第 32 卷第 4 期
2016 年 4 月

FOOD & MACHINERY 食品与机械

Vol. 32, No. 4
Apr. 2016

DOI:10.13652/j.issn.1003-5788.2016.04.021

基于支持向量机的野生蘑菇近红外识别模型

Study on recognition model for wild mushroom based on support vector machine of near infrared spectral diagnosis

刘 洋[1] 王 涛[1] 左月明[2]

LIU Yang[1] *WANG Tao*[1] *ZUO Yue-ming*[2]

(1. 乌兰察布职业学院机电技术系,内蒙古 集宁 012000;2. 山西农业大学工学院,山西 太谷 030801)

(1. *Electromechanical Technology Department of Wulanchabu Vocational College, Jining, Inner Mongolia* 012000, *China*; 2. *Engineering College of Shanxi Agricultural University, Taigu, Shanxi* 030801, *China*)

摘要:提出一种应用近红外光谱技术对野生蘑菇识别的新方法。使用 FieldSpec3 便携式近红外光谱仪对包括野生蘑菇在内的 13 种蘑菇进行漫反射光谱采集,通过对光谱的分析,得出 1 483,1 727,1 930,2 100,2 180,2 310 nm 为样品中多糖、蛋白质、脂肪的特征吸收峰,且野生蘑菇中 3 种成分的相关基团近红外特征吸收峰显著地高于其它蘑菇。将以上 6 处波长的吸光率作为输入变量,建立支持向量机识别模型,取 RBF 核,优化后惩罚因子 $C=2\ 048.0$,核参数 $\gamma=0.031\ 25$,该模型对未知样本正确识别率为 95.3%。结果表明,利用近红外光谱少量波长处的特征吸收差异,建立支持向量机识别模型可以很好地对野生蘑菇进行鉴别。

关键词:蘑菇;识别;光谱分析;支持向量机

Abstract: A new method for discrimination of wild Mushroom by means of near infrared spectroscopy (NIRS) was established. Visible and near infrared reflectance spectra of 13 different types of mushroom including 3 wild mushrooms were collected with a portable near infrared spectrometer (ASD Fieldspec3). Based on the analysis of spectrum, are the characteristics absorption peaks of polysaccharides, protein, fat of the mushroom are1 483, 1 727, 1 930, 2 100, 2 180, 2 310 nm, which of wild mushroom in are significantly higher than other mushrooms. With these 6 wavelengths of absorbance as input variables to establish recognition model of support vector machine, which takes RBF kernel, the parameters of $C = 2\ 048.0$, gamma $= 0.031\ 25$, the model for unknown sample recognition correct was 95.3%. Results show that, using the mushroom sample in a few wavelengths of near infrared spectral characteristics of absorption, establish support vector machine recognition model can be identified very well.

Keywords: mushroom; discrimination; near infrared spectroscopy (NIRS); support vector machine (SVM)

野生蘑菇富含人体所需的各种氨基酸和矿物质,堪称绿色、有机、保健食品[1],较人工种植蘑菇价格昂贵,以至于市场上频频出现人工栽培蘑菇冒充野生蘑菇的不法现象。长期以来对野生蘑菇的鉴定主要是依靠专业人员的感官经验,而常人难以鉴别野生蘑菇的真假,故有必要开发出既快速又可靠的鉴别手段和方法[2-5]。

近红外光谱技术通过有机化合物对不同光谱的吸收,再结合相关模型校正方法可以实现对物质的定性和定量分析,较常规检测技术有操作简单、效率高、成本低及无污染等特点,现已被广泛应用于食品分析检测中[6-7]。新鲜蘑菇中水分约占 90%,干物质主要成分为蛋白质、碳水化合物(主要为多糖)和脂肪[8],其中 C—H、O—H、N—H 和 S—H 等含氢基团,在近红外光区产生的振动谐波代表了蘑菇中有机成分的化学信息,决定了近红外光谱的吸收带[9]。多糖主要表现为 C—H 和 O—H 键的吸收,蛋白质分子主要是特征结构肽键(CONH)及其和肽键相关基团的吸收,脂肪主要体现在羧基中 C═O 键和不饱和脂肪酸分子链中—CH_2、—CH_3基的 C—H 键的吸收。如碳水化合物、脂类、蛋白质这三大类物质在不同种类蘑菇中的含量和结构有差异,再结合相关农产品品种鉴别的判别模型[10],就有可能通过这些成分在近红外光谱上的某个敏感吸收波长点或几个敏感吸收波长组合的光谱差异来区分不同种类的蘑菇。野生蘑菇的营养成分含量普遍要高于人工栽培品种[11-12],那么相关基团因敏感吸收产生的差异也应在近红外光谱上得到响应。本研究以 5 种野生蘑菇和 8 种人工栽培蘑菇为代表,试图分析出 13 种蘑菇化学成分中相关基团的近红外敏感吸收波长,将这些波长的吸光率做模型变量输入,建立定性识别模型,对蘑菇的品种识别问题加以研究,旨在寻找一种快速有效的野生蘑菇

作者简介:刘洋,女,乌兰察布职业学院副教授,博士。

通讯作者:左月明(1954-),男,山西农业大学教授,硕士。
E-mail:zuoyueming@163.com

收稿日期:2016-02-01

鉴别方法。

1 试验设计

1.1 试验设备

近红外光谱仪：FieldSpec3 型，美国 ASD(Analytical Spectral Device)公司。波长范围：350～2 500 nm；采样间隔：1.4 nm(区间 350～1 000 nm)，2 nm(区间 1 000～2 500 nm)；

数据分析软件为 MATLAB7.3；

高速中药粉碎机：QE-02 型，浙江屹立工贸有限公司。

1.2 样品来源及光谱获取

共 13 种干样样品：山西五台县产高、中、低档 3 种不同价格的野生台蘑；内蒙古根河大兴安岭林区产的两种野生蘑菇：榛蘑、黄油菇；8 种山西晋中产人工栽培蘑菇：人工台蘑、滑子菇、茶树菇、鸡腿菇、牛肝菌、花菇、杏鲍菇、香菇。分两次购进样品：第一次购进以上 13 种蘑菇后，每种样品各取 10 个样本，共 130 个样本作为建模集；第二次再购进与第一次所购不同批次的以上 13 种蘑菇，每种样品各取 5 个样本，共 65 个样本作为预测集。

13 种蘑菇粉碎后过 60 目筛，分别置于直径 90 mm 的培养皿，将样品表面刮平。测定前进行系统配置优化和白板校正，探头视场角为 10°，入射角 45°，光谱仪每次扫描时间为 0.1 s，输出的光谱线为 10 条原始扫描光谱自动平均所得，所需时间为 1.0 s。

1.3 支持向量机

支持向量机[13](Support Vector Machine，SVM)为 Corinna Cortes 和 Vapnik 于 1995 年提出，该方法建立在统计学习理论的 VC 维理论和结构风险最小化原理基础之上，在确定网络结构、过学习与欠学习、局部极小点等问题上较神经网络有较大改进，在解决小样本、非线性及高维模式识别中表现出许多特有的优势，能够推广应用到函数拟合等其他机器学习问题中。

SVM 在线性可分的情况下寻求最优分类面，该分类面不但能将两类无错误地分开，而且可以使两类的分类间隔最大。其原理为：

设有 n 个样本 x_i 及其所属类别 y_i：(x_i, y_i)，$x_i \in R^N$，$y_i \in \{1, -1\}(i=1,\cdots,n)$。

超平面方程 $W \cdot X + b = 0$，能将两类样品正确区分，并使分类间隔最大的优化问题可用在 $y_i[W \cdot X + b] \geqslant 0(i=1,\cdots,n)$ 的约束下求寻优目标函数 $\varphi(W)$ 的最小值，即：

$$\varphi(W) = \frac{1}{2}||\omega||^2 = \frac{1}{2}(\omega \cdot \omega)。 \quad (1)$$

考虑到一些样本有可能不被正确地分类，所以引入松弛因子 $\xi_i \geqslant 0(i=1,\cdots,n)$，以保证分类的准确性，这时它的约束条件变成为：$y_i[W \cdot X + b] - 1 + \xi_i \geqslant 0(i=1,\cdots,n)$。

这时寻优目标函数 $\varphi(W)$ 最小值变为：

$$\varphi(W) = \frac{1}{2}(\omega \cdot \omega) + C\sum_{i=1}^{n}\xi_i 。 \quad (2)$$

式(2)惩罚因子 C 起到对错分样本惩罚程度控制的作用，实现在错分样本的比例和算法复杂程度之间的“折衷”。对于非线性问题，若在原始空间中利用分类面不能获得满意的结果，这就需要通过非线性变换将非线性问题转化为线性问题，变换后再求最优分类面，将样本进行分类。SVM 经过几十年的理论准备和算法研究，被成功的应用到身份识别、图像识别、智能控制、故障诊断、光谱分析等诸多领域。

2 结果与分析

2.1 原始光谱

光谱仪所采集各样品在 350～2 500 nm 的吸收光谱图(见图 1)。由图 1 可知，13 种蘑菇近红外光谱的波形趋势基本一致，在 1 483，1727，1 930，2 100，2 180，2 310 nm 处出现较强吸收峰。在 350～1 400 nm 时，各样品谱线交叉较多，在 370 nm 处达到光谱最大吸收峰值后，降低趋势逐渐由急到缓，到 1 330 nm 处为最低。此后光谱曲线起伏较大，但总体呈上升趋势。到 1 400～2 500 nm 时，5 种野生蘑菇的各处光谱吸收普遍高于其他人工栽培蘑菇，为数学建模鉴别野生蘑菇提供了依据。

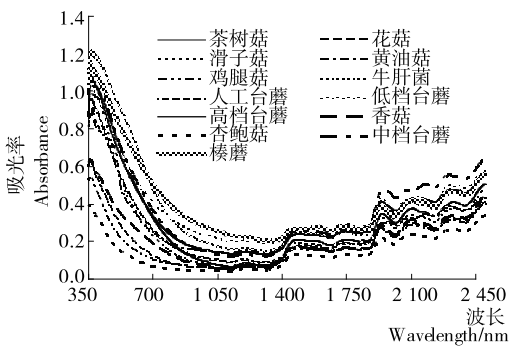

图 1 13 种蘑菇的可见光－近红外光谱图

Figure 1 Visible and near infrared spectrum of 13 types of mushroom

2.2 特征吸收波长分析

蘑菇中各主要成分所含基团在近红外漫反射光谱中有特定的吸收峰，基团不同，吸收峰位置也不同，峰高也因样品而异。结合相关文献[14]分析，样品中光谱特征吸收峰对应的分子结构见表 1。

敏感吸收波长的提取是建立在有机物化学成分敏感光谱的吸收率与该成分含量的相关关系基础上的，这些化学成分中相关基团的化学键在一定辐射水平的照射下发生振动，引起对某些波长的光谱吸收产生不同的吸收率差异，且该波长处光谱吸收率的变化对该化学成分的含量非常敏感。为了进一步比较野生蘑菇和人工栽培蘑菇样品中多糖、蛋白质、脂肪相关基团在近红外吸收上的差异，选取了分别代表多糖、蛋白质、脂肪 3 种物质特征吸收峰的 1 930，2 180，2 310 nm处吸光率进行了差异显著性分析，结果见表 2。由表 2 可知：3 处波长吸光率由大到小均为中档台蘑、低档台蘑、高档台蘑、榛蘑、黄油菇、牛肝菌、茶树菇、滑子菇、人工台蘑、香菇、花菇、鸡腿菇、杏鲍菇。其中，中档台蘑与其他12 种蘑菇差异显著($P<0.01$)；低档台蘑与全部 8 种人工栽培蘑菇差异显著($P<0.01$)，但与高档台蘑、榛蘑在 $P=0.01$ 水平

表 1　样品中光谱特征吸收峰与分子结构对照表†

Table 1　The relation of spectrum characteristic absorption peaks and molecular structure in the sample

波长/nm	官能团	光谱(结构)	物质类型
1 483	酰胺(·NH 或·NH_2)的 N—H	N—H(2ν),$CONH_2$	酰胺/蛋白质
1 727	亚甲基(·CH_2)的 C—H 反对称振动	亚甲基 C—H 反对称(2ν)	烃,脂肪烃
1 930	O—H(·O—H 和 HOH)	O—H 伸缩和 HOH 变形的组合频	多糖
2 100	聚合体(·O—H 和·C—O)O—H/C—O	O—H 伸缩和 C—O 伸缩组合频	多糖
2 180	蛋白质的 N—H	N—H(3δ)	蛋白质/氨基酸
2 310	C—H(·C—H 弯曲)	C—H(3δ)	油脂

†　2ν 表示伸缩振动基频的一级倍频谱带;3δ 弯曲振动基频的二级倍频普带。

表 2　13 种蘑菇在 3 处波长吸光率的 Duncan's 新复极差多重比较†

Table 2　Duncan's multiple-range test result for absorbance of 13 mushrooms in 3 wavelengths

样品	1 930 nm	2 180 nm	2 310 nm
中档台蘑	$0.469\ 7^{A}$	$0.508\ 1^{A}$	$0.556\ 5^{A}$
低档台蘑	$0.424\ 3^{B}$	$0.435\ 1^{B}$	$0.499\ 7^{B}$
高档台蘑	$0.411\ 7^{BC}$	$0.425\ 2^{BC}$	$0.475\ 5^{B}$
榛蘑	$0.389\ 6^{BCD}$	$0.417\ 9^{BC}$	$0.469\ 5^{BC}$
黄油菇	$0.374\ 6^{CD}$	$0.390\ 5^{CD}$	$0.457\ 8^{BC}$
牛肝菌	$0.356\ 5^{DE}$	$0.359\ 9^{DE}$	$0.429\ 1^{CD}$
茶树菇	$0.348\ 9^{DE}$	$0.357\ 5^{DE}$	$0.409\ 7^{DE}$
滑子菇	$0.326\ 3^{EF}$	$0.343\ 3^{EF}$	$0.383\ 7^{EF}$
人工台蘑	$0.300\ 0^{FG}$	$0.323\ 7^{EF}$	$0.373\ 5^{EF}$
香菇	$0.290\ 1^{FG}$	$0.305\ 3^{FG}$	$0.341\ 3^{FG}$
花菇	$0.272\ 6^{G}$	$0.278\ 5^{G}$	$0.318\ 3^{G}$
鸡腿菇	$0.266\ 7^{G}$	$0.274\ 0^{G}$	$0.313\ 7^{G}$
杏鲍菇	$0.221\ 2^{H}$	$0.229\ 1^{H}$	$0.266\ 3^{H}$

†　大写字母表示在 P=0.01 水平差异显著。

上差异不显著;高档台蘑与榛蘑、黄油菇在 P=0.01 水平上差异不显著,但与其他 8 种人工栽培蘑菇差异显著(P<0.01)。可见,对单一波长吸光率用常规统计方法处理后,不能将 13 种已知样品完全区分,这就需要引入其他数学方法对近红外光谱数据进行处理。

在可见光区 350～700 nm 处,光谱吸收应该受蘑菇自身色泽的影响较大。凭人的视觉感观,样品牛肝菌、榛蘑颜色最深,其吸收率在可见光区最高,杏鲍菇、鸡腿菇、花菇、香菇颜色较浅,其谱线在该区也最低。各样品谱线在 350～700 nm可见光区内交叉频繁,经 700～1 400 nm 的过渡调整,到 1 400 nm 后各谱线层次才稳定。由于可见光区域的光谱吸收不能反映野生蘑菇化学成分较高于其他蘑菇的特点,所以不宜从该区提取敏感吸收波长参与建模。另外,储存条件、时间等都可能影响蘑菇的颜色发生变化,若在这些条件下引用该区的有关敏感吸收波长建立识别模型,可能使模型精度下降,误判增加。

2.3　支持向量机模型的建立及预测结果

将建模集样品中多糖、蛋白质、脂肪的敏感吸收峰:1 483,1 727,1 930,2 100,2 180,2 310 nm 处的吸光率作为支持向量机识别模型的输入变量。经交叉验证的网格搜索法优化后(交叉验证精度 CV Rate=84.074 1%),得惩罚因子 $C=2\ 048.0$,核参数 $\gamma=0.031\ 25$,使得 SVM 分类器的分类效果达最优。训练后,每个类的 SVM 的个数为:4,2,6,3,5,4,8,3,8,5,7,2,5,支持向量总个数 62。所建模型对预测集 65 个样本的正确识别率为 95.3%。SVM 模型的核函数有多种,在本试验中选用 RBF 核函数,主要考虑到:① RBF 核函数可以将样本非线性地映射到高维的空间中,从而解决类标签和属性之间非线性关系的问题。② Sigmoid 核函数取某些特定参数时性能和 RBF 相近。③ 多项式核函数数目比 RBF 核函数数目多,模型选择上更为复杂。此外,试验还对不同敏感波段个数做 SVM 模型输入的识别正确率做了比较。当从 6 个波长中任选 1～5 个敏感吸收建模时,对 65 个未知样本的识别正确率最高分别为 52.3%,60.1%,75.6%,80.6%,85.7%,说明随着输入模型波长数的增加,反映相关基团敏感吸收的化学信息越充分,SVM 模型的识别正确率也相应提高。

3　结论

本研究对野生台蘑等 13 种蘑菇样品的近红外光谱信息进行了分析,得出 1 483,1 727,1 930,2 100,2 180,2 310 nm 是样品中多糖、蛋白质、脂肪相关基团的敏感吸收波长点。将这 6 处表征样品化学成分组成及含量的近红外特征吸收的吸光率作为支持向量机的输入,建立识别模型,对未知的 65 个样本识别正确率为 95.3%。结果表明,利用支持向量机建立的近红外蘑菇品种识别模型可以作为野生蘑菇的低成本、高可靠性的鉴别方法。

参考文献

[1] 陈巧玲,李忠海,陈素琼. 5 种地产食用菌氨基酸组成比较及营养评价[J]. 食品与机械,2014,30(6):43-46.

[2] 初洋,倪新江,姜海华,等. 侧耳属 3 种食用菌解剖学性状比较[J]. 中国食用菌,2010,29(2):9-11.

[3] 刘剑虹,刘刚,鼎珊,等. 几种牛肝菌显微结构的扫描电镜观察[J]. 电子显微学报,2007,26(1):74-77.

[4] 傅安涛,宋爱荣,田雪梅,等. 同工酶技术及其在我国食用菌研究中的应用[J]. 菌物研究,2006,4(4):57-61.

(下转第 112 页)

3.2.1 物料颗粒的功率消耗 由上述知，分离盘转速 $n=1\ 400$ r/min，则分离盘回转 1 周所需时间为 $\Delta t=3/70$ s，按最小生产率为 2 000 kg/h 计算得分离盘回转 1 周所需分离的物料质量为 $m_w=[(5/9)\cdot\Delta t]$ kg。待分离的物料颗粒是通过位于分离盘上方的进料斗缓慢落入到分离盘中心处的，即分离盘每回转 1 周，需要将质量为 m_w 的物料颗粒从初速度为 0 加速到绝对速度为 v_a，其动能为 $E_w=m_w\cdot {v_a}^2/2$。因此，需要电动机对其提供的最小功率 P_w[11] 11-32 为：

$$\begin{cases} P_w = T_w \cdot n/9\ 550; \\ T_w = E_w。 \end{cases} \tag{15}$$

3.2.2 分离盘和主轴的功率消耗 分离盘采用 1Cr18Ni9Ti 制造，壁厚为 5 mm；主轴采用 45 钢制造，直径为 50 mm，长度为 450 mm。由于分离盘和主轴做匀速转动，故其转距的计算公式[11] 12-20 为：

$$T_{pz}=m_{pz}\cdot g\cdot r_{pz}, \tag{16}$$

式中：

T_{pz}——负载作用分离盘和主轴上的转矩，N·m；

m_{pz}——分离盘和主轴总质量，kg；

r_{pz}——分离盘和主轴的等效半径，m。

根据机械产品做旋转运动时，各旋转零部件功率的计算公式，得出需要电动机对分离盘和主轴提供的最小功率 P_{pz}[11] 11-32 为：

$$P_{pz}=T_{pz}\cdot n/9\ 550, \tag{17}$$

式中：

P_{pz}——分离盘和主轴所需功率，kW；

n——分离盘和主轴转速，r/min。

根据式(15)～(17)计算得电动机的总功率为：

$P=P_w+P_{pz}=14.15$ kW。

考虑到一定的功率储备，查有关资料选用型号为 Y160L-4 的 Y 系列电动机，其功率为 15 kW，转速为 1 460 r/min，与要求转速 $n=1\ 400$ r/min 接近，故基本能够满足使用要求。

4 结论

(1) 本试验利用豌豆脱皮后豆胚与豆皮的假密度差异设计了一种离心式豌豆胚皮分离机，假密度较大的豆胚抛出较远而落入外面的接胚筒内，而假密度较小的豆皮则落入里面的接皮筒内。调节可调支座即可方便地改变接皮筒相对于分离盘的高度，进而改变落入接皮筒内物料的数量，达到改变豆胚中的含皮率和豆皮中的含胚率的目的，以适应不同企业、不同豌豆品种对分离率的不同要求。本研究解决了现有分离设备存在的由于豆胚与豆皮的真密度差异小、分离效果差和结构复杂、工作可靠性差等问题，完全能够满足规模淀粉企业的使用要求。

(2) 机理分析结果表明，影响豌豆胚皮分离效果的因素主要有豆胚和豆皮的假密度、分离盘锥角 α、分离盘角速度 ω 以及分离盘与混合物之间的摩擦系数 μ 等，因此对这些参数进行合理优化，即可得到较为理想的分离效果。

(3) 目前，本研究设计的离心式豌豆脱皮分离机还处于设计阶段，其处理量和分离效果有待制作样机后通过试验获得。

参考文献

[1] 崔再兴，李玲. 豌豆的特征特性及开发利用价值[J]. 杂粮作物，2010，30(2)：154-155.

[2] Ratnayake W S，Hoover R，Warkentin T. Pea starch：composition，structure and properties：a review[J]. Starch/Starke，2002(54)：217-234.

[3] Hoover R，Sosulski F W. Composition，structure，functionality，and chemical modification of legume starches[J]. Can. J. Physiol. Pharm.，1991(69)：79-92.

[4] Barron C，Buleon A，Colonna P，et al. Structural modifications of low hydrated pea starch subjected to high thermomechanical processing[J]. Carbohy-drate Polymer，2000(43)：171-181.

[5] 张志杉，孙传祝. 豆类脱皮机理及螺旋揉搓式豆类脱皮机设计研究[J]. 农机化研究，2013，35(9)：104-107.

[6] 潘光杰，孙传祝，张志杉，等. 揉搓式豌豆脱皮机研究与设计[J]. 食品与机械，2012，28(4)：62-66.

[7] 吴建章，朱永义. 气固流态化用于谷物风选的研究[J]. 粮食与饲料工业，2002(6)：11-13.

[8] 褚良银，陈文梅，李晓钟，等. 水力旋流器结构与分离性能研究(三)—锥段结构[J]. 化工装备技术，1998，19(5)：1-4.

[9] 薛立桥，孙传祝，潘光杰. 旋流式豌豆胚皮分离机研究与设计[J]. 农机化研究，2012，34(12)：132-135.

[10] 董焕俊，孙传祝，薛立桥. 复杂摆动式豌豆胚皮分离机设计[J]. 农机化研究，2013，35(12)：112-114，158.

[11] 闻邦椿. 机械设计手册[M]. 5 版. 北京：机械工业出版社，2010.

(上接第 94 页)

[5] 吕长武，吕杰，陈恒雷，等. RAPD 分子标记在食用菌研究中的应用[J]. 中国生物工程杂志，2006，26(1)：77-80.

[6] 吴晨，何建国，刘贵珊，等. 基于近红外高光谱成像技术的马铃薯干物质含量无损检测[J]. 食品与机械，2014，30(4)：133-136.

[7] 张令标，何建国，刘贵珊，等. 基于可见/近红外高光谱成像技术的番茄表面农药残留无损检测[J]. 食品与机械，2014，30(1)：82-85.

[8] 刘宏. 食用菌营养价值及开发利用[J]. 中国食物与营养，2007(12)：25-27.

[9] 严衍录，赵龙莲，韩东海，等. 近红外光谱分析基础与应用[M]. 北京：中国轻工业出版社，2005：31-40.

[10] 程文宇，管骁，刘静. 近红外光谱技术检测液态奶中微量三聚氢氨的可行性研究[J]. 食品与机械，2015，31(1)：71-74.

[11] 贺沛芳，杨怀民，张治家，等. 五台山野生食用菌资源营养价值及展望[J]. 中国食用菌，2010，29(3)：7-9.

[12] 史琦云，邵威平. 八种食用菌营养成分的测定与分析[J]. 甘肃农业大学学报，2003，38(3)：336-339.

[13] Burge C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery，1998，2(2)：121-169.

[14] Workman J，Weyer J L. Practical guide to interpretive near-infrared spectroscopy[M]. London：Taylor & Francis Group，2008：18-19.