近红外光谱结合 GA-LSSVR 分析烟草尼古丁含量

郭志明[1,2]　陈立平[1,2*]　黄文倩[1,2]　张　驰[1,2]

([1] 北京农业智能装备技术研究中心，北京 100097；[2]国家农业智能装备工程技术研究中心，北京 100097)

摘要　为了提高近红外光谱快速检测烟草尼古丁含量的精度和稳定性，利用近红外光谱结合遗传算法-最小二乘支持向量回归(GA-LSSVR)建立了回归预测模型。在 LSSVR 模型建立过程中，采用遗传算法对 LSSVR 参数进行自动优化。相比于利用常规最小二乘支持向量机和遗传偏最小二乘法等建立的回归预测模型，GA-LSSVR 法建立的回归预测模型泛化能力更强，预测效果更好，验证集相关系数 R^2 为 0.9766，预测均方根误差为 0.1065。研究结果表明，GA-LSSVR 是一种快速准确的建模方法，为烟草尼古丁含量的近红外测定和近红外光谱数据的处理提供了新的方法与途径。

关键词　测量；近红外光谱；遗传算法；最小二乘支持向量回归；烟草；尼古丁

中图分类号　O657.33　　**文献标识码**　A　　**doi**：10.3788/LOP49.021201

Application of Genetic Algorithm-Least Squares Support Vector Regression with Near Infrared Spectroscopy for Prediction of Nicotine Content in Tobacco

Guo Zhiming[1,2]　Chen Liping[1,2]　Huang Wenqian[1,2]　Zhang Chi[1,2]

([1] *Beijing Research Center of Intelligent Equipment for Agriculture*, *Beijing* 100097, *China*
[2] *National Research Center for Intelligent Agricultural Equipment*, *Beijing* 100097, *China*)

Abstract　In order to improve the detecting precision and robustness in determination of the content of nicotine in tobacco by near infrared spectroscopy, a predictive model is established by genetic algorithm combined least square support vector regression (GA-LSSVR). The practical use of support vector machine is limited because of its set of parameters to be defined by the user. For this reason, a genetic algorithm is utilized to approach LSSVR parameter optimization in the calibration model. To highlight the superiority of GA-LSSVR algorithm, it is compared with traditional LSSVR and genetic algorithm-partial least square. The correlation coefficient R^2 and root mean square error of prediction (RMSEP) for the test set are used as evaluation parameters for the model. The optimal model is obtained with R^2 of 0.9766 and RMSEP of 0.1065. Generally, in the context of performance and robustness, the results demonstrate that GA-LSSVR is a good method for the analysis and modelling of near infrared data.

Key words　measurement; near infrared spectroscopy; genetic algorithm; least square support vector regression; tobacco; nicotine

OCIS codes　120.4290; 300.6340; 200.3050

1　引　　言

烟草是一种重要的经济作物，其尼古丁含量是衡量烟草和卷烟质量的重要指标之一[1,2]。尼古丁对人体产生毒副作用，最大的危害是吸食它会产生依赖性以至成瘾，其燃烧后大部分变成去甲烟碱亚硝胺，是一种强烈的致癌物质。在烟草行业中，尼古丁是影响卷烟口感的主要因素之一，当某种烟叶中尼古丁含量较低时，需要向烟叶中喷施一定量的尼古丁。因此，对烟草中尼古丁含量进行定量分析具有十分重要的实用价值。但常规化学分析过程比较复杂，费时、费力、成本高。近红外(NIR)光谱法因其快速、简便、低成本、非破

收稿日期：2011-09-23；**收到修改稿日期**：2011-10-10；**网络出版日期**：2011-12-07

基金项目：国家 863 计划(2011AA100703)和国家自然科学基金(31071324)资助课题。

作者简介：郭志明(1982—)，男，工程师，主要从事近红外光谱及高光谱图像技术在农产品质量安全检测中的应用等方面的研究。E-mail：guozm@nercita.org.cn

* **通信联系人**。E-mail：chenlp@nercita.org.cn

坏性和多组分同时测定等优点受到人们的重视，在烟草成分分析中得到越来越广泛的应用[3~5]。

应用近红外光谱分析技术的关键是建立稳健的定标模型。近红外光谱分析中常用的定标方法有多元线性回归、主成分回归和偏最小二乘回归，这几种算法适用于线性和近似线性的数据处理，但对非线性数据的建模能力较弱。人工神经网络处理非线性数据具有一定优势，但训练中容易出现“过拟合”现象，导致拟合结果较好，而泛化能力弱。支持向量机(SVM)是在高维特征空间使用线性函数来假设空间的学习系统，通过最优化理论的学习算法训练，已经成功用于解决非线性函数的逼近问题。在烟草尼古丁含量的近红外光谱分析中，光谱参数和尼古丁的化学值存在一定的非线性关系。本文采用的最小二乘支持向量回归(LSSVR)是在经典 SVM 基础上的改进，是解决多元建模的一种快速求解方法，具有全局最优及更好的泛化能力[6,7]。目前，LSSVR 参数大多凭经验选取，无法获得最佳的函数逼近效果[8]。为了能够自动地获得最佳的 LSSVR 参数，本文研究基于自适应遗传算法的 LSSVR(GA-LSSVR)参数优选方法，从而确保 LSSVR 参数选择的准确性，以提高烟草尼古丁含量模型的预测精度，同时与遗传最小二乘法和常规 LSSVR 法相比较，评价模型优劣。

2 材料和方法

2.1 仪器设备及光谱采样方法

试验用烟叶为某烟厂提供，均产于 2009 年，共 110 个样本。从中随机选取 75 个作为训练集，余下的 35 个作为验证集，然后将它们分别编号后置于 4 ℃冰柜中贮藏。光谱采集试验在温度稳定的试验室内(温度控制为 25 ℃)进行。试验前，将冰柜中取出的烟叶置于试验室中 12 h，以使烟叶整体温度与环境温度达到一致。试验时，将烟叶样本放于样品杯中，采用 Antaris Ⅱ近红外光谱仪(美国 Thermo Scientific 公司)进行光谱扫描，扫描波数范围为 10000～4000 cm^{-1}，扫描次数 32 次，分辨率 4 cm^{-1}。样本采集完一次后倒出样品杯再重新倒入，重复采集三次，取三次的平均光谱作为整个样本的原始光谱。图 1 为训练集烟样的近红外光谱原始图。烟叶样品采用 AA3 连续流动分析仪(德国 BRAN LUEBBE 公司)测定其尼古丁含量，方法参考 ISO 15152-2005[9]。

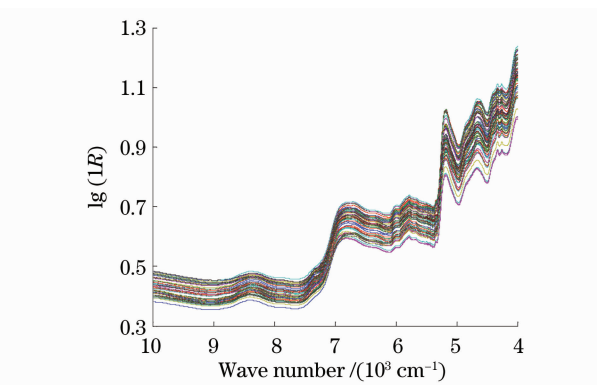

图 1 烟草样本的近红外光谱图

Fig. 1 Near-infrared spectra of tobacco samples

2.2 遗传-最小二乘支持向量回归

支持向量机具有完备的统计学习理论基础和很强的学习性能，可用于小样本问题的学习，计算速度快，预测能力强。LSSVR 方法采用最小二乘线性系统作为损失函数，将标准 SVM 中的不等式约束改成等式约束，并把经验风险由误差的一范数改为二范数，求解二次优化的问题就转化成了求解一次线性方程组问题，它能够进行线性和非线性的多元建模，并且降低计算的复杂性，提高了算法收敛速度[10]。由于径向基(RBF)核函数作为非线性函数能够减少训练过程中计算的复杂性，本文选择 RBF 核函数进行 LSSVR 计算。

如何确定 LSSVR 模型中的正则化参数 γ 和核函数参数 σ^2，一直是提高最小二乘 SVM 学习和泛化能力的主要研究问题之一。目前许多学者做了这方面的研究，这些研究将调整超参数方法归纳为两类，分别为基于分析的方法和基于启发式的搜索方法。第一类方法通过推广误差的梯度来确定参数，但不能保证搜索方向为推广性能度量的最优值。本研究采用现代启发式算法中的遗传算法优选 LSSVR 参数[11,12]，实现过程如图 2 所示。遗传算法能解决传统搜索方法难以解决的复杂的非线性问题，不需要目标函数明确的数学方程和倒数表达式，是一种全局寻优算法，避免了传统算法易陷入局部最优解，寻优效率高。具体步骤如下：

1) 选定 LSSVR 的训练样本和校验样本，设定正则化参数 $\lg\gamma$ 的选择范围为 1～10，RBF 核函数参数 $\lg\sigma^2$ 的范围为 0～6，从而产生 LSSVR 参数初始群；

2) 设定基因交换概率 0.8，变异概率 0.01，群体规模 30，遗传迭代次数 100；

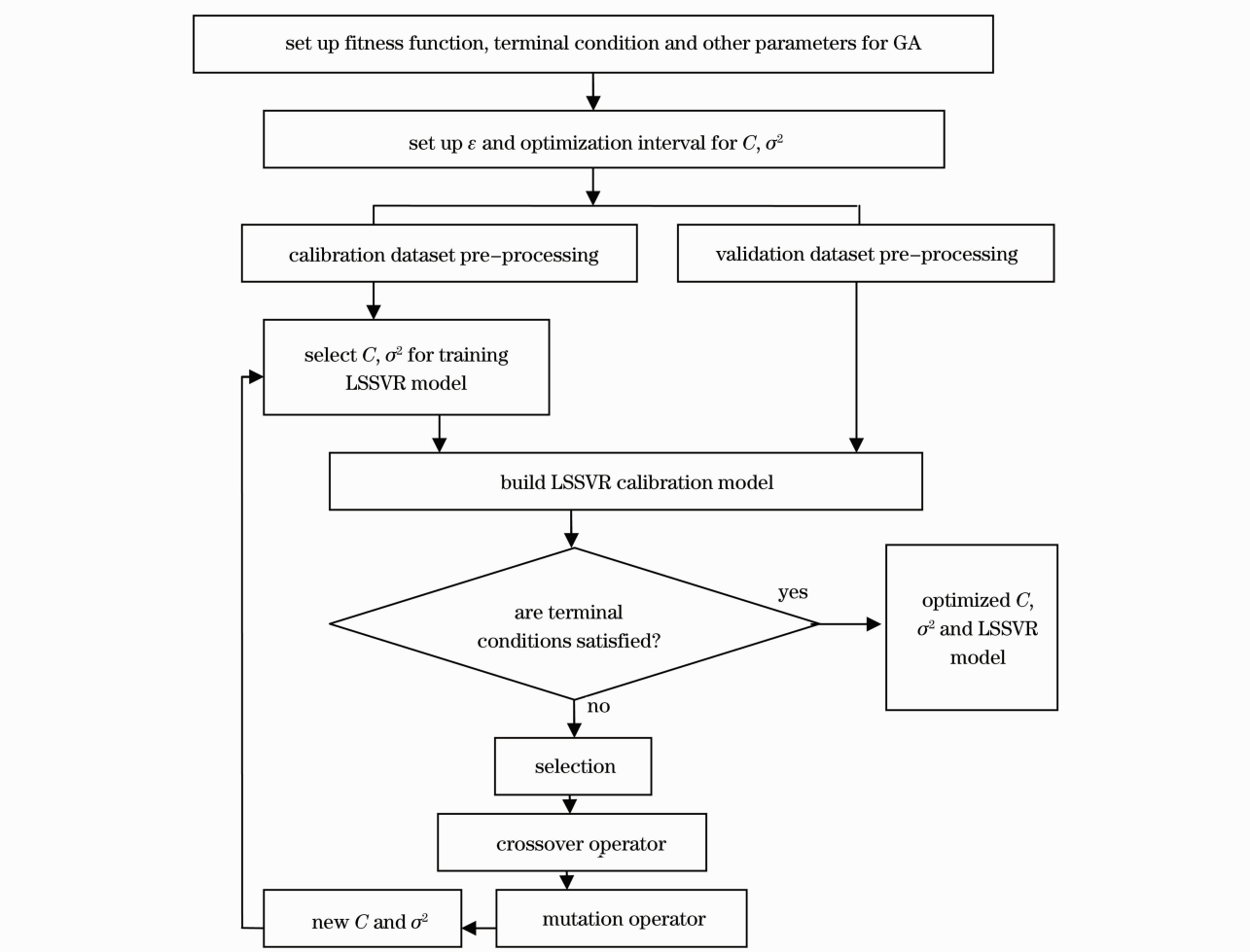

图 2 GA-LSSVR 参数优选流程图

Fig. 2 Flow chart for GA-LSSVR parameters optimization

3）进行 LSSVR 训练；

4）计算遗传算法适应度函数值 $f=\frac{\sum_{i=1}^{N}\left|\frac{y_i'-y_i}{y_i}\right|}{N}\times 100\%$，式中 N 表示训练样本数，y_i'表示第 i 个样本的实测值，y_i 表示第 i 个样本的预测值；

5）根据计算 GA 适应度，对群体进行选择、变异、交叉操作，产生下一代参数群体；

6）若满足 GA 训练停止条件（训练误差小于等于 0.001 或达到迭代次数）则停止训练，跳转 7），否则转向 3）；

7）结束遗传训练，得到最终的 LSSVR 参数向量，建立 GA-LSSVR 模型。

3 结果与分析

3.1 光谱数据预处理

控制光谱采集在相同的条件下进行，样品的装样厚度、装样的紧密性和颗粒均匀度等都力求一致，但在采集近红外光谱时，有许多高频随机噪声、基线漂移、样本颗粒大小和光散射等噪声信息夹入[13]。这将干扰近红外光谱与样品内有效成分含量间的关系，并直接影响所建立模型的可靠性和稳定性。为消除光谱数据中夹入的噪声信息，首先采用 Savitzky-Golay 平滑法，平滑点数取为 5，用于消除光谱曲线中的高频成分，保留有用的低频信息，然后进行多元散射校正（MSC）处理，以降低烟叶样本的散射作用对光谱数据造成的影响。从图 1 可以看出，在 10000～8000 cm^{-1}区间内曲线平滑，吸光度变化不大，在 4500～4000 cm^{-1}区间存

在较强的末端吸收。为提高测量数据的信噪比，选取 8000～4500 cm^{-1} 的光谱数据作为下一步建模数据的输入。

3.2 GA-LSSVR 参数优化与建模

LSSVR 建模中需确定正则化参数 γ 和核函数参数 σ^2。γ 为 RBF 核函数的正则化参数，决定了结构风险最小化和经验风险最小化之间的平衡，对模型非常重要。σ^2 则控制函数回归误差，并且直接影响初始的特征值和特征向量。σ^2 过小会导致大量回归量的产生，并且最终导致过拟合。相反，σ^2 过大会导致回归量的减少，模型过于简单，从而影响预测精度。此外，σ^2 还关系到 LS-SVR 模型对输入变量噪声的灵敏度。合理选择参数能够解决“欠学习”和“过学习”的问题，同时，能够使机器学习的泛化能力提高。由于遗传搜索过程具有较强的随机性，为了防止算法运行过程中随机性对结果的影响，研究采用多次运算相比较，选择预测性能最好的模型，结果如表 1 所示。

表 1 各次 GA-LSSVR 建模的结果

Table 1 Results of GA-LSSVR prediction model for each operation procedure

No.	γ	σ^2	R^2 for calibration	RMSE	R^2 for validation	RMSEP
1	69.52	5.57	0.9837	0.0821	0.9766	0.1065
2	58.45	6.83	0.9802	0.0896	0.9707	0.1161
3	87.22	3.54	0.9795	0.0910	0.9681	0.1205
4	63.58	6.27	0.9810	0.0884	0.9754	0.1092
5	71.23	5.08	0.9788	0.0927	0.9672	0.1215

研究以相关系数(R^2)、交叉验证的均方根误差(RMSE)和预测均方根误差(RMSEP)评价 GA-LSSVR 模型的稳健性和预测性能。经遗传训练后，最优的 GA-LSSVR 模型中正则化参数 γ 为 69.52，核函数参数 σ^2 为 5.57。烟草尼古丁含量的 GA-LSSVR 评价模型的结果如图 3 所示。训练集样本预测值与实测值之间的相关系数为 0.9837，交叉验证的均方根误差为 0.0821；验证集样本预测值与实测值之间的相关系数为 0.9766，预测均方根误差为 0.1065。

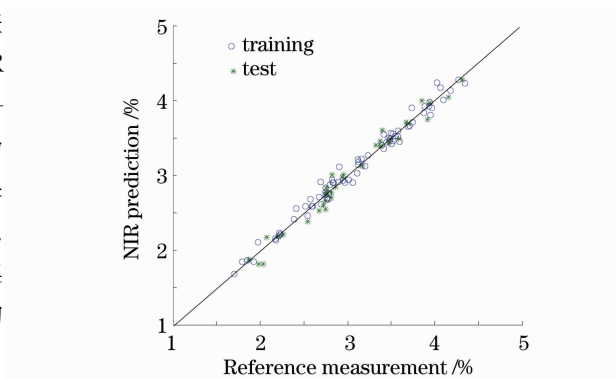

图 3 烟草尼古丁测量值与近红外预测值散点图

Fig. 3 Reference determination versus NIR prediction for nicotine

3.3 多元校正方法结果比较

LSSVR 算法中，采用人工多次搜索人为选取正则化参数和核函数参数，以同样的光谱数据和参考值，根据误差结果选择最优的参数，正则化参数 γ 为 90.83，核函数参数 σ^2 为 6.55。以同样的光谱数据和参考值，采用遗传偏最小二乘法(GA-PLS)建立烟草尼古丁的近红外光谱定标模型并进行优化，运行 5 次所得模型结果变化很小，模型主因子数均为 6，其中最好模型的验证集样本验证值与实测值之间的相关系数为 0.9113，预测均方根误差为 0.2058。三种方法模型结果如表 2 所示，可明显看出 GA-LSSVR 的效果最好，优于传统 LSSVR 、GA-PLS 等算法。

表 2 烟草尼古丁模型的多元校正方法比较

Table 2 Result and comparison of nicotine model using different multivariate calibration methods

Multivariate calibration method	Calibration set		Validation set	
	R^2	RMSE	R^2	RMSEP
GA-LSSVR	0.9837	0.0821	0.9766	0.1065
Traditional LSSVR	0.9696	0.1120	0.9432	0.1563
GA-PLS	0.9391	0.1587	0.9113	0.2058

相比于以经验人工选择 LSSVR 的模型参数,全局搜索的遗传算法选取的正则化参数和核函数参数建立的 LSSVR 模型取得了较好的结果,预测精度和泛化能力均有所提高。可见,GA-LSSVR 是一种行之有效的非线性多元校正方法。SVM 学习算法是基于结构风险最小化准则,能同时兼顾经验风险和置信范围,在样本有限的情况下,取得实际风险的最小值。烟草中尼古丁含量和光谱数据之间存在一定的非线性关系,因此 GA-PLS 法建模结果不理想。

4 结 论

利用遗传算法全局寻优的功能,对 LSSVR 算法参数进行寻优,构建基于 GA-LSSVR 算法的烟草尼古丁含量的预测模型。训练集样本预测值与实测值之间的相关系数为 0.9837,交叉验证的均方根误差为 0.0821;验证集样本预测值与实测值之间的相关系数为 0.9766,预测均方根误差为 0.1065。相比于 LSSVR、GA-PLS 等多元校正模型,GA-LSSVR 模型对烟草尼古丁含量的近红外光谱分析预测能力更强,既提高了预测精度,又降低了预测差错率。因此,在近红外光谱分析中采用 GA-LSSVR 算法,既可以处理非线性问题、抑止过拟合现象,又能自动优化参数,建立的数学模型更为可靠,得到的结果更为准确。

参 考 文 献

1 K. Torikaiu, Y. Uwano, T. Nakamori *et al.*. Study on tobacco components involved in the pyrolytic generation of selected smoke constituents [J]. *Food and Chemical Toxicology*, 2005, **43**(4): 559～568

2 Yong Zhang, Qian Cong, Yunfei Xie *et al.*. Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine [J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2008, **71**(4): 1408～1413

3 Shi Xue, Cai Wensheng, Shao Xueguang. A local regression method for near infrared spectral quantitative analysis of tobacco samples [J]. *Spectroscopy and Spectral Analysis*, 2008, **28**(11): 2561～2564
石 雪,蔡文生,邵学广. 局部建模方法用于烟草样品的近红外光谱定量分析[J]. 光谱学与光谱分析, 2008, **28**(11): 2561～2564

4 Liu Xu, Chen Huacai, Liu Taiang *et al.*. Application of PCA-SVR to NIR prediction model for tobacco chemical composition[J]. *Spectroscopy and Spectral Analysis*, 2007, **27**(12): 2460～2463
刘 旭,陈华才,刘太昂 等. PCA-SVR 联用算法在近红外光谱分析烟草成分中的应用[J]. 光谱学与光谱分析, 2007, **27**(12): 2460～2463

5 Zhihui He, Wenliu Lian, Mingjian Wu *et al.*. Determination of tobacco constituents with acousto-optic tunable filter-near infrared spectroscopy[J]. *Journal of Near Infrared Spectroscopy*, 2006, **14**(1): 45～50

6 Hongdong Li, Yizeng Liang, Qingsong Xu. Support vector machines and its applications in chemistry [J]. *Chemometrics and Intelligent Laboratory Systems*, 2009, **95**(2): 188～198

7 Sun Guangming, Liu Fei, Zhang Fan *et al.*. Determination of proline in herbicide-stressed oilseed rape leaves based on near infrared spectroscopy[J]. *Acta Optica Sinica*, 2010, **30**(4): 1192～1196
孙光明,刘 飞,张 帆 等. 基于近红外光谱技术检测除草剂胁迫下油菜叶片中脯氨酸含量的方法[J]. 光学学报, 2010, **30**(4): 1192～1196

8 B. Ustun, W. J. Melssen, M. Oudenhuijzen *et al.*. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization[J]. *Analytica Chimica Acta*, 2005, **544**(2): 292～305

9 International Standard Orgnizaton. Tobacco-Determination of the Content of Total Alkaloids as Nicotine-Continuous-Flow Analysis Method [S]. ISO 15152-2005, 2005

10 Sun Jun, Mao Hanping, Yang Yiqing. Prediction of nitrogen content rate of paddy rice leaf based on GA-LS-SVM [J]. *J. Jiangsu University: Natural Science Edition*, 2010, **31**(1): 6～10
孙 俊,毛罕平,羊一清. 基于 GA-LS-SVM 的水稻叶片含氮率预测[J]. 江苏大学学报(自然科学版), 2010, **31**(1): 6～10

11 Z. Yang, X. S. Gu, X. Y. Liang. Genetic algorithm-least squares support vector regression based predicting and optimizing model on carbon fiber composite integrated conductivity[J]. *Materials and Design*, 2010, **31**(3): 1042～1049

12 Kuanyu Chen. Forecasting systems reliability based on support vector regression with genetic algorithms [J]. *Reliability Engineering and System Safety*, 2007, **92**(4): 423～432

13 Zhao Jiewen, Guo Zhiming, Chen Quansheng *et al.*. Feasibility study for the use of near infrared spectroscopy in the quantitative analysis of catechins in green tea [J]. *Acta Optica Sinica*, 2008, **28**(12): 2302～2306
赵杰文,郭志明,陈全胜 等. 近红外光谱法快速检测绿茶中儿茶素的含量[J]. 光学学报, 2008, **28**(12): 2302～2306