文章编号：1672-8785(2014)12-0041-04

不同产地苹果的近红外光谱分类识别法

李　敏

(乐山师范学院物理与电子工程学院，四川 乐山 614000)

摘　要：以山东和陕西两地产的红富士苹果作为实验对象，提出了一种不同产地苹果的分类识别法。首先对苹果的近红外光谱数据进行小波软阈值预处理，去除噪声和冗余；再采用主成分分析法 (Principle Component Analysis, PCA) 进行降维；然后应用 Fisher 判决 (Fisher Discriminant Analysis, FDA) 进一步提取特征；最后使用 K_ 近邻法进行分类识别 (K_near neighbor classification, KNN) 。通过实验比较，本文提出的方法能很好地实现不同产地苹果无损、快速和准确分类识别，识别正确率达到 97.5% 。

关键词：不同产地；近红外光谱；主成分分析； Fisher 判决； K_ 近邻分类

中图分类号：S123　　**文献标志码**：A　　**DOI**: 10.3969/j.issn.1672-8785.2014.12.008

Near Infrared Spectral Classification Method of Apples from Different Regions

LI Min

(School of Physics and Electrical Engineer of Leshan Normal University, Leshan 614000, China)

Abstract: Taking the red Fuji apples produced in Shandong and Shanxi Provinces as the experimental objects, a classification method for identifying the apples produced in different regions is proposed. Firstly, the near infrared (NIR) spectra of the apples are preprocessed by a wavelet soft-threshold method, removing the noise and redundancy. Then, a Principal Component Analysis (PCA) method is used to reduce the dimension of the NIR data. Secondly, a Fisher Discriminant Analysis (FDA) method is used to further extract the features from the data. Finally, a K_near Neighbor Classification (KNN) method is used for the classification and identification of the apples. The experimental result shows that the proposed method can well realize the nondestructive, fast and accurate classification and identification of the apples produced in different regions. Its identification accuracy is up to 97.5%

Key words: different region; near infrared spectroscopy; principal component analysis; fisher decision; K_near neighbor classification

0 引言

山东省和陕西省是我国红富士苹果生产的主要地区，两地产的红富士都具有色泽鲜艳红润、味道清脆香甜的特征。但受地理位置、气候条件和土壤等因素影响，两地红富士又存在不同的特征。与陕西红富士相比，山东红富士一般个头更大，果肉的硬度更大，含糖量更高。也就是说山东红富士更脆更甜，更受消费者亲睐，所以价格更高。目前市场上还没有一种方法能无损、快速、准确地鉴别两地的红富士苹果，仅

收稿日期：2014–10–27

基金项目：四川省教育厅重点项目 (12ZA070)

作者简介：李敏 (1977-)，女，四川汉源人，副教授，硕士研究生，研究方向为模式识别、信号处理、近红外光谱分析和小波分析等。 E-mail:cassie_li@163.com

凭经验和目测很难准确鉴别两地的红富士。所以存在很多以假乱真、以次充好等混乱现象。因此急需研发出一种快速、准确、无损鉴别不同产地红富士苹果的方法。

近红外光谱分析具有非破坏性、检测快速高效和成本低等特点，近年来在农产品和食品检测领域得到很快的发展，例如应用于蓝莓营养成分的检测、苹果擦伤检测、杏果品质检测和桃子成熟度检测等 [1−3] 。也有学者将近红外光谱技术用于山药、辣椒和红景天等品种的鉴别 [1] 。但目前还没有学者将它用于不同产地苹果的分类鉴别。

PCA 分析属于非监督学习法，是特征提取和数据降维的常用方法 [4] 。 Fisher 判别又称为线性判别分析，是一种有效的特征提取方法。KNN 是基于统计的分类方法，该算法根据待测样本在特征空间的 K 个最近邻样本中的多数样本的类别来进行分类，具有直观、无师学习和无需先验统计知识等特点，是一种重要的非参数分类方法。本文以山东和陕西两地产的红富士为研究对象，采用 PCA 分析和 Fisher 判别对两类苹果的近红外光谱进行降维和特征提取，采用 KNN 法进行分类鉴别，对两类苹果的正确识别率达到 97.5% 。

1 分类识别方法简介

1.1 主成分分析

PCA 是一种把多个指标化为几个综合指标的统计方法。它将多维光谱数据沿着协方差最大的方向向低维数据空间投影，达到数据降维的目的。各不同主成分向量之间相互正交。通过合理选择主成分向量既可以避免建模的数据冗余，又不会过多地丢失光谱信息 [5] 。

1.2 Fisher 判别分析

FDA 是统计模式识别的基本方法之一。其基本思想是将原来高维的模式样本投影到最佳鉴别向量空间，达到抽取分类信息和压缩特征空间维数的效果。投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离。这种方法的关键是求解最佳鉴别向量。若 $X=\{x_i\}, i=1,2,3,\cdots,n$ 是 l 维的样本，$X_1,X_2,\cdots,X_N$ 是已知的 N 类模式， $x_i \in X_j$, $j \in \{1,2,\cdots,N\}$ 。各类间离散度矩阵 S_b 、类内离散度矩阵 S_w 和总体离散度矩阵 S_t 可分别定义如下 [6] ：

类间离散度矩阵

$$S_b=\sum_{i=1}^{N}P_i(m_i-\bar{m})(m_i-\bar{m})^T \tag{1}$$

类内离散度矩阵

$$S_w=\sum_{i=1}^{N}P_iE\{(x-m_i)(x-m_i)^T|X_i\} \tag{2}$$

样本总体离散度矩阵

$$S_t=S_b+S_w=E\{(x-\bar{m})(x-\bar{m})^T\} \tag{3}$$

式中， $m_i, i\in\{1,2,\cdots,N\}$ 是第 i 类样本的均值向量； $\bar{m}$ 是整体样本的均值； P_i 是 X_i 的先验概率。一般情况下， $P_i=1/N$ 。

为了在样本数据投影后使类间离散度变得尽可能大，使类内离散度变得尽可能小，可定义 Fisher 判别准则函数如下：

$$J(\varphi)=\frac{\varphi^TS_b\varphi}{\varphi^TS_w\varphi} \tag{4}$$

式中， φ 是 1 维空间 φ_1^l 中的任一向量。对其采用拉格朗日乘子法，可得：

$$S_b\varphi^*=\lambda S_w\varphi^* \tag{5}$$

式中， φ^* 是 $J(\varphi)$ 取最大值时的 φ 。如果 S_w 是非奇异的，可得：

$$S_w^{-1}S_b\varphi^*=\lambda\varphi^* \tag{6}$$

求解式 (6) 也就是求解矩阵 $S_w^{-1}S_b$ 的特征向量与特征值。若样本的个数大于样本的维数即为大样本，此时类内离散度矩阵 S_w 是非奇异矩阵，求解 φ^* 较容易。但实际应用中，往往样本个数远远小于样本维数即为小样本，此时 S_w 可能是奇异矩阵，求解 φ^* 很困难。因此，需要先通过 PCA 对光谱数据降维，使降维后的维数 l 满足下式 [6] ：

$$N\le l\le min(rank(S_w),n-1) \tag{7}$$

然后对降维的数据用 Fisher 判别分析进行特征提取。

1.3 K 近邻分类算法

KNN 是一种非参数分类算法，现已广泛应用于数据挖掘和模式识别等领域。其基本思想是将一个待分类样本 X 分成训练集和测试集，从训练集里找出与测试集最接近的或最相似的 K 个样本，然后根据这 K 个样本确定测试集样本的类别。算法步骤如下 [7]：(1) 构建训练集和测试集；(2) 确定一个 K 的初始值，然后根据实验结果调整，直到最优；(3) 在训练集中找出与测试集最相似的 K 个样本。相似程度由欧式距离来度量。假定样本点 x 属于 n 维空间 R^n，设第 i 个样本 $x_i=(x_1^i,x_2^i,\cdots,x_n^i)\in R^n$，那么两个样本 x_i 和 x_j 之间的欧式距离定义为

$$d(x_i,x_j)=\sqrt{\sum_{l=1}^{n}(x_l^i-x_l^j)^2} \tag{8}$$

若 x_q 为一待分类的样本，$x_1, x_2,\cdots, x_K$ 为与 x_q 距离最接近的 K 个样本。设离散的目标函数为 $f{:}R^n\to\nu_i$，ν_i 表示第 i 个类别的标签，标签集合为 $V=\{\nu_1,\nu_2,\cdots,\nu_s\}$，

$$\tilde{f}(x_q)=\arg\max_{\nu\in V}\sum_{i=1}^{K}\delta(\nu,f(x_i)) \tag{9}$$

$\tilde{f}(x_q)$ 是对 $f(x_q)$ 的估计，当 $a=b$ 时，$\delta(a,b)=1$。否则，$\delta(a,b)=0$。就是待测样本 $\tilde{f}(x_q)$ 的类别。

2 实验分析

选用大小均匀、无损伤的山东红富士 (这里用“hfsd”表示) 和陕西红富士 (这里用“hfsx”表示) 各 60 个作为实验样品。样品在室温为 20~25° 的实验室内存放 12 h 待测。

光谱采集器采用赛默飞世尔 (Thermo Fisher) 公司生产的 Antaris II 近红外光谱分析仪。该红外仪器性能高，能节省大量的检测成本。

采集光谱时，先把近红外光谱分析仪预热 1 h，采用反射积分球模式采集苹果的近红外光谱，光谱扫描的波数为 4000~10000 cm^{-1}，扫描间隔为 3.856 cm^{-1}。为减少误差，近红外光谱分析仪沿每个样品赤道轨迹扫描 3 次，取其平均值作为最终的样本数据。每个样本是一个 1557 维的数据。最终获得两类苹果 120 个 1557 维的光谱数据。两类苹果的近红外光谱如图 1 所示。

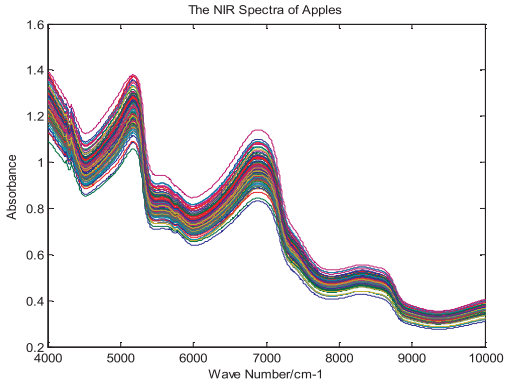

图 1　两个产地红富士苹果的近红外光谱

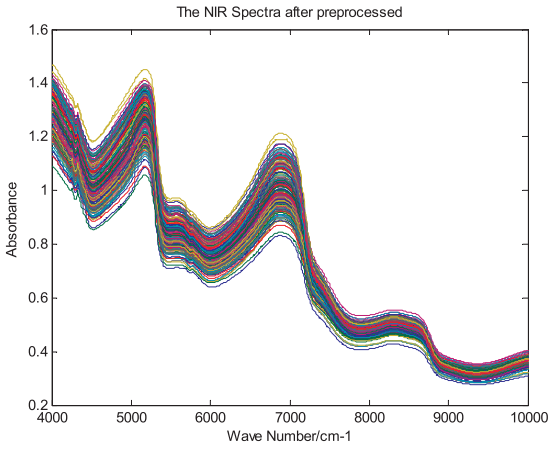

图 2　经小波软阈值预处理后的光谱

光谱数据不经预处理和采用多元散射校正 (MSC) 或小波软阈值法进行预处理，对苹果的正确分类识别率的影响明显不同。光谱数据不经预处理时，正确识别率最低；而小波软阈值法预处理的效果明显优于多元散射预处理，其正确识别率最高，达到 97.5%。小波软阈值处理后的光谱如图 2 所示。经过不同预处理方法后获得的苹果正确分类识别率见表 1。

光谱数据经预处理后，采用 PCA 分析进行降维，然后取两类苹果光谱数据的前 40 个样本组成训练集，后 20 个样本组成测试集，这样训

练集共 80 个样本，测试集共 40 个样本。接着采用 Fisher 判决进行特征提取。最后采用 KNN 算法进行分类鉴别，选取不同的 K 值，直到得到最高的正确识别率。 K=4 时，正确识别率达到最高，为 97.5% ，分类结果如图 3 所示。不同 K 值，对应的正确分类识别率见表 2 。

表 1 不同预处理方法的分类识别结果

处理方法	K 的取值	正确识别样本数	正确识别率 (%)
无预处理	4	30	75
多元散射校正 (MSC)	4	35	87.5
小波软阈值	4	39	97.5

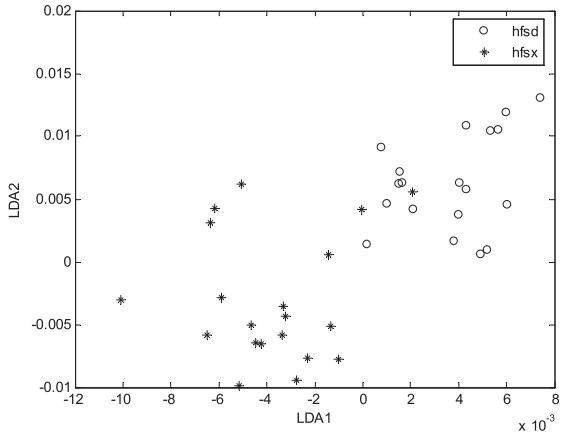

图 3 两地红富士的最佳分类结果 (K=4)

表 2 K 的取值对分类识别的影响

K 取值	正确识别样本数	正确识别率 (%)
3	35	87.5
4	39	97.5
5	38	95
6	37	92.5

3 结论

提出了一种结合 PCA 分析、 Fisher 判别和 KNN 分类的不同产地红富士苹果分类识别算法。通过实验，光谱数据经预处理后，正确分类识别率能得到大大提高。小波软阈值预处理方法优于多元散射校正 (MSC) 预处理法。采用 KNN 分类识别时， K 的取值也会影响正确分类识别率。实验表明，当 K=4 时，两类苹果的正确识别率达到最高，为 97.5% 。本文提出的方法能对两地红富士苹果进行快速、无损、准确的分类识别，为近红外光谱分析技术提供了一种新思路。

参考文献

[1] 王敏，付荣，赵秋菊，等．近红外光谱技术在果蔬品质无损检测中的应用 [J].**中国农学通报**,2010.**26**(5):174–178.

[2] 褚小立，袁洪福，陆婉珍．近红外分子中光谱预处理及波长选择方法进展与应用 [J].**化学进展**,2004,**17**(4):528–542.

[3] 高荣强，范世福，严衍禄，等．近红外光谱的数据预处理研究 [J].**光谱学与光谱分析**,2004,**24**(12):1563–1565.

[4] 何勇，李晓丽，邵咏妮。基于主成分分析和神经网络的近红外光谱苹果品珍鉴别方法 [J].**光谱学与光谱分析**,2006,**26**(5):850–853.

[5] 陈全胜，赵杰文，张海东，等．基于支持向量机的近红外光谱鉴别茶叶的真伪 [J]. **光学学报**,2006,**6**(26):933–937.

[6] 卜锡滨，武斌，贾红雯．苹果近红外光谱的特征提取和分类研究 [J].**计算机工程与应用**,2013,**49**(2):170–173.

[7] Sang Y B. Research of Classification Algorithm Based on K Nearest Neighbor [D]. Chongqing University,2009.