第42卷，第7期 光 谱 学 与 光 谱 分 析 Vol. 42，No. 7，pp2025-2032
2022年7月 Spectroscopy and Spectral Analysis July，2022

基于NIR和SOM-RBF网络的兰州百合关键营养物质定量分析方法

廉小亲[1,2]，陈 群[1,2]，汤燊淼[1,2]，吴静珠[1,2]，吴叶兰[1,2]，高 超[1,2]

1. 北京工商大学人工智能学院，北京 100048
2. 北京工商大学中国轻工业工业互联网与大数据重点实验室，北京 100048

摘 要 为了实现兰州百合关键营养物质蛋白质和多糖的快速无损检测，在12 000～4 000 cm^{-1}光谱范围内采集了59份兰州百合粉的近红外光谱(NIRS)。首先运用SG、Normalize、SNV、MSC、Detrend、OSC、SG+1D、SG+Normalize、SG+SNV和SG+Detrend十种预处理方法对原始光谱数据进行处理，确定蛋白质的最佳预处理方法为SG+Detrend、多糖的最佳预处理方法为Detrend；然后运用CARS、SPA和PCA三种算法对预处理的光谱数据进行特征波长筛选，确定蛋白质和多糖的最佳特征波长提取方法均为SPA算法；最后采用PLSR法建立了兰州百合关键营养物质蛋白质和多糖含量的预测模型，结果显示，经过SG+Detrend_SPA处理所建立的蛋白质PLSR模型中，预测集相关系数R_p为0.810 6，预测集均方根误差RMSEP为1.195 3；经过Detrend_SPA处理所建立的多糖PLSR模型中，预测集相关系数R_p为0.810 9，预测集均方根误差RMSEP为2.0946。考虑到经典PLSR无损预测模型精度的限制，在该研究中提出SOM-RBF神经网络无损预测模型。首先利用SOM网络对数据样本进行聚类，然后将得到的聚类类别数和聚类中心作为RBF网络的隐层节点个数和隐层节点数据中心，以此来优化RBF的结构参数。在建立的蛋白质SOM-RBF神经网络模型中，预测集相关系数R_p为0.866 6，预测集均方根误差RMSEP为1.038 5；建立的多糖SOM-RBF神经网络模型中，预测集相关系数R_p为0.868 1，预测集均方根误差RMSEP为1.799 4。比较PLSR和SOM-RBF两种模型对两种物质的预测结果，确定了SOM-RBF神经网络模型为最优建模方法，最终确定在蛋白质检测中，最优模型为基于SG+Detrend_SPA_SOM-RBF建立的模型，模型的预测集相关系数较PLSR高5.6%，预测集均方根误差较PLSR低0.156 8；在多糖检测中，确定的最优模型为基于Detrend_SPA_SOM-RBF建立的模型，模型的预测集相关系数较PLSR高5.72%，预测集均方根误差较PLSR低0.295 2。研究结果表明，运用NIR和SOM-RBF技术可以实现对兰州百合关键营养物质蛋白质和多糖的快速无损检测，为今后快速无损检测兰州百合营养物质提供理论依据。

关键词 兰州百合；蛋白质；多糖；近红外光谱；无损检测；SOM-RBF神经网络

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)07-2025-08

引 言

兰州百合作为甘肃省兰州市特产，是中国国家地理标志产品。其为多年生麟茎类草本植物，茎块由数十瓣鳞片相叠抱合，有百片合成之意而得名。具有很高的食用、药用和观赏价值，是我国卫生部首批审批通过的药食两用植物之一。研究表明，兰州百合的地下麟茎的糖分和蛋白质含量要明显高于其他产地和品种的百合，这两种生物活性物质最能代表兰州百合的营养价值[1]。

目前常见兰州百合蛋白质和多糖的检测方法有比色测定法和凯氏定氮法，这些方法虽然精确，但缺点是样本的检测耗时长、操作复杂繁琐、对样本具有破坏性且需要专业的技术人员，因此需要一种快速无损、便捷有效的检测方法[2-3]。近红外光谱(NIRS)技术由于具有快速、精确、高灵敏度和绿色无污染等优点，已经成功应用于食品追溯、石油检测、农业产品鉴别等多方面的研究[4]。近些年来，国内外学者建立了基于近红外光谱的灵芝、大米、藜麦、马铃薯以及兰州百

收稿日期：2021-02-08，**修订日期**：2021-04-02

基金项目：国家自然科学基金项目(61807001)和北京工商大学研究生培养-研究生教育质量提升计划项目(19008020144)资助

作者简介：廉小亲，女，1967年生，北京工商大学人工智能学院教授 e-mail：lianxq@263.net

合等作物品质的评价模型，用以快速实现蛋白质和多糖等含量的检测。在对各类产品蛋白质和多糖检测方面，赖长江生等[5]利用近红外分析方法构建了灵芝多糖含量快速预测模型，结果显示在 5 折交互检验优化参数下，最优模型的测试集相关系数为 0.851 6，验证均方根差为 0.023 6；李路等[6]采用近红外光谱技术对大米蛋白质和总糖等物质进行了检测，结果表明采用 BP 网络方法对蛋白质的预测精度为 91.2%，采用 PLS 法对总糖的预测精度为 91.89%；石振兴等[7]开展了藜麦的蛋白质等物质的近红外反射光谱预测建模研究，结果表明采用最佳预处理方法 FD+MSC 处理后所建立的蛋白质模型预测精度为 95.88%；蒙庆琰等[8]基于近红外光谱实现马铃薯蛋白质的无损检测，结果表明在最优预处理方法 MSC 下，采用 PLSR 系数法进行特征波长提取，最终所建立的蛋白质检测模型预测精度为 97.79%；廉小亲等[9]基于近红外利用 OSC 最佳预处理建立蛋白质和多糖的 PLS 模型，结果显示，蛋白质和多糖 PLS 模型 R_p 分别为 0.924 和 0.920，RMSEP 分别为 0.878 和 1.898。以上研究表明，近红外光谱技术结合不同的预处理和特征波长提取方法在物质的蛋白质和多糖检测方面具有较好的效果。

本研究采用近红外光谱技术采集 12 000～4 000 cm^{-1}波数范围内的兰州百合近红外光谱信息，通过光谱预处理和特征波长提取，减少光谱信息中的冗余信息，提升光谱信息的有效性。在研究分析兰州百合关键营养物质蛋白质和多糖方面，通过利用百合光谱特征波长建立的 PLSR 无损检测模型，发现其预测精度未能达到预期的 85%以上，且预测均方根误差未低于 2.0，这在实际应用中是不可取的。

虽然在参数预测领域 BP 网络应用的较多，但由于 BP 网络自身存在易形成局部最小、训练次数多、收敛速度慢等缺陷，因此其应用受到一定限制。而 RBF 由于其自身具有较强的泛化、信息处理及非线性映射等能力，且算法简单，分析能力很强，在处理非线性问题方面有较为广泛的应用。但对于 RBF，其隐层节点的数目、隐层径向基函数的中心和宽度却难以确定，这就需要一种方法来确定上述参数。SOM 以其清晰的聚类原则和简单的结构设计正好弥补了 RBF 自身的缺点[10]。因此利用基于 SOM 改进的 RBF 网络，来预测兰州百合关键营养物质蛋白质和多糖的含量。该方法在目前的近红外定量模型构建和兰州百合的关键营养物质快速预测领域还鲜有报道。SOM-RBF 法可以实现兰州百合蛋白质和多糖含量的快速预测，为研发兰州百合关键营养物质快速无损检测设备提供了技术支撑。

1 实验部分

1.1 材料

新鲜百合可食用的部分进行鳞片选料、清洗、焯水、烘干等步骤，粉碎制成百合粉(过 40 目筛)，共计获得 59 个样本，分别装入袋中进行编号，放置在 20 ℃室温环境中，等待采集近红外光谱信息。

1.2 仪器

光谱采集用仪器为 BRUKER 公司生产的 Vertex70 型傅里叶变换近红外光谱仪，其外观如图 1 所示。

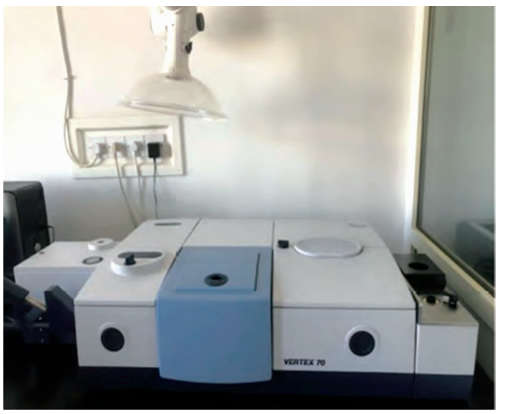

图 1 VERTEX70 型傅里叶变换近红外光谱仪

Fig. 1 Vertex70 Fourier transform infrared spectrometer

1.3 近红外光谱采集

光谱仪的扫描波长范围为 12 000～4 000 cm^{-1}，分辨率为 8 cm^{-1}，扫描次数为 64 次[11]。将百合样本置于石英杯，采用大样品杯旋转采样方式进行光谱扫描。通过重新加载将每个样品连续扫描 3 次，并将获得的 3 条光谱曲线的平均值作为兰州百合的最终光谱数据。

1.4 兰州百合蛋白质和多糖含量测定结果

根据国标法 GB/T 5009.5—2016《食品安全国家标准 食品中蛋白质的测定》测定百合粉中蛋白质的含量，根据国标法 DB12/T 884—2019《百合鳞茎中多糖的含量测定 紫外/可见分光光度法》测定百合粉中多糖的含量，实验重复 3 次取平均值。采用 K-S 算法按照 3∶1 原则将样本划分为建模集和预测集，其中 45 个样本用于定量建模，14 个样本用于模型预测。兰州百合样本的基本统计值如表 1 所示。

表 1 兰州百合样本基本统计表

Table 1 Basic statistics of Lanzhou lily samples

营养物质	样本集	样本个数	最小值 g/100 g	最大值 g/100 g	平均值 g/100 g	标准偏差
蛋白质	建模集	45	4.93	14.20	9.42	1.86
	预测集	14	4.93	12.10	9.61	1.82
多糖	建模集	45	16.78	32.57	21.86	3.16
	预测集	14	17.35	29.65	21.64	3.57

1.5 光谱数据预处理与建模分析

为了减少背景环境噪声、基线漂移和样本不均匀等影响，需要对原始光谱进行预处理[12]。使用常用的光谱预处理方法包括卷积平滑(S-G)、归一化(Normalize)、标准正态变换(SNV)、多元散射校正(MSC)、去趋势(Detrend)、正交信号校正(OSC)以及这几种方法与一阶微分(FD)的组合方法对原始光谱进行预处理，以选出最合适的预处理方法。进一步采用竞争性自适应重加权采样(CARS)法、连续投影法[13-15](SPA)和主成分分析(PCA)分别进行特征波长筛选，去除全光谱中冗余和无用的信息，降低模型复杂度，提高建模效率。最后应用 PLSR 和 SOM-RBF 网络法分别对兰州百

合关键营养物质蛋白质和多糖的含量进行建模分析。

1.6　模型评价

在确定最佳预处理方法和最优特征波长提取方法中，通过建立 PLSR 模型，根据训练集相关系数(correlation coefficient of calibration，R_c)、训练集交叉验证相关系数(correlation coefficient of cross validation，R_v)、预测集相关系数(correlation coefficient of prediction，R_p)、建模均方根误差(root mean squared error of calibration，RMSEC)、预测均方根误差(root mean squared error of prediction，RMSEP)和交叉验证均方根误差(root mean squared error of cross calibration，RMSECV)等模型的评价系数确定最佳预处理方法和特征波长提取方法。相关系数 R 和均方根误差 RMSEC 的计算如式(1)和式(2)所示

$$R=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

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

其中，x_i 为第 i 个样本预测的化学值，y_i 为第 i 个样本实测的化学值，$\bar{x}$ 为全部样本预测的化学值的平均值，$\bar{y}$ 为全部样本实测的化学值的平均值，$\hat{y}_i$ 为第 i 个样本预测值，n 为样本个数。

在建立 PLSR 和 SOM-RBF 网络模型中，根据 R_p 和 RMSEP，来评价模型的预测精度。评价依据为：当 R_p 越接近 1 且 RMSEC、RMSEP、RMSECV 越接近于 0，表明所建模型的预测精度越高[10]。兰州百合光谱数据的预处理工作在 The Unscrambler X 10.4 软件中进行，特征波长的筛选和模型的建立在 MATLAB R2017a 中进行。

2　结果与讨论

2.1　光谱数据预处理

将近红外光谱仪测得的兰州百合光谱数据导入软件 The Unscrambler X 10.4 中进行分析，得到兰州百合的近红外光谱原始图谱。由于原始光谱中不仅包含有用信息，还包含噪声信号，同时还可能存在基线平移和漂移等问题，为了消除这些干扰，对原始光谱采用 SG、Normalize、SNV、MSC、Detrend、OSC、SG＋1D、SG＋Normalize、SG＋SNV 和 SG＋Detrend 十种方法进行预处理。为了研究不同预处理方法对建模的影响，分别对全波段下预处理的光谱曲线建立蛋白质和多糖含量的 PLSR 模型，建模结果如表 2 和表 3 所示。

表 2　不同光谱预处理的兰州百合蛋白质的 PLSR 建模结果

Table 2　PLSR modeling results of Lanzhou lily protein using the spectra pretreated by different methods

处理参数	预处理方法	R_c	RMSEC	R_v	RMSECV	R_p	RMSEP
蛋白质	None	0.824 8	0.910 7	0.706 0	1.157 9	0.771 8	1.424 4
	SG	0.824 1	0.912 3	0.705 9	1.158 0	0.771 7	1.424 3
	Normalize	0.895 3	0.719 6	0.835 5	0.891 8	0.846 8	1.130 5
	SNV	0.864 2	0.850 7	0.748 4	1.134 6	0.848 6	0.992 5
	MSC	0.801 8	0.920 8	0.666 5	1.163 6	0.759 0	1.212 4
	Detrend	0.849 9	0.912 7	0.778 2	1.094 2	0.822 4	1.256 1
	OSC	0.996 0	0.166 9	0.870 5	0.917 0	0.346 2	1.890 6
	SG＋1D	0.965 0	0.461 7	0.672 2	1.307 8	0.778 1	1.181 2
	SG＋Normalize	0.808 6	1.046 3	0.693 2	1.294 7	0.796 9	1.221 9
	SG＋SNV	0.749 4	1.067 1	0.599 7	1.306 8	0.814 2	1.138 1
	SG＋Detrend	0.915 3	0.699 4	0.827 5	0.989 0	0.870 1	1.081 1

表 3　不同光谱预处理的兰州百合多糖的 PLSR 建模结果

Table 3　PLSR modeling results of Lanzhou lily polysaccharide using the spectra pretreated by different methods

处理参数	预处理方法	R_c	RMSEC	R_v	RMSECV	R_p	RMSEP
多糖	None	0.888 9	1.446 8	0.770 4	2.038 6	0.777 0	2.317 5
	SG	0.887 4	1.455 8	0.771 1	2.036 9	0.777 0	2.316 8
	Normalize	0.877 8	1.512 5	0.736 2	2.181 8	0.740 7	2.438 6
	SNV	0.926 7	1.186 7	0.826 1	1.787 5	0.949 6	1.623 2
	MSC	0.941 0	0.919 8	0.833 7	1.511 3	0.901 8	1.648 8
	Detrend	0.966 7	0.697 3	0.903 1	1.171 7	0.921 6	1.692 1
	OSC	0.999 3	0.117 0	0.966 6	0.809 8	0.835 9	2.131 1
	SG＋1D	0.953 6	0.941 7	0.518 3	2.696 6	0.452 2	1.241 4
	SG＋Normalize	0.917 9	1.225 2	0.838 8	1.690 2	0.632 1	2.819 4
	SG＋SNV	0.946 5	0.877 3	0.850 2	1.438 7	0.914 3	1.611 4
	SG＋Detrend	0.933 2	1.135 0	0.846 6	1.688 1	0.909 8	1.914 7

由表 2 和表 3 可知，不同的光谱预处理方法对模型的建立有明显的影响。蛋白质和多糖建立的 PLSR 模型中，蛋白质含量模型的最佳预处理方法为 SG＋Detrend，其 R_c＝0.915 3，RMSEC＝0.699 4，R_v＝0.827 5，RMSECV＝0.989 0，R_p＝0.870 1，RMSEP＝1.081 1；多糖含量模型的最佳预处理方法为 Detrend，其 R_c＝0.966 7，RMSEC＝0.697 3，R_v＝0.903 1，RMSECV＝1.171 7，R_p＝0.921 6，RMSEP＝1.692 1。因此，采用 SG＋Detrend 作为蛋白质含量模型的预处理方法，Detrend 作为多糖含量模型的预处理方法，后续特征波长以及预测模型的建立都是基于这两种最优预处理之上。59 组原始光谱经 SG＋Detrend 和 Detrend 预处理后得到的光谱如图 2 和图 3 所示。可以看到经 SG＋Detrend 和 Detrend 处理的光谱信息在保留原有光谱主要信息的同时，有效的消除了噪声和漂移的影响。

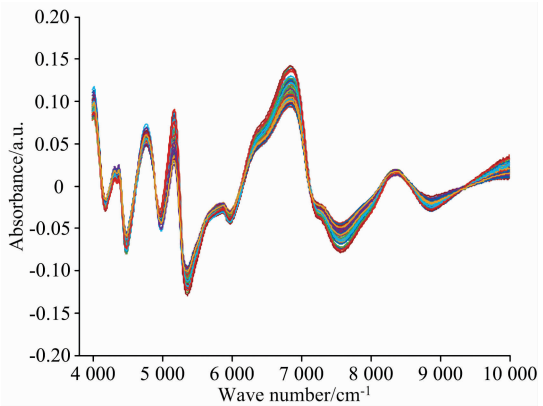

图 2 SG＋Detrend 方法处理的近红外光谱图

Fig. 2 NIR spectra of the SG＋Detrend method

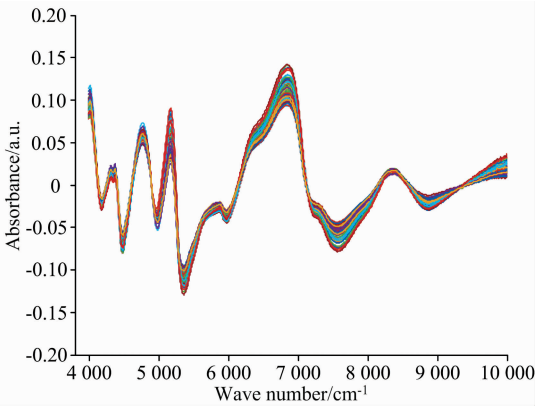

图 3 Detrend 方法处理的近红外光谱图

Fig. 3 NIR spectra of the Detrend method

2.2 波长的确定

近红外光谱主要由有机分子中含氢官能团的倍频和合频吸收峰组成，但是由于这些吸收峰强度低、灵敏性弱、吸收带较宽、重叠区较为严重，因此利用全波段建模会引入冗余信息和共线性变量[14]。通过对全波段的特征提取，不仅可以降低模型复杂度，提高模型的运行速度，节省时间，还可以用最少的有用信息表征全波段信息，从而提高模型的预测精度。

目前常用的特征波长提取方法有 CARS、SPA、PCA、无信息变量消除法(UVE)、随机蛙跳法(RF)等。CARS 是近些年来提出的新型特征信息筛选方法，该方法是通过自适应重加权采样技术筛选出 PLS 模型中回归系数绝对值大的波长点，去除权重小的波长点，交叉验证得到均方根误差 RMSECV 最小的最优波段子集。SPA 是一种前向循环选择方法，是从一个波长开始，每次循环计算它在未选入波长上的投影，将投影向量最大的波长引入到波长组合，直到循环 N 次，它通过选择含有最少冗余信息的波长变量组合以最小化信息重复叠加[15]。PCA 是用原数据在主元空间上的映射来表示原数据矩阵，因为在主元空间上可以用更少的量来表示，从而实现了数据的降维，消除随机噪声，保留了主要信息。利用 CARS、SPA 和 PCA 方法分别提取经过相同预处理的光谱特征波段，然后对所提取的特征波长建立 PLSR 模型，结果如表 4 所示。

表 4 基于 CARS、SPA 和 PCA 方法的 PLSR 建模结果对比

Table 4 Comparison of PLSR modeling results based on CARS, SPA and PCA methods

处理参数	预处理	特征提取法	特征波长数	R_c	RMSEC	R_v	RMSECV	R_p	RMSEP
蛋白质	SG＋Detrend	None	831	0.915 3	0.699 4	0.825 7	0.989 0	0.870 1	1.081 1
	SG＋Detrend	CARS	4	0.882 6	0.774 8	0.846 7	0.880 8	0.867 2	1.144 9
	SG＋Detrend	SPA	2	0.884 5	0.722 8	0.853 8	0.807 6	0.810 6	1.195 3
	SG＋Detrend	PCA	3	0.787 5	1.084 1	0.741 5	1.180 4	0.816 4	1.218 3
多糖	Detrend	None	831	0.966 7	0.697 3	0.903 1	1.171 7	0.921 6	1.692 1
	Detrend	CARS	13	0.929 1	1.000 1	0.874 4	1.329 3	0.885 1	1.668 1
	Detrend	SPA	14	0.967 8	0.632 7	0.920 3	0.988 6	0.810 9	2.094 6
	Detrend	PCA	3	0.737 2	1.896 2	0.501 5	2.487 7	0.225 3	4.402 6

由表 4 可知，与全波段的建模比较发现，经过特征提取之后所建立的蛋白质和多糖模型，预测集相关系数和预测均方根误差与全光谱建模相差并不大。但由于全波段的建模包含了更多的冗余信息，数据处理较慢，运行较为耗时，而特征波长提取的方法可有效的去除这些冗余信息，使模型的性能得到优化。通过比较 CARS、SPA 和 PCA 三种方法可以发现，SPA 无论是在相关系数，还是均方根误差均优于其他两种方法。对于蛋白质而言，R_c＝0.884 5，RMSEC＝0.722 8，

R_v=0.853 8，RMSECV=0.807 6，R_p=0.810 6，RMSEP=1.195 3，选择的特征波长数为 2 个；对于多糖而言，R_c=0.967 8，RMSEC=0.632 7，R_v=0.920 3，RMSECV=0.988 6，R_p=0.810 9，RMSEP=2.094 6，选择的特征波长数为 14 个。故选用 SPA 作为特征波长提取方法。

2.3　预测模型的建立与比较

利用所建立的 PLSR 和 SOM-RBF 模型预测测试集中的 14 份兰州百合样品的蛋白质和多糖含量，同时与用标准方法测定的蛋白质和多糖的标准理化值进行比较，以此验证兰州百合关键营养物质蛋白质和多糖预测模型的精度。

2.3.1　PLSR 预测模型的建立

首先利用 PLSR 分别建立 SG+Detrend_SPA_PLSR 蛋白质模型和 Detrend_SPA_PLSR 多糖模型。蛋白质和多糖含量的建模结果对比如表 4 所示，预测结果如图 4 和图 5 所示。蛋白质和多糖的相关系数 R 分别为 0.810 6 和 0.810 9。预测均方根误差 RMSEP 分别为 1.195 3 和 2.094 6。

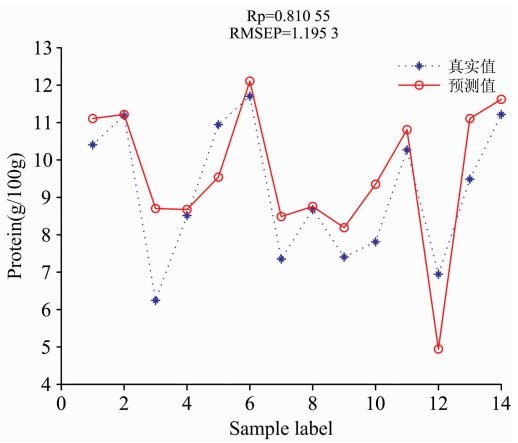

图 4　SG+Detrend_SPA_PLSR 法对蛋白质的预测结果图

Fig. 4　SG+Detrend_SPA_PLSR method for the prediction results of protein

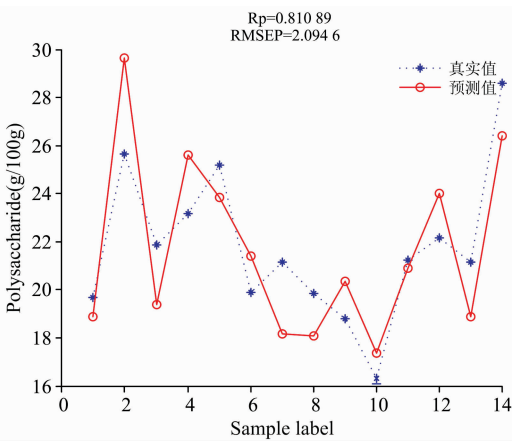

图 5　Detrend_SPA_PLSR 法对多糖的预测结果图

Fig. 5　Detrend_SPA_PLSR method for the prediction results of polysaccharide

2.3.2　SOM-RBF 预测模型的建立

利用 SOM 自组织聚类的特点以及 RBF 非线性逼近能力，设计 SOM-RBF 网络模型用于建立兰州百合的 SG+Detrend_SPA_SOM-RBF 蛋白质模型和 Detrend_SPA_SOM-RBF 多糖模型。模型建立主要分为两个步骤：

首先，利用 SOM 网络分别对提取特征波长后的蛋白质和多糖样本进行聚类训练，确定出聚类类别数，即聚类中心的个数。同时得到聚类中心向量，即各获胜神经元节点与样本输入节点相连的权值向量，亦称神经元获胜节点的内星权向量。然后计算各获胜神经元的内星权向量和映射到该获胜神经元的所有样本之间的欧式距离，将其中最小欧式距离判定为 SOM 网络聚类中心的半径。蛋白质和多糖的 SOM 神经网络结构拓扑图如图 6 所示。

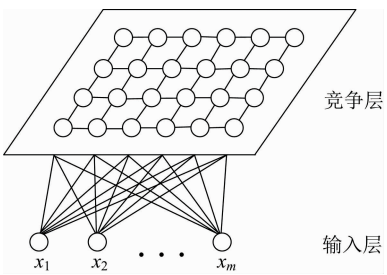

图 6　蛋白质和多糖的 SOM 神经网络结构

Fig. 6　SOM neural network structure of protein and polysaccharide

其次，待 SOM 网络训练完成后，将得到的聚类中心个数作为 RBF 网络隐层节点的个数；将得到的各聚类中心向量作为 RBF 网络隐层节点的中心向量；并将计算出的各聚类中心半径作为隐层节点中心的宽度。蛋白质和多糖的 RBF 网络结构拓扑图如图 7 所示。其中输入样本与 SOM 网络相同；RBF 网络隐层节点的径向基函数采用最常用的高斯函数；输出层节点采用线性函数；输出即为预测的蛋白质或多糖值。

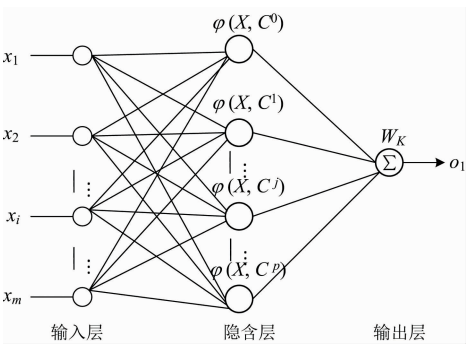

图 7　蛋白质和多糖的 RBF 网络拓扑结构

Fig. 7　RBF network topology of protein and polysaccharide

兰州百合的蛋白质和多糖预测算法流程图如图 8 所示。算法中 SOM 网络聚类类别数 g=4，即 RBF 网络隐层节点个数为 4；各聚类中心对应的连接权值向量 W_j^* 作为 RBF 网络隐层节点径向基函数的数据中心 c_j；SOM 网络聚类得到的

聚类半径 r_i 作为 RBF 隐层节点径向基函数数据中心的宽度 δ_j，W_j 为第 j 个神经元的内星权向量的初始值；η 为网络的学习效率；k_j 为第 j 个竞争层节点内星权向量的更新系数；当 $\min\{s_i\}>l$ 时，表示聚类结束，l 表示映射到 SOM 各聚类节点 s_i 最少样本数的阈值，考虑到训练样本为 45 组，本文取 $l=3$。

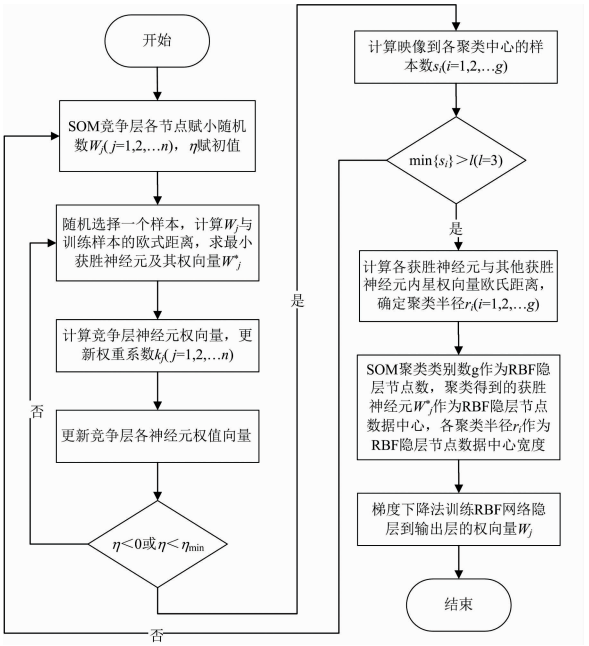

图 8 蛋白质和多糖的 SOM-RBF 预测算法流程图[10]

Fig. 8 SOM-RBF prediction algorithm flow chart for protein and polysaccharide[10]

SOM-RBF 神经网络对兰州百合关键物质蛋白质和多糖进行预测，得到的结果如图 9 和图 10 所示。蛋白质和多糖的相关系数 R 分别为 0.866 6 和 0.868 1。预测均方根误差 RMSEP 分别为 1.038 5 和 1.799 4。比较可得出，SOM-RBF 的预测能力要优于 PLSR。

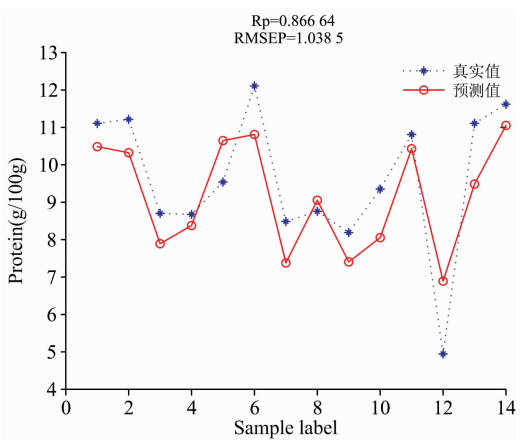

图 9 SOM-RBF 网络对蛋白质的预测结果图

Fig. 9 SOM-RBF network prediction results of protein

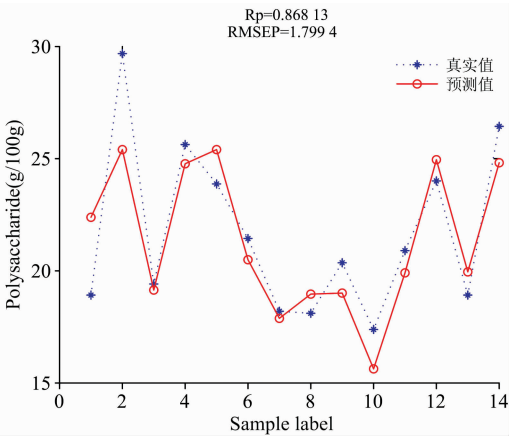

图 10 SOM-RBF 网络对多糖的预测结果图

Fig. 10 SOM-RBF network prediction results of polysaccharide

2.3.3 预测模型的比较

为了便于将经典的 PLSR 法和提出的 SOM-RBF 神经网络方法进行比较，将这两种预测模型的结果汇总如表 5 所示。

表 5 PLSR 和 SOM-RBF 预测模型建模结果比较

Table 5 Comparison of modeling results between PLSR and SOM-RBF prediction models

营养物质	预测方法	R_p	RMSEP
蛋白质	PLSR	0.810 6	1.195 3
	SOM-RBF	0.866 6	1.038 5
多糖	PLSR	0.810 9	2.094 6
	SOM-RBF	0.868 1	1.799 4

由表 5 中可得，对于蛋白质而言，采用 SOM-RBF 法要比 PLSR 法的 R_p 高出 5.6%、RMSEP 低 0.156 8；对于多糖而言，采用 SOM-RBF 法要比 PLSR 法的 R_p 高出 5.72%、RMSEP 低 0.295 2，因此提出的 SOM-RBF 神经网络预测模型在本研究中是可行的。

3 结 论

利用近红外光谱技术在 12 000～4 000 cm^{-1} 波段范围内对兰州百合关键营养物质蛋白质和多糖含量进行无损检测研究。建立了基于近红外光谱技术的蛋白质和多糖含量检测方法和模型。首先利用 10 种不同预处理方法对原光谱进行处理，确定出蛋白质最佳预处理方法为 SG＋Detrend，多糖最佳预处理方法为 Detrend；然后利用特征波长提取方法 SPA 提取原光谱的特征波长；最后利用 PLSR 法和 SOM-RBF 法构建了蛋白质和多糖含量的预测模型。通过比较得出，PLSR 法对蛋白质含量预测中 R_p 和 RMSEP 分别为 0.810 6 和 1.195 3；对多糖含量预测中 R_p 和 RMSEP 分别为0.810 9 和 2.094 6；SOM-RBF 法对蛋白质含量预测中 R_p 和 RMSEP

分别为 0.866 6 和 1.038 5；对多糖含量预测中 R_p 和 RMSEP 分别为 0.868 1 和 1.799 4。

实验结果表明，采用 SOM-RBF 预测模型具有较高的预测精度和较低的预测均方根误差，该方法可以实现对兰州百合内部关键营养物质蛋白质和多糖的快速无损预测，为快速检测兰州百合内部营养物质含量提供新思路。但由于目前对百合样本的获取需经过采集、风干、碾碎和营养物质的理化分析等一系列操作，蛋白质和多糖理化值的测量周期较长，因此所获得的样本数量较少，品种也不多，并且模型的精度有限，在后续的研究中将会尽力扩大样本数量并研究新模型以期获得更高的预测精度。

References

[1] Meng Xiuhua, Li Na, Zhu Hongtao, et al. Journal of Agricultural and Food Chemistry, 2019, 67(19): 5318.

[2] Hua Cuiping, Wang Yajun, Xie Zhongkui, et al. Sciences in Cold and Arid Regions, 2018, 10(2): 159.

[3] Yang Yue, Zhang Jing, Wang Jing, et al. Scientia Horticulturae, 2020, 261: 108928.

[4] LI Xin-xing, YAO Jiu-bin, CHENG Jian-hong, et al(李鑫星, 姚久彬, 成建红, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(1): 189.

[5] LAICHANG Jiang-sheng, ZHOU Rong-rong, YU Yi, et al(赖长江生, 周融融, 余　意, 等). China Journal of Chinese Materia Medica(中国中药杂志), 2018, 43(16): 3243.

[6] LI Lu, HUANG Han-ying, ZHAO Si-ming, et al(李　路, 黄汉英, 赵思明, 等). Journal of the Chinese Cereals and Oils Association(中国粮油学报), 2017, 32(7): 121.

[7] SHI Zheng-xing, ZHU Ying-ying, YANG Xiu-shi, et al(石振兴, 朱莹莹, 杨修仕, 等). Cereals & Oils(粮食与油脂), 2017, 30(12): 55.

[8] MENG Qing-yan, HE Jian-guo, LIU Gui-shan, et al(孟庆琰, 何建国, 刘贵珊, 等). Food Science and Technology(食品科技), 2016, 40(3): 287.

[9] LIAN Xiao-qin, TANG Shen-miao, WU Jing-zhu, et al(廉小亲, 汤燊淼, 吴静珠, 等). Food Science and Technology(食品科技), 2020, 45(7): 298.

[10] LIAN Xiao-qin, CHEN Qun, WANG Li-wei, et al(廉小亲, 陈　群, 王俐伟, 等). Computer Simulation(计算机仿真), 2020, 37(9): 194.

[11] GAO Tong, WU Jing-zhu, MAO Wen-hua, et al(高　彤, 吴静珠, 毛文华, 等). Transactions of the Chinese Society for Agricultural Machniery(农业机械学报), 2019, 50(S1): 399.

[12] MA Wen-qiang, ZHANG Man, LI Yuan, et al(马文强, 张　漫, 李　源, 等). Transactions of the Chinese Society for Agricultural Machniery(农业机械学报), 2019, 50(S1): 374.

[13] FU Dan-dan, WANG Qiao-hua(付丹丹, 王巧华). Food Science(食品科学), 2016, 37(22): 173.

[14] FENG Dou, CAI A-min, XUE Xiao, et al(冯　豆, 蔡阿敏, 薛　宵, 等). Chinese Journal of Animal Nutrition(动物营养学报), 2019, 31(1): 452.

[15] SUN Jun, JIANG Shu-ying, MAO Han-ping, et al(孙　俊, 蒋淑英, 毛罕平, 等). Transactions of the Chinese Society for Agricultural Machniery(农业机械学报), 2016, 47(1): 234.

Quantitative Analysis Method of Key Nutrients in Lanzhou Lily Based on NIR and SOM-RBF

LIAN Xiao-qin[1,2], CHEN Qun[1,2], TANG Shen-miao[1,2], WU Jing-zhu[1,2], WU Ye-lan[1,2], GAO Chao[1,2]

1. School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
2. China Light Industry Key Laboratory of Industrial Internet and Big Data, Beijing Technology and Business University, Beijing 100048, China

Abstract In order to realize the rapid and nondestructive detection of key nutrients protein and polysaccharide of Lanzhou lily, near infrared spectroscopy (NIRS) of 59 Lanzhou lily powder samples were collected in the range of 12 000～4 000 cm^{-1}. Firstly, ten pretreatment methods of SG, Normalize, SNV, MSC, Detrend, OSC, SG+1D, SG+Normalize, SG+SNV and SG+Detrend were used to process the original spectral data, and the optimal pretreatment method was SG+Detrend, Detrend was the best pretreatment method for polysaccharide. Then, CARS, SPA and PCA were used to screen the characteristic wavelength of the preprocessed spectral data. Finally, the SPA algorithm was used to determine the best extraction method for protein and polysaccharide's characteristic wavelength. The results showed that the correlation coefficient R_p of the prediction set was 0.810 6, and the root mean square error of the prediction set RMSEP was 1.195 3 in the protein PLSR model established by SG+Detrend_SPA treatment. In the polysaccharide PLSR model established by the Detrend SPA treatment, the correlation coefficient R_p of the prediction set was 0.810 9, and the root means square error RMSEP of the prediction set was 2.094 6. Considering the limitation of precision of the classical PLSR nondestructive prediction model, SOM-RBF neural network nondestructive prediction model is proposed in this paper. Firstly, the SOM network is used to cluster the data samples, and then the number of clustering categories and clustering center obtained is used as the number of hidden layer nodes and the data center of hidden layer nodes of the RBF network to optimize the structural parameters of RBF. In the established protein SOM-RBF neural network model, the correlation coefficient R_p of the prediction set is 0.866 6, and the root means square error of the prediction set RMSEP is 1.038 5. In the SOM-RBF neural network model established for polysaccharides, the correlation coefficient R_p of the prediction set was 0.868 1, and the root means square error RMSEP of the prediction set was 1.799 4. Comparing-PLSR and SOM-RBF prediction results, the SOM-RBF neural network model was determined as the optimal modeling method. Finally, the optimal model was established based on SG+Detrend_SPA_SOM-RBF in protein detection. The correlation coefficient of the prediction set of the model was 5.6% higher than that of PLSR, and the root means square error of the prediction set was 0.156 8 lower than that of PLSR. In the detection of polysaccharides, the optimal model was established based on Detrend_SPA_SOM-RBF, and the correlation coefficient of the model was 5.72% higher than that of PLSR, and the root means square error of the model was 0.295 2 lower than that of PLSR. The results showed that NIR and SOM-RBF techniques could be used for the rapid and non-destructive detection of key nutrients, proteins and polysaccharides, and the results could provide a theoretical basis for the future rapid and non-destructive detection of nutrients in Lily of Lanzhou.

Keywords Lanzhou lily; Protein; Polysaccharide; Near infrared spectroscopy; Nondestructive testing; SOM-RBF neural network

(Received Feb. 8, 2021; accepted Apr. 2, 2021)