

最小二乘支持向量机用于 人参皂苷近红外光谱分析研究

丁海泉^{1,2}, 温江北¹, 高洪智¹, 刘振尧¹, 卢启鹏²

(1. 广东星创众谱仪器有限公司, 广东 广州 510663;

2. 中国科学院长春光学精密机械与物理研究所/应用光学国家重点实验室, 吉林 长春 130033)

摘要: 人参作为名贵药材, 具有较高的药用价值。其中皂苷是人参中主要的活性药用成分, 在治疗疾病及保健功能中发挥重要作用。人参皂苷含量高低会因不同条件的变化而有较大差异。皂苷传统测定方法是紫外可见分光光度计法, 该方法操作过程繁琐、耗时, 且破坏样品。近红外光谱技术具有快速、无损、绿色无污染等优点, 为人参皂苷成分的测定进而快速评价人参的质量提供了依据。本文利用近红外技术对人参皂苷含量进行测定分析, 利用最小二乘支持向量机中常用的核函数对人参皂苷进行定量建模分析, 得出不同核函数的最佳模型定标效果, 进一步提高了人参皂苷近红外快速定标模型分析的稳健性及精确性。最终为改善传统近红线性分析模型的局限性, 提高人参皂苷近红外光谱定量分析精度, 增强模型泛化能力, 深度研究了最小二乘支持向量机定量分析方法。首先对采集的人参原始光谱进行多元散射校正消除颗粒度影响, 接下来重点讨论了最小二乘支持向量机算法中常用的3种核函数模型的参数优化及调节过程, 建立了人参皂苷含量的定量分析模型。最优模型相关系数达到0.98, 交叉检验均方根偏差为0.14%, 获得了较高的分析精度。表明近红外分析技术结合最小二乘支持向量机方法在人参皂苷含量预测方面有较好的应用前景。

关键词: 近红外光谱; 人参皂苷; 最小二乘支持向量机; 核函数; 参数优化

中图分类号: O657.3 **文献标识码:** A **文章编号:** 1673-2154 (2021) 05-0017-06

0 引言

人参作为名贵药材之一, 具有较高的药用价值。其中皂苷是人参中主要的活性药用成分, 在治疗不同的疾病及保健功能中发挥着重要作用^[1-3]。皂苷含量的高低会因产地、气候、栽培技术等条件的变化而有较大差异。人参皂苷的传统测定方法是紫外可见分光光度计法, 该方法虽具有较高的精度, 但过程繁琐、耗时、破坏样品且分析成本高。

近红外光谱技术是近年来发展迅速的一种分析技术, 具有快速、无损、绿色无污染等优点, 目前已作为一种成熟的分析技术在农业、石油、医药等领域广泛应用^[4-6]。近红外技术的特点为人参皂苷成

分测定进而快速评价人参的质量提供了依据, 并为人参样品成分实时在线分析提供了可能。近年来, 近红外光谱技术在人参皂苷检测研究方面也有一些积极的探索^[7-8]。汪静静等^[9]利用偏最小二乘方法 (Partial Least Squares, PLS) 建立了人参样品中人参皂苷的近红外定量分析模型; 肖雪等^[10]利用近红外技术结合偏最小二乘方法对人参皂苷中的主要元素成分进行定量建模分析, 均取得不错的效果。上述应用的建模方法均是常用的线性回归方法, 而线性建模方法应用的基础是光谱和参比值之间要有固有的线性相关性, 否则线性模型很难得到良好的预测结果。而随着非线性技术的不断完善, 非线性模型使得近红外预测能力和泛化能力均有了一定的提高。

收稿日期: 2021-08-11

基金项目: 国家重点研发计划项目 (2018YFD0401000); 广东省科技计划项目 (2017B090907013); 广州开发区院士专家创新创业项目 (2017-L5); 国家自然科学基金面上项目 (61475155)。

作者简介: 丁海泉 (1982—), 男, 博士, 副研究员, 主要研究方向为现代光谱技术与仪器。

通讯作者: 卢启鹏 (1964—), 男, 研究员, 主要研究方向为光谱技术与仪器。E-mail: luqipeng@126.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

因此针对人参皂苷的近红外建模方法中非线性方法研究值得尝试,具有重要的应用价值。

最小二乘支持向量机 (Least Squares Support Vector Machines, LS-SVM) [11-13] 是支持向量机 (Support Vector Machine, SVM) [14-15] 在二次损失函数形式下的一种扩展,通过求解一组线性方程代替支持向量机中复杂的二次优化问题的非线性建模方法。该方法通过结构风险最小化原理可提高模型泛化能力,在解决小样本、非线性、高维数等问题中有着较好的优势。

本文利用近红外技术对人参皂苷含量进行测定分析,采用 LS-SVM 方法对人参皂苷进行定量建模分析,利用 LS-SVM 中常用的线性核函数、多项式核函数、高斯径向基核函数分别对人参皂苷含量进行建模比较,得出不同核函数的最佳模型定标效果,进一步提高了人参皂苷近红外快速定标模型分析的稳健性及精确性。

1 LS-SVM模型及模型评价

LS-SVM的基本原理是运用核函数使原始自变量数据映射到某一高维特征空间进而使因变量与自变量存在某一线性关系。

对于训练集样本 $\{x_i, y_i\}_{i=1}^N \in R^M \times R$, 其中 x_i 表示第 i 个样本的多维自变量, y_i 表示第 i 个样本的因变量。利用非线性变化 $\varphi(\cdot)$ 将原自变量数据映射到更高维特征空间中,在此空间中构造最优线性回归函数 $f(x) = \omega\varphi(x) + b$, 通过对线性回归核函数的优化最终得出回归预测模型。

$$\hat{c}_j = \sum_{i=1}^m a_i K(x_j, x_i)$$

式中:

$\hat{c}_j$ ——第 j 个样本的模型预测值;

a_i ——拉格朗日乘子, $a_i \in R$;

$K(x_j, x_i)$ ——取决于样本光谱数据核函数的计算结果;

$x_i (i=1, 2, \dots, l)$ ——训练集样本光谱。

$$a_i = \left((A_i)^T A_i + \frac{1}{2\gamma} \right)^{-1}$$

式中:

Γ ——正则化参数;

A_i ——通过对因变量数据进行加权处理,所有自

变量样本光谱数据的线性组合。

LS-SVM 模型效果取决于模型中核函数及相应核参数的选择,当确定某一类核函数后,通过其对应核参数进行调试,及结合正则化参数 γ 的变化即可完成对整个 LS-SVM 模型的优化。其中正则化参数 γ 用于调节模型拟合的效果,核函数中参数用于调节样本数据在高维空间重分布的复杂度。

为获得较好的 LS-SVM 建模效果,本文对比分析了 LS-SVM 中常用的 3 种核函数构建回归模型。

线性核函数: $K(x_j, x_i) = x_j \cdot x_i + b$

多项式核函数: $K(x_j, x_i) = (x_j \cdot x_i + b)^d$

高斯径向基核函数:

$$K(x_j, x_i) = \exp(-\|x_j - x_i\|^2 / 2\sigma^2)$$

式中:

d ——多项式的次数;

σ^2 ——高斯径向基的核宽度,用于调节不同核函数下模型泛化能力。

当多项式次数 $d=1$ 时,多项式核函数演化成线性核函数,因此对 LS-SVM 模型优化时重点讨论多项式核函数及高斯径向基核函数。

模型评价体系中利用模型相关系数 (R) 交互检验均方根偏差 (RMSECV) 作为模型参数优选及模型评价的依据。

2 实验样品与方法

选用具有一定梯度的人参皂苷含量样品 29 份,样品由旋风磨粉碎,过 1 mm 孔筛后分成 2 份,1 份用于实验室化学分析,1 份用于近红外光谱数据的采集。

人参皂苷含量的化学值通过紫外可见分光光度计法测定,皂苷成分浓度范围为 2.17% ~ 5.67%。样本光谱采集实验仪器采用 FOSS XDS 光栅型近红外光谱分析仪,光谱采集范围为 1 100 ~ 2 500 nm,光谱分辨率为 2 nm,环境温度为室温 $26 \text{ }^{\circ}\text{C} \pm 1 \text{ }^{\circ}\text{C}$ 。样品装入培养皿后由专用装置固定,光谱采集方式为漫反射,测量时每隔 120° 旋转培养皿,每个位置扫描 32 次漫反射光谱并进行自动保存。每个样本重复采集 3 次光谱数据,最终以 3 次光谱数据的平均值作为建模的光谱数据。

3 结果与讨论

29 份人参粉末样品近红外光谱如图 1 所示。由人参样品的光谱图发现, 由于此次收集到的人参样本水分含量的不同所导致人参样本硬度不同, 在进行人参样品粉碎时难以确保相同的颗粒度状态, 从人参的原始吸光度数据中看出人参样品光谱间由于颗粒度的不同而导致样品光谱图基线平移, 故对原始光谱进行多元散射校正 (MSC), 经过多元散射校正后的数据可有效地消除原始数据由于颗粒度等因素所导致的基线平移和非线性差异的影响, 经过多元散射校正后的人参样本吸光度如图 2 所示。

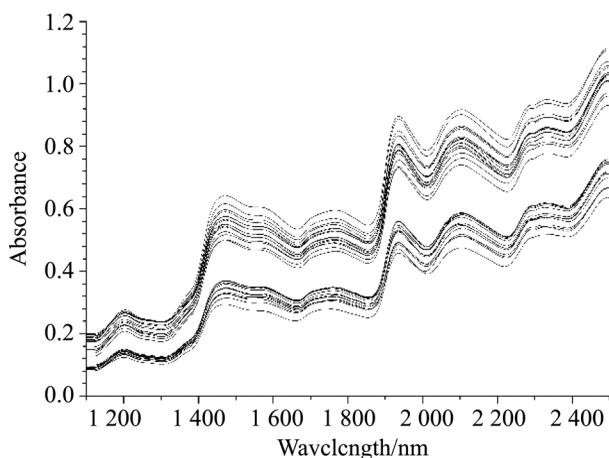


图 1 人参样品原始光谱图

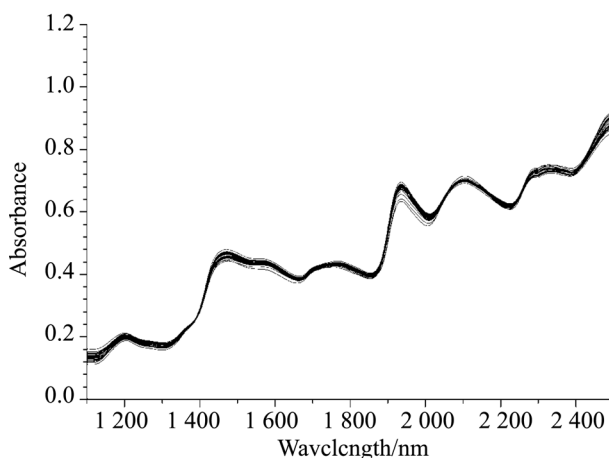


图 2 人参样品多元散射光谱图

利用多元散射校正后的光谱数据结合人参皂苷的理化值, 分别采用不同的核函数, 建立并比较了人参皂苷含量的 LS-SVM 模型, 通过对正则化参

数 γ 及核函数参数的调节, 完成对人参皂苷模型的优化。因多项式核函数包含线性核函数 ($d=1$), 故仅对多项式核函数及径向基核函数中核参数调整结合正则化参数 γ 的变化对 LS-SVM 回归模型效果进行比较讨论。对核函数参数及正则化参数 γ 的优选采取常用的网格搜索结合交叉检验对目标函数进行寻优, 直至找到最佳的模型参数。以模型的交叉检验均方根偏差 (RMSECV) 为优选依据, 选取合适的 RMSECV 确定模型的最佳参数进而建立最佳模型。

3.1 多项式核函数

首先, 讨论多项式核函数的 LS-SVM 建模结果。其中正则化参数 γ 取值一般参照指数 $2^{\pm 2}$ 增长方式进行取值, 参数调节中 γ 在 $2^{-11}, 2^{-9} \dots 2^{-1}, 2^1 \dots 2^{45}, 2^{47}$ 界限范围内取值, 共 30 个参数值, 多项式次数 d 取 0, 1, 2, 3, 参数调整组合共计 120 个。为进一步保证模型训练的高效及准确, 当全局最优值处于搜索空间的边界区域, 也就是说最优的惩罚参数 γ 为 2^{-11} 或 2^{47} , 接着按相同的乘积步长 ($2^{\pm 2}$) 增加元素, 直到模型不能优化为止, 模型的比较结果如图 3 所示。

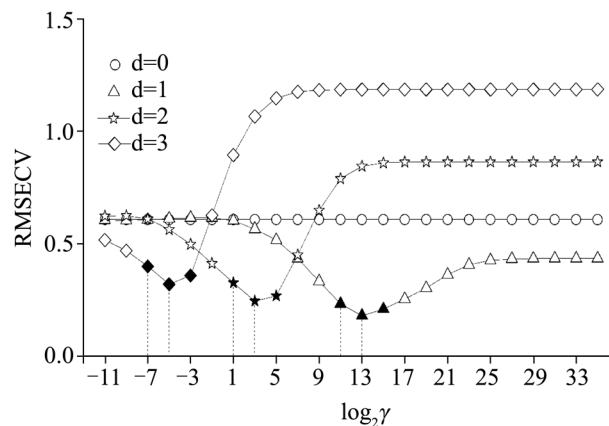


图 3 固定多项式次数下不同惩罚参数 RMSECV 值比较

比较图 3 中的结果发现, 当多项式次数 $d=0$ 时, 模型的 RMSECV 是保持不变的, 也就意味着当使用固定的常数项作为核函数进行计算时, 模型的效果不受正则化参数 γ 的影响, 模型起不到优化的作用。另外线性核函数 ($d=1$), 二次核函数 ($d=2$), 三次核函数 ($d=3$) 均能在本次参数优选的界限中找到

优化的正则化参数值(图3中实心点),利用多项式核函数中最佳的模型参数对LS-SVM模型进行计算,模型效果如表1所示,不同多项式次数下的最优模型散点图如图4所示。

表1 多项式核函数优化模型的比较

d	γ	RMSECV	R_c
1	2^{15}	0.171	0.983
2	2^5	0.241	0.966
3	2^{-3}	0.270	0.957

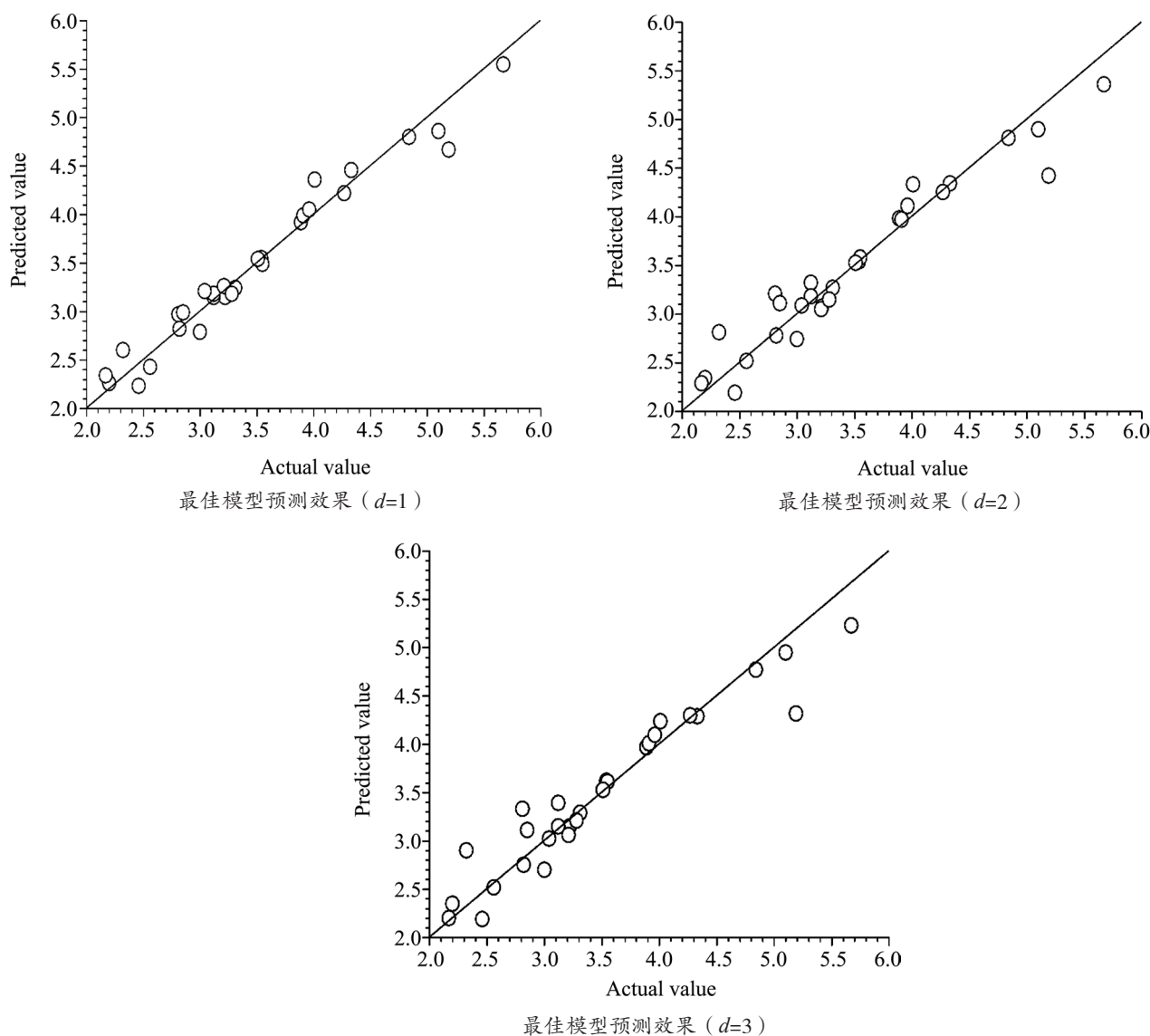


图4 不同次数下多项式核函数优化模型的比较

3.2 高斯径向基核函数

针对高斯径向基核函数的讨论中,其中高斯径向基核参数 σ^2 一般同样是参照指数增长的方式进行取值,这里与正则化参数 γ 在集合 2^{-11} , 2^{-9} ... 2^{-1} ,

2^1 ... 2^{45} , 2^{47} 中取值,并结合模型的效果对其参数进行扩展计算,共计900组参数的调整,针对900组参数组合的模型效果,选取其中高斯径向基核参数 σ^2 中较优的模型进行比较,比较结果如图5所示。

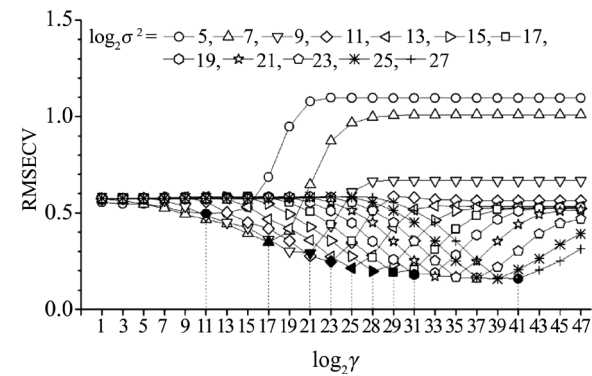


图 5 固定 σ^2 下不同惩罚参数的 RMSECV 比较

比较图 5 的结果发现，随着高斯径向基核函数中参数 σ^2 的增大，其相应模型的最佳正则化参数 γ 也同样的变大，另外根据图 5 中不同 σ^2 参数下的最优模型数据点 (图 5 中实心点) 构成的曲线呈现出凹形的下降趋势，因此可确定高斯径向基核函数的最佳模型参数在本次模型界限调整范围内。其中优选的模型参数建模结果如表 2 所示，高斯径向基核函数最优模型参数建模效果如图 6 所示。

表2 优选的模型参数建模结果			
σ^2	γ	RMSECV	Rc
2^{13}	2^{27}	0.200	0.977
2^{15}	2^{29}	0.199	0.977
2^{17}	2^{31}	0.199	0.977
2^{19}	2^{33}	0.200	0.977
2^{21}	2^{35}	0.203	0.976
2^{23}	2^{37}	0.206	0.975
2^{25}	2^{41}	0.137	0.989
2^{27}	2^{43}	0.140	0.988
2^{29}	2^{45}	0.140	0.988

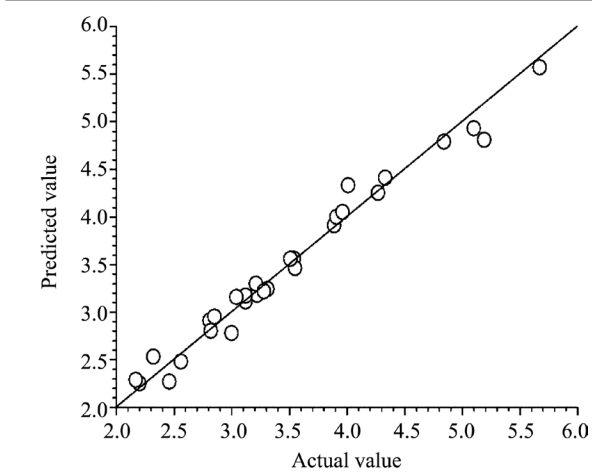


图 6 高斯径向基核函数的优化模型

在多项式核函数及高斯径向基核函数参数优选

过程中发现，对于多项式核函数而言，随着模型的复杂程度的提高 (即多项式次数 d 变大)，用于人参皂苷的模型效果反而有所下降。其中最佳线性核函数的 RMSECV 能够达到 0.17%，也进一步说明了人参皂苷成分模型有较高的线性相关性。对于高斯径向基核函数而言，最优的模型参数效果 RMSECV 能够达到 0.14%。综合比较各最优模型的散点图效果，其中高斯径向基核函数回归模型精度较高。

4 结论及展望

人参具有很高的药用价值，对多种常见疾病均具有治疗及控制作用，其中皂苷的含量更是在人参中发挥着重要作用。而皂苷的常规检测方法会带来一些弊端，如破坏样品、耗时，做到对人参皂苷含量无损、快速的检测具有明显的现实意义。

本文利用近红外光谱技术结合 LS-SVM 方法用于人参皂苷成分的无损、快速检测，将 LS-SVM 3 种常用核函数应用于模型优化中，采用网格及步进式阶梯搜索完成核参数及正则化参数的优选过程，分别讨论了线性、多项式、高斯径向基核函数在核变换中输出的最优定标效果，其中最优模型的交叉检验均方根偏差 RMSECV 达到 0.14%，具有较高的检测精度。实验验证了近红外分析技术结合 LS-SVM 方法用于人参皂苷含量预测的可行性，采用的参数调整及模型优化方法在近红外非线性建模算法的拓展中有着重要的参考意义。

参考文献

[1] 刘铁城. 中国西洋参 [M]. 北京: 人民卫生出版社, 1995.

[2] 国家药典委员会. 中华人民共和国药典 3 部 [M]. 北京: 中国医药科技出版社, 2010.

[3] 程海涛, 许永华, 郭爽, 等. 人参光环境研究进展 [J]. 人参研究, 2010, 22 (3): 27-30.

[4] 陆婉珍. 现代近红外光谱分析技术. 第 2 版 [M]. 北京: 中国石化出版社, 2007.

[5] 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.

[6] DOWNEY G. "Book-review" near-infrared technology in the agricultural and food industries[J]. Irish Journal of Food Science and Technology, 1988, 12(1/2): 103-104.

[7] 芦永军, 曲艳玲, 冯志庆, 等. 人参总皂甙近红外光谱及定标建模分析 [J]. 光谱学与光谱分析, 2007, 27 (3): 490-493.

[8] 张瑞, 曲正义, 张浩, 等. 人参产品中总皂苷、水分、灰分近红外光谱快速分析技术研究 [J]. 特产研究,

- 2015 (3): 43-46.
- [9] 汪静静, 闫述模, 杨滨. 近红外光谱技术对人参药材人参皂苷含量测定及产地识别的研究[J]. 光谱学与光谱分析, 2015, 35 (7): 1885-1888.
- [10] 肖雪, 李军山, 张博, 等. 近红外光谱快速测定红参提取过程中5种人参皂苷成分含量[J]. 南开大学学报(自然科学版), 2017 (3): 46-50.
- [11] 安欣, 徐硕, 张录达, 等. 多因变量LS-SVM回归算法及其在近红外光谱定量分析中的应用[J]. 光谱学与光谱分析, 2009, 29 (1): 127-130.
- [12] CHAUCHARD F, COGDILL R, ROUSSEL S, et al. Application of ls-svm to non-linear phenomena in nir spectroscopy: development of a robust and portable sensor for acidity prediction in grapes[J]. Chemometrics and Intelligent Laboratory Systems, 2004, 71(2): 141-150.
- [13] 王克奇, 杨少春, 戴天虹, 等. 采用遗传算法优化最小二乘支持向量机参数的方法[J]. 计算机应用与软件, 2009, 26 (7): 109-111.
- [14] XIAOLIN HUANG, LEI SHI, JOHAN A.K. SUYKENS. Asymmetric least squares support vector machine classifiers[J]. Computational Statistics and Data Analysis, 2014, 70: 395-405.
- [15] THISSEN U, PEPERS M, ÜSTÜN B, et al. Comparing support vector machines to pls for spectral regression applications[J]. Chemometrics and Intelligent Laboratory Systems, 2004, 73(2): 169-179.

The Least Square Support Vector Machine is Used in the Near Infrared Spectroscopy Experimental Study of Ginsenoside

Ding Haiquan^{1,2}, Wen Jiangbei¹, Gao Hongzhi¹, Liu Zhenyao¹, Lu Qipeng²

(1. Guang Dong Xinxtra Instrument Co., Ltd., Guangzhou 510663, China; 2. State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics/Chinese Academy of Sciences Changchun, Changchun 130033, China)

Abstract: Ginseng is a valuable medicinal material with high medicinal value. Ginsenoside is the main active pharmaceutical ingredient in ginseng and plays an important role in the treatment of diseases and health care functions. The content of ginsenoside can vary greatly due to different conditions. The traditional method of the determination of ginsenoside is the ultraviolet-visible spectrophotometric method, which is tedious, time-consuming and destroys the samples. Near infrared spectroscopy has the advantages of fast, nondestructive, green and pollution-free. It has become a mature analytical technique widely used in many fields. This technique is characterized by the determination of ginsenoside components and the rapid evaluation of the quality of ginseng, and provides the possibility for real-time online analysis of ginseng sample components. The application of near-infrared linear modeling method is based on the inherent linear correlation between spectral data and reference ratio, otherwise the linear model can hardly get good prediction results. With the continuous improvement of nonlinear technology, the nonlinear model has improved the near-infrared prediction ability and generalization ability. It is worthwhile to study the near-infrared nonlinear modeling method for ginsenoside. The least squares support vector is an extension of support vector machine in the form of quadratic loss function, by solving a set of linear equations instead of the support vector machine in the complex quadratic optimization problem of nonlinear modeling method. This method can improve the generalization ability of the model through the principle of structural risk minimization, and has a good advantage in solving the problem of small sample, non-linearity and high dimension. By using near infrared technology for determination of ginsenoside content analysis, using the commonly kernel function of the least squares support vector machine to perform the quantitative modeling analysis of ginsenoside, obtain the best model effect of the different kernel functions, to further improve the near-infrared ginsenosides fast calibration model robustness and accuracy. In order to improve the quantitative analysis accuracy of the ginsenoside content by using near infrared spectroscopy and enhance the generalization ability of the model. In this paper, the LS-SVM quantitative analysis was studied in depth. Firstly, eliminating the influence of particle impaction by using the multiple scattering correction with the original spectrum of the ginseng, then the paper focuses on discussing the parameters optimization and adjustment process of the three commonly kinds of kernel function model used in the LS-SVM algorithm, and establishes quantitative analysis models of the ginsenoside content. In the end, the correlation coefficient R of the optimal model was 0.98, and the cross-validation root mean square deviation error RMSECV was 0.14%, which obtained a higher analytical precision. It is indicated that near infrared analysis technique combined with LS-SVM method has a better application prospect in the prediction of ginsenoside content.

Key words: near-infrared spectroscopy; ginsenoside; LS-SVM; kernel function; parameter optimization

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. <http://www.cnki.net>