨论基于 $k=5$，10，15，20四种不同子波段数量的区间划分数据的潜变量转换模式，优选FT-NIR光谱特征子波段，并讨论优选连续子波段的组合建模情况。结果表明，不同阈值范围下的FT-NIR半定性判别模型的预测准确率差别较大，但不同阈值范围的最优判别模型的预测准确率均在75%以上，各个区间划分的优选子波段或合并子波段的判别准确率均达到了90%以上，可以满足不同程度的应用水平。利用PLS结合iPCA-LR将定量预测转换为半定性判别的方法能够应用于土壤总氮的FT-NIR光谱分析，能够解决常规PLS定量分析容易过拟合和过于理想化的问题，半定性判别结果更符合实际，有利于光谱技术的在线应用。

关键词　土壤总氮；FT-NIR；波段优选；iPCA-LR模型；半定性判别

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2020)02-0562-05

引　言

土壤肥力是农业可持续发展的基础。土壤总氮含量是衡量土壤肥力的重要指标之一[1]。传统的土壤总氮的检测一般是在化学实验室进行，需要采用化学反应，费时费力且操作繁琐[2]。利用近红外光谱对土壤总氮的含量实现直接快速定量分析具有十分重要的意义[3-4]。傅里叶近红外(FT-NIR)光谱分析可从大量的实验数据中提取样品中的待测成分信息，具有快速简便、无试剂、非破坏性、过程无污染等特点。近年来，随着信息技术和化学计量学的发展，FT-NIR光谱分析在食品、农业、环境、生物医学等众多领域得到广泛的应用[5-7]。

偏最小二乘法(PLS)是FT-NIR光谱常用的定量分析方法[8-9]。由于近红外光谱信号重叠严重，没有明显的波峰能够反应单一待测成分的信息，而且容易造成数据过拟合[10]，在此基础上建立的定标预测模型有可能被理想化，不利于在线检测的实际应用。因此，我们考虑给定量分析模型添加容错机制，将FT-NIR定量分析转化为半定性判别分析，以加强光谱模型的应用能力。

逻辑回归(LR)一种常用的定性分析方法，采用二分类模式进行定性建模和预测[11]。考虑采用潜变量分析技术[12]结合LR回归建立FT-NIR半定性判别模型，为PLS回归提供定量容错机制，有望可以避免数据过拟合现象，提供更为稳定的FT-NIR定标方案。主成分分析(PCA)被视为最简单有效的潜变量分析技术，合理选择恰当的主成分数是PCA技术的关键，能够有效降低光谱噪声和充分利用光谱特征信息[13-14]。

另一方面，由于特定的待测组分会在某一特定的光谱区域内形成较强的光谱响应信息[15]，考虑采用区间间隔波段搜索模式[16-17]，寻找土壤总氮的FT-NIR光谱信息子波段，在每一个子波段中利用PCA进行潜变量提取，进一步和LR

收稿日期：2018-12-09，**修订日期**：2019-04-13

基金项目：国家自然科学基金项目(61505037)，广西自然科学基金项目(2016GXNSFBA380077，2018GXNSFAA050045)资助

作者简介：辜　洁，1995年生，桂林理工大学统计学硕士研究生　　e-mail：82223404@qq.com

*通讯联系人　　e-mail：hzchengut@foxmail.com

回归建立能够实现对土壤总氮半定性判别的区间间隔 PCA 逻辑回归(iPCA-LR)模型。在此之前，采用标准正态变换(SNV)完成对测量光谱的降噪处理[18]，采用常规 PLS 算法做初步的定量预测，并调试预测容错百分比，设定半定性判别标记。

1　实验部分

1.1　材料和测量方法

采集 135 份广西土壤样本，经过风干、碾磨并过 2 mm 筛，在实验室采用凯氏定氮法[19]测定样品中的总氮含量，作为光谱分析的参考化学值。全体样品的参考化学值最大值、最小值、平均值和标准偏差分别为 0.289，0.056，0.133，0.045(%)。采用 Perkin-Elmer 公司的 Spectrum One NTS 傅里叶变换近红外光谱仪检测样本光谱，如图 1 所示。光谱扫描区域设定为 10 000～4 000 cm^{-1}，每个样本经由系统自动扫描 64 次，输出平均光谱。实验保持恒温恒湿环境，温度为(25±1) ℃，湿度为(46%±1%)RH。

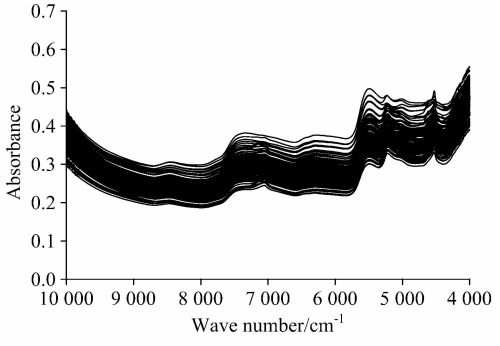

图 1　135 个土壤样品的 FT-NIR 光谱

Fig. 1　FT-NIR spectra of 135 soil samples

1.2　PLS 半定性转换机制

采用常规 PLS 算法建立 FT-NIR 光谱定量分析模型，对所有样本的待测成分含量进行先验预测，并将预测结果转换为半定性判别模式。设定 PLS 半定性判别机制的阈值范围 r(一般 $r\leqslant 0.2$)，根据光谱建模预测值是否落在参考化学值的阈值范围内来赋予先验判别标记(M)，如果光谱建模预测值落在参考化学值的阈值范围内，则认为半定性阈值先验判别准确(标记为 $M=1$)，否则认为先验判别不准确(标记为 $M=0$)，即

$$M_i=\begin{cases}0 & |y'_i-y_i|\geqslant r\cdot y_i\\ 1 & |y'_i-y_i|<r\cdot y_i\end{cases}$$

其中，M_i 为第 i 个样本的 PLS 半定性阈值先验判别结果；y_i 为第 i 个样本的参考化学值；y'_i 为对应的预测值。

1.3　iPCA-LR 建模方法

iPCA-LR 方法的核心思想是采用区间间隔搜索模式寻找 FT-NIR 光谱信息子波段，利用 PCA 算法在待测子波段中提取潜变量信息，结合 LR 回归分析对既有的 PLS 半定性先验判别标记进行建模预测。将整个光谱扫描区域划分为 k 个等宽子波段，每个子波段的数据 X 包含波长点数量为 $t=[p/k]$，p 为全谱段波长点个数。在每一个子波段中对光谱数据提取潜变量 V，结合 PLS 先验判别标记 M 建立 iPCA-LR 模型，利用交叉检验模式完成建模训练和参数优化，进一步对测试集样本进行判别预测。

对输入的光谱潜变量值 V 寻找线性划分边界 $Z=\theta^{\mathrm{T}}V$，基于 logistic 回归方法构造预测函数

$$h_\theta(Z)=h(\theta^{\mathrm{T}}V)=\frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}V}}$$

函数 $h(\cdot)$的值表示 iPCA-LR 预测判别结果为 1 的概率 $P(y=1|V;\theta)$，即

$$P(M'=1\mid V;\theta)=h_\theta(Z)$$
$$P(M'=0\mid V;\theta)=1-h_\theta(Z)$$

其中 M'为 iPCA-LR 模型对每个输入样本的半定性判别的预测标记。根据预测判别标记 M'和先验判别标记 M 构建模型评价指标。

1.4　模型评价指标

建立 PLS 半定性机制结合 iPCA-LR 的 FT-NIR 光谱分析模型，利用交叉验证的方式拟合建模系数，进而对每个土壤样本总氮含量的半定性判别准确率进行评价，通过构造混淆矩阵表来判断模型的预测准确率，能够更详细地分析模型的预测性能。混淆矩阵的结构如表 1 所示，表 1 中的 TP，FN，FP 和 TN 4 个计数值分别用来记录模型预测判别准确与否。进一步利用混淆矩阵中的数值计算 FT-NIR 光谱结合 iPCA-LR 半定性判别方法的预测准确率，计算公式如下

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

表 1　判别预测准确率的混淆矩阵

Table 1　The confusion matrix for discriminant accuracy

		The posterior M'	
		1	0
The prior M	1	TP	FN
	0	FP	TN

注：TP 表示真正值计数，即 $M=1$ 且 $M'=1$ 的数量；FN 表示假负值计数，即 $M=1$ 且 $M'=1$ 的数量；FP 表示假正值计数，即 $M=0$ 且 $M'=1$ 的数量；TN 表示真负值计数，即 $M=0$ 且 $M'=0$ 的数量。

2　结果与讨论

土壤 FT-NIR 光谱全扫描波段为 10 000～4 000 cm^{-1}，光谱分辨率为 4 cm^{-1}，形成 1 512 个波数点。为了降低因固体颗粒大小、表面散射效应和光程变化而形成的噪音干扰，利用 SNV 方法对光谱数据进行预处理，将预处理后的数据用于光谱建模半定性判别分析。

采用常规 PLS 算法建立 FT-NIR 光谱定量分析模型，对 135 个土壤样本的总氮含量进行初步预测，结合半定性机制，针对阈值范围 r 的三个不同取值(0.05，0.10 和 0.15)分别确定半定性先验判别标记，标记样本数量如表 2 所示。根据 PLS 半定性先验标记进一步讨论 iPCA-LR 建模的定性判别。采用区间间隔搜索模式寻找土壤总氮的 FT-NIR 光谱信息子

波段，将整个光谱扫描区域划分成 k 个等宽子波段，分别取 $k\in\{5, 10, 15, 20\}$；不同 k 值对应的每个子波段范围如表 3 所示。在每一个子波段中利用 PCA 算法完成潜变量提取，结合 LR 回归对三个不同的半定性阈值范围(0.05，0.10 和 0.15)所对应的先验判别标记建立土壤总氮的 FT-NIR 光谱 iPCA-LR 模型进行判别预测。

表 2 三个不同阈值范围对应的 PLS 半定性先验判别标记

Table 2 The quasi-qualitative prior discriminant mark of PLS regression corresponding to three different thresholds

The threshold r	Number of samples	
	$M=1$	$M=0$
0.05	58	77
0.10	98	37
0.15	121	14

表 3 不同子波段数(k 值)对应的波段划分结果

Table 3 The waveband division corresponding to the different numbers of wavebands (k)

Number of wavebands (k)	Waveband/cm^{-1}
5	10 000～8 806，8 803～7 607，7 603～6 407，6 403～5 208，5 204～4 008
10	10 000～9 406，9 402～8 806，8 803～8 207，8 203～7 607，7 603～7 007，7 003～6 407，6 403～5 808，5 804～5 208，5 204～4 608，4 604～4 008
15	10 000～9 609，9 605～9 212，9 208～8 814，8 810～8 417，8 413～8 020，8 016～7 623，7 619～7 226，7 222～6 828，6 824～6 431，6 427～6 034，6 030～5 637，5 633～5 240，5 236～4 842，4 838～4 445，4 441～4 048
20	10 000～9 708，9 704～9 410，9 406～9 112，9 108～8 814，8 810～8 517，8 513～8 219，8 215～7 921，7 917～7 623，7 619～7 325，7 321～7 027，7 023～6 729，6 725～6 431，6 427～6 133，6 129～5 835，5 831～5 538，5 534～5 240，5 236～4 942，4 938～4 644，4 640～4 346，4 342～4 048

阈值范围(r)是影响 iPCA-LR 模型性能的一个关键参数，阈值赋值越小，所允许的定量容差范围越小，转换为半定性判别分析之后的准确性要求越强，预测准确率会相对较低。讨论 r 对建模效果的影响，对每一个固定的 r 值，选择使用不同的子波段建立 iPCA-LR 模型进行预测，比较各波段的预测准确率，选择这个固定的 r 值所对应的最佳子波段，对应的 PCA 因子数优选结果如图 2 所示，阈值 $r=0.05$ 的最佳子波段为 6 030～5 637 cm^{-1}，其预测准确率随着因子数的增加基本呈上升趋势，后期略有下降，当因子数是 27 获得最高准确率 75.6%；阈值 $r=0.10$ 的最佳子波段为 6 824～6 431 cm^{-1}，其预测准确率也是随着因子数的增加呈上升趋势，当因子数取值≥26 时准确率达到了 80%以上；阈值 $r=0.15$ 的最佳子波段为 8 413～8 020 cm^{-1}，其预测准确率基本稳定在 90%附近。由此可见，给定阈值范围越大，调试 PCA 因子数越大，半定性预测效果越好。因此，在线检测过程中，如果环境条件允许，可以选择更宽泛的阈值范围以提高光谱实时快检的准确率；如果现场条件比较苛刻，我们只能选择比较小的阈值，其预测准确率也能够达到 75%，可以满足部分在线分析的需求。

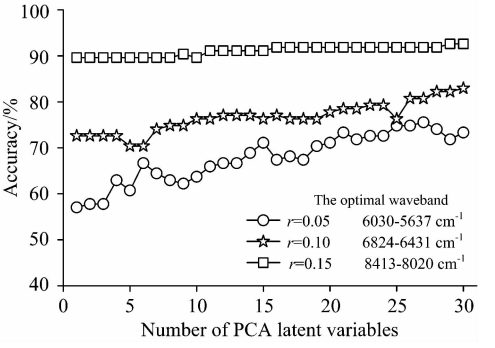

图 2 不同阈值范围对应最优波段的 PCA 潜变量优选结果

Fig. 2 The optimal predictive results of the optimal waveband based on PCA latent variable extraction, corresponding to the three designated thresholds

针对表 3 中不同的 k 值划分的每一个子波段，比较不同阈值范围，选择预测准确率最高值，得到每一个子波段的最优预测准确率如图 3 所示。由图 3 可知，所有子波段的最优准确率均大于 88%，依此选择最优子波段为 6 129～5 835 cm^{-1}($k=20$ 划分的一个子波段)和 5 633～5 240 cm^{-1}($k=15$ 划分的一个子波段)，其对应最高预测准确率达到 93.3%。此外，从次优准确率取值(92.5%)可选择次优子波段为 6 824～6 034 cm^{-1}($k=15$ 划分的连续两个子波段)、8 203～7 007 cm^{-1}($k=10$ 划分的连续两个子波段)和6 403～5 208 cm^{-1}($k=5$ 划分的一个子波段)。

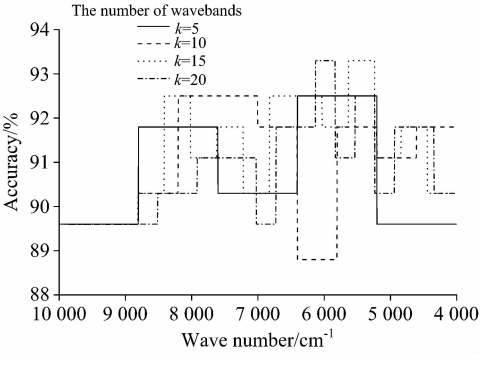

图 3 不同子波段对应最佳准确率分布

Fig. 3 The best predictive accuracy corresponding to different wavebands

依据上述优选的几个波段建立潜变量逻辑回归半定性判别模型，特别针对连续两个波段的情况进行波段合并，结合 PCA 潜变量技术，重新建模确定判别准确率，结果如表 4 所

示。由表 4 可以看出，针对不同的 k 值均能得到优选子波段或合并子波段；尽管合并波段的预测准确率比单个波段有所下降，但仍然保持在 90%以上。结果表明，本半定性判别 iPCA-LR 建模方法应用于土壤总氮含量的 NIR 光谱定量预测能够获得较高的预测准确率。图 4 表示 SNV 方法预处理之后的光谱曲线，并在图中把几个光谱特征波段标记出来。

表 4　优选(组合)波段的 iPCA-LR 建模结果

Table 4　The iPCA-LR modeling results based on the optimal selected wavebands or waveband combinations

k	Optimal waveband/cm^{-1}	Combination	Latent variables	Accuracy /%
5	6 403～5 208	No	29	92.5
10	8 203～7 007	Yes	20	91.1
15	6 824～6 034	Yes	27	91.8
15	5 633～5 240	No	28	93.3
20	6 129～5 835	No	26	93.3

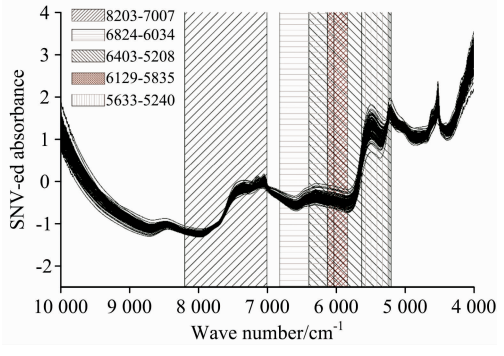

图 4　经过 SNV 预处理的土壤 FT-NIR 光谱波段选择

Fig 4　The optimal wavebands highlighted for the full-range SNV-pretreated FT-NIR spectra

3　结　论

采用 FT-NIR 光谱检测土壤中的总氮含量。首先利用 PLS 方法定量预测 135 个土壤样本中的总氮含量，通过设定 r=0.05，0.10，0.15 三个不同的容错阈值范围，给样本赋予先验判别标记，将定量分析模式转换为 LR 半定性判别模式，结合采用 iPCA 的区间间隔波段搜索潜变量提取方法，经过样本训练，建立土壤总氮近红外 iPCA-LR 半定性判别模型。虽然不同阈值范围下的 FT-NIR 半定性判别模型的预测准确率差别较大，阈值 0.15 的预测准确率基本达到 90%以上，而阈值 0.10 的预测准确率最优可达 80%以上，阈值 0.05 的最优模型也可获得大于 75%的预测准确率，可以满足不同程度的应用水平。另一方面，对比讨论了选择 k=5，10，15，20 四种不同子波段数量区间划分的 iPCA-LR 建模判别准确率，优选近红外光谱特征子波段，并讨论优选连续子波段的组合建模情况，优选的子波段或合并子波段的判别准确率均达到了 90%以上。结果表明，利用 PLS 结合 iPCA-LR 将定量预测转换为半定性判别的方法能够应用于土壤的 FT-NIR 光谱分析中，能够解决常规 PLS 定量问题中容易出现的数据过拟合问题，定标判别结果更符合实际，有利于光谱技术在线检测的应用推广。

References

[1] LI Min-zan, ZHENG Li-hua, AN Xiao-fei, et al(李民赞, 郑立华, 安晓飞, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2013, 44(3): 73.

[2] LU Shan, MAO Cai-yun, XIAO He-xia, et al(鲁　珊, 毛彩云, 肖荷霞, 等). Journal of Anhui Agricultural Sciences(安徽农业科学), 2014, 42(18): 5789.

[3] Chen H Z, Feng Q X, Jia Z, et al. Asian Journal of Chemistry, 2014, 26(15): 4839.

[4] ZHANG Juan-juan, XIONG Shu-ping, SHI Lei, et al(张娟娟, 熊淑萍, 时　雷, 等). Soils(土壤), 2015, 47(4): 653.

[5] LU Wan-zhen(陆婉珍). Modern Near-Infrared Spectroscopy Analytical Technology(现代近红外光谱分析技术). 2nd ed(第 2 版). Beijing: China Petrochemical Press(北京：中国石化出版社), 2007.

[6] Chen H Z, Liu Z Y, Gu J, et al. Analytical Methods, 2018, 10: 5004.

[7] WANG Fan, LI Yong-yu, PENG Yan-kun, et al(王　凡, 李永玉, 彭彦昆, 等). Chinese Journal of Analytical Chemistry(分析化学), 2018, 49(9): 1424.

[8] Sampaio P S, Soares A, Castanho A, et al. Food Chemistry, 2018, 242: 196.

[9] WANG Chang, HUANG Chi-chao, YU Guang-hui, et al(王　昶, 黄驰超, 余光辉, 等). Acta Pedologica Sinica(土壤学报), 2013, 50(5): 881.

[10] Dong X L, Sun X D. Journal of Food Measurement and Characterization, 2013, 7: 141.

[11] MAO Yi, CHEN Wen-lin, GUO Bao-long, et al(毛　毅, 陈稳霖, 郭宝龙, 等). Acta Automatic Sinica(自动化学报), 2014, 40(1): 62.

[12] Zhou L, Chen J, Yao L, et al. Chemometrics and Intelligent Laboratory Systems, 2017, 161: 88.

[13] Sharifzadeh S, Ghodsi A, Clemmensen L H, et al. Engineering Applications of Artificial Intelligence, 2017, 65: 168.

[14] JIANG Xiao-qing, SONG Jiang-feng, LI Da-jing, et al(姜晓青, 宋江峰, 李大婧, 等). Modern Food Science and Technology(现代食品科技), 2013, 29(8): 2020.

[15] Goodarzi M, Saeys W. Talanta, 2016, 146: 155.

[16] Borille B T, Marcelo M C A, Ortiz R S, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2017, 173: 318.

[17] PENG Hai-gen, PENG Yun-fa, ZHAN Ying, et al(彭海根, 彭云发, 詹 映, 等). Food Science and Technology(食品科技), 2014, 39(6): 276.
[18] Bi Y M, Yuan K L, Xiao W Q, et al. Analytica Chimica Acta, 2016, 909: 30.
[19] HJ/T 717—2014. National Environmental Protection Standards of the People's Republic of China(中华人民共和国国家环境保护标准).

FT-NIR Spectroscopy Quasi-Qualitative Determination Applied to the Waveband Selection for Soil Nitrogen

GU Jie[1], CHEN Hua-zhou[1,2*], CHEN Wei-hao[1], MO Li-na[1], WEN Jiang-bei[2]
1. College of Science, Guilin University of Technology, Guilin 541004, China
2. Guangdong Spectrastar Instruments Co. Ltd., Guangzhou 510663, China

Abstract Nitrogen is an important component to measure soil fertility. The traditional chemical method for detecting soil nitrogen content is complex and time-consuming. Fourier transform near infrared (FT-NIR) technology is utilized for direct and rapid quantitative determinationof soil nitrogen. Nevertheless, the calibration models always perform too ideally well to believe when established by the linear analytical methods, like partial least squares (PLS). That is not convinced for the practical application in on-line detection. In this paper, we proposed a fault-tolerant mechanism to be plug-into the quantitative analytical model, transforming the FT-NIR quantitative mode into a quasi-qualitative discriminant mode. In this way, the application ability of the calibration model can be enhanced. A new discriminant method was proposed for quasi-qualitative determination by combining the interval search principal component analysis algorithm with logistic regression (iPCA-LR). The nitrogen contents of soil samples were firstly predicted based on the common PLS regression. The fault-tolerant threshold was set as three different values of 0.05, 0.10 and 0.15, respectively. The samples were marked as accurately or non-accurately discriminated according to the priori predictive values and the thresholds, so that the original quantitative calibration method was transformed into a new quasi-qualitative discriminant method. The iPCA-LR method was applied for the FT-NIR quasi-qualitative discrimination of soil nitrogen. In the same process, we also discussed the latent variable extraction based on different wavebands that were generated by tuning the waveband division number as 5, 10, 15 and 20. Some informative FT-NIR wavebands were selected with optimal discriminant accuracy. And some combination of informative wavebands were also tested. Results showed that the FT-NIR quasi-qualitative discriminant predictive accuracy varied significantly for different thresholds, but fortunately the worst optimal accuracy climbed tothe level slightly above 75%. And the test of different informative wavebands or the combination of informative wavebands output optimal calibration models with the accuracy above 90%. These results were able to meet some practical cases of online detection. In the application of FT-NIR prediction of nitrogen content in soil samples, the proposed method of iPCA-LR manage to transform the common quantitative prediction problem into the quasi-qualitative discriminant problem when combined with the priori PLS prediction. The newly proposed method deals with the disadvantages of overfitting and overidealistic modeling that always appears in common PLS quantitative analysis. In comparison, the quasi-qualitative discriminant mode is more suitable for actual cases in field detection, more beneficial for real-time application of spectroscopy technology.

Keywords Soil nitrogen; FT-NIR; Waveband selection; iPCA-LR model; Quasi-qualitative determination

(Received Dec. 9, 2018; accepted Apr. 13, 2019)

* Corresponding author