研究报告

支持向量机结合近红外光谱法测定杉木木质素的含量

HUANG An-min

黄安民[1]，焦淑菲[2]，任海青[1]，相玉红[2]，张卓勇[2*]

（1. 中国林业科学研究院 木材工业研究所，北京 100091；
2. 首都师范大学 化学系，北京 100037）

摘 要： 采用支持向量机(SVM)结合近红外光谱(NIR)技术建立测定杉木中木质素的定量分析模型。以47个杉木样品作为实验材料，用常规方法测定了样品中木质素的含量，用近红外光谱仪采集相应的光谱，对光谱数据进行平滑、求导、小波压缩以及归一化，结合支持向量机，以径向基(RBF)作为核函数，建立了测定杉木中木质素含量的模型。校正相对误差的平方和为0.007 433，预测相对误差的平方和为0.001 219。结果表明，该方法测量比较准确，可以用于杉木中木质素含量的预测。

关键词： 支持向量机；近红外光谱；杉木；木质素

中图分类号：TQ 351.013　　**文献标识码：**A　　**文章编号：**0253－2417(2009)05－0001－05

Determination of Lignin Content of Chinese Fir Wood by Support Vector Machine and Near-infrared Spectrometry

HUANG An-min[1], JIAO Shu-fei[2], REN Hai-qing[1], XIANG Yu-hong[2], ZHANG Zhuo-yong[2]

(1. Research Institute of Wood Industry, Chinese Academy of Forestry, Beijing 100091, China;
2. Department of Chemistry, Capital Normal University, Beijing 100037, China)

Abstract: Via near-infrared spectroscopy (NIR) combined with support vector machine (SVM), a model for determining lignin content of Chinese fir wood was established. Forty-seven Chinese fir samples were used as experimental material. Lignin content of samples was measured according to wet-chemical method. The spectra of Chinese fir samples were recorded by near infrared spectrometer. The spectral data were pretreated by smoothing, derivating, compressing and scaling. Radial basis function (RBF) was used as kernel function to establish a model for determing lignin content. The square sum of relative calibration error and relative prediction error were 0.007 433 and 0.001 219, respectively, which demonstrated that this method is precise for determination of lignin content of Chinese fir wood.

Key words: support vector machine; near-infrared spectroscopy; Chinese fir; lignin

木材的主要成分木质素与木材的其他性质以及木材的加工利用密切相关。在造纸工业中，木质素的含量是制定蒸煮和漂白工艺条件的重要依据。快速、准确、低成本地评价木材中木质素的含量成为充分、合理利用木材资源的一个重要研究内容。测量造纸原料中木质素含量的方法主要是依据国标GB/T 2677.8－1994。该方法操作手续繁琐，成本高，无法实现大批量样品的快速测定。近红外光谱技术的主

收稿日期：2008－10－23
基金项目：北京市自然科学基金资助项目(6092021)；“十一五”国家科技支撑计划资助(2006BAD03A15)；北京市教育委员会科技发展项目(KM200710028009)
作者简介：黄安民(1975－)，男，河北邢台人，副研究员，博士，主要研究方向为木材及制品性质分析与评价；
E-mail：hbham2000@sina.com
*通讯作者：张卓勇，教授，博士，博士生导师，研究领域为光谱分析；**E-mail：**gusto2008@vip.sina.com。

要特点是分析速度快、可同时测定样品的多个组成或性质、测量过程简便且不会产生二次污染、便于实现在线监测,已在农业、食品、医药、石油化工等行业中得到广泛的应用[1-3]。近年来也在木材科学研究方面得到了越来越多的重视和运用。有关的研究表明,近红外光谱和纤维素、淀粉、木质素以及其他碳水化合物,如葡萄糖、果糖等的含量都具有很强的相关性[4],同时,近红外光谱还含有大量的物质结构方面的信息。国外将偏最小二乘回归[5-6]、偏最小二乘[7]、近红外光谱[8-14]等技术应用到木材的化学性质[5,8-9]、物理性质[6-7,10-13]、力学性质[8,14]等方面研究,并取得了一系列成果。中国林业科学研究院已有预测木材密度[15]、水分[16]、纤维素结晶度[17]、腐朽类型[18],以及毛竹中综纤维素含量[19]等的研究报道,他们主要采用的是偏最小二乘法[15,17,19]结合近红外光谱的方法。由于近红外光谱吸收峰重叠严重,因此在进行定性和定量分析中必须使用化学计量学技术与计算机数据处理来提取相关信息。小波变换是在傅里叶变换基础上发展起来的数学方法,它可以对近红外光谱数据压缩,以提高计算速度。支持向量机(SVM)是 Vapnik 等根据统计学理论提出的一种专门研究小样本情况下统计估计和预测的问题,探索在有限样本的情况下如何得到最优解的通用学习方法。它是建立在统计学理论的 VC 维理论和结构风险最小原理基础上的,体现了兼顾经验风险和置信范围的一种折中的思想,能够较好地解决小样本、非线性、高维数和局部极小点等众多实际问题[20-21]。本研究利用近红外漫反射光谱技术,运用支持向量回归算法建立数学模型,建立了杉木中木质素的测定方法。该方法简便、可靠,具有较大的实际应用价值。

1 实验部分

1.1 支持向量回归(SVR)的原理

支持向量机是 Vapnik 等首先提出并应用的一种新型学习算法[22]。最初,该算法是用来解决模式识别中的二分类问题,在引入 Vapnik 提出的 ε 不敏感损失函数后,支持向量机也可用来解决非线性的回归问题。对于回归建模问题,传统的化学计量学算法在拟合训练样本时,将有限样本数据中的误差也拟合进数学模型了,而支持向量回归采用 ε 不敏感损失函数,即对于用 $f(x)$ 拟合目标值 y 时,$f(x) = w^{\mathrm{T}}x_i + b$,目标值 y_i 拟合在 $|y_i - w^{\mathrm{T}}x_i - b| \leqslant \varepsilon$ 时,认为进一步拟合是无意义的。这样拟合得到的不是唯一解,而是一组无限多个解。这一求解策略使过拟合受到限制,显著提高了数学模型的预报能力。SVR 方法是在一定约束条件下,以 $\| \omega_2 \|$(表示两类不同样本间隔的函数)取最小的标准为适应样本集的非线性。SVR 通过非线性映射将数据映射到高维的特征空间中,在其中进行线性回归。通过运用一个非敏感性损耗函数,非线性 SVR 的解可参照文献[23]求出。

1.2 试剂与仪器

亚氯酸钠(74.6 %),工业纯;苯、乙醇、冰醋酸、丙酮、硫酸均为分析纯。

杉木试材采自安徽黄山林场,从不同海拔高度伐取 6 株,每株从胸高处开始,每隔 2 m 截取一个 6 cm 圆盘,气干后分心材和边材分别劈成小薄片,经过粉碎,筛分,最后选取 380 μm 筛孔与 250 μm 筛孔标准筛之间的木粉,供分析和采谱用。

美国 ASD 公司生产的 LabSpec ® Pro 近红外光谱仪,采用漫反射积分球附件,扫谱范围:350~2 500 nm,低噪声 512 阵元 PDA;光谱采样间隔:1.4 nm@ 350~1 050 nm, 2 nm@ 1 000~2 500 nm。

1.3 常规方法测定

杉木中木质素含量的测定按照国标 GB/T 2677.8 - 1994 进行。同时做两份平行样测定,取其算术平均值作为测定值。

1.4 近红外光谱(NIR)采集

近红外光谱(NIR)仪和光谱采集均在装有空调的恒温(20 ℃ ±2 ℃)室内进行。用杯光源检测器对商用聚四氟乙烯白板进行空白校准后,再对样品的近红外光谱在全光谱范围内(350~2500 nm)进行采集,扫描 30 次并自动平均为一个光谱,每个样品采谱 3 次,取平均值。选取 804~2504 nm近红外光谱区作为测定杉木中木质素的波长范围,每隔 10 nm 记录一个点,共得 171 个点的吸光度值。

1.5　数据处理

光谱数据的平滑处理可有效平滑高频噪声,提高信噪比,减小运算量,所以首先对光谱数据进行卷积平滑(Savitzky-Golay)并进行一阶和二阶求导,消除斜坡背景和基线的影响。小波变换是在傅里叶变换基础上发展起来的数学方法,可用于数据压缩、平滑滤噪、基线校正、多组分重叠信号解析和图像处理等。为提高处理速度,利用小波函数 db_1 对求导后的数据进行压缩,选择压缩次数以压缩后能保证特征谱图为依据,最终将光谱中的 171 个数据压缩为 86 个数据。最后,将数据按(0,1)范围归一化。

2　结果与讨论

2.1　核函数类型的选择

支持向量机建立模型优先解决的问题是核函数的选择。常见的核函数有线性函数、多项式函数、径向基函数和 Sigmoid 函数等 4 种类型,选择不同的核函数对所建立模型的性能影响很大。一般在没有先验知识指导的情况下,用径向基函数往往能够得到较好的结果,因为径向基函数可以将非线性样本数据映射到高维特征空间,处理具有非线性关系的样本数据[24]。另外径向基函数取值($0 < K \leqslant 1$)要比多项式函数取值($0 < K$ 或 $\infty > K > 1$)简单,而且计算速度明显优于 Sigmoid 函数。因此,实验采用径向基函数作为核函数。

2.2　参数的优化

确定了核函数后,就要对计算用参数进行优化,对模型会产生影响的参数有 ε 不敏感损失函数、径向基系数(γ)和惩罚参数(C)。

实验数据共包含 47 个样本,随机抽取出 8 个样本作外部验证,用留一法进行内部交叉验证,所谓留一法就是留一个样本作为预测样本,其他样本作为训练样本,重复此过程,直到每个样本都作为预测样本 1 次,作为训练样本 $n-1$ 次。用相对误差平方和来评价模型。

2.2.1　*ε 不敏感损失函数的选择*　ε 不敏感损失函数的含义是,当 x 点的观察值(y)与预测值($f(x)$)之差不超过事先给定的 ε 时,则认为在该点的 $f(x)$ 是无损失的,尽管 $f(x)$ 和 y 可能并不完全相等。ε 控制着误差的边界,ε 太小易产生过拟合现象,ε 太大易产生欠拟合现象。ε 的大小通过试验来确定,固定 C 为 10, γ 为 1 调整 ε 的值,结果如表 1 所示。

表 1 表示了相对误差平方和随 ε 的变化情况。从表中数据可以看出,随着 ε 的减小,相对误差平方和也逐渐减小,当 ε 减小到 0.01 后,相对误差平方和变化不再明显。

2.2.2　*径向基系数(γ)的选择*　参数 γ 控制着支持向量机对输入量变化的敏感程度。γ 的大小也通过试验来确定,固定 C 为 10, ε 为 0.001 调整 γ 的值,结果亦列入表 1。表中表示了相对误差平方和随 γ 的变化情况。从表中数据可以看出,过大的 γ 使支持向量机反应迟钝,而过小的 γ 对输入过于敏感,致使样本预测结果也不好。

表 1　ε 和 γ 值对支持向量回归模型分析结果的影响

Table 1　Influences of ε and γ on the result of support vector regression model

ε	相对误差平方和 square sum of relative error	γ	相对误差平方和 square sum of relative error
1.5×10^{-1}	1.443×10^{-2}	1.0×10^{-3}	1.024×10^{-2}
1.0×10^{-1}	1.425×10^{-2}	5.0×10^{-3}	8.397×10^{-3}
5.0×10^{-2}	1.385×10^{-2}	1.0×10^{-2}	7.433×10^{-3}
1.0×10^{-2}	1.359×10^{-2}	5.0×10^{-2}	8.171×10^{-3}
5.0×10^{-3}	1.354×10^{-2}	1.0×10^{-1}	9.271×10^{-3}
1.0×10^{-3}	1.352×10^{-2}	5.0×10^{-1}	9.725×10^{-3}
5.0×10^{-4}	1.351×10^{-2}	1.0	1.352×10^{-2}
1.0×10^{-4}	1.351×10^{-2}	5.0	1.539×10^{-2}

2.2.3　*惩罚参数(C)的选择*　C 是与核函数及样本的非线性有关的参数,它控制着经验风险,过大或过小都将使误差增大,一般都取一个稍大的数用来降低误差,以取得对训练样本较好的拟合。在确定了

ε 为 0.001、γ 为 0.01 之后,改变 C 的值,当 C 在较大的范围内变动时,相对误差平方和没有变化,这也表明 ε 和 γ 的选取是合适的。但是,最好在此基础上适当减小 C 的值,避免过大的 C 值引起经验误差,导致泛化能力下降。最终确定 C 为 10。

2.3 实验结果

用杉木的近红外光谱数据建立的测定木质素含量的模型,校正相对误差的平方和为 0.007 433,预测相对误差的平方和为 0.001 219。表 2 表示了预测结果。预测模型的相对误差最大为 1.110 5 %,最小为 -1.970 8 %,其绝对值均小于 5 %,结果比较好。

表 2 采用支持向量回归建立模型对木质素的预测结果

Table 2 Predicted results of lignin content with support vector regression model

样品序号 samples	实测值/% determined values	预测值/% predicted values by NIR	绝对误差/% absolute deviation	相对误差/% relative deviation
1	34.1800	33.6969	-0.4831	-1.4134
2	33.2800	33.6496	0.3696	1.1105
3	34.1000	33.6060	-0.4940	-1.4487
4	33.6300	33.9592	0.3291	0.9789
5	34.0900	33.7253	-0.3647	-1.0697
6	33.4100	33.4265	0.0165	0.0495
7	34.3200	33.6436	-0.6764	-1.9708
8	33.0500	33.3582	0.3082	0.9324

3 结 论

对支持向量回归的基本原理进行了简单介绍,以 47 个杉木样品作为实验材料,用 SVM-近红外光谱法(NIR)建立了测定杉木中木质素的定量分析模型,以径向基函数作为核函数,采用留一法对计算用参数进行了优化,建立了最佳模型。校正相对误差的平方和为 0.007 433,预测相对误差的平方和为 0.001 219。结果表明,该方法可以用于杉木中木质素含量的预测。

参考文献:

[1]冯海,徐光明,刘迪霞,等. 近红外光谱法同时测定多种雌、孕激素[J]. 分析化学,2001,29(2):175-177.

[2]严衍禄. 近红外光谱分析基础与应用[M]. 北京:中国轻工业出版社,2005.

[3]徐广通,沈师孔,陆婉珍,等. 近红外光谱在清洁汽油生产控制分析中的应用 [J]. 石油炼制与化工,2001,32(6):51-55.

[4]黄安民,江泽慧. 近红外光谱技术在木材性质预测中的应用研究进展[J]. 世界林业研究,2007,20(1):49-54.

[5]FACKLER K, SCHWANNINGER M, GRADINGER C, et al. Qualitative and quantitative changes of beech wood degraded by wood-rotting basidiomycetes monitored by Fourier transform infrared spectroscopic methods and multivariate data analysis[J]. Fems Microbiology Letters, 2007,271(2):162-169.

[6]JONES P D, SCHIMLECK L R, PETER G F, et al. Nondestructive estimation of wood chemical composition of sections of radial wood strips by near infrared spectroscopy[J]. Wood Sci Technol, 2006,40(8):709-720.

[7]HODGE G R, WOODBRIDGE W C. Use of near infrared spectroscopy to predict lignin content in tropical and sub-tropical pines[J]. J Near Infrared Spectrosc, 2004,12(6):381-390.

[8]KELEEY S S, RIALS T G, SNELL R, et al. Use of near infrared spectroscopy to measure the chemical and mechanical properties of solid wood [J]. Wood Sci Technol, 2004,38(4):257-276.

[9]YAMADA T, YEH T F, CHANG H M, et al. Rapid analysis of transgenic trees using transmittance near-infrared spectroscopy (NIR)[J]. Holzforschung, 2006,60(1):24-28.

[10]POKE F S, RAYMOND C A J. Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in *Eucalyptus globulus*[J]. Wood Chem Technol, 2006,26(2):187-199.

[11]POKE F S, WRIGHT J K, RAYMOND C A. Predicting extractives and lignin contents in *Eucalyptus globulus* using near infrared reflectance analysis[J]. J Wood Chem Technol, 2004,24(1):55-67.

[12]HOFFMEYER P, PEDERSEN J G. Evaluation of density and strength of Norway spruce wood using near infared reflectance spectroscopy[J]. Holz als Roh-und Werkstoff, 1995,53:165-170.

[13]SCHIMLECK L R,ROBERT E,JUGO I. Application of near infrared spectroscopy to a diverse range of species demonstrating wide density and stiffness variation[J]. IAWA Journal,2001,22(4):415-429.
[14]JONES P D,SCHIMLECK L R,PETER G F,et al. Nondestructive estimation of *Pinus taeda* L. wood properties for samples from a wide range of sites in Georgia[J]. Can J Forest Res,2005,35(1):85-92.
[15]江泽慧,黄安民,王斌. 木材不同切面的近红外光谱信息与密度快速预测[J]. 光谱学与光谱分析,2006,26(6):1034-1037.
[16]江泽慧,黄安民. 木材中的水分及其近红外光谱分析[J]. 光谱学与光谱分析,2006,26(8):1464-1468.
[17]江泽慧,费本华,杨忠. 光谱预处理对近红外光谱预测木材纤维素结晶度的影响[J]. 光谱学与光谱分析,2007,27(3):435-438.
[18]杨忠,江泽慧,费本华,等. SIMCA 法判别分析木材生物腐朽的研究[J]. 光谱学与光谱分析,2007,27(4):686-690.
[19]江泽慧,李改云,王戈,等. 近红外光谱法测定毛竹综纤维素的含量研究[J]. 林产化学与工业,2007,27(1):15-18.
[20]张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-42.
[21]崔伟东,周志华,李星. 支持向量机研究[J]. 计算机工程与应用,2001,37(1):58-61.
[22]张录达,苏时光,王来生,等. 支持向量机(SVM)在傅里叶变换近红外光谱分析中的应用研究[J]. 光谱学与光谱分析,2005,25(1):33-35.
[23]陆文聪,陈念贻,叶晨洲,等. 支持向量机算法和软件 ChemSVM 介绍[J]. 计算机与应用化学,2002,19(6):697-702.
[24]叶美盈,汪晓东. 混沌光学系统辨识的支持向量机方法[J]. 光学学报,2004,24(7):953-956.