第25卷,第9期
2005年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol.25,No.9,pp1400-1403
September, 2005

# SVM回归法在近红外光谱定量分析中的应用研究

张录达[1], 金泽宸[2], 沈晓南[1], 赵龙莲[2], 李军会[2], 严衍禄[2]

1. 中国农业大学理学院, 北京 100094
2. 中国农业大学信息学院, 北京 100094

**摘 要** 研究了基于统计学习理论的支持向量机(SVM)回归法在近红外光谱定量分析中的应用。以66个小麦样品为实验材料, 由33个小麦样品作为校正样品, 采用4种不同核函数方法对小麦样品蛋白质含量与小麦样品近红外光谱进行SVM回归建模。以所建4种不同SVM回归模型对33个小麦预测样品的蛋白质含量进行了预测; 不同回归模型的预测结果与凯氏定氮法确定的蛋白质含量的标准化学值间的相关系数均在0.97以上, 平均绝对误差小于0.32。为了考察SVM回归校正模型的预测效果, 同所建PLS回归模型的预测结果进行了比较, 表明所建预测小麦样品蛋白质含量的SVM回归模型亦可通过近红外光谱进行实际样品的定量分析, 且有较好的分析效果。

**主题词** 支持向量机回归; 近红外光谱; 定量分析
**中图分类号:** O657.3 **文献标识码:** A **文章编号:** 1000-0593(2005)09-1400-04

近红外光(Near-infrared, NIR)为700～2 500 nm的光谱区[1], 习惯上又将近红外区划分为近红外短波(700～1 100 nm)和近红外长波(1 100～2 500 nm)两个区域。最小的氢原子(H)引起近红外波段的吸收。氢不仅能引起分子的基频吸收, 还会产生非简谐振动并引起较强的倍频吸收。H的基频和倍频与分子中的其他振动耦合产生了遍布整个近红外区域的合频吸收。这样每个分子都会有许多个吸收带, 且吸收带的强弱受分子浓度的影响。NIR作为一种分析手段, 可以测定有机物以及部分无机物。这些物质分子中化学键结合的各种基团(如 C═C, N═C, O═C, O═H, N═H) 的伸缩、振动、弯曲等运动都有它固定的振动频率。当分子受到红外线照射时, 被激发产生共振, 同时光能量的一部分被吸收, 测量其吸收光, 可以得到极为复杂的图谱, 这种图谱表示被测物质的特征。不同物质在近红外区域有丰富的吸收光谱, 每种成分都有特定的吸收特征, 这就为近红外光谱定量分析提供了基础。近红外光谱技术是20世纪90年代以来发展最快、最引人注目的光谱分析技术[2], 测量信号的数字化和分析过程的绿色化使该技术具有典型的时代特征。研究内容增多、范围拓宽, 在谷物产品(小麦、糙米、大豆、玉米)、食品、饲料、油脂工业等领域得到应用, 测定的成分也越来越多, 包括: 蛋白质、脂肪、水、淀粉、纤维、氨基酸、脂肪酸、矿物质含量、物理特性等[3]。化学计量学[4, 5]将计算机、数学及统计分析技术应用于化学与分析化学, 在20世纪70年代后形成了化学的一门独立分支学科。近红外光谱分析较早地应用了化学计量学方法。随着计算机技术和化学计量学的发展, 多元信息处理的理论与技术的发展, 解决了NIR谱区吸收弱和重叠的困难。近红外技术依据某一化学成分对近红外区光谱的吸收特性进行定量测定, 应用NIR光谱进行检测的技术关键就是在两者之间建立一种定量的函数关系, 且数学模型的建立方法是重要的研究工作。目前主要有多元线性回归(MLR)、逐步回归(SMR)、主成分分析(PCA)、主成分回归(PCR)与偏最小二乘法(PLS)、人工神经网络(ANN)等[6, 7], 其中PLS的最显著特点就是利用全部光谱信息, 选择为数不多, 且与待测质量参数相关的独立主成分变量建立回归方程, 因此具有好的分析效果。

支持向量机[8](support vector machine: SVM)方法是建立在SLT(statistic learn theory)的VC(vapnik-chervonenkis)维理论和结构风险最小原理基础上, 根据有限样本信息在模型的复杂性和学习能力之间寻求最佳折衷, 以期获得最好的推广能力。SVM目前已成功地推广到了函数逼近、信息融合等领域。最小二乘估计作为函数回归最基本的工具之一, 如果能将最小二乘问题转化为SVM形式的问题加以解决, 可以保证得到的函数具有最小的预测风险。支持向量机回归建模将低维非线性的输入映射到高维线性的输出, 模型简单, 具有良好的应用前景。SVM回归基于最小化结构风险, 而不是传统意义上的经验风险最小化, 方法的基本思想是通过一个

**收稿日期:** 2004-03-31, **修订日期:** 2004-07-28
**基金项目:** 国家高技术研究发展计划(863计划)项目(2002AA248051)(2002AA243011), "十五"国家科技攻关项目(2004BA210A03), 国家重大基础研究前期研究专项(2002CCA00800)和农业科技成果转化资金项目(02EFN216900720)资助
**作者简介:** 张录达, 1953年生, 中国农业大学理学院教授

非线性映射，将数据样本映射到高维特征空间，并在这个空间进行线性回归。由于 SVM 的理论较新，目前，多数有关支持向量机的研究仅仅局限于理论和仿真，而应用于实际较少。因此将理论应用于解决实际问题的研究具有重要意义。本文将 SVM 回归法应用于近红外光谱定量分析，建立定标模型，取得了较好的效果。

## 1　仪器与实验材料

### 1.1　仪器

实验所用仪器为 IA450 近红外光谱仪(Bran + Luebbe 公司)。该仪器具有从 1 445 nm 到 2 348 nm 之间的 19 个分隔的滤光片，如表 1 所示。

**Table 1　The wavelengths of 19 light filter glasses**

| 滤光片号 | 波长/nm | 滤光片号 | 波长/nm |
|---|---|---|---|
| 2 | 2 336 | 12 | 1 818 |
| 3 | 2 348 | 13 | 1 778 |
| 4 | 2 310 | 14 | 2 100 |
| 5 | 2 270 | 15 | 1 759 |
| 6 | 2 230 | 16 | 1 940 |
| 7 | 2 208 | 17 | 1 734 |
| 8 | 2 190 | 18 | 1 722 |
| 9 | 2 139 | 19 | 1 445 |
| 10 | 2 180 | 20 | 1 680 |
| 11 | 1 982 | | |

### 1.2　实验材料

66 个小麦样品由中国农业科学院品种资源所提供。样品由中国农业科学院品种资源所通过国标凯氏定氮法测定其蛋白质含量(称为标准值)。66 个小麦样品被碾磨成粉状，过孔径为 0.423 mm 筛，在 IA450 型近红外光谱仪上扫描近红外漫反射光谱，测定样品在 19 个分隔的滤光片对应波长处的光谱信息。

## 2　支持向量机(SVM)回归的原理[8]

用线性回归函数 $f(x)=w\cdot x+b$ 拟合数据$\{x_i, y_i\}$，$i=1,\cdots,n$，$x_i\in R^d$，$y_i\in R$ 的问题，根据 SVM 理论，若采用线性 $\varepsilon$ 不敏感损失函数

$$|f(x)-y|_\varepsilon=\begin{cases}0 & |f(x)-y|\leqslant\varepsilon\\ |f(x)-y|-\varepsilon & \text{其他}\end{cases}$$

并引入松弛因子 $\xi_i\geqslant 0$ 和 $\xi_i^*\geqslant 0$，则问题为在约束条件

$$\begin{cases}y_i-w\cdot x_i-b\leqslant\varepsilon+\xi_i\\ w\cdot x_i+b-y_i\leqslant\varepsilon+\xi_i^*\end{cases}\quad i=1,2,\cdots,n \tag{1}$$

下，最小化目标函数

$$\Phi(w,\xi_i,\xi_i^*)=\frac{1}{2}|w|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{2}$$

常数 $C>0$ 控制对超出误差 $\varepsilon$ 的样本的惩罚程度。采用优化方法可以得到其对偶问题，即在约束条件

$$\begin{cases}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ 0\leqslant\alpha_i,\ \alpha_i^*\leqslant C\end{cases}\quad i=1,2,\cdots,n \tag{3}$$

下，对 Lagrange 因子 $\alpha_i$，$\alpha_i^*$ 最大化目标函数

$$W(\alpha,\alpha^*)=-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(x_i\cdot x_j) \tag{4}$$

从而得回归模型

$$f(x)=(w\cdot x)+b=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)(x_i\cdot x)+b^* \tag{5}$$

其中 $\alpha_i$，$\alpha_i^*$ 不为 0，对应的样本就是支持向量。如果用核函数 $K(x_i, x_j)$ 替代 (4)，(5)中的内积运算就可以确定非线性拟合函数

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i\cdot x_j)+b^* \tag{6}$$

式(5)，(6)中的 $b^*$，取在边界上的一点，即可确定。

关于非线性核函数的种类较多，常用的有 Poly 函数 $k(x_i, x)=[(x\cdot x_i)+1]^{p_1}$；RBF 函数 $k(x_i, x)=e^{-\frac{|x-x_i|^2}{2p_1^2}}$，Sigmoid 函数 $k(x_i,x)=\tanh(p_1\cdot(x\cdot x_i)+p_2)$ 等多种核函数的形式。

## 3　结果与分析

本研究将 66 个小麦样品分为两组，第一组 33 个样品用于 SVM 回归建模校正；第二组 33 个样品用于样品蛋白质含量的预测。取线性 $\varepsilon$ 不敏感损失函数，采用 4 种不同核函数：linear 核函数、poly 核函数、Sigmoid 核函数与 RBF 核函数进行 SVM 回归建模，校正模型的各有关参数见表 2。以所建 SVM 回归模型对 33 个预测样品的蛋白质含量进行了预测，结果列于表 3。4 种 SVM 回归模型的预测结果与凯氏定氮法确定的蛋白质含量的化学值的相关系数均在 0.97 以上，平均绝对误差均小于 0.32。为了考察 SVM 回归校正模型的预测效果，同 PLS 回归模型预测的结果进行了比较，结果列于表 3。由表 3 表明所建的 SVM 回归模型也可进行实际样品的定量分析，且有较好的分析效果。

**Table 2　The parameters table of SVM regression**

| 参数 | Linear 核函数 | Poly 核函数 | Sigmoid 核函数 | RBF 核函数 |
|---|---|---|---|---|
| $P_1$ | | 1.3 | 0.018 | 3.95 |
| $P_2$ | | | 0.01 | |
| $C$ | $10^6$ | $10^6$ | $10^7$ | $10^7$ |
| $\varepsilon$ | 0.1 | 0.1 | 0.1 | 0.1 |
| 支持向量数 | 24(72.7%) | 23(69.7%) | 25(75.8%) | 24(72.7%) |

**Table 3 The protein content table of 33 wheat predicting samples estimated by four SVM regression models and PLS regression models**

| 序号 | 标准值 | PLS 回归预测 | Linear 核函数建模预测 | Poly 核函数建模预测 | Sigmoid 核函数建模预测 | RBF 核函数建模预测 |
|---|---|---|---|---|---|---|
| 1 | 19.34 | 19.511 35 | 19.373 7 | 18.958 5 | 19.394 7 | 18.877 4 |
| 2 | 16.49 | 16.582 79 | 16.506 4 | 16.119 | 16.605 8 | 16.287 1 |
| 3 | 17.37 | 17.065 11 | 17.551 2 | 17.296 6 | 17.341 1 | 17.346 3 |
| 4 | 14.89 | 14.735 78 | 15.434 2 | 15.175 7 | 15.045 | 15.133 2 |
| 5 | 16.85 | 16.875 39 | 16.846 6 | 16.583 7 | 16.752 5 | 16.612 1 |
| 6 | 15.38 | 15.463 97 | 15.312 | 15.017 3 | 15.368 7 | 15.065 8 |
| 7 | 18.33 | 17.966 67 | 18.294 8 | 17.946 6 | 17.935 7 | 17.975 6 |
| 8 | 13.01 | 12.766 56 | 12.934 2 | 12.587 1 | 13.137 5 | 12.633 9 |
| 9 | 14.35 | 14.204 02 | 14.658 3 | 14.391 3 | 14.432 2 | 14.436 1 |
| 10 | 13.73 | 13.885 34 | 13.778 3 | 13.261 1 | 13.712 | 13.229 4 |
| 11 | 15.07 | 15.829 02 | 15.189 2 | 14.798 5 | 15.606 7 | 14.812 4 |
| 12 | 13.74 | 13.798 43 | 14.017 3 | 13.706 2 | 14.161 7 | 13.811 9 |
| 13 | 14.71 | 14.924 14 | 14.929 4 | 14.523 7 | 14.853 2 | 14.454 4 |
| 14 | 15.03 | 14.749 3 | 15.075 6 | 14.749 1 | 14.734 | 14.681 7 |
| 15 | 16.88 | 16.601 62 | 16.997 6 | 16.790 8 | 17.054 9 | 16.927 8 |
| 16 | 16.06 | 16.075 71 | 15.804 6 | 15.486 9 | 15.918 1 | 15.499 5 |
| 17 | 15.78 | 16.147 88 | 15.828 3 | 15.497 4 | 16.168 5 | 15.503 3 |
| 18 | 16.19 | 16.717 61 | 15.738 1 | 15.527 9 | 16.639 6 | 15.735 7 |
| 19 | 14.18 | 14.587 56 | 13.825 2 | 13.421 5 | 14.405 7 | 13.614 6 |
| 20 | 18.32 | 18.807 82 | 18.715 | 18.304 1 | 18.812 4 | 18.353 8 |
| 21 | 14.9 | 15.061 32 | 14.804 5 | 14.569 5 | 15.196 8 | 14.724 8 |
| 22 | 15.18 | 15.043 29 | 15.421 7 | 15.173 5 | 15.275 6 | 15.217 2 |
| 23 | 13.09 | 13.101 26 | 13.693 3 | 13.370 9 | 13.525 3 | 13.32 6 |
| 24 | 13.32 | 13.397 67 | 13.362 5 | 13.073 9 | 13.706 1 | 13.224 1 |
| 25 | 13.54 | 14.462 07 | 14.771 | 14.361 9 | 14.557 8 | 14.281 |
| 26 | 16.08 | 16.388 95 | 16.906 7 | 16.655 6 | 16.700 5 | 16.741 4 |
| 27 | 14.56 | 14.160 57 | 14.846 3 | 14.539 3 | 14.494 2 | 14.580 5 |
| 28 | 18.48 | 17.386 92 | 17.807 1 | 17.455 | 17.300 5 | 17.479 |
| 29 | 14.26 | 14.608 98 | 14.639 7 | 14.245 3 | 14.691 | 14.227 1 |
| 30 | 15.47 | 15.837 2 | 15.736 7 | 15.371 | 15.852 3 | 15.386 4 |
| 31 | 15.08 | 14.887 12 | 15.483 3 | 15.135 4 | 15.187 8 | 15.087 |
| 32 | 13.67 | 13.516 55 | 13.780 5 | 13.308 8 | 13.527 4 | 13.207 5 |
| 33 | 16.96 | 16.568 3 | 17.634 1 | 17.395 6 | 17.290 6 | 17.454 9 |
| 与标准值的相关系数 $R$ | | 0.972 845 | 0.975 686 | 0.974 537 | 0.975 411 | 0.976 355 |
| 平均绝对误差 $S$ | | 0.294 | 0.286 | 0.317 6 | 0.298 4 | 0.294 6 |

## 4 讨 论

SVM 回归法在函数拟合方面，尚属于原理性研究。一般研究认为，和其他方法相比 SVM 具有适应性强、效率高的特点。但由于 SVM 是一种新技术，应用研究工作尚处于起步阶段，在各领域中的应用研究均属于探索性研究。本文采用了 SVM 回归法进行建模分析，将 SVM 回归技术应用于近红外光谱定量分析，也作了一次探索性的尝试。初步结果表明了这一应用的可行性，从而为化学计量学增补了新的定量分析建模算法。SVM 回归选择不同的核函数实际是对信息“主成分”(PC)不同的非线性提取方法的应用。有关这方面的研究称为 K-PCA 法，即提取非线性核-主成分也是 SVM 研究的热点。当分析体系具有非线性特征时，采用非线性 K-PCA 法建立 SVM 回归模型将有望获得好的定量分析模型，进而获得好的模型预测效果。这在原理上和应用上很有实际意义。近年来，有关近红外光谱的研究报道比较多，例如可参阅文献[9]。

## 参　考　文　献


[1] TIAN Di, JIN Qin-han(田　地, 金钦汉). Analysis Instrument(分析仪器), 2001, (3): 39.

[2] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, QIANG Dong-mei(陆婉珍, 袁洪福, 徐广通, 强冬梅). Technology of Modern NIR Spectral Analysis(现代近红外光谱分析技术). Beijing: China Petroleum Publishing House(北京: 中国石油出版社), 2000. 4.

[3] YAN Yan-lu, ZHAO Long-lian, YANG Shu-ming, et al(严衍禄, 赵龙莲, 杨曙明, 等). Foundation of NIR Spectral Analysis and Its Application(近红外光谱分析基础与应用). Beijing: Light Industry Publishing House(北京: 中国轻工业出版社), 2005.

[4] XU Guang-tong, YUAN Hong-fu, LU Wan-zhen(徐广通, 袁洪福, 陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(2): 134.

[5] SHI Yong-gang, FENG Xin-lu, LI Zi-cun, et al(史永刚, 冯新泸, 李子存, 等). Chinese Journal of Spectroscopy Laboratory(光谱实验室), 2002, 19(2): 201.

[6] YU Ru-qin(俞汝勤著). Introduction to Chemometrics(化学计量学导论). Changsha: Hunan Education Publishing House(长沙: 湖南教育出版社), 1991. 5.

[7] REN Rui-xue, TANG Zhen, LIU Fu-qiang, GOU Yu-hui, REN Yu-lin(任瑞雪, 汤　真, 刘福强, 苟玉彗, 任玉林). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2001, 21(4): 521.

[8] Vapnik V N. The Nature of Statistical Learning Theory. NY: Springer-Verlag, 1995.

[9] WANG Duo-jia, ZHOU Xiang-yang, JIN Tong-ming, et al(王多加, 周向阳, 金同铭, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(4): 447.


# Applied Study on Support Vector Machine (SVM) Regression Method in Quantitative Analysis with Near-Infrared Spectroscopy


ZHANG Lu-da[1], JIN Ze-chen[2], SHEN Xiao-nan[1], ZHAO Long-lian[2], LI Jun-hui[2], YAN Yan-lu[2]

1. College of Science, China Agricultural University, Beijing　100094, China

2. College of Information, China Agricultural University, Beijing　100094, China



**Abstract**　This paper introduced the application of support vector machines(SVM) regression method based on statistics study theory to the quantitative analysis with near-infrared (NIR) spectroscopy. Sixty-six wheat samples were used as experimental materials, and thirty-three of them were used as calibration samples. The protein contents and NIR spectra of the calibration samples were used to build SVM regression models by four different kernel functions. The protein content of the predicting samples are estimated by four different SVM regression models. All of the correlation coefficients between the estimated values by different SVM regression models and the standard chemical values of protein content by Kjeldahl's method are more than 0.97. The average absolute error is less than 0.32. To investigate the predicting effect, it is compared with PLS regression models. The result suggested that the SVM regression, which was built to estimate the protein content of wheat samples, can also be used in the quantitative analysis of real samples by NIR.



**Keywords**　Support vector machine regression; Near-infrared spectroscopy; Quantitative analysis

(Received Mar. 31, 2004; accepted Jul. 28, 2004)