第 28卷第 6期
2009年 12月

红 外 与 毫 米 波 学 报
J. Infrared Millim. Waves

Vol. 28, No. 6
December, 2009

文章编号: 1001 - 9014(2009)06 - 0423 - 05

基于无信息变量消除法和连续投影算法的可见/近红外光谱技术白虾种分类方法研究

吴 迪[1], 吴洪喜[2, 3*], 蔡景波[2, 3], 黄振华[2, 3], 何 勇[1*]

(1. 浙江大学 生物系统工程与食品科学学院,浙江 杭州 310029;
2. 浙江省海洋水产养殖研究所,浙江 温州 325005;
3. 浙江省近岸水域生物资源开发与保护重点实验室,浙江 温州 325005)

摘要:应用无信息变量消除法结合连续投影算法对可见/近红外光谱区进行有效波长的选择,选择后的波长作为输入变量建立最小二乘-支持向量机模型,对白虾属中三种典型种,脊尾白虾、秀丽白虾和东方白虾进行鉴别分类. 实验采用 Kennard-Stone算法选取 150个样本作为建模集, 50个样本作为预测集,通过 UVE-SPA优选了数值分别为 392、431、517、551、595、627、676、734、760、861、943和 1018 nm的 12个波长为 LS-SVM的输入变量,建立了白虾种分类模型. 该模型对 50个预测集样本检验的准确率达到了 92.00%. 结果表明,采用可见/近红外光谱对白虾种进行鉴别是可行的, UVE-SPA能够有效地进行波长选择,使 LS-SVM模型获得最优的分类结果.

关 键 词:可见/近红外光谱;无信息变量消除;连续投影算法;最小二乘-支持向量机
中图分类号: O657.33 **文献标识码**: A

CLASSIFYING THE SPECIES OF EXOPALAEMON BY USING VISIBLE AND NEAR INFRARED SPECTRA WITH UNINFORMATIVE VARIABLE ELIMINATION AND SUCCESSIVE PROJECTIONS ALGORITHM

WU Di[1], WU Hong-Xi[2, 3*], CAI Jing-Bo[2, 3], HUANG Zhen-Hua[2, 3], HE Yong[1*]

(1. School of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China;
2. Zhejiang Mariculture Research Institute, Wenzhou 325005, China;
3. Zhejiang Key Lab of Exploitation and Preservation of Coastal Bio-resource, Wenzhou 325005, China)

Abstract: Using visible-near infrared spectra to classify different species of exopalaemon was studied. Successive projections algorithm (SPA) combined with uninformative variable elimination (UVE) were used to select effective wavelengths from visible and near infrared (Vis-NIR) bands. The selected effective wavelengths were set as inputs of least square-support vector machine (LS-SVM) for the classification of three typical exopalaemon species, namely, E. carincauda, E. modestus and E. orientis. Kennard-Stone algorithm was used to select 150 samples for calibration and the remaining 50 samples for prediction. Twelve effective wavelengths were selected by UVE-SPA, and they were 392, 431, 517, 551, 595, 627, 676, 734, 760, 861, 943 and 1018 nm. The correct rate is 92.00% for classifying samples in prediction set by LS-SVM model based on these twelve effective wavelengths. The overall results demonstrate that it is feasible to utilize Vis-NIR spectroscopy to classify different species of exopalaemon, and UVE-SPA can extract the most effective wavelengths to build the LS-SVM model with an optimal classification result.

Key words: visible-near infrared spectroscopy; uninformative variable elimination (UVE); successive projections algorithm (SPA); least square-support vector machine (LS-SVM)

收稿日期: 2008-10-22,**修回日期**: 2009-05-22 **Received date:** 2008-10-22, **revised date:** 2009-05-2
基金项目:国家科技支撑项目 (2006BAD10A04);国家高技术研究发展计划 (863计划)项目 (2006AA10Z234);浙江省自然科学基金项目 (Y506152);浙江省台州市重大科技招标项目 (20071ZB02)
作者简介:吴 迪 (1984-),男,浙江杭州人,博士生,主要从事数字农业和农产品光谱检测技术研究
* **通讯联系人**:何 勇: yhe@zju.edu.cn;吴洪喜: whxchina@126.com

引言

白虾属是十足目真虾次目长臂虾科下的一个较小的属.目前,我国已鉴定出 4个种,分别为脊尾白虾 (E. carinicauda)、秀丽白虾 (E. modestus)东方白虾 (E. orientis,俗称"绿籽虾")和安氏白虾 (E. annandalei),皆为经济虾类.该属种类体长一般不超过 90mm,甲壳薄而透明,微带蓝褐或红色点,死后个体呈白色,因此不易区别,尤其是均生活在海水中的脊尾白虾和东方白虾更易混淆.在浙江,若在脊尾白虾养殖塘混入了东方白虾,东方白虾极易形成优势种群,使养殖效益明显降低,给白虾围塘养殖业带来重大的经济损失,引起了当地有关部门的高度重视.

可见近红外光谱 (Visible and near infrared, Vis-NIR)分析技术,是一种快速、无损、低成本、无污染的分析方法,它可以对物质的品质、种类、成分等进行定性和定量分析[1,2],但在虾种分析中的应用还较少.由于可见近红外光谱信息重叠严重,利用全波段进行建模分析时,光谱中的大量冗余信息及噪声等使模型的性能受到影响.如何在纷繁复杂的光谱信息中提取出有用信息,提高模型校正的速度和建模效率是本文研究的重点内容.

常用的波长选择方法,如相关系数法[3]、载荷值法[4]、回归系数法[4]等大多根据主观经验进行阈值选择,而退火算法[5]和遗传算法[6]的搜寻过程非常耗时,且不稳定.无信息变量消除算法 (uninformative variable elimination, UVE)是基于偏最小二乘 (partial least squares, PLS)的回归系数建立的波长选择算法,用于消除不提供信息的变量,减少模型变量,降低模型的复杂性[7].但是有时 UVE得到的变量数依然较多,因此仍需从 UVE得到的变量中进一步选择有效变量.连续投影算法 (successive projections algorithm, SPA)是一种新的变量提取方法[8,9],它能够利用向量的投影分析,寻找含有最低限度的冗余信息的变量组,并使变量之间的共线性达到最小,同时能大大减少建模所用变量的个数,提高建模的速度和效率.

本文应用 UVE-SPA提取的有效波长作为最小二乘支持向量机 (least square-support vector machine, LS-SVM)[10]的输入,建立 UVE-SPA变量选择模型,应用于白虾品种鉴别的研究.并将此结果与分别采用全波段波长及仅应用 UVE得到的有效波长作为 LS-SVM模型输入的结果进行比较.

1 材料与方法

1.1 仪器设备

实验使用美国 ASD (Analytical Spectral Device, Boulder, USA)公司的 Handheld FieldSpec光谱仪,其光谱范围为 325～1075nm,探头视场角为 15°,光谱扫描次数设为 30次,光源采用 14.5 V卤素灯;试验采用漫反射模式;分析软件采用 Unscrambler V9.7 (CAMO AS, Oslo, Norway)和 Matlab V7.6 (The MathWorks, Natick, USA).

1.2 样本制备及样本集划分

实验选用我国白虾属中的脊尾白虾、秀丽白虾和东方白虾进行鉴别分析.试验中将同种类的虾均匀平铺在光谱仪采集视角范围内,样本覆盖背景,进行光谱采集.脊尾白虾和秀丽白虾各采集 50个样本,东方白虾采集了 100个样本.采用 Kennard-Stone算法[11]选取 150个样本作为建模集,剩余 50个样本作为预测集.建模集样本用于模型建立,预测集样本用于对模型的预测性能进行检验.

1.3 无信息变量消除算法原理及算法实现

在 PLS模型中,光谱矩阵 X和浓度矩阵 Y存在如下的关系:

$$Y = Xb + e \quad , \tag{1}$$

其中,b是系数向量,e是误差向量.UVE就是把一定变量数目的随机变量矩阵加入光谱矩阵中,然后通过交叉验证建立 PLS模型,得到系数矩阵 B,分析系数向量 b的平均值和标准偏差的商 C的稳定性,即:

$$C_i = \frac{\mathrm{mean}(b_i)}{S(b_i)} \quad , \tag{2}$$

其中,mean(b)表示系数向量 b的平均值,$S(b)$表示系数向量 b的标准偏差,i表示光谱矩阵中的第 i列向量.根据 C_i的绝对值大小确定是否把第 i列变量用于最后的 PLS模型中.具体算法见参考文献[7].

1.4 连续投影算法实现

SPA算法的简要介绍如下[12]:

记 $x_{k(0)}$为初始迭代向量,N为需要提取的变量个数,光谱矩阵为 j列.

(1) 迭代开始前,任选光谱矩阵的 1列 j,把建模集的第 j列赋值给 x_j,记为 $x_{k(0)}$;

(2) 把未选入的列向量位置的集合记为 s, $s = \{j, 1 \leq j \leq J, j \notin \{k(0), \cdots, k(n-1)\}\}$;

(3) 分别计算 x_j对剩余列向量的投影: $Px_j = x_j - (x_j^T x_{k(n-1)}) x_{k(n-1)} (x_{k(n-1)}^T x_{k(n-1)})^{-1}$, $j \in s$;

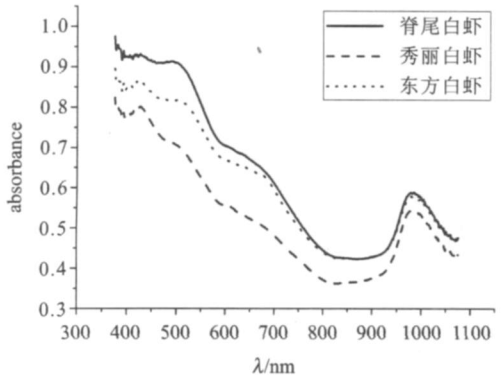

图 1　三种典型白虾的原始可见 近红外吸收光谱

Fig 1　Original absorbance spectra of exopalaemon from three typical species

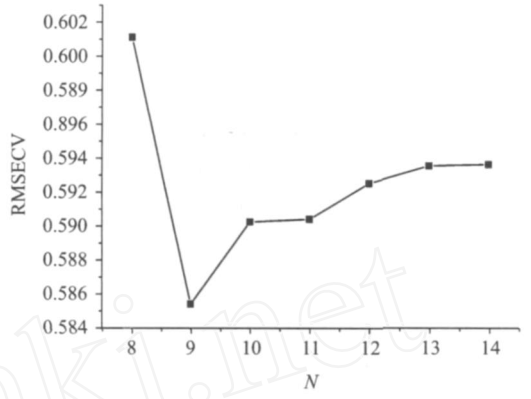

图 2　不同主成分数的 UVE得到的 RMSECV分布图

Fig 2　RMSECV plot of different numbers of principal components by UVE, where N is number of principal components

(4) 记 $k(n) = \arg(\max \| Px_j \|)$, $j \in s$;

(5) 令 $x_j = Px_j$, $j \in s$;

(6) $n = n + 1$,如果 $n < N$,回到 b循环计算.

最后,提取出的变量为 $\{x_{k(n)} = 0, \cdots, N - 1\}$. 对应于每一个 $k(0)$和 N,循环一次后进行多元线性回归分析 (MLR),得到验证集的预测标准偏差 (RMSEV),最小的 RMSEV值对应的 $k(0)$和 N就是最优值.

2　试验结果与分析

2 1　不同白虾种的可见 近红外光谱图

由于测量到的光谱在 325～375nm范围内存在较大的噪声,因此选用 376～1075nm波长范围内共计 700个变量进行分析. 不同种类白虾的典型可见 - 近红外吸收光谱图如图 1所示,图中横坐标为波长,纵坐标为吸光度. 从图 1可以看出,不同白虾种的光谱曲线的趋势非常相似,在 990nm附近有明显的水的吸收峰. 然而只从光谱特征上,难以区分不同种类的白虾. 因此需要用化学计量学建模方法对光谱数据进行处理.

2 2　应用 UVE选取有效波长

采用平滑、SNV、求导等对全波长光谱进行预处理,然后作为 LS-SVM的输入建立模型. 从表 1可以看出,采用原始光谱全波长作为输入变量时的结果最好,预测集的正确率达到 90%. 采用原始光谱数据的结果优于平滑、求导等处理. 这可能是因为在预处理过程中,引入或产生了新的噪声,或者是处理后的有用信息量下降,或者两者并存,使得所建模型的预测性能下降. 虽然采用全波长作为输入变量已经达到了较好的分类结果,但是 LS-SVM模型的输入变量有 700个,部分变量可能仍包含无用的或者不相关的信息,并且上百个输入变量造成模型冗余、复杂[13]. 为此采用 UVE对未经预处理的全波长 700个光谱变量进行选择.

UVE中产生的随机变量个数同样设置为 700个. UVE中的最优主成分数是经内部交叉验证得到的. 从图 2可以看到当主成分数为 9时,UVE选择变量的分类效果最好. 主成分为 9时的 UVE变量选择结果见图 3. 黑色竖线的左侧为 700个变量的稳定性 C分布曲线,右侧为 UVE中产生的 700个随机变量的稳定性 C分布曲线. 两条水平虚线表示变量选择的阈值上下限,在虚线外的数值对应的变量被保留,在虚线内的数值对应的变量不用于建模. 阈值

表 1　基于不同预处理和变量选择方法的 LS-SVM模型的三种典型白虾分类结果

Table 1　Classification results of exopalaemon from three typical species by LS-SVM models by using different pretreatment and variable selection methods

pretreatment	variable selection methods	variables	calibration set(n = 150) correct rate(%)	prediction set(n = 50) correct rate(%)
raw	none	700	99. 33	90. 00
SG + SNV	none	700	100. 00	88. 00
1 - Der	none	700	96. 67	84. 00
2 - Der	none	700	86. 67	78. 00
raw	UVE	291	100. 00	88. 00
raw	UVE-SPA	12	98. 67	92. 00

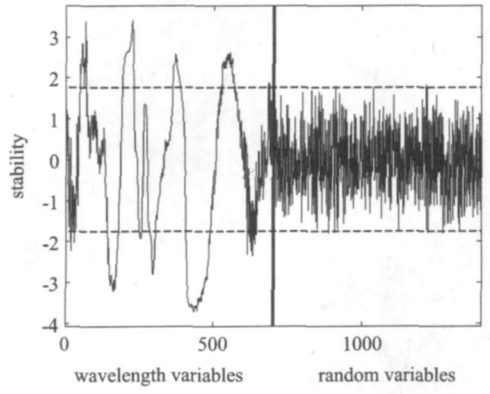

图3 主成分个数为9时的UVE稳定性分布曲线. 两条水平虚线表示变量选择的阈值上下限

Fig. 3 Stability distribution of UVE when nine components were considered. Two horizontal dot lines represent the threshold boundaries

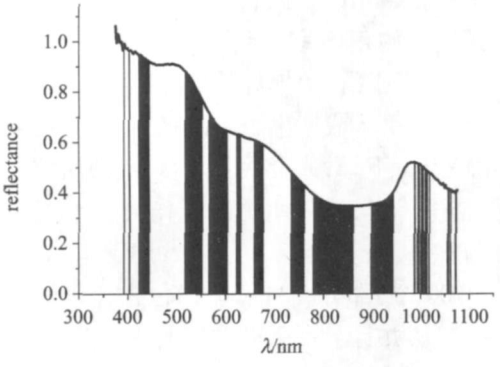

图4 UVE选择的291个波长分布图. 竖线为选中的波长

Fig. 4 Plot of 291 selected wavelengths by UVE. Columns represent selected wavelength

的选择标准为随机变量稳定性最大值的99%. 经过UVE的变量选择,最后得到了291个波长,其分布情况见图4,其中竖线表示选中的波长. 可以看到,UVE选择的波长在423～443nm的紫光部分、517～551nm和566～603nm的绿光部分、659～676nm和734～760nm的红光部分以及780～860nm和1000nm附近的近红外范围都有一定的分布. 说明无论是可见光还是近红外光谱都包含关于白虾种分类的有用信息. 将得到的这291个波长作为LS-SVM的输入变量,预测结果见表1. 可以看到虽然模型的分类正确率和700个变量作为输入时相比差不多,但是输入变量数却从700个减少到291个,模型得到了优化.

2.3 应用UVE-SPA选取有效波长

经过UVE选择过的变量个数仍过于庞大,因此采用SPA对经过UVE选择后的291个变量做进一步选择,得到共线性最小的有效波长. 对以上291个

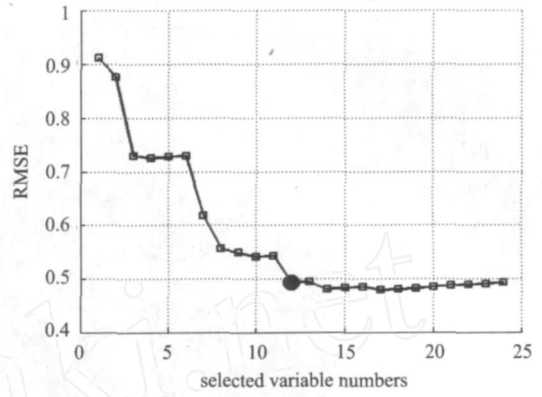

图5 SPA选择的不同变量数的RMSE分布

Fig. 5 RMSE plot of number of selected variables by SPA

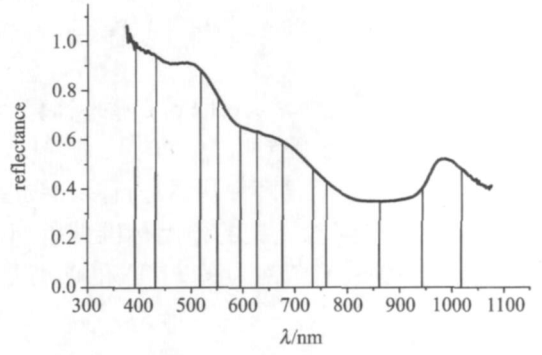

图6 UVE-SPA选择的12个波长(竖线表示)

Fig. 6 Twelve wavelengths selected by UVE-SPA. Columns represent selected wavelengths

变量进行建模,经过SPA进一步选择的不同波长数的RMSE分布如图5所示,图中黑色实心圆点表示所选到的波长数. 由图5可以看出,选取3个波长时,RMSE有一个下降过程,然后到选择8个波长时,RMSE又有一个下降过程,到提取出12个有效波长时RMSE到达一个低点,以后RMSEV趋于平坦,因此确定从291个波长中选择出其中的12个波长. 由UVE-SPA得到的这12个波长的分布情况如图6所示,其中竖线表示选中的波长. 通过UVE-SPA提取的12个波长分别为392, 431, 517, 551, 595, 627, 676, 734, 760, 861, 943和1018nm. 将这12个波长作为输入变量建立白虾品种的UVE-SPA-LS-SVM模型,分类结果如表1所示. 从表1中可以看出,通过UVE-SPA得到的12个波长所建立的LS-SVM模型的建模集分类正确率略低于采用未经预处理的全波段700个波长以及仅仅采用UVE选择的291个波长建立的LS-SVM模型,但是预测集的分类正确率却有一定的提高. 虽然提高的幅度不大,但是UVE-SPA-LS-SVM模型仅采用12个有效波长

建模.说明采用 UVE-SPA所提取的有效波长能充分代表原始光谱的有效信息,可以作为波长提取的有效手段.选择的 12个波长均匀地分布在可见和近红外光谱范围内,说明对于白虾种的分类不能仅仅考虑某个范围的波长.结果表明,采用 UVE-SPA对可见/近红外光谱进行有效波长选择后建立的 LS-SVM模型对白虾种的分类是可行的,并且获得了满意的准确度.

3　结论

采用 UVE-SPA对可见/近红外光谱进行有效波长选择,选择后的波长作为输入变量建立 LS-SVM模型对白虾种进行分类.通过 UVE-SPA从全波长700个变量中选择得到了 12个最能够反应光谱信息的波长.预测集样本的分类准确率达到了 92.00%,优于采用全波段 700个波长以及仅仅采用 UVE选择的 291个波长作为输入变量的 LS-SVM模型.结果表明,应用可见/近红外光谱技术对白虾种进行分类是可行的,采用 UVE-SPA-LS-SVM组合模型能获得满意的分类结果,为白虾种的鉴别提供了方法依据.

REFERENCES

[1] YAN Yan-Lu, ZHAO Long-Lian, HAN Dong-Hai, *et al*. *The Foundation and Application of Near Infrared Spectroscopy Analysis*[M]. Beijing: China Light Industry Press(严衍禄,赵龙莲,韩东海,等.近红外光谱分析基础与应用.北京:中国轻工业出版社), 2005.

[2] WU Di, FENG Lei, ZHANG Chuan-Qing, *et al*. Early detection of gray mold (Cinerea) on eggplant leaves based on Vis/Near infrared spectra[J]. *J. Infrared Millim. Waves* (吴迪,冯雷,张传清,等.基于可见/近红外光谱技术的茄子叶片灰霉病早期检测研究.红外与毫米波学报), 2007, **26**(4): 269—273.

[3] Min M, Lee W S. Determination of significant wavelengths and prediction of nitrogen content for citrus[J]. *Trans ASAE*, 2005, **48**(2): 455—461.

[4] Wu D, He Y, Feng S. Short-wave near-infrared spectroscopy analysis of major compounds in milk powder and wavelength assignment[J]. *Anal Chim. Acta*, 2008, **610**(2): 232—242.

[5] Kalivas J H, Roberts N, Sutter J M. Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry[J]. *Anal Chem.*, 1989, **61**(18): 2024—2030.

[6] Jouan Rimbaud D, Massart D L, Leardi R, *et al*. Genetic algorithms as a tool for wavelength selection in multivariate calibration[J]. *Anal Chem.*, 1995, **67**(23): 4295—4301.

[7] Centner V, Massart D L, De Noord O E, *et al*. Elimination of uninformative variables for multivariate calibration[J]. *Anal Chem.*, 1996, **68**(21): 3851—3858.

[8] Araújo M C U, Saldanha T C B, Galvão R K H, *et al*. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis[J]. *Chemom. Intell Lab Syst*, 2001, **57**(2): 65—73.

[9] Galvão R K H, Araújo M C U, Fragoso W D, *et al*. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm[J]. *Chemom. Intell Lab Syst*, 2008, **92**(1): 83—91.

[10] WU Di, HE Yong, FENG Shui-Juan, *et al*. Application of infrared spectra technique based on LS-SVM to the non-destructive measurement of fat content in milk powder[J]. *J. Infrared Millim. Waves* (吴迪,何勇,冯水娟,等.基于LS-SVM的红外光谱技术在奶粉脂肪含量无损检测中的应用.红外与毫米波学报), 2008, **27**(3): 180—184.

[11] Macho S, Lusa R, Callao M P, *et al*. Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy standardisation of the calibration model[J]. *Anal Chim. Acta*, 2001, **445**(2): 213—220.

[12] CHEN Bin, MENG Xiang-long, WANG Hao. Application of successive projections algorithm in optimizing near infrared spectroscopic calibration model[J]. *Journal of Instrumental Analysis* (陈斌,孟祥龙,王豪.连续投影算法在近红外光谱校正模型优化中的应用.分析测试学报), 2007, **26**(1): 66—69.

[13] 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]. *Chemom. Intell Lab Syst* 2004, **71**(2): 141—150.