第 26 卷 第 6 期
2006 年 6 月

光 学 学 报
ACTA OPTICA SINICA

Vol. 26, No. 6
June, 2006

文章编号：0253-2239(2006)06-0933-5

基于支持向量机的近红外光谱鉴别茶叶的真伪*

陈全胜[1] 赵杰文[1] 张海东[1,2] 王新宇[1]

(1 江苏大学食品与生物工程学院，镇江 212013
2 云南农业大学工程技术学院，昆明 650201)

摘要： 快速准确地鉴别名优茶的真伪是当前茶叶行业亟待解决的一项重大课题。针对这一现状，提出了一种快速准确鉴别名优茶真伪的新思路。试验中，以碧螺春茶为研究对象，利用近红外光谱分析技术结合支持向量机(SVM)模式识别原理建立碧螺春茶真伪鉴别模型。试验结果显示，通过标准归一化(SNV)预处理，选取 6500～5500 cm^{-1}波长范围内的光谱经过主成分分析后，提取 11 个主成分，选用径向基函数(RBF)作为核函数建立的模型最佳。对训练集中的 138 个茶叶样本，模型的回判鉴别率达到 93.48%；对 90 个独立样本进行预测时，模型的预测鉴别率达到 84.44%。研究结果表明基于支持向量机的近红外光谱鉴别名优茶真伪的方法是可行的。

关键词： 光谱学；近红外光谱；支持向量机；鉴别；茶叶

中图分类号：O657.33　　文献标识码：A

Indentification of Authenticity of Tea with Near Infrared Spectroscopy Based on Support Voctor Machine

Chen Quansheng[1] Zhao Jiewen[1] Zhang Haidong[1,2] Wang Xinyu[1]

(1 *School of Food and Biological Engineering, Jiangsu University, Zhenjiang* 212013
2 *College of Engineering and Technology, Yunnan Agricultural University, Kunming* 650201)

Abstract: It is urgent to think up a quick and precise method in the discrimination of the famous tea. A new discriminate method of tea is proposed. In this study, Biluochun tea serves as the target and the model for discriminating the authenticity of tea is built up using near-infrared spectroscopy combined with the pattern recognition of support vector machine (SVM). The experimental result shows that in the spectra region between 6500 cm^{-1} and 5300 cm^{-1}, the best model is built by the standard normal variate (SNV) preprocessing, when 11 principal components are selected and the radial basis function (RBF) is used as the kernel function. The discriminate rate of the model for 138 samples in training set is 93.48%, and 84.44% for 90 samples in predicting set. The research shows that it is feasible to apply near-infrared spectroscopy to discriminate the authenticity of the famous tea based on SVM.

Key words: spectroscopy; near-infrared spectroscopy; support vector machine (SVM); identification; tea

1 引　　言

传统的茶叶鉴别方法是感官评定法和化学方法。其中，感官评定的结果受人为因素和外界环境的干扰很大，影响到结果的客观性；化学方法虽然能够准确地鉴别茶叶，但是繁琐的步骤和昂贵的费用使它不能应用到茶叶的快速识别上。近红外漫反射光谱(NIR)分析具有速度快、成本低以及结果重现性好等优点[1]。国内外学者先后利用近红外光谱方法定性和定量地分析了茶叶中蛋白质、咖啡碱、氨基酸、多酚类以及水分的含量[2~4]，但是近红外光谱分析技术在茶叶真伪的快速鉴别上的应用研究还很少。

支持向量机(Support vector machine)是 20 世纪 90 年代形成的一种新的模式识别方法，它已表现出许多优于其它模式识别方法的性能。支持向量机

* 国家自然科学基金(30370813)和国家术 863 计划(2002AA248051)资助课题。

作者简介：陈全胜(1973～)，男，安徽桐城人，江苏大学博士研究生，主要从事近红外光谱分析与图像处理研究工作。E-mail：chenjiang0518@yahoo.com.cn

收稿日期：2005-07-18；收到修改稿日期：2005-10-28

方法将待解决的模式识别问题转化成为一个二次规划寻优问题，在理论上，保证了全局最优解，避免了局部收敛现象。近红外光谱结合支持向量机的模式识别方法在中草药[4]和石油[6]的分类和鉴别上得到了成功的应用。鉴于此，本文提出了基于支持向量机的近红外光谱快速鉴别茶叶真伪这一思路。

2 支持向量机原理[7,8]

2.1 线性可分问题

支持向量机的理论最初来自两类线性可分的数据处理。设 X 为输入空间，Y 表示输出域，通常模式集合 $X=\{x_i\}\in R^n$ 由两类点组成，即 $Y=\{-1,1\}$。对于 n 个样本组成的训练集

$$S=[(x_1,y_1),\cdots,(x_n,y_n)]\subseteq(X\times Y)^n,$$

根据结构风险最小原理，构造一个目标函数，寻找一个满足要求的分割超平面，并使训练集中的点距离分割超平面尽可能的远。在二维空间中如图 1 所示。其中，实心圆和和空心圆分别代表两类样品，H 为最优分类超平面，H_1，H_2 分别为过各类样本中离分类超平面最近且平行于分类超平面的超平面。如果将分类超平面记为(1)式，则归一化之后，对于线性可分样本集 (x_i,y_i) 的分类超平面应满足(2)式，此时，H_1，H_2 上的训练样本点称为支持向量。

$$w\cdot x_i+b=0, \tag{1}$$

$$y_i(\langle w\cdot x\rangle+b)-1\geqslant 0, \tag{2}$$

其中 w 是垂直于分类超平面的法向量；margin 等于 $2/\|w\|$ 为区域间隔距离。

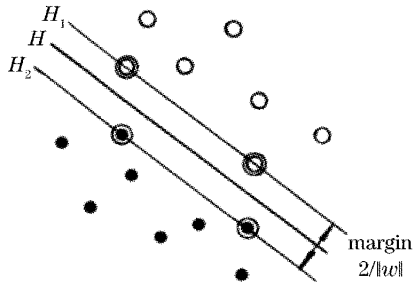

图 1 线性可分情况下的最优分类面

Fig. 1 Optimal plane of linearly separable samples

2.2 非线性可分问题

当问题为线性不可分时，可以利用核函数 $K(x_i,x)$ 实现非线性变换，将线性不可分问题转换为另一个高维空间中的线性可分问题，并在该高维空间中寻求最优分类面。当然，采用不同的核函数，所得到的分类结果也不一样。目前，研究较多的核函数有以下三种形式：

1) 多项式(Poly)形式的核函数：

$$K(x_i,x)=[(x_i\cdot x)+1]^\delta, \tag{3}$$

2) 径向基函数(RBF)形式的核函数：

$$K(x_i,x)=\exp\left(-\frac{\|x-x_i\|}{2\delta^2}\right), \tag{4}$$

3) Sigmoid 函数形式的核函数：

$$K(x_i,x)=\tanh[\delta(x_i\cdot x)+\gamma]. \tag{5}$$

3 材料与方法

3.1 试验材料

试验中所用的正品碧螺春样本的出产日期基本一致，都为 2004 年 5 月份，初产地为江苏；伪品碧螺春样本是模仿碧螺春的加工工艺制作而成的，所有伪品碧螺春样本的出产日期集中在 2004 年 5～7 月份，初产地分别为安徽、湖南和江西等地。为使取样均匀，试验前先将每一种茶叶分别用咖啡粉碎机粉碎并过 40 目的筛，然后在每一个品种的茶叶中，按照四分法原则，随机称取 5g 作为一个样本。一共选取 228 个茶叶样本，其中，138 个样本作为训练集包括 30 个正品碧螺春样本，剩余的 90 个样本作为预测集包括 25 个正品碧螺春样本。

3.2 试验方法

试验所用的近红外检测系统主要是近红外光谱仪(Nexus 670 FT-IR，美国 Nicolet 公司)。扫描范围：11000～3800 cm^{-1}；扫描次数：64 次；分辨力：4 cm^{-1}。试验时，保持室内的温度和湿度基本一致，将样本倒入样品杯中，充分压实。每一个样本在不同时间，不同位置分别采集 4 次，取 4 次采集的平均值作为该样本的原始光谱数据。

所有的数据分析都是基于 TQ Analysis V6 (Nicolet 近红外系统自带)、Matlab V6.5 的软件平台。

4 结果与讨论

4.1 光谱的预处理

试验中，虽然茶叶样本是经过粉碎后过 40 目筛得到的，但是样本粒径的大小和均匀度不能保证完全一致，这些都对近红外光的漫反射产生一定影响；同时样本的密实度也影响了光在茶叶样本中的传播。因此，需要对样本的原始光谱数据进行预处理。光谱的预处理方法很多，本试验运用了多元散射校正(MSC)、标准归一化(SNV)、一阶导数和二阶导数等 4 种预处理方法，通过分析对比发现，标准归一化的预处理效果明显优于一阶导数和二阶导数，略

优于多元散射校正。因此，本试验最终采用了标准归一化预处理方法。

4.2 光谱区域的选择

虽然真伪碧螺春茶的外形可以达到完全相似。但是它们内部多数有机物的（如多酚类、植物碱类、氨基酸、蛋白质以及纤维素等）含量与比例受到诸如地理位置、气候土壤状况以及采摘时间等因素的影响。因此，在内部有机物含量及比例上，真伪碧螺春茶总是存在一定的差别。而这些有机物的含氢基团（如 C—H、O—H、S—H 和 N—H 等）在近红外区域都能产生倍频与合频吸收，它们的一级倍频近红外光谱带位于 7200～5500 cm^{-1} 处；二、三、四级倍频位于 12800～8300 cm^{-1} 处；合频位于 5000～4000 cm^{-1} 处[9]。如果茶叶内部有机物含量和比例不同，那么它们在近红外光谱上就表现出不同的吸收信号，本文正是基于这一原理，利用支持向量机模式识别方法对真伪碧螺春茶进行鉴别的。

图 2 是正品和几种伪品碧螺春样本的原始光谱图 2(a)和一阶导数光谱图 2(b)，从图 2(a)中可以看出茶叶的原始近红外光谱在 5155 cm^{-1} 和 6944 cm^{-1} 附近有一个明显的吸收峰，一阶导数光谱附近有明显的波动。因为纯水中的 O—H 伸缩振动的一级倍频位于6944 cm^{-1} 附近，它的一个合频区位于 5155 cm^{-1} 附近，在这两个波长附近是水分吸收的敏感区[9]。从图 2 中可以看出在这两个区域，干茶中的水分对近红外光谱的吸收峰影响很大。试验中，干茶的含水率在 5%左右，为了减少水分的影响，分析时，选择光谱波长范围尽量避开水分吸收峰的特征波长区。本文有比较地选用了各段的波长进行了分析，结果显示在一级倍频区选用 6500～5300 cm^{-1} 范围内的光谱数据既避开了水分的影响又取得了较好的试验结果。

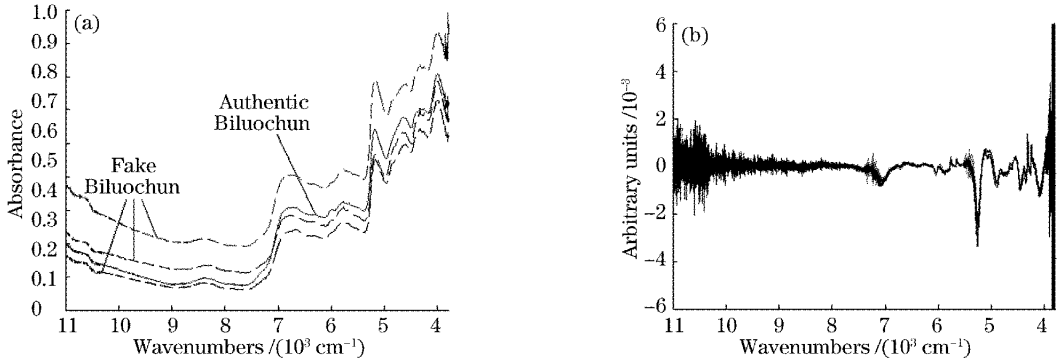

图 2　茶叶的原始光谱(a)和一阶导数光谱(b)

Fig. 2　Raw spectra (a) and the first derivative spectra (b) of tea

4.3 模型的建立

4.3.1 核函数类型和参量的确定

支持向量机建立鉴别模型优先解决的问题是核函数的选择。如前所述，常见的核函数有多项式、径向基和 Sigmoid 函数等三种类型，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，用径向基往往能够得到较好的拟合结果，因为径向基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[8]，另外径向基取值($0<K\leqslant 1$)要比多项式取值($0<K$ 或 $\infty>K>1$)简单。对于 Sigmoid 核函数，计算速度明显低于前两种，在实际中很少应用，因此本试验用径向基作为支持向量机的核函数。确定了核函数后，就要对核函数的参量进行优化，核函数参量对模型也会产生一定的影响，径向基核函数所要确定的参量有惩罚系数 C 和核函数的宽度参量 δ，对于这些参量的选择，目前尚无比较成熟的方法，一般要靠多次试验确定[7]。本实验通过多次尝试比较发现，在惩罚系数 $C=100$，径向基核函数的宽度参量 $\delta=0.5$ 的条件下，建立的模型识别效果最佳。

4.3.2 主成分数的确定

从前面的分析可知，茶叶中许多有机物的含氢基团都能在近红外区域产生倍频与和频的吸收，因此，茶叶样本的近红外光谱数据间存在大量的相关性，造成大量的信息冗余。在建立模型中，这些冗余信息的介入会使模型的预测性能降低。主成分分析（PCA）是把多个指标化为几个综合指标的一种统计方法，它沿着协方差最大方向由多维光谱数据空间向低维数据空间投影，各主成分向量之间相互正交。通过选择合理的主成分既可以避免建模中的信息冗余，又不会过多地丢失光谱信息，同时在分析数据中也达到简化的目的。模型在训练过程中，主成分数的多少对模型

的预测性能有一定的影响。

试验中，将训练和预测时的误判数作为衡量模型优劣的一个指标，主成分数对模型的影响如图 3 所示。从图 3 可以看出，模型在主成分数等于 11 之前，训练和预测时模型的误判数都随着主成分数的增加而减少。但是，当主成分数增加到 11 以后，再随着主成分数的增加，训练时模型的误判数基本不变，而在预测时误判数却有上升趋势。因为主成分数达到 11 时，此时的累计方差贡献率为 99.86%，这 11 个主成分因子已经几乎能全部反映光谱的总体信息。因此，由 11 主成分因子建立的模型最佳。

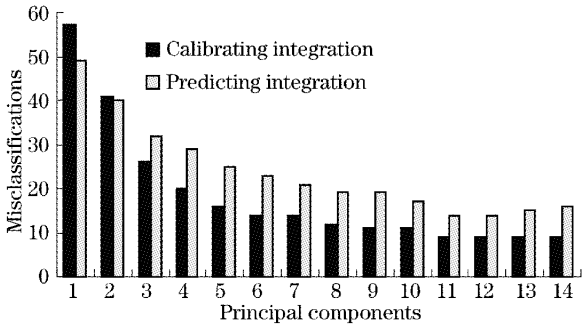

图 3 训练和预测时模型的误判数与主成分数的关系

Fig. 3 Relationship between misclassifications and principal components in the training and predicting course

4.4 模型的训练和预测

根据以上的分析，选用 6500～5300 cm^{-1} 范围内的光谱数据，经过标准归一化预处理，提取 11 个主成分因子，以 138 个样本的 11 个主成分信息构成 $X_{138\times11}$ 的矩阵作为支持向量机的输入因子。在训练算法的设计过程中，根据 4.3.1 节的分析，本文选用径向基函数作为支持向量机的核函数，在惩罚系数 $C=100$，核函数的宽度参量 $\delta=0.5$ 的条件下，进行训练以建立鉴别模型。该模型对训练集样本的回判鉴别率达到 93.48%，对预测集样本的预测鉴别率达到 84.44%。同时，为了说明支持向量机模型的优越性，将其与常规的贝叶斯(Bayes)判别分析模型以及反向传播神经网络(BP-ANN)模型相比较，其结果如表 1 所示。从表 1 中可以看出，支持向量机和反向传播神经网络模型的训练结果优于贝叶斯判别模型。造成这种现象的原因是因为支持向量机和反向传播神经网络同属于非线性识别模型，而贝叶斯判别模型为线性识别模型，又因为非线性模型在训练过程中自学习能力明显强于线性模型，所以从训练的结果看贝叶斯判别模型的回判鉴别率仅有 73.19%，明显低于支持向量机模型(93.48%)和反向传播神经网络模型(94.93%)。尽管反向传播神经网络的训练结果与支持向量机相差不大甚至略高它，但是从预测结果看，支持向量机模型(84.44%)明显优于反向传播神经网络模型(63.33%)。因为反向传播神经网络采用经验最小化原则，在样本数据有限时，容易在训练过程中产生局部最优解，即“过训练”现象，使训练的结果远好于预测的结果[10]。支持向量机模型是基于结构风险最小化原则建立模型，其优点在于同时兼顾了经验风险和置信范围这两个最小化风险的泛函问题，训练得到的结果通常也是全局最优解[11]。因此，支持向量机模型要好于其它两种模型。

表 1 基于支持向量机和其它判别模型的训练和预测结果

Table 1 Results of training and predicting based on SVM and other discriminate model

Discriminate models	Principal component factors	Discriminate rate in training /%	Discriminate rate in test /%
SVM	11	93.48	84.44
Bayes	8	73.19	68.89
BP-ANN	11	94.93	63.33

5 结 论

本文对支持向量机的基本原理进行了简单地介绍，并在主成分分析的基础上，利用基于径向基核函数的支持向量机模式识别原理，建立了碧螺春茶真伪的近红外光谱鉴别模型，该模型基本能正确鉴别碧螺春茶叶的真伪，对模型的训练和预测结果进行了分析性的说明。试验结果表明基于支持向量机模式识别原理，利用近红外光谱分析技术鉴别名优茶的真伪这一思路是可行的。同时为支持向量机模式识别原理在近红外光谱分析中的进一步应用奠定了基础。

参 考 文 献

1 Wenliang Chen, Rong Liu, Houxin Cui *et al.*. Application of transcutaneous diffuse reflectance spectroscopy in the measurement of blood glucose concentration[J]. *Chin. Opt. Lett.*, 2004, **2**(7): 411～413

2 J. Lupaert, M. H. Zhang, D. L. Massart. Feasibility study for the using near infrared spectroscopy in the qualitative and quantitative of green tea, Camellia sinensis (L)[J]. *Analytica Chemica Acta*, 2003, **487**(2): 303～312

3 M. H. Zhang, J. Lupaert, Q. S. Xu *et al.*. Determination of total antioxidant capacity in green tea by NIRS and multivariate calibration [J]. *Talanta*, 2004, **62**(1): 25～35

4 H. Schulz, U. H. Engelhardt, A. Wengent *et al.*. Application of NIRS to the simultaneous prediction alkaloids and phenolic substance in green tea leaves[J]. *J. Agric Food Chem.*, 1999, **47**(12): 5064～5067

5 Zhang Luda, Su Shiguang, Wang Laisheng *et al.*. Study on application of Fourier transformation near-infrared spectroscopy

analysis with support vector machine (SVM)[J]. *Spectroscopy and Spectral Analysis*, 2005, **25**(1): 33～35
张录达,苏时光,王来生 等. 支持向量机在傅立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

6 Yao Xiaogang , Dai Liankui , Fang Jun. Diesel cetane number measurement with NIR spectral analysis using LS-SVM[J]. *Control and Instruments in Chemical Industry*, 2004, **31**(2): 48～51 (in Chinese)
姚肖刚,戴连奎,方　骏. 基于支持向量机的柴油十六烷值近红外光谱测量方法[J]. 化工自动化及仪表, 2004, **31**(2): 48～51 (in Chinese)

7 Nello Cristianini, John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* [M]. Li Guozheng, Wang Meng, Zeng Huajun transl. Beijing: Publishing House of Electronic Industry, 2004. 8～149 (in Chinese)
Nello Cristianini, John Shawe-Taylor. 支持向量机导论[M]. 李国正, 王　猛, 曾华军 译. 北京: 电子工业出版社, 2004. 8～149

8 Ye Meiying, Wang Xiaodong. Identification of chaotic optical system based on support vector machine[J]. *Acta Optica Sinica*, 2004, **24**(7): 953～956 (in Chinese)
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

9 Lu Wanzhen, Yuan Hongfu, Xu Guangtong *et al*.. *The Modern Analysis Technique for Near-Infrared Spectra* [M]. Beijing: Chinese Oil and Chemical Press, 2000. 19～36 (in Chinese)
陆婉珍, 袁洪福, 徐广通 等. 现代近红外光谱分析[M]. 北京: 中国石化出版社, 2000. 19～36

10 Qi Feng , Liu Wenqing , Zhou Bin *et al*.. Improving DOAS system measurement precision with artificial neural networkmethod[J]. *Acta Optica Sinica*, 2002, **22**(11): 1345～1349 (in Chinese)
齐　锋, 刘文清, 周　斌 等. 利用人工神经网络方法提高差分光学吸收光谱系统测量精度研究[J]. 光学学报, 2002, **22**(11): 1345～1349

11 Meiying Ye. Improving linearity of position-sensitive detector using support vector machines[J]. *Chin. Opt . Lett.*, 2005, **3**(4): 205～207