DOI:10.13526/j.issn.1006-6144.2014.03.007

近红外光谱结合 LSTSVM 算法判别不同等级烟叶

宋相中[1],赖衍清[1],李祖红[2],郑 波[2],李倩倩[1],吴丽君[1],
张录达[1],熊艳梅*[1],闵顺耕*[1]
(1.中国农业大学理学院,北京 100193;
2.云南省烟草公司曲靖市公司,云南曲靖 655000)

摘 要:本文用近红外光谱结合最小二乘双胞胎支持向量机(LSTSVM)算法建立了烟叶等级分类模型。从三个等级共 210 个烟叶样品中,取出 120 个样品作为建模集,剩余 90 个样品作为预测集。为了建立最优模型,对光谱预处理方法和模型参数进行筛选优化,最优模型对预测集样品的平均识别率为 95.56%,结果表明该方法可以作为烟叶等级分类的一种有效方法。此外,将该算法与 SIMCA、PLS-DA、SVM 等三种常见的模式识别算法进行了比较,结果表明基于样品的原始光谱,同等条件下,LSTSVM 算法的预测效果优于其他三种算法。

关键词:近红外光谱;最小二乘双胞胎支持向量机算法;烟叶;等级分类
中图分类号:O657.33 **文献标志码**:A **文章编号**:1006-6144(2014)03-327-05

在烟草行业中,烟叶等级的分类工作是一项具有重要意义的工作,烟叶等级不同,其品质存在着较大的差异。近红外光谱作为一种无需样品前处理的分析技术被广泛应用于烟草行业中,近年来近红外光谱结合模式识别算法也开始越来越多地应用于烟叶等级分类工作中。杜文等利用近红外光谱结合 SIMCA 算法建立了烟叶的产地、等级以及品种识别模型,取得了较为理想的识别效果[1]。束茹欣等报道了 NIR-PCA-SVM 联用技术在烤烟烟叶产地模式识别中的应用情况,结果表明 NIR-PCA-SVM 联用技术可成功识别烟叶样品的产地[2]。

最小二乘双胞胎支持向量机(Least Square Twins Support Vector Machine,LSTSVM)算法是在双胞胎支持向量机(Twins Support Vector Machine,TSVM)算法的基础上开发出的一种新算法,与传统的支持向量机算法相比,LSTSVM 算法具有计算速度较快的优点[3]。支持向量机算法一般适用于建立两类分类模型,目前尚未有将 LSTSVM 算法应用于多分类工作中的报道。本文采用 LSTSVM 算法结合"one VS the rest"(一对其余)思想建立了不同等级烟叶的近红外多分类模型[4]。为了评价 LSTSVM 算法建立烟叶等级识别模型的优劣,还将它与其他几种算法进行了比较,包括 SIMCA、PLS-DA 和 SVM 算法。

1 LSTSVM 算法原理

LSTSVM 算法首先将输入的数据点通过核函数映射到更高维的特征空间,在这个特征空间中,对应于非线性划分的平面执行一个线性分类的过程[5]。

$$K(x',C')\mu^{(1)}+\gamma^{(1)}=0 \text{ 和 } K(x',C')\mu^{(2)}+\gamma^{(2)}=0 \quad (1)$$

其中,$C=\begin{bmatrix}A\\B\end{bmatrix}$,$A$、$B$ 分别为两类样品光谱矩阵,K 是一个任意的人工核函数。对应于公式(1)得到的最小

收稿日期:2013-03-19 修回日期:2013-24-24
基金项目:中国烟草公司资助项目(No.2010YN65)
* 通讯作者:闵顺耕,男,硕士,教授,博士研究生导师,主要从事近红外/红外光谱分析、化学计量学等方面的研究工作.
熊艳梅,女,博士,副教授,硕士研究生导师,主要从事近红外光谱分析方面的研究工作.

二乘双胞胎分类支持向量机的方程表达式为:

$$\underset{\mu^{(1)},\gamma^{(1)}}{\text{Min}}\ \frac{1}{2}\|K(A,C')\mu^{(1)}+e\gamma^{(1)}\|^2+\frac{c_1}{2}\|y\|^2,\ s.t.\ -(K(B,C')\mu^{(1)}+e\gamma^{(1)})+y=e \tag{2}$$

$$\underset{\mu^{(2)},\gamma^{(2)}}{\text{Min}}\ \frac{1}{2}\|K(B,C')\mu^{(2)}+e\gamma^{(2)}\|^2+\frac{c_2}{2}\|y\|^2,\ s.t.\ K(A,C')\mu^{(2)}+e\gamma^{(2)}+y=e \tag{3}$$

其中,c_1、c_2 为待优化参数。将约束条件代入目标函数,解出两个超平面系数 $\mu^{(1)}$、$\gamma^{(1)}$、$\mu^{(2)}$、$\gamma^{(2)}$。计算出每个样本点分别距两个超平面的距离,根据距离大小对样品的归属进行分类。样品距离哪一个超平面的距离小,就属于该超平面所代表的那一类样品集。

2　材料与方法

2.1　仪器与样品

实验所用仪器为 MATRIX-I 型傅里叶变换近红外光谱仪(德国,Bruker 公司),带有漫反射积分球和样品旋转器采样附件。

210 个烟叶样品收集于 2009 年,由曲靖烟草公司提供。烟叶样品按照外观形态可分为上部叶(B2F)、中部叶(C3F)和下部叶(X2F)三个等级,每个等级分别有 70 个样品。所有样品在 40 ℃烘箱中干燥 0.5 h,粉碎后,过 60 目筛。

2.2　光谱采集

光谱扫描范围为 4 000～10 000 cm^{-1},光谱分辨率为 4 cm^{-1},重复扫描次数为 64 次,参比采用仪器内置背景,测量过程中温度、湿度等环境条件尽量保持一致。

2.3　数据处理软件

LSTSVM 算法和几种预处理算法均为作者自行采用 MATLAB 软件(Ver. R. 2012. a, The Math Work, USA)编写完成。SVM 算法采用台湾大学林智仁教授在互联网上公开的免费工具箱(LibSVM-3. 14)完成[6]。SIMCA 和 PLS-DA 算法则采用 The Unscrambler 软件(Ver. 9. 7, CAMO(Computer Aided Modelling, Trondheim, Norway))完成。

3　结果与讨论

3.1　烟叶样品光谱分析

三个等级烟叶样品的原始近红外光谱与样品的平均光谱分别如图 1 中(a)、(b)所示,由图 1 可以发现下部叶(X2F)样品的近红外光谱与其他两类烟叶样品有明显的差异,而上部叶(B2F)和中部叶(C3F)样品光谱重叠严重,其平均光谱也几乎叠合在一起。可能的原因是同一株烟草植株的上部叶(B2F)和中部叶(C3F)的化学成分差异本身较小,加之不同植株生长状况不同,导致不同植株间的上部叶(B2F)和中部叶(C3F)的化学成分差异更加不显著。

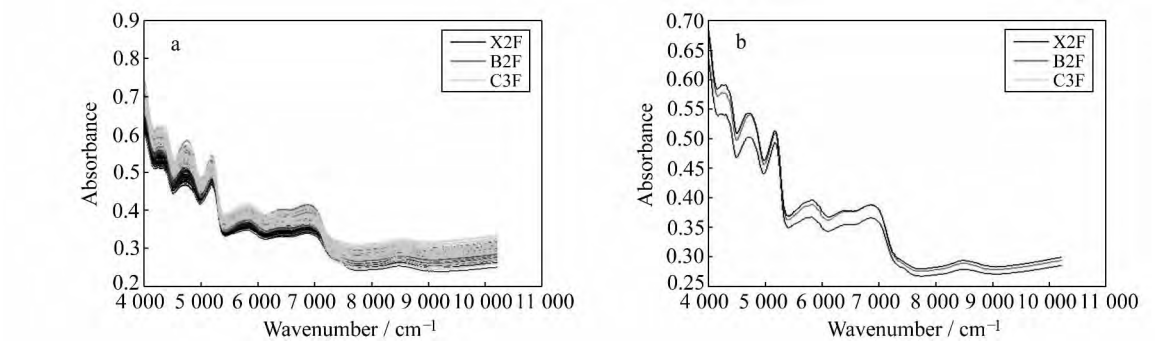

图 1　三个等级烟叶样品的原始光谱(a)和平均光谱(b)

Fig. 1　Original spectra(a)and average spectra (b)of tobacco leaf samples with three grades

3.2　光谱预处理方法筛选

近红外光谱包含了样品大量的物理、化学信息,但由于信号重叠、强度较弱以及噪声干扰等因素,也在

一定程度上降低了近红外光谱的可利用性，因此有必要对采集得到的近红外光谱进行预处理，以提高模型的预测性能。常用的光谱预处理方法有平滑、一阶导数、二阶导数、矢量归一化(Standard Normal Variate Tranformation，SNV)、多元散射校正(Multiplicative Scatter Correction，MSC)等[7-9]。本文对上述预处理算法进行了筛选，各种预处理方法建立模型的结果如表 1 所示。

3.3　核函数选择

在支持向量机方法中，可供选择的核函数 $K(x_i,x)$ 主要有多项式(Poly)函数、径向基函数(RBF)、Sigmoid 函数等形式，采用不同的核函数会产生不同的支持向量机算法[10]。本文采用径向基函数(RBF)作为核函数，即核函数为：

$$K(x_i,x)=e^{-\frac{\|x_i-x\|^2}{2p^2}} \tag{4}$$

式(4)中 p 为参数。

3.4　参数优化

从三个等级烟叶样品，取出一个等级烟叶作为 $+1$ 类样品，剩下的两个等级烟叶作为 -1 类样品。随机从两类样品中分别选出 40 和 80 个样品作为训练集，剩余的 90 个样品作为检验集。将训练集样品平均划分为五组，每次留一个组作为检验集，交叉验证，优化建模参数 c_1、c_2、p，参数的优化范围均为 2^{-9} 到 2^9。利用最优参数结合全部训练集样品建立第一个等级样品的识别模型，并对检测集 90 个样品的归属进行预测，将第一个等级的样品与其余两个等级样品区分开。依此类推，可以分别建立了另两个等级的识别模型。统计未知样品的归属正确率作为评价该模型的指标，利用优化出的最优参数建立模型结果如表 1 所示。

表 1　光谱预处理方法筛选和参数优化

Table 1　Optimization of spectral pretreatment methods and modeling parameters

Spectral pretreatment methods	Grades	Modeling parameters	Recognition number	Test set number	Recognition accuracy(%)	Average recognition accuracy(%)
No	X2F	$c_1=0.0625, c_2=0.0625, p=4$	22	30	73.33	77.78
	C3F	$c_1=0.125, c_2=0.125, p=4$	24	30	80.00	
	B2F	$c_1=1, c_2=0.0625, p=0.5$	24	30	80.00	
Savitzky-Golay smoothing,	X2F	$c_1=0.0625, c_2=1, p=16$	24	30	80.00	87.78
9 points	C3F	$c_1=0.0625, c_2=0.5, p=4$	26	30	86.67	
	B2F	$c_1=0.0625, c_2=0.5, p=16$	29	30	96.67	
First order derivative,	X2F	$c_1=0.25, c_2=0.25, p=1$	29	30	96.67	95.56
9 points	C3F	$c_1=0.125, c_2=0.0625, p=0.5$	27	30	90.00	
	B2F	$c_1=0.125, c_2=2, p=1$	30	30	100.00	
Second order derivative,	X2F	$c_1=0.125, c_2=0.125, p=0.125$	29	30	96.67	85.56
9 points	C3F	$c_1=0.0625, c_2=4, p=0.0625$	28	30	93.33	
	B2F	$c_1=0.125, c_2=0.0625, p=0.125$	20	30	66.67	
SNV	X2F	$c_1=0.0625, c_2=0.25, p=8$	19	30	63.33	63.33
	C3F	$c_1=0.125, c_2=0.5, p=16$	23	30	76.67	
	B2F	$c_1=16, c_2=0.0625, p=1$	15	30	50.00	
MSC	X2F	$c_1=0.0625, c_2=0.0625, p=16$	23	30	76.67	77.78
	C3F	$c_1=0.0625, c_2=0.125, p=2$	24	30	80.00	
	B2F	$c_1=0.125, c_2=0.5, p=16$	23	30	76.67	

3.5　模型结果

从表 1 中首先可以发现，采用不同的预处理方法，LSTSVM 算法建模优化出的最优参数存在着较大的差异。这是因为采用不同预处理方法之后，样品点在特征空间的分布情况发生了变化。同时还可以发现，采用的预处理方法不同，模型识别率差异较大。卷积平滑、一阶导数、二阶导数等预处理方法可以明显提高模型识别率，采用多元散射校正的模型识别率则没有显著性变化，而采用矢量归一化的模型识别率变

得更差。本实验中最优的光谱预处理方法为一阶导数,最优参数为 X2F($c_1=0.25$,$c_2=0.25$,$p=1$),B2F($c_1=0.125$,$c_2=2$,$p=1$),C3F($c_1=0.125$,$c_2=0.125$,$p=0.125$)。利用最优模型对各等级未知样品的识别率分别为 X2F(29/30,96.67%),B2F(27/30,90.00%),C3F(30/30,100.00%),平均识别率为 95.56%。

3.6　模型评价

为了评价采用 LSTSVM 算法建立烟叶等级识别模型的优劣,本研究将 LSTSVM 算法与几种常见的模式识别分类算法(SIMCA、PLS-DA、SVM 算法)进行了比较[11-13]。各种算法采用的光谱均为原始光谱,训练集和检验集样品划分方式也完全一致,四种算法建模的结果如表 2 所示。结果表明,采用 SIMCA 、PLS-DA 和 SVM 算法建立的识别模型,对下部叶(X2F)的识别效果较好,对其他两个等级烟叶的识别效果却不是十分理想,而采用 LSTSVM 算法建立的识别模型对于三个等级的烟叶样品均能取得较为理想识别效果。因此,基于烟叶的原始光谱,采用 LSTSVM 算法建立识别模型相对于其他三种算法具有一定的优势。

表 2　四种模式识别算法比较

Table 2　Comparison of four pattern recognition algorithms

Methods	Grades	Recognition number	Test set number	Recognition accuracy(%)	Average recognition accuracy(%)
LSTSVM	X2F	22	30	73.33	77.78
	C3F	24	30	80.00	
	B2F	24	30	80.00	
SVM	X2F	29	30	96.67	67.78
	C3F	11	30	36.67	
	B2F	21	30	70.00	
PLS-DA	X2F	26	30	86.67	76.67
	C3F	23	30	76.67	
	B2F	20	30	66.67	
SIMCA	X2F	17	30	56.67	28.89
	C3F	1	30	3.33	
	B2F	8	30	26.67	

4　结论

本研究采用一种新近发展的 LSTSVM 算法结合“one VS the rest”(一对其余)思想,建立了烟叶等级的近红外光谱识别模型,并成功实现了不同等级烟叶的判别。通过光谱预处理方法的筛选,发现光谱预处理方法的选用对识别模型结果的影响极为显著。本实验中最优的光谱预处理方法为一阶导数。将 LSTSVM 算法与几种常见的模式识别算法(SIMCA、PLS-DA、SVM 算法)进行了比较,发现基于烟叶原始光谱建立烟叶等级识别模型,LSTSVM 算法具有一定的优势。

参考文献:

[1] DU Wen(杜　文), YI Jian-hua(易建华), TAN Xin-liang(谭新良), LIU Jin-yun(刘金云). Acta Tabacaria Sinica(中国烟草学报)[J], 2009, **5**: 1.

[2] SHU Ru-xin(束茹欣), SUN Ping(孙　平), YANG Kai(杨　凯), ZHANG Jian-ping(张建平), LIU Tai-ang(刘太昂). Tobacco Science & Technology(烟草科技)[J], 2011, **292**: 50.

[3] SONG Xiang-zhong(宋相中), CHEN Chang-zhou(陈昌洲), MIN Shun-geng(闵顺耕), HE Xiong-kui(何雄奎), LI Zheng(李　铮), MI Jin-rui(米津锐), ZHANG Lu-da(张录达). Chinese Journal of Analytical Chemistry(分析化学)[J], 2012, **6**: 950.

[4] Brereton Richard G, Lloyd Gavin R. Analyst[J], 2010, **135**: 230.

[5] Arun Rumar M, Gopal M. Expert Systems with Applications[J], 2009, **36**: 7535.

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm(台湾大学林智仁教授在互联网上公开的免费工具箱(libsvm)网址).

[7] Arruabarrena J, Coello J, Maspoch S. Journal of Pharmaceutical and Biomedical Analysis[J], 2012, **60**: 59.

[8] ZHAO Ling-zhi, DOU Ying, MI Hong, REN Mei-yan, REN Yu-lin. Spectrochimica Acta Part A[J], 2007, **66**: 1327.
[9] LI Hua(**李　华**), WANG Ju-xiang(**王菊香**), GUO Heng-guang(**郭恒光**), TAO Yang(**陶　杨**), LIU Jie(**刘　洁**). Journal of Analytical Science(**分析科学学报**)[J], 2010, **26**: 551.
[10] ZHANG Lu-da(**张录达**), SU Shi-guang(**苏时光**), WANG Lai-sheng(**王来生**), LI Jun-hui(**李军会**), YANG Li-ming(**杨丽明**). Spectroscopy and Spectral Analysis(**光谱学与光谱分析**)[J], 2005, **25**: 33.
[11] Krämer K, Ebel S. Analytica Chimica Acta[J], 2000, **420**: 155.
[12] Vitale Raffaele, Bevilacqua Marta, Bucci Remo, Magrì, Magrì Antonio L, Marini Federico. Chemometrics and Intelligent Laboratory Systems[J], 2013, **121**: 90.
[13] ZHAO Jie-wen, LIN Hao, CHEN Quan-sheng, HUANG Xing-yi, SUN Zong-bao, ZHOU Fang. Journal of Food Engineering[J], 2010, **98**: 408.

Grading Tobacco Leaves Based on Near Infrared Spectroscopy Combined with Least Square Twins Support Vector Machine

SONG Xiang-zhong[1], LAI Yan-qing[1], LI Zu-hong[2], ZHENG Bo[2], LI Qian-qian[1], WU Li-jun[1], ZHANG Lu-da[1], XIONG Yan-mei*[1], MI Shun-geng*[1]
(1. *College of Science, China Agricultural University, Beijing* 100193;
2. *Yunnan Tobacco Company Qujing Branch, Qujing* 655000)

Abstract: In this work, we built a model to grade tobacco by least square twins support vector machine (LSTSVM) algorithm based on near infrared spectroscopy. We took 120 samples out of 210 samples from 3 grades as training set, and the rest 90 samples as testing set. In order to establish a optimal model, spectral pretreatment methods and model parameters were optimized. The average recognition rate of optimal model for testing set was 95.56%, which demonstrated that it could be a useful method to classify tobacco leaf grades. In addition, the LSTSVM algorithm was compared with three other kinds of general pattern recognition algorithms included SIMCA、PLS-DA and SVM, and the results indicated that the LSTSVM model based on original spectra had better prediction ability than others at the same conditions.

Keywords: Near infrared spectroscopy; Least square twins support vector machine (LSTSVM) algorithm; Tobacco leaf; Grades classification