文章编号：1672-8785(2021)01-43-06

一种用于预测蚕丝含量占比的近红外光谱分析方法

范雅婷[1]　刘　胜[2]

(1.东华理工大学长江学院，江西 抚州　344000；
2.北京林业大学理学院，北京　100083)

摘　要：针对近红外光谱分析技术中未充分利用预测模型光谱数据的问题，提出了一种可充分利用光谱数据和有效预测蚕丝含量占比的新方法。以 5 种类型共 145 个样本的蚕丝含量占比以及相应的所有蛋白质基光谱数据为研究对象，将这些样本分别划分为校正集和验证集，并采用偏最小二乘回归(Partial Least Squares Regression，PLSR)方法和提出的偏最小二乘回归多模型(multi-model Partial Least Squares Regression，multi-PLSR)方法建立了预测模型。然后对比和观察了两种方法的预测效果。以类型 2 的蚕丝样本为例，选用 13 个主成分并对比两种模型后发现，multi-PLSR 模型的相关系数由 0.594 增至 0.9784，平均相对误差由 0.4866 降至 0.1384。实验结果表明，新方法充分利用了光谱数据中的信息，提高了蚕丝含量占比预测模型的精度，为建立近红外光谱预测模型提供了一种新思路。

关键词：近红外光谱；蚕丝含量；多模型；偏最小二乘回归

中图分类号：O657.33　　**文献标志码**：A

DOI：10.3969/j.issn.1672-8785.2021.01.009

A Near Infrared Spectroscopy Method for Predicting the Percentage of Silk Content

FAN Ya-ting[1]，LIU Sheng[2]

(*1. Yangtze River College，East China Institute of Technology，Fuzhou 344000，China；*
2. College of Science，Beijing Forestry University，Beijing 100083，China)

Abstract：Aiming at the problem that the spectral data of prediction model are not fully utilized in near infrared spectroscopy，a new method which can make full use of spectral data and effectively predict the proportion of silk content is proposed. The proportion of silk content in 145 samples of 5 types and the corresponding spe-

收稿日期：2020-08-19

基金项目：国家自然科学基金项目(61571002；61179034)

作者简介：范雅婷(1993-)，女，江西抚州人，硕士，主要研究方向为概率论与数理统计。
E-mail：fan_yating@163.com

ctral data of all protein bases are taken as the research objects. The samples are divided into calibration and verification sets respectively. The partial least squares regression method and the partial least squares regression multi-model method are respectively used to establish the prediction model, and the prediction effects of the two methods are compared and observed. Taking the silk samples of type 2 as an example, 13 principal components are selected and the two models are compared. It is found that the correlation coefficient of multi-PLSR increases from 0.594 to 0.9784, and the average relative error decreases from 0.4866 to 0.1384. The experimental results show that the new method makes full use of the information in the spectral data and improves the accuracy of the prediction model of silk content proportion, which provides a new thought for the establishment of the near infrared spectrum prediction model.

Key words: near infrared spectrum; silk content; multiple model; partial least squares regression

0 引言

在近红外光谱分析技术中，建立可靠的预测模型是一个非常关键的环节。利用近红外光谱技术进行定量分析时，PLSR 是一种很常用的建模方法。近些年，采用 PLSR 方法进行近红外光谱定量分析的研究成果表明[1-3]，该方法预测效果较好，可用于农业、林业、食品以及木材等领域的成分快速检测。此外，多模型建模方法也被用来快速、有效地预测木材成分含量[4-6]。

从以往的研究中发现，多模型方法擅长处理低维度数据且预测效果好，但也存在光谱数据未得到充分利用的不足。从这些研究中可以看出，所用光谱数据越多，模型预测效果越好[7]。为了充分利用光谱数据，我们尝试在多模型建模方法的基础上进行改造。PLSR 方法能够解决数据量大的问题，适合处理高维度数据，且模型预测效果较好。本文利用 PLSR 方法的优势，并结合多模型建模思想提出了一种新方法——multi-PLSR 方法。采用基于该方法的近红外光谱预测模型来研究蚕丝含量占总质量的百分比，以解决之前研究中光谱数据未得到充分利用的问题。研究结果表明，提出的 multi-PLSR 模型能够充分利用光谱数据，并可有效预测蚕丝含量占总质量的百分比。

本文研究的出发点是探讨与尝试解决多模型建模方法中光谱数据未能得到充分使用的问题。研究目的是探索用于预测蚕丝含量占总质量百分比的有效近红外光谱分析模型。研究意义是在探索过程中对建模方法进行创新，并结合两种建模方法的优势，用数据验证这种建模方法的有效性。这意味着本文可能为建立某些成分含量的近红外光谱预测模型提供了一种新思路。

1 材料获取方法

1.1 样本制备方法

所收集的蚕丝来自于广东省、四川省和浙江省，其生产年份为 2017 年、2016 年或 2015 年。所收集的涤纶和锦纶布样来自于河北省、浙江省、江苏省和广东省，其生产年份为 2015 年或 2014 年。用植物粉碎机将收集到的蚕丝、涤纶布样和锦纶布样打成粉末，使其可以通过八十目的筛子。然后依据实验设计，使用精度为万分之一的天平来称取每次所需的某种粉末，并将不同的粉末按预定比例进行混合，共制备了 145 个含蚕丝的样本和 155 个不含蚕丝的样本。样本的蚕丝含量(或涤纶、锦纶含量)由实际称取的数值给出。

1.2 仪器设备与光谱数据采集方法

采集光谱所用的仪器是日本日产公司生产的 UH4150 近红外分光光度计。光谱采集方法如下：将制备的每个含蚕丝样本(或不含蚕丝的样本)放进样本池进行扫描；波长范围为 800～2500 nm，扫描速度为 1200 nm/min，分辨率设为 5 nm；去除本底光谱，最终得到建模所需的近红外光谱。

2 建模方法

2.1 将样本分为校正集与验证集

将蚕丝样本分为 5 类，并以类型 2 的蚕丝样本为例介绍建模方法。其余 4 个类型的蚕丝样本采用相同方法来建模。类似于文献[6]中的分组方法，根据蚕丝含量占总质量的百分比数值，对类型 2 中的 31 个蚕丝样本进行从小到大排序。采用四分之一取法将其中 7 个样本放入验证集，并将其余 24 个样本放入校正集。为便于后续建模，将取出的校正集和验证集样本重新排序，具体如下：将校正集中 24 个样本的蚕丝含量占比用 Y_1～Y_{24}标记，并记 $Y_J=(Y_1,Y_2,\cdots,Y_{24})$，用于表示校正集样本的蚕丝含量占比矩阵。将验证集中 7 个样本的蚕丝含量占比用 Y_{25}～Y_{31}标记，并记 $Y_Y=(Y_{25},Y_{26},\cdots,Y_{31})$。

对于蚕丝样本的光谱数据，考虑到建立偏最小二乘回归子模型的需要(偏最小二乘回归模型研究的是两个矩阵之间存在的关系)以及噪声因素，作出以下筛选：校正集的每个蚕丝样本的光谱数据中都有 360 个波长，每个波长形成一个维度为 24 的分辨率向量，从而构成一个 360 行 24 列的光谱矩阵(校正集光谱矩阵，记为 $X_J=(a_{ij})$)。其中，元素 a_{ij}表示校正集光谱数据中第 i 个波长处第 j 个样本的分辨率。而对于验证集蚕丝样本的光谱数据，每个波长形成一个 7 维的分辨率向量，360 个波长构成一个 360 行 7 列的光谱矩阵(验证集光谱矩阵，记为 $X_Y=(b_{ij})$)。其中，元素 b_{ij}表示验证集光谱数据中第 i 个波长处第 j 个样本的分辨率。

将校正集蚕丝样本的光谱数据与蚕丝含量占比用于建立校正模型，并将验证集数据用于检测模型预测精度。

2.2　建立 multi-PLSR 模型

采用校正集蚕丝光谱矩阵和蚕丝含量占比分别建立了 PLSR 和 multi-PLSR 预测模型。然后将验证集数据代入这两个模型，并比较了它们的预测效果。将相关系数和平均相对误差作为预测模型评价指标。建立 multi-PLSR 模型的具体步骤如下：

(1)第一步，依次建立偏最小二乘回归子模型。在建立子模型前，需要先准备各个子模型所需的光谱矩阵。以文献[6]中建立子模型时对光谱数据的处理方法为指导，将校正集光谱矩阵 X_J均匀分成 10 个子光谱矩阵，并将其记为 X_{J1}～X_{J10}。对验证集的光谱矩阵 X_Y 作同样的处理，得到 10 个子光谱矩阵 X_{Y1}～X_{Y10}。具体的操作在 MATLAB 软件中实现。将得到的 10 对子光谱矩阵分别用于建立 10 个偏最小二乘回归子模型。

以建立第一个偏最小二乘回归子模型为例，详细给出建模过程。类似地，可以建立其余 9 个子模型。利用第一个子模型的校正集光谱矩阵 X_{J1}、验证集光谱矩阵 X_{Y1}以及校正集与验证集的蚕丝含量占比向量 Y_J与 Y_Y来建立第一个偏最小二乘回归子模型。算法公式如下：

$$X_{J1}=T\cdot P^{\mathrm{T}}+E_1 \tag{1}$$

$$Y_J=U\cdot Q^{\mathrm{T}}+E_2 \tag{2}$$

$$U=T\cdot B \tag{3}$$

$$X_{Y1}=T^*\cdot P^{\mathrm{T}} \tag{4}$$

$$U^*=T^*\cdot B \tag{5}$$

$$Y_{Y1}^*\approx U^*\cdot Q^{\mathrm{T}}=T^*\cdot B\cdot Q^{\mathrm{T}}=X_{Y1}\cdot P\cdot B\cdot Q^{\mathrm{T}} \tag{6}$$

式中，T 和 U 分别为 X_{J1}和 Y_J的得分矩阵；P 和 Q 分别为 X_{J1}和 Y_J的载荷矩阵；E_1 和 E_2 分别为 X_{J1}和 Y_J在提取主成分时所得的残差矩阵[8]；B 为回归系数矩阵；T^* 为 X_{Y1}的得分矩阵；U^* 为验证集蚕丝含量占比预测值 Y_{Y1}^*的得分矩阵。利用 X_{Y1}、P、Q 以及 B 就可求出 Y_{Y1}^*。

对 X_{J1}和 Y_J进行偏最小二乘回归处理。经过上述计算得到 Y_{Y1}^*，然后利用式(7)计算出预测值与实际值之间的相关系数 R_1。接着采用类似方法计算出 Y_{Y2}^*～Y_{Y10}^*以及相关系数 R_2～R_{10}。

$$R_1=\frac{7\sum_{i=1}^{7}Y_Y(i)Y_{Y1}^*(i)-\sum_{i=1}^{7}Y_Y(i)\sum_{i=1}^{7}Y_{Y1}^*(i)}{\sqrt{7\sum_{i=1}^{7}Y_Y(i)^2-(\sum_{i=1}^{7}Y_Y(i))^2}\sqrt{7\sum_{i=1}^{7}Y_{Y1}^*(i)^2-(\sum_{i=1}^{7}Y_{Y1}^*(i))^2}} \tag{7}$$

(2)第二步，子模型加权。按照式(8)对 10

个子模型分别得到的验证集蚕丝含量占比的预测值进行加权平均，获得采用 multi-PLSR 方法时蚕丝含量占比的最终预测值。

$$Y_Y^* = \sum_{k=1}^{10} q_k Y_{Yk}^* \tag{8}$$

第 k 个子模型权重系数 q_k 的计算公式为

$$q_k = \frac{(1-R_k^2)^{-20}}{\sum_{j=1}^{10}(1-R_j^2)^{-20}} \tag{9}$$

(3)第三步，评估验证集模型的预测效果。采用 Y_Y^* 与 Y_Y 之间的相关系数和平均相对误差两个指标来评估 multi-PLSR 方法所建模型的预测效果。平均相对误差的计算公式为

$$ARE = \frac{1}{7}\sum_{i=1}^{7}\frac{|Y_Y^*(i)-Y_Y(i)|}{Y_Y(i)} \tag{10}$$

3 结果与分析

3.1 校正集与验证集样本的蚕丝含量占比分析

表 1 列出了校正集与验证集样本分布数据。可以看出，类型 2 蚕丝校正集样本的蚕丝含量占比值的分布范围比验证集样本更大，而标准差大致相同，说明分组后的校正集样本具有一定的代表性。

3.2 PLSR 与 multi-PLSR 模型的预测效果对比

对于类型 2 的蚕丝样本，分别采用 PLSR 和 multi-PLSR 方法建立了蚕丝含量占比的校正模型，并将验证集数据代入该模型。表 2 对比了两种建模方法的预测效果。可以

表 1 校正集与验证集样本分布

蚕丝含量占比	平均值	最大值	最小值	标准差
校正集	0.4837	0.9592	0.0412	0.2826
验证集	0.5627	0.9091	0.2037	0.2550

表 2 两种建模方法的预测效果对比

建模方法	主成分数	相关系数	平均相对误差
PLSR	6	0.815	0.3211
	7	0.908	0.2895
	8	0.8982	0.3448
	9	0.8822	0.3436
	10	0.8981	0.2633
	11	0.8908	0.2694
	12	0.8906	0.2697
	13	0.594	0.4866
	14	0.5833	0.4956
	15	0.5437	0.496
multi-PLSR	6	0.921	0.2599
	7	0.9636	0.1649
	8	0.9653	0.1342
	9	0.911	0.3031
	10	0.9727	0.2712
	11	0.9722	0.3176
	12	0.966	0.2703
	13	0.9784	0.1384
	14	0.9779	0.1478
	15	0.8225	0.2015

看出，当主成分数取 6～15 时，PLSR 模型的相关系数在 0.5437～0.9080 范围内，平均相对误差在 0.2633～0.496 范围内；而 multi-PLSR 模型的相关系数在 0.8225～0.9784 范围内，平均相对误差在 0.1342～0.3176 范围内。这说明两个模型的预测效果都比较好。注意到采用不同的主成分数时，multi-PLSR 模型的相关系数都比 PLSR 模型高，而平均相对误差都比 PLSR 模型小，说明 multi-PLSR 模型的预测精度高于 PLSR 模型。当主成分数为 13 时，multi-PLSR 模型的相关系数为 0.9784(达到最高)，模型预测效果最好。

采用 multi-PLSR 模型且主成分数为 13 时，蚕丝含量占比的实际值与预测值如表 3 和图 1 所示。

表 3 蚕丝含量占比的实际值与预测值

序号	实际值	预测值
42	0.2037	0.335
43	0.3402	0.3577
44	0.4455	0.4344
45	0.5446	0.5659
46	0.68	0.5942
47	0.8155	0.7936
48	0.9091	0.8586

3.3　5 种类型蚕丝样本的预测结果对比

对于其他 4 种类型的蚕丝样本，使用相同方法来划分校正集与验证集，并采用 PLSR 和 multi-PLSR 方法分别进行建模，得到相应蚕丝含量占比的预测值。主成分数取 13。表 4 列出了两种方法所建模型对 5 种类型蚕丝样本的预测数据。

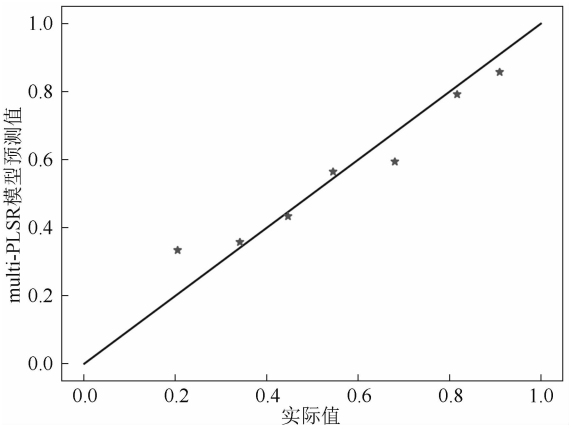

图 1 蚕丝含量占比的实际值与预测值的散点图

由表 4 可知，在预测 5 种不同类型的蚕丝含量占比时，multi-PLSR 方法对类型 2 样本的预测效果提高最为明显，相关系数由 0.594 增至 0.9784，平均相对误差由 0.4866 降至 0.1384；类型 1、3 和 5 蚕丝样本的 multi-PLSR 模型的相关系数都更高，平均相对误差都更小，模型预测效果都得到了提高。结果表明，在预测蚕丝含量占比时，提出的 multi-PLSR 建模方法比 PLSR 法有效且预测效果更好。

4　结束语

在预测 5 种不同类型的蚕丝含量占比时，提出的 multi-PLSR 建模方法比传统 PLSR 方法的相关系数高、平均相对误差小，并且预测效果更好。以类型 2 的样本为例，当主成分数为 13 时，multi-PLSR 模型的预测效果最好。与 PLSR 模型相比，其相关系数由 0.594 增至 0.9784，平均相对误差由 0.4866 降至 0.1384。

表 4 两种模型对 5 种类型蛋白样本的预测效果

项目	PLSR		multi-PLSR	
	相关系数	平均相对误差	相关系数	平均相对误差
类型 1	0.9038	0.1750	0.9050	0.1726
类型 2	0.594	0.4866	0.9784	0.1384
类型 3	0.7705	0.8846	0.924	0.3449
类型 4	0.6024	0.5005	0.7697	0.7244
类型 5	0.7286	0.8353	0.9003	0.3114

multi-PLSR 方法将全部光谱数据用于建模，不会丢失数据中的有效信息，并能够利用 PLSR 方法可提取光谱数据矩阵主成分的优势，提高了模型的运算速度与预测精度。multi-PLSR 方法可以实现对蚕丝含量占比值的快速有效预测。这意味着本文为建立蚕丝含量的近红外光谱预测模型提供了一种新的有效方法。该方法或许有望用于其他材料成分含量的近红外光谱快速预测分析。

致谢

感谢张勇老师、姚胜博士对本文工作提供的帮助！文中所用数据来源于浙江理工大学材料与纺织学院，在此致以感谢！

参考文献

[1] 赵明慧，胡广，刘洋. 浅析近红外光谱分析技术在聚酯纤维/氨纶面料成分检测中的应用 [J]. **中国纤检**，2019，**40**(7)：70-72.

[2] 买书魁，杨洋，赵小波，等. 基于 NIR 的白酒酿酒高粱中关键指标的定量分析 [J]. **食品科技**，2019，**44**(2)：301-307.

[3] 沈乐丞，刘书航，邓海玲，等. 近红外光谱结合偏最小二乘法快速测定糖果中水分含量 [J]. **食品工业科技**，2018，**39**(7)：255-258.

[4] 刘胜. 相思树酸溶木素含量近红外光谱分析的新方法 [J]. **光谱学与光谱分析**，2014，**34**(1)：69-72.

[5] 刘栋梁. 近红外光谱法建立欧美杨木质素和戊聚糖含量数学模型 [D]. 北京：北京林业大学，2013.

[6] 范雅婷，刘胜. 用多模型方法预测相思树苯醇抽提物含量 [J]. **中国农业科技导报**，2017，**19**(2)：131-138.

[7] 刘胜，范雅婷. 基于近红外光谱分析的多模型建模方法研究 [J]. **林业科技**，2014，**39**(2)：20-24.

[8] 范雅婷. 毛白杨与相思树的近红外光谱分析数学模型 [D]. 北京：北京林业大学，2016.