高维特征选择方法在近红外光谱分类中的应用

秦玉华[1,2],丁香乾[1],宫会丽[1]

(1. 中国海洋大学 信息科学与工程学院,山东 青岛 266100;
2. 青岛科技大学 信息科学技术学院,山东 青岛 266061)

摘 要:针对卷烟近红外光谱高噪和高冗余特点,提出了一种基于随机森林(RF)和主成分分析(PCA)的特征优选方法 RF-PCA,建立了 5 种不同质量级别卷烟的分类模型,并和其他方法进行了比较。该方法能够有效地对高维数据样本进行分类,用于甄别卷烟品质真伪。特征选择可以过滤与分类不相关的特征,而通过 PCA 方法可以消除冗余特征的不良影响,并可进一步降低特征维数。实验表明:RF-PCA 方法能有效地剔除近红外光谱数据中的噪声特征和冗余特征,提高了分类效率。

关键词:近红外光谱; 特征选择; 随机森林; 主成分分析; 卷烟

中图分类号:O433.4 **文献标志码**:A **文章编号**:1007-2276(2013)05-1355-05

High dimensional feature selection in near infrared spectroscopy classification

Qin Yuhua[1,2], Ding Xiangqian[1], Gong Huili[1]

(1. College of Information Science and Engineering, China Ocean University, Qingdao 266100, China;
2. College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: With regard to the large number of irrelevant and redundant features exist in the near infrared spectra, a novel feature selection method based on random forest and principal component analysis (RF-PCA) was proposed in this paper. By using the RF-PCA, a classification model of cigarettes qualitative evaluation was developed and also compared with other methods. The result shows that RF-PCA effectively classifies the samples of high dimensional data and can be used to evaluate quality and authenticity of the cigarettes. RF feature selection removes irrelevant features of the classification, while PCA further eliminates the influence of redundant features and also reduces the feature dimensionalities. The experiments show that RF-PCA effectively removes noise and redundant features in the NIR spectra and the classification accuracy is improved as well.

Key words: NIR spectra; feature selection; RF; PCA; cigarettes

收稿日期:2012-09-17; **修订日期**:2012-10-08

基金项目:科技部创新基金(06C26213710334)

作者简介:秦玉华(1971-),女,副教授,硕士,主要从事近红外光谱学、数据挖掘、模式识别方面的研究。Email:yuu71@163.com

0 引 言

近红外光谱技术具有检测速度快、成本低、效率高、非破坏性等优点,目前在烟草界及各行各业已得到了广泛应用。采用近红外技术,可以定量分析烟草中的常规化学成分[1],识别不同产地风格的烟叶[2],进行卷烟真伪鉴别[3]。

在实际应用中,近红外光谱分析数据往往具有高维和小样本特点,样本数目一般为几十或上百,而特征波长往往高达几千维,其中含有大量与分类无关的噪声特征;相邻的特征波长高度相关,包含了大量冗余信息。噪声和信息冗余等因素会降低分类器的分类性能。利用原始数据构建分类器,不但会花费大量时间,还会降低分类效果。因此,如何从上千个特征中选出与分类密切相关的特征就显得尤为重要。目前解决光谱数据信息冗余问题的常用方法有主成分分析(PCA)和偏最小二乘法(PLS)[4],但它们仍然需要预先去掉与分类无关的噪声特征[5]。随机森林(RF)[6]是一种利用多个分类树对数据进行判别或分类的方法,在对数据进行分类的同时,还可以给出各个特征在分类过程中的重要性评分,具有精度高、减少过拟合、使用方便等特点,已成功应用于微阵列数据[7]、光谱数据[8]等。

文中针对卷烟近红外光谱提出一种基于 RF 和 PCA 的特征优选方法 RF-PCA,建立了 5 个不同级别卷烟的分类模型,能够有效对高维样本进行分类,用于甄别卷烟品质真伪及质量控制。实验表明:该方法能有效地剔除近红外光谱数据中的噪声特征和冗余特征,提高了分类效率。

1 实验与方法

1.1 样品的制备

选择国内不同卷烟厂市售的卷烟样品 129 个,其中一档烟 35 个样品,二档烟 19 个样品,三档烟 20 个样品,四档烟 35 个样品,五档烟 15 个,分别用类别标签 1、2、3、4、5 进行标定。每 2 小盒(40 支)卷烟为一个检测单位,将卷烟烟支剥开,取出烟丝,粉碎,过 60 目筛。

1.2 卷烟近红外光谱的采集

采用 Foss DS2500 近红外光谱仪,光谱扫描范围为 400~2 500 nm,扫描次数 32 次,数据间隔为 0.5 nm,室温保持在 18~22 ℃,将样品置于样品池中,用压样器压实样品,置于近红外光谱仪中扫描。为消除样品均匀性不一致等因素的影响,每个样品均重复装样测定 3 次,然后计算其平均值。

1.3 RF 算法

RF 是 Breiman[6]于 2001 年提出的一种组合分类器算法,是由许多棵分类回归树组成,最后通过投票法决定最终分类结果。建立 RF 的基本思想是,通过自助法(bootstrap)重采样技术[9],选择样本子集构建多个树分类器。在每棵树的生长过程中,假设一共有 M 个特征,从 M 个特征中随机选取 m 个特征,按照 Gini 指数最小原则[10]从 m 个特征中筛选出最佳分裂点进行分支生长。随机树充分生长,不进行剪枝操作。在整个森林构建过程中,m 值保持不变。最后利用投票综合各分类器的结果得到最终的结果。根据参考文献[11]的研究,文中选取 $m=\sqrt{M}$。

由于采用 bootstrap 抽样法,原始训练集中约有 1/3 的样本可能未被抽中,这些样本可以作为测试数据对该分类树的泛化性能进行估计,这部分数据称为袋外数据(OOB),这种估计方法称为 OOB 估计。当 RF 中树的数目足够多以使测试误差收敛时,OOB 估计为无偏估计[6]。RF 算法的分类误差计算如下:

$$ER \approx ER^{OOB}=N^{-1}\sum_{i=1}^{N} I(Y^{OOB}(X_i)\neq Y^i) \tag{1}$$

式中:N 为森林中树的个数;$I(\cdot)$为指示函数;$Y^{OOB}(X_i)$为 X_i 做为袋外数据的估计值。

RF 的另一个重要特性是能计算单个特征的重要性。原理是将待某一特征的值替换,通过替换前后分类器的分类正确率变化来判断其重要性。如果给某特征随机地加入噪声后,OOB 准确率大幅度下降,则表明该特征的重要程度较高。利用 RF 计算特征重要性的特点,可将其应用于特征选择。另外作为一种分类准确度高的 wrapper 特征选择方法,它的可调节参数少,使用时只需设定 3 个参数:终结点规模,每个分支点的变量个数及森林中树分类器的个数。而且它还可以选择有分类意义的特征波长,并且计算量比 wrapper 方法少。

1.4 主成分分析

除了与分类无关的噪声,光谱数据中还有大量的冗余特征,这些特征也会影响到分类的精确程度。

冗余特征的处理方法有多种,PCA 法就是其中之一。PCA[12]是一种经典的数据变换方法,它可以将分散在一组变量上的信息转换为几个综合指标, 使变换后的新分量正交且互不相关, 既保留了绝大多数的原始信息, 又能很好地降低维数。因此, 文中使用 PCA 来处理数据集中的冗余特征。

1.5 RF-PCA 方法

用近红外光谱仪测量 **m** 个样本, 每个样本测量 **n** 个点(波长变量),因此,一组样本的光谱数据可由自变量矩阵 $X=(x_{mn})$ 和对应的类别的因变量矩阵 $Y=(y_m)$ 组成, 建立分类模型实际上就是建立矩阵 X 和 Y 的关系。

文中提出的 RF-PCA 方法如下。

(1) 采用 Bootstrap 技术在训练集上随机构造 **N** 个训练数据子集 $\mathbf{F}_i(i=1,2,\cdots,\mathbf{N})$,按照 RF 算法中 Gini 指数最小原则计算各特征波长的重要性, 构建 **N** 个特征选择器, 然后集成各特征选择器的结果得到最终的优选的特征波长子集。

例如得到的特征组集合为 $\mathbf{F}=\{\mathbf{F}_1,\mathbf{F}_2,\cdots,\mathbf{F}_N\}$,其中 $\mathbf{F}_i$ 提供的特征波长为 $\mathbf{F}_i=(f_i^1,\cdots,f_i^l)$,最终综合各组的特征评分可按照下式得出:

$$f^l=\sum_{i=1}^{N}\omega f_i^l \qquad (2)$$

式中:ω 为权,可用于特征波长的重要性的调整参数或基于分类器分类效能相联系的指标, 最简单的情况取等权 $\omega=1$。可根据具体研究问题精细的情况选择等权或非等权,文中选取等权情况来研究。

(2) 把得到的特征子集评分结果按降序排列,从中依次剔除排在后面的不重要的特征波长,用训练集数据训练分类器,并用测试集数据记录各分类误差。

(3) 重复步骤(2),直到特征子集中无剩余特征。

(4) 选取分类误差最小的特征数作为最小特征子集。

(5) 对得到的最小特征子集进行 PCA 降维以去掉冗余特征。

2 结果与分析

2.1 卷烟的近红外光谱及预处理

对 129 个卷烟样品进行光谱扫描,选取1 120~2 500 nm 为建模波长, 图 1 为所有样本的原始光谱图。由图 1 可以看到,原始谱图中光谱在吸光度轴上差异较大,因此模型建立前,为了消除测定过程中的干扰,谱图基线漂移等对模型的影响,需要对其进行预处理。

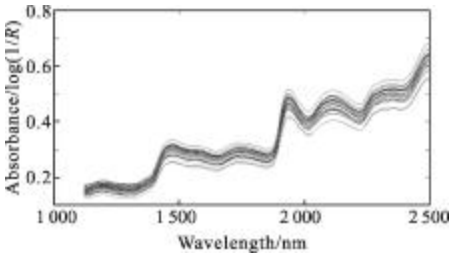

图 1 卷烟原始光谱图

Fig.1 Original NIR spectra of cigarettes

经过比较各种预处理方法, 文中选用标准正态变量变换(SNV)和二阶导数作为预处理方法,可消除原始谱图的漂移现象,突出有用信息。

2.2 特征选择

全光谱含有 2 760 个特征波长,其中包含大量与分类无关的噪声特征和冗余信息, 如果采用全部特征进行分类,速度较慢且分类准确度不高。

首先采用随机森林算法来度量各个特征波长的重要性。用 Bootstrap 抽样法,在训练集样本上生成 6 000 棵树, 按照 RF 算法中 Gini 指数最小原则计算各特征波长的重要性评分,重复实验 100 次,按照公式(2)计算每个特征波长的综合得分。各特征波长对分类重要性的统计结果见图 2。

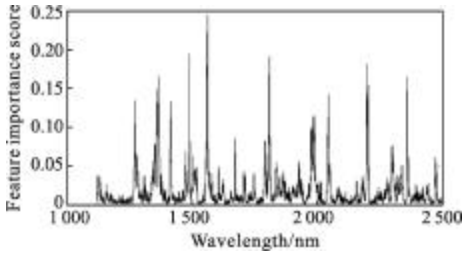

图 2 特征波长重要性度量结果

Fig.2 Result of feature importance measure

按照上述得到的重要性综合评分结果对各特征波长进行排序, 选取 2/3 的样本做为训练集,1/3 的样本做为测试集, 使用不同个数的优选特征进行建模,选出最佳模型,图 3 为优选出的特征个数与测试集分类误差曲线 (分类误差=误判个数/测试样品个数)。可以看出,采用特征选择算法,只需 185 个特

征,就可达到最低的分类误差。而继续增加选取特征的个数,由于噪声信息和冗余信息的影响,分类器的误差会有所增加。

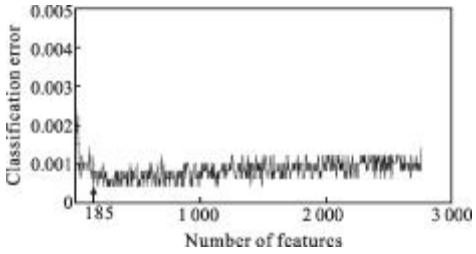

图 3 不同的特征选择与分类误差

Fig.3 Different selected features and classification errors

特征选择可明显地消除与分类无关的噪声特征,由图 2 特征重要性度量结果可知,重要性评分高的特征波长,其相邻谱带也有较高的重要性评分,因为它们表示原始光谱图中同一吸收峰,因此,筛选出的特征波长还存在一定的冗余特征, 这些也会影响到分类的精确程度。冗余特征的处理方法有多种, PCA 法就是其中之一。通过 PCA 方法可以消除冗余特征对分类的不良影响,并可进一步降低特征维数,文中选取保留 90%的原始信息。

2.3 性能比较

一些使用 RF 算法度量特征重要性的研究[11]同时也以 RF 做为分类器, 都得到了比较好的分类结果。文中同时选取 RF 和适合高维小样本的支持向量机(SVM)做为分类模型,分别依据以下 4 个特征进行分类分析: (1) 全部特征;(2) PCA 降维后的特征;(3) 基于随机森林特征重要性度量后的优选特征;(4) RF-PCA 方法后的特征。

图 4 是使用 RF 作为分类器分别训练 1~800 棵树的分类误差, 可以看出理想的森林中分类树的数目为 181,这时的分类误差最低且趋向于稳定。因此,以 RF 作为分类器模型选取森林中树的个数为 181。

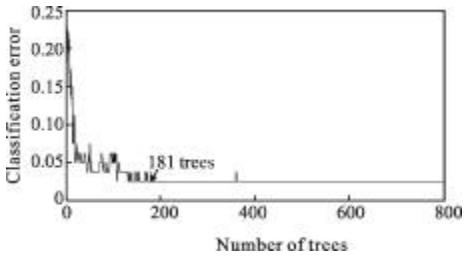

图 4 树的数量与分类误差

Fig.4 Classification errors with different number of trees

表 1 为不同方法的分类误差的比较结果。

表 1 不同方法的分类误差比较(%)

Tab.1 Comparison of different classification methods(%)

Classification model	All features	PCA	RF feature selection	RF-PCA
SVM	12.82	10.25	8.16	5.12
RF	15.38	12.82	7.69	2.56

从表 1 的实验结果可以看出:SVM 分类器使用全部特征和 PCA 降维后的分类误差均低于 RF 分类器。主要因为 RF 是一个组合分类器算法,采用重采样技术和随机选择特征生成多棵分类树,然后采用简单多数投票法决定最终的分类结果,其中分类效果很差的树会降低总体分类准确率。如果能挑选出少数性能较好的分类树可提高分类效果。

使用 RF-PCA 选择特征变量,SVM 分类误差均高于 RF 分类,这主要因为 RF 特征优选本质是一个基于 RF 算法的 Wrapper 方法,它是将分类算法嵌入到特征选择过程中,不仅分类准确度高,还可以选出有分类意义的特征波长, 所以用随机森林模型建模比 SVM 模型表现出更好的性能。

总体看来,RF-PCA 方法在两种分类模型上都获得了最小的分类误差,RF 特征优选的性能次之,使用 PCA 降维后性能再次之,而使用全部特征进行分类结果最差, 这也说明了高维近红外光谱数据含有较多的噪声信息。RF-PCA 在此基础上又采用 PCA 法消除了冗余信息, 大大减少了训练及分类所需的时间及消耗的资源,且提高了分类正确率,实验结果表明了文中方法对不同卷烟级别近红外光谱数据分类的有效性。

3 结 论

文中根据卷烟近红外光谱的特点, 提出了一种新的基于 RF 特征重要性和 PCA 的高维特征优选方法 RF-PCA, 具有更强的剔除卷烟近红外光谱数据中噪声特征和冗余特征的能力。对 5 种不同级别的卷烟识别,RF-PCA 具有更高的分类正确率。后续工作试图根据筛选出的重要性评分高的特征波长,探索对分类有主要贡献的化合物种类, 对卷烟的质量

品质进行宏观的把握。同时也可以分析卷烟中主要化学成分的吸收波长与筛选的特征波长的对应关系,为定量分析打下基础。

参考文献:

[1] Liu Xu, Chen Huacai, Liu Taiang. Application of PCA-SVR to NIR prediction model for tobacco chemical composition [J]. Spectroscopy and Spectral Analysis, 2007, 27(12): 2460-2463. (in Chinese)
刘旭, 陈华才, 刘太昂. 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.

[2] Hana M, Mcclure W F, Whitaker T B. Applying artificial neural networks II. Using near infrared data to classify tobacco types and identify native grown tobacco [J]. Journal of Near Infrared Spectroscopy, 1997, 5: 19-25.

[3] Tang Xuemei, Zhang Wei, Li Hui. Application of near-infrared reflectance spectroscopy in discriminating counterfeit cigarettes from genuine[J]. Tobacco Science and Technology, 2008, 11: 5-8. (in Chinese)
唐雪梅, 张薇, 李慧. 卷烟真伪鉴别的近红外定性分析方法[J]. 烟草科技, 2008, 11: 5-8.

[4] Bylesjo M, Rantalainen M, Nicholson J K, et al. K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space [J]. BMC Bioinformatics, 2008, 9(1): 106-112.

[5] Boaz Nadler, Coifman Ronald R. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration [J]. Journal of Chemometrics, 2005, 19(2): 107-118.

[6] Leo Breiman. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[7] Statnikov A, Wang L, Aliferis C F. A comprehensive comparison of random forests and support vector machines for microarray based cancer classification [J]. BMC Bioinformatics, 2008, 9: 319-323.

[8] Menze B H, Petrich W, Hamprecht F A. Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy [J]. Analytical and Bioanaytical Chemistry, 2007, 387(5): 1801-1807.

[9] Efron B, Tibshirani R J. Bootstrap measures for standard errors, confidence interval and other measures of statistical accuracy[J]. Statistical Science, 1986, 1(1): 54-74.

[10] Menze B H, Kelm B M, Masuch R, et al. A comparison of random forest ant its Gini importance with standard chemometric methods for the feature selection and classification of spectral data[J]. BMC Bioinformatics, 2009, 10: 1-16

[11] Simon Bernard, Laurent Heutte, Sebastien Adam. Influence of hyperparameters on random forest accuracy [J]. MCS, 2009: 171-180.

[12] Jolliffe I T. Principal Component Analysis [M]. New York: Springer-Verlag, 1986.