傅里叶变换红外光谱的土壤团聚体有机碳和全氮含量估测

刘翠英[1]，张津瑞[2]，曾　涛[1]，樊建凌[2*]

1. 南京信息工程大学应用气象学院，江苏省农业气象重点实验室，江苏 南京　210044
2. 南京信息工程大学环境科学与工程学院，江苏省大气环境监测与污染控制高技术研究重点实验室，江苏 南京　210044

摘　要　土壤团聚体是土壤生态系统的重要组成部分，其碳氮含量及动态决定着土壤碳氮循环过程、稳定性及肥力。由于团聚体分级方法的差异，不同研究所获得的团聚体粒径也不尽相同，应用红外光谱对土壤团聚体性质进行建模预测时若对不同粒径团聚体分别建模需要大量样本且难以对所有组分同时进行合理预测。该研究对不同粒径团聚体样本进行综合建模预测，探寻一种高效可行的不同粒径团聚体性质的综合预测方法。采集了内蒙古淡栗钙土土壤样本进行傅里叶变换红外光谱分析，用遗传算法对特征波长进行了选择，基于偏最小二乘法(PLSR)、支持向量机(SVM)、人工神经网络(ANN)和随机森林(RF)等方法建立了不同粒径团聚体土壤有机碳(SOC)、全氮(TN)和红外光谱吸光度之间的估测模型。结果表明，基于遗传算法筛选的特征光谱区间构建的土壤团聚体 SOC 和 TN 含量的 ANN 模型的预测能力均是最好的(RPD＞2)，显著优于 PLSR、SVM 及 RF 模型；基于全谱数据的 ANN 模型对土壤团聚体 SOC 和 TN 的预测效果均低于基于 GA 选择的特征光谱区间的 ANN 模型，说明基于 GA 的特征光谱区间选择不仅可以简化模型结构，剔除无关的信息，而且可以提高模型的精度和预测效果。该研究将不同粒径土壤团聚体 FTIR 数据混合建模，通过遗传算法筛选特征光谱，发现人工神经网络模型可以很好地对土壤团聚体碳氮含量进行预测，且不会受团聚体粒径的影响，主要由于在遗传算法选择特征光谱时已将某些反映土壤矿物、粘粒等特征的波长区间包含在内，而人工神经网络所建立的模型可能已包含了不同粒径对土壤碳氮含量的影响，该结果表明基于遗传算法筛选特征波长区间并采用人工神经网络可以将不同粒径土壤团聚体统一建模，用于团聚体土壤有机碳和全氮含量的估测。

关键词　淡栗钙土；有机碳；全氮；团聚体；红外光谱

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2020)12-3818-07

引　言

土壤团聚体是土壤的基本单位，是土壤矿物通过胶结团聚过程形成的具有一定大小的土壤基本结构单元，具有水稳性、力稳性、多孔性等特点，对土壤的理化性质有着重要的影响[1]。此外，土壤团聚体的形成过程及其稳定机制受土壤有机质，特别是其中碳氮的影响，团聚体与有机质含量是土壤结构状况和肥力水平的重要评判依据。然而，影响土壤有机碳氮在团聚体内的含量与分布的因素非常复杂[2]，快速、精确地了解土壤有机碳氮含量在团聚体内部的分布与变化将有助于了解土壤碳氮交换量、储量变化及土壤养分的管理。

近几十年来，人们对红外光谱法预测土壤性质进行了大量的尝试，利用红外光谱可以根据样本的光谱特征确定其中特定成分的含量，整个测量过程简单、快速且无破坏性。红外光谱法已被用于对土壤有机碳、粘粒含量、全氮含量、pH、可提取态 P、K、Fe、Ca、Mg 及 CEC 等指标的分析[3]。由于光谱数据的复杂性，主成分回归、多元线性回归(MLR)、偏最小二乘法回归(PLSR)等多元数据分析技术常被用于红外光谱预测土壤性质分析建模，随着数据挖掘技术的发展，支持向量机(SVM)、随机森林(RF)、人工神经网络(ANN)等机器学习方法也逐渐被应用于红外光谱数据建模。然而，针对多元数据处理算法的比较研究还较缺乏，各种算法的优劣及适用范围尚不清楚。最近，Moura-Bueno 等[4]基

收稿日期：2020-06-14，**修订日期**：2020-09-26

基金项目：国家重点基础研究发展计划项目(2014CB954002)，江苏省“六大人才高峰”项目(JNHB-061)，南京市留学人员科技创新项目(R2019LZ09)，南京信息工程大学人才启动基金项目资助

作者简介：刘翠英，1982 年生，南京信息工程大学应用气象学院副教授　　e-mail：002263@nuist.edu.cn

＊通讯联系人　　e-mail：jlfan@nuist.edu.cn

于红外光谱，采用 PLSR、MLR、SVM 和 RF 方法建立了土壤有机碳的预测模型，结果表明，预测效果最好的是 PLSR 模型；基于随机森林建立的模型预测效果不够突出。另一方面，已有研究多针对全土性质进行建模分析，用红外光谱对土壤团聚体有机碳(SOC)和全氮(TN)进行预测的研究还不多见。本研究以我国内蒙古淡栗钙土为研究对象，采用遗传算法(genetic algorithm，GA)进行特征波长选择，采用 PLSR、SVM、ANN 和 RF 等方法建立了不同粒径团聚体 SOC、TN 和红外光谱吸光度之间的关系模型，进而评价不同模型预测土壤团聚体中 SOC 和 TN 含量的潜力，探寻有效预测土壤团聚体性质的最佳算法，为实时、快速分析土壤团聚体有机碳和全氮含量提供技术支撑。

1　实验部分

1.1　土壤样本

土壤样品采自内蒙古自治区乌兰察布市四子王旗，是典型的内蒙古短花针茅荒漠草原带，共采集 24 个土壤样品，土壤类型为淡栗钙土。样品运回实验室后在 4 ℃保存，采用湿筛法[5]对土壤团聚体进行分级，共分为>2，0.25～2，0.053～0.25 及<0.053 mm 四级团聚体，共得到 96 个团聚体样品。所得各级团聚体样品经冷冻干燥后研磨过 100 目筛，用稀 HCl 溶液(1 mol·L^{-1})去除土壤样品中的无机碳。将土壤再次研细，过 100 目，用元素分析仪(Vario EL IIII，Elementar，Germany)测定团聚体土壤有机碳(SOC)和全氮(TN)含量，每个样品重复测定三次并求平均值。

1.2　红外光谱测定与光谱数据处理

团聚体样品红外光谱用傅里叶变换红外光谱仪(Nicolet iS5，Thermo Fisher Inc.)测定，光谱范围为 400～4 000 cm^{-1}，分辨率为 4 cm^{-1}，扫描次数为 32 次。将所得光谱数据转换为吸光度[log(1/反射率)]，并校正基线。采用 Savitzky-Golay 平滑法对光谱数据进行平滑处理，然后取一阶微分，预处理后的光谱如图 1 所示。

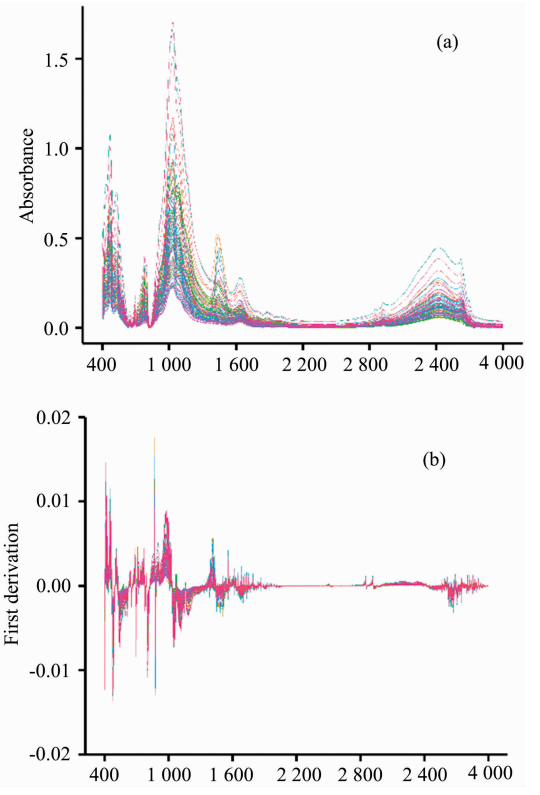

图 1　经 S-G 平滑(a)及一阶微分(b)处理后的团聚体 FTIR 光谱图

Fig. 1　FTIR spectra of soil aggregates after Savitzky-Golay smoothing (a) and first order differential pretreatments (b)

1.3　校正数据集与预测数据集的划分

在建立模型前，需要将样品红外光谱及对应的 SOC、TN 数据划分为建模数据集和验证数据集，采用 Kennard-Stone 算法进行划分，通过计算样本光谱变量之间的欧氏距离，在样本特征空间里均匀地选取建模样本。划分所得建模数据集与验证数据集的统计结果如表 1 所示。

表 1　建模数据集与验证数据集的样本统计结果

Table 1　Statistic summary of the modeling and validating data subsets

	样本数	SOC				TN			
		均值	标准差	最小值	最大值	均值	标准差	最小值	最大值
建模集	70	1.938	1.278	0.639	7.492	0.177	0.093	0.051	0.537
验证集	26	1.666	0.574	0.877	2.668	0.177	0.075	0.078	0.334

1.4　光谱变量选择与建模方法

1.4.1　确定区间大小

采用遗传算法(genetic algrothim)[6]进行特征波长选择，由于 FTIR 光谱波长数目较大(每一光谱包含 7 468 个数据)，直接采用原始数据会使 GA 的优化搜索空间过大，因此需要按一定波长区间对光谱数据进行划分。按适宜波长间隔将整个光谱均分为 x 个区间，然后求取各区间数据的标准偏差，平均标准偏差值越小说明区间数据越相近，因此在保障区间数据相近性的前提下应该尽量减少变量数目以优化计算效率。

1.4.2　特征光谱的确定

经 S-G 平滑及一阶微分处理后的数据，按适宜的波长区间对光谱进行均分，以各区间的平均谱数据为自变量，以交叉校验均方根误差 RMSECV 为适应度函数，采用遗传算法(GA)进行最优光谱波长的选择，设定种群大小为 30，最大繁殖代数为 100，交叉概率为 0.5，变异概率为 0.01[7]，五次重复遗传算法后，确定特征光谱。本研究使用“caret”包进行 GA 分析。

1.4.3 偏最小二乘法回归(PLSR)

偏最小二乘回归(PLSR)是一种广泛用于土壤光谱定量分析的线性回归模型，使用潜变量方法对预测变量和观察变量的两个投影空间中的协方差结构进行建模。因此，PLSR 模型克服了变量之间的共线性问题。此外，PLSR 模型在选择特征向量时，强调自变量对因变量的解释和预测，消除了无用噪声对回归的影响，并最大程度地减少了模型中包含的变量数量。本研究使用“pls”包进行 PLSR 建模。

1.4.4 支持向量机(SVM)

支持向量机(SVM)是一种对数据进行二元分类的模型，SVM 的基本模型是找到一个可以分隔正反数据的超平面，选择的超平面需要离训练集数据尽可能远；支持向量机的关键在于核函数，这是一个非线性的分类器。它的学习策略就是间隔最大化，找到距离超平面间隔最小的样本点，然后将其间隔最大化。本研究使用“e1071”包进行 SVM 建模。

1.4.5 人工神经网络(ANN)

人工神经网络是基于生物学中神经网络的基本原理，模仿生物的神经网络来对复杂的外界信息进行处理，它是对生物神经元网络的简易化模仿。人工神经网络模型可以并行分布地去处理问题，拥有很高的容错性，把信息的加工和记忆能力结合一起。它实际上是一个复杂网络，连接大量的简单元件，因此 ANN 具有很高的非线性，能够进行复杂的逻辑操作，处理非线性关系的样本数据。本研究使用“neuralnet”包进行 ANN 建模。

1.4.6 随机森林(RF)

随机森林模型通过自助重采样技术，连续生成训练样本和测试样本，并从训练样本中生成多个分类树以形成随机森林。它通过对决策树进行平均来降低过拟合的风险。即使新的数据点出现在数据集中，整个算法也不会受到太大的影响，只会影响一个决策树，并且很难影响所有决策树。本研究使用“randomForest”包进行 RF 建模。

1.5 模型评价

主要采用决定系数(R^2)、均方根误差(RMSE)以及相对分析误差(RPD)对模型进行评价。R^2 越大且 RMSE 越小说明模型精度越好、其预测效果越好。R^2 的判断标准为：$R^2>0.90$ 表示预测结果出色；R^2 在 0.81～0.90 之间表示预测结果很好；R^2 在 0.66～0.80 之间为预测结果一般；$R^2<0.66$ 表示预测结果很差。PRD 的判断标准为：RPD>2 表明模型具有极好的预测能力；1.4<RPD<2 表明模型可对样品作粗略估测；RPD<1.4 表示模型无法对样品进行预测[8]。本研究相关计算及绘图均采用 R 语言编程完成。

2 结果与讨论

2.1 适宜波长区间

由图 2 可见，波长间隔越小，样本的平均标准偏差越低，但同时区间数目越多。在波长间隔大小为 6 cm^{-1}时最大标准偏差出现一个低谷，此时区间数为 600 个，因此在兼顾数据相近性与变量少量性的前提下，确定 6 cm^{-1}作为划分区间的适宜大小，将整个光谱分为 600 个区间，然后将区间的平均光谱值作为自变量形成光谱曲线。

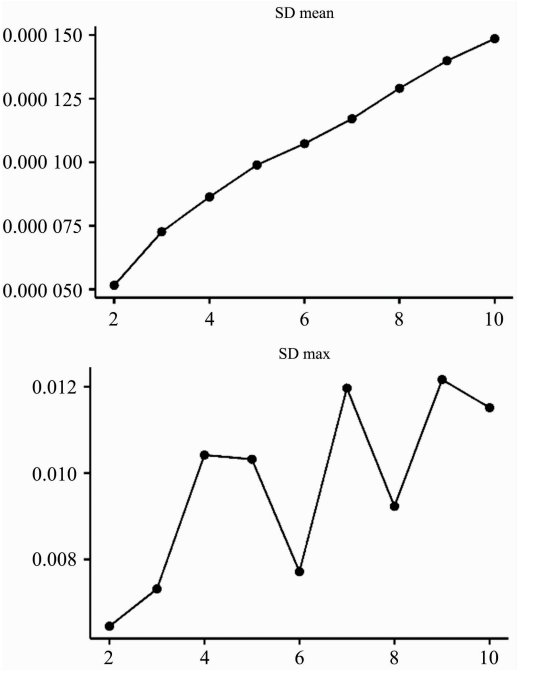

图 2 平均标准偏差和最大标准偏差与波长区间大小的关系

Fig. 2 Relations between the mean SD, maximum SD and wavelength interval

2.2 土壤团聚体特征光谱区间

采用 GA 算法并重复 5 次后，筛选出团聚体 SOC 特征光谱区间共计 303 个。由团聚体 SOC 特征光谱区间的分布(图 3)可以看出，SOC 的特征光谱覆盖范围较广，但在 2 200～2 700及＞3 700 cm^{-1}以上波长范围内选取的特征光谱区间较少。此外，在 GA 筛选过程中 1 132～1 138，1 372～1 378，1 630～1 636，1 810～1 816，1 864～1 870

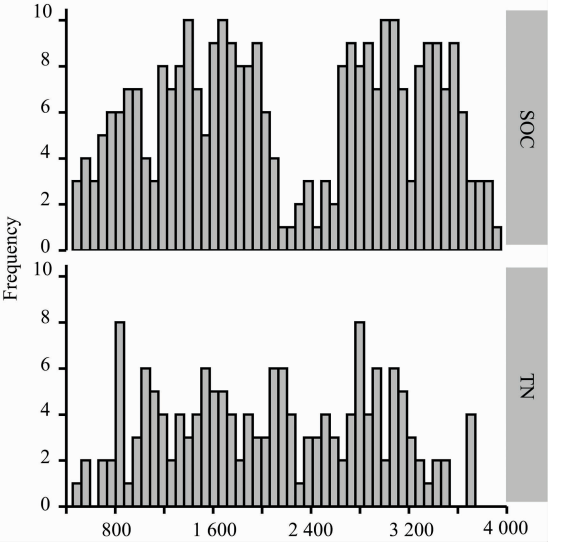

图 3 团聚体 SOC 和 TN 特征光谱区间分布

Fig. 3 Histrogam of charasteristic wavelengths of SOC and TN in soil aggregates

cm^{-1}这五个光谱区间在每次重抽样中均被选为特征光谱，表明这些光谱区间对团聚体 SOC 含量较为敏感。

对于土壤团聚体 TN，GA 算法共筛选出特征光谱 161 个。由团聚体 TN 特征光谱区间的分布(图 3)可以看出，TN 的特征光谱覆盖范围较平均，但在＞3 200 cm^{-1}以上波长范围内选取的特征光谱区间较少。在 GA 筛选过程中 1 114～1 120，1 258～1 264，1 264～1 270，1 276～1 282 和 1 408～1 414 cm^{-1}这五个光谱区间在每次重抽样中均被选为特征光谱，表明这些光谱区间对团聚体 TN 含量较为敏感。

2.3　不同建模方法对团聚体 SOC 预测结果比较

基于特征光谱区间，采用 PLSR 或 ANN 对团聚体 SOC 的建模结果均非常出色(R^2＞0.90，表 2)，同时这两种模型对验证样本的预测结果也很好(R^2＞0.80)。然而 PLSR 模型对验证样本 SOC 含量较低时出现了较明显的高估(图 4)，导致模型预测结果在低值区与实测值偏差较大，RMSE 值较高，使得 PLSR 整体预测能力较为一般(PRD＜2)。采用 SVM 模型对团聚体 SOC 的建模结果也较好，但模型对样本的预测能力却较差(R^2＜0.66，PRD＜2)。然而，RF 对团聚体 SOC 的建模及预测结果均较差，基本无法对样本进行预测(PRD＜1.4)。总之，四种模型中 ANN 的建模及预测结果均是最好的，其对验证样本的预测除有少数点偏离外，其余点基本均在 1∶1 线的附近(图 4)，其 RMSE 值在四种模型中最低(RMSE＝0.227)。此外，ANN 在对团聚体 SOC 预测时并没有受团聚体粒径的影响，整体表现出极好的预测能力(RPD＞2)。然而，本研究样品数量相对较少，若增加建模样品数量可能会建立效果更好的模型。

表 2　不同模型对 SOC 和 TN 预测效果比较

Table 2　Comprison of SOC and TN predicting results by different models

		建模样本		验证样本		PRD
		R^2	RMSE	R^2	RMSE	
SOC	PLSR	0.911	0.379	0.827	0.300	1.88
	SVM	0.831	0.674	0.610	0.354	1.59
	ANN	0.999	0.025	0.850	0.227	2.48
	RF	0.681	0.782	0.546	0.413	1.36
TN	PLSR	0.852	0.036	0.778	0.040	1.84
	SVM	0.862	0.040	0.670	0.046	1.60
	ANN	0.999	0.029	0.781	0.036	2.05
	RF	0.650	0.557	0.681	0.046	1.60

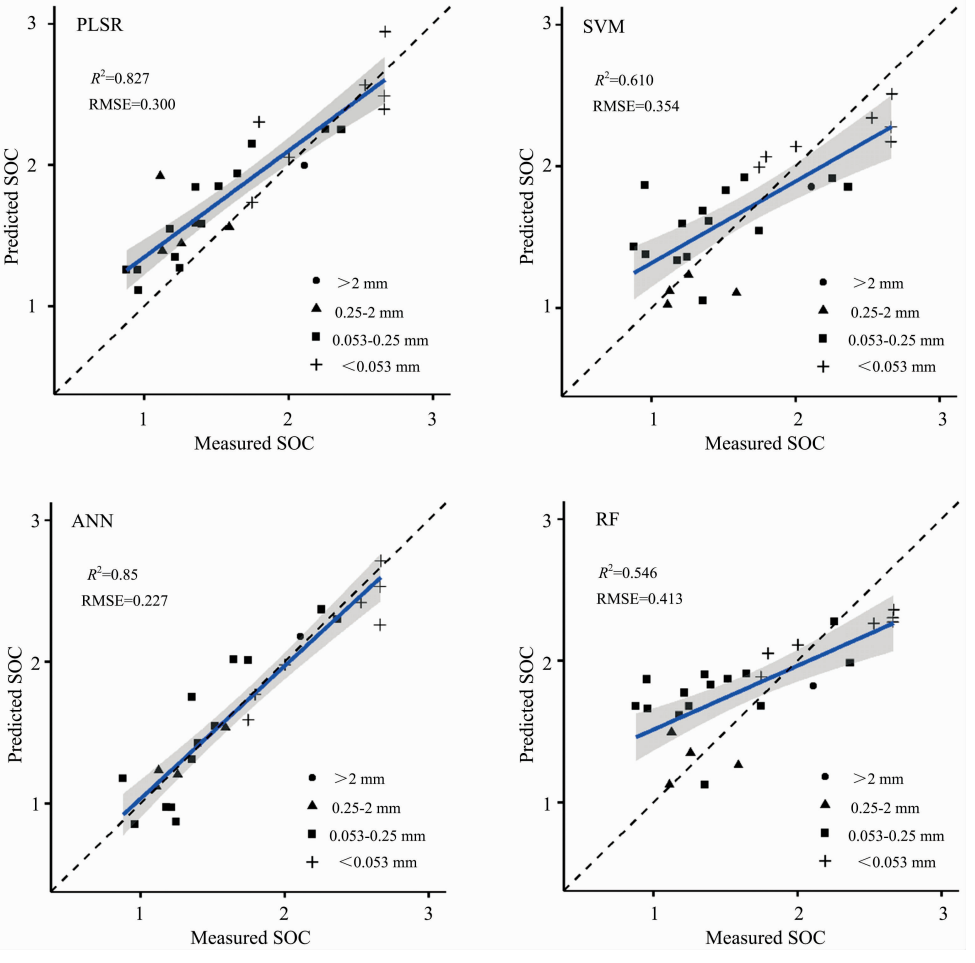

图 4　不同模型对团聚体 SOC 的预测值与实测值对比

Fig. 4　Comparison of predicted SOC and measured SOC in soil aggregates by different models

2.4 不同建模方法对团聚体 TN 预测结果比较

基于特征光谱区间，采用 PLSR 或 SVM 对团聚体 TN 的建模结果均较好($R^2>0.80$，表 2)，然而这两种模型对验证样本的预测结果却一般($0.66<R^2<0.80$)。RF 对团聚体 TN 的建模及预测结果均较差，基本无法对样本进行预测。ANN 对团聚体 TN 的建模结果出色($R^2>0.90$)，虽然对验证样本的预测结果一般，但仍表现出了极好的预测能力(RPD>2)，这主要是因为 ANN 模型对团聚体 TN 的预测没有出现明显的高估或低估，整体预测点均在 1∶1 线的附近(图 5)，其 RMSE 值在四种模型中最低(RMSE=0.036)。因此，随着建模样品数量的增加，可以使用 ANN 建立效果更好的团聚体 TN 预测模型。

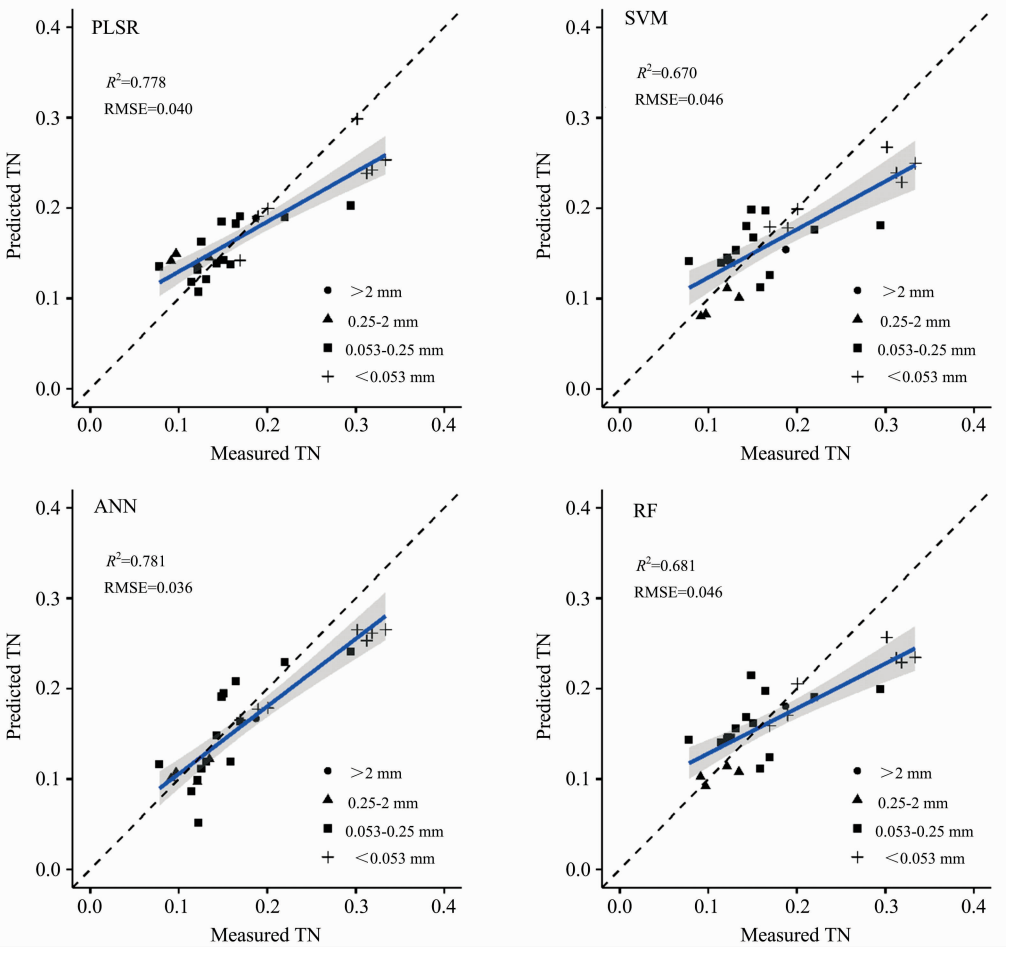

图 5 不同模型对团聚体 TN 的预测值与实测值对比

Fig. 5 Comparison of predicted TN and measured TN in soil aggregates by different models

2.5 基于全谱的团聚体 SOC 和 TN 估测

为了验证特征光谱选择对团聚体 SOC 和 TN 估测的影响，运用上述结果中表现最好的 ANN 模型对 FTIR 全谱数据进行了建模预测。结果表明，采用全谱数据对验证样本 SOC 的预测结果较好，对 TN 的预测结果却一般(图 6)。此外，采用全谱数据对 SOC 和 TN 的预测效果均低于基于 GA 选择的特征光谱区间的 ANN 模型(图 4 和图 5)。可见，采用 GA 进行特征光谱区间的选择不仅可以简化模型的结构，剔除光谱中无关的信息，而且可以提高模型的精度和预测效果。

特征波长选择是红外光谱预测土壤性质的一个重要步骤，通过对特征波长的选择可以剔除无关的信息，提高模型的预测性能及计算效率。由于红外光谱数据量大，常采用相关系数法、连续投影算法、遗传算法、粒子群算法、蚁群算法等[5-6]来选取特征波长。如刘振尧等[9]基于随机森林优选信息波长，选取了 215 个波长信息，建立了土壤有机质的预测模型。本研究选用遗传算法对特征光谱区间进行筛选，充分利用了土壤的光谱信息，采用人工神经网络建模对团聚体 SOC 和 TN 的预测也取得了较满意的结果。因此，遗传算法从整体最优化的角度筛选了土壤团聚体 SOC 和 TN 的特征波段，简化了模型的结构并提高了模型的精度，结合人工神经网络建模，更适用于团聚体土壤有机碳和全氮的光谱特征分析与估测模型的构建。

红外光谱由于测量过程简单、快速且无破坏性，已被广泛用于对土壤有机碳、粘粒含量、全氮含量、pH、可提取态 P、K、Fe、Ca、Mg 及 CEC 等指标的分析[3, 10]。Erktan 等[11]采集了法国南部 75 个荒地土壤样品并将其分为<1，1～2 和 3～5 mm 三级团聚体，利用中红外-近红外光谱对团聚体稳

定性进行了预测。Shi 等[12]采集了 83 个比利时土壤样本并将其分为＞250，63～250 和＜63 μm 三级团聚体，利用可见-近红外光谱建立了团聚体稳定性的偏最小二乘法定量模型。可见，已有研究往往针对不同粒径组分分别建立模型进行预测，该过程要求样本量大且很难对所有组分同时进行合理预测。此外，不同研究者所选用的土壤团聚体粒径分级方案存在差异，获得的团聚体粒径也不尽相同，不易针对不同粒径进行建模分析，严重制约了红外光谱法在分析预测土壤团聚体理化性质中的应用。本研究发现，虽然不同粒径土壤团聚体在不同波长的吸光度存在较大差异，但其 FTIR 光谱的整体趋势是一致的，因此可以考虑将不同粒径土壤团聚体 FT-IR 数据混合建模。通过遗传算法筛选特征光谱，我们发现人工神经网络模型可以很好地对土壤团聚体碳氮含量进行预测，且不会受团聚体粒径的影响。这可能由于在遗传算法选择特征光谱时已将某些反映土壤矿物、粘粒等特征的波长区间包含在内(图 3)，如 780～800 cm^{-1}处是石英矿物中 Si—O 键伸缩振动、910 cm^{-1}为高岭石和三水铝石等粘土矿物中 O—H 的弯曲振动、1 034 cm^{-1}为高岭石矿物中 Si—O 键的伸缩振动、3 600～3 700 cm^{-1}为粘土矿物中 O—H 键的伸缩振动[10, 13]。由于人工神经网络的复杂性和高度的非线性，其所建立的模型可能已包含了不同粒径对土壤碳氮含量的影响。因此，我们认为基于遗传算法筛选特征波长区间并采用人工神经网络可以将不同粒径土壤团聚体统一建模，用于团聚体土壤有机碳和全氮含量的估测。

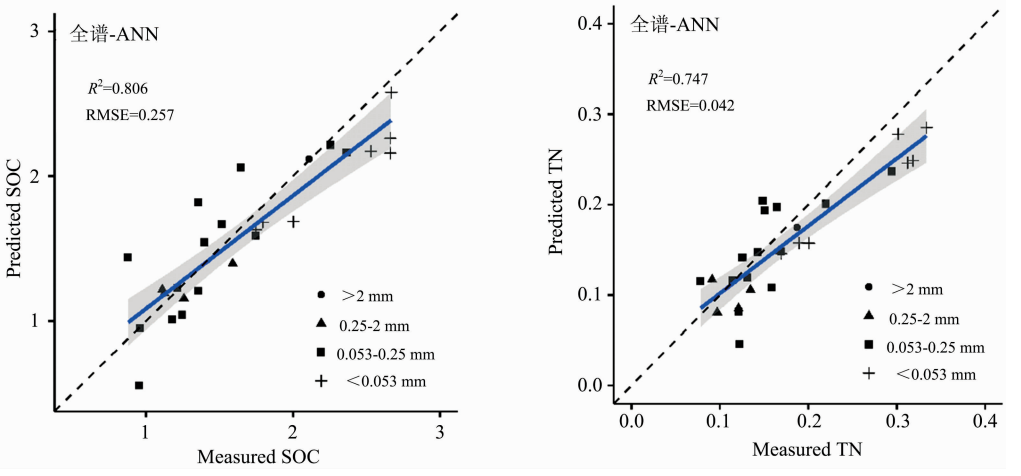

图 6　基于全谱数据对团聚体 SOC 和 TN 的预测值与实测值对比

Fig. 6　Comparison of predicted TN and measured TN in soil aggregates by different models

3　结　论

(1)采用 Savitzky-Golay 平滑并取一阶微分对 FTIR 光谱数据进行预处理后，在波长间隔大小为 6 cm^{-1}时最大标准偏差出现一个低谷，同时平均标准偏差较小，区间数量适中，可以用于划分 FTIR 光谱区间。

(2)基于遗传算法筛选的特征光谱区间构建的土壤团聚体 SOC 和 TN 含量的 ANN 模型建模及预测能力均是最好的(RPD＞2)，显著优于 PLSR、SVM 及 RF 模型。

(3)基于全谱数据的 ANN 模型对土壤团聚体 SOC 和 TN 的预测效果均低于基于 GA 选择的特征光谱区间的 ANN 模型。结果表明，采用 GA 进行特征光谱区间的选择不仅可以简化模型结构，剔除无关的信息，而且可以提高模型的精度和预测效果。

(4)将不同粒径土壤团聚体 FTIR 数据混合建模。通过遗传算法筛选特征光谱，发现人工神经网络模型不仅可以很好地对土壤团聚体 SOC 和 TN 含量进行预测，而且不会受团聚体粒径的影响，表明该方法可以用于团聚体土壤有机碳和全氮含量的估测。

References

[1] Totsche K U, Amelung W, Gerzabek M H, et al. Journal of Plant Nutrition and Soil Science, 2017, 181(1): 104.
[2] Yu H, Ding W, Chen Z, et al. Scientific Reports, 2015, 5: 13804.
[3] Jaconi A, Poeplau C, Ramirez-Lopez L, et al. European Journal of Soil Science, 2018, 70(1): 127.
[4] Moura-Bueno J M, Dalmolin R S D, ten Caten A, et al. Geoderma, 2019, 337(5): 565.
[5] Lin Y, Ye G, Kuzyakov Y, et al. Soil Biology and Biochemistry, 2019, 134(7): 187.
[6] Tsakiridis N L, Tziolas N V, Theocharis J B, et al. European Journal of Soil Science, 2019, 70(3): 578.
[7] CHEN Hong-yan, ZHAO Geng-xing, ZHANG Xiao-hui, et al(陈红艳，赵庚星，张晓辉，等). Chinese Agricultural Science Bulletin(中国农学通报), 2015, 31(2): 209.
[8] LÜ Mei-rong, REN Guo-xing, LI Xue-ying, et al(吕美蓉，任国兴，李雪莹，等). Spectroscopy and Spectral Analysis(光谱学与光谱分

析)，2020，40(4)：1082.
[9] LIU Zhen-yao，WEN Jiang-bei，GAO Hong-zhi，et al(刘振尧，温江北，高洪智，等). Modern Agricultural Equipment(现代农业装备)，2017，6：37.
[10] Parikh S J，Goyne K W，Margenot A J，et al. Advances in Agronomy，2014，126(4)：1.
[11] Erktan A，Legout C，De Danieli S，et al. Geoderma，2016，271(11)：225.
[12] Shi P，Castaldi F，van Wesemael B，et al. Geoderma，2020，357(1)：113958.
[13] HAO Xiang-xiang，HAN Xiao-zeng，ZOU Wen-xiu(郝翔翔，韩晓增，邹文秀). Chinese Journal of Analytical Chemistry(分析化学)，2018，46(4)：616.

Determination of Soil Organic Carbon and Total Nitrogen Contents in Aggregate Fractions From Fourier Transform Infrared Spectroscopy

LIU Cui-ying[1]，ZHANG Jin-rui[2]，ZENG Tao[1]，FAN Jian-ling[2]*

1. Jiangsu Key Laboratory of Agricultural Meteorology，College of Applied Meteorology，Nanjing University of Information Science and Technology，Nanjing 210044，China
2. Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control，School of Environmental Science and Engineering，Nanjing University of Information Science and Technology，Nanjing 210044，China

Abstract Soil aggregates are the main soil components，in which carbon (C) and nitrogen (N) content and dynamics significantly influence the soil Cand Ncycle process，stability and soil fertility. Due to the difference of aggregates fractionation methods，the size of aggregate fractions obtained from different studies was not the same. Therefore，a large quantity of aggregates samples was required when using infrared spectroscopy to predict the properties of soil aggregates，while it is difficult to reasonably predict each fraction. Comprehensive modeling and prediction of samples from different aggregate fractions were conducted. Fourier-transform infrared spectroscopy analysis was carried out on the soil samples of the light chestnut soil in Inner Mongolia，using a genetic algorithm to select the characteristic wavelength. Prediction models of soil organic carbon (SOC) and total nitrogen (TN) in aggregate fractions were established based on partial least squares (PLSR)，support vector machine (SVM)，artificial neural network (ANN) and random forest (RF) methods. Based on the characteristic spectral interval screened by genetic algorithm，the ANN model showed the best modeling and prediction abilities of SOC and TN content in soil aggregates(RPD>2)，which is significantly better than PLSR，SVM and RF models. The prediction ability of the ANN model based on full-spectrum data is lower than that of the ANN model based on GA-selected characteristic spectral intervals. The results indicated that the selection of GA-based characteristic spectral intervals could not only simplify the model structure and eliminate irrelevant information but also improve the accuracy and prediction ability of the model. In the present study，FTIR data from different aggregate fractions were mixed for modeling. By using a genetic algorithm to filter the characteristic spectrum，we found that the artificial neural network model can reliably predict the SOC and TN contents in soil aggregates，which was not affected by aggregate size. This might be mainly due to the fact that some wavelength ranges reflecting soil minerals，clay particles，etc. have been included in the selection of characteristic spectra by genetic algorithms，and that the effect of particle size on the SOC and TN might have already been included in the ANN model. The result highlights that the screening of characteristic wavelength intervals based on genetic algorithms and the use of artificial neural networks can model soil aggregates of different particle sizes in a unified manner，and can be used to estimate SOC and TN contents of aggregates.

Keywords Light chestnut soil；Soil organic carbon；Total nitrogen；Soil aggregates；FTIR spectroscopy

(Received Jun. 14，2020；accepted Sep. 26，2020)

* Corresponding author