文章编号： 1672-8785(2015)05-0043-04

乐山茶叶的近红外光谱分类识别

李敏

（乐山师范学院物理与电子工程学院，四川乐山， 614000）

摘 要：以乐山产正品竹叶青、劣质竹叶青和峨眉山毛峰为研究对象，提出了一种基于近红外光谱的不同茶叶品种分类识别算法。该算法采用多元散射校正 (Multiplicative Scatter Correction, MSC) 对 3 种茶叶的近红外光谱数据进行预处理，最大限度地扣除光谱数据中的随机变异；再采用主成分分析算法 (Principal Component Analysis, PCA) 对预处理后的光谱数据进行降维，去除冗余；接下来进行线性判别分析 (Linear Discriminant Analysis, LDA)，进一步提取特征；最后采用 K_ 近邻算法 (K_Nearest Neighbor, KNN) 对 LDA 结果的前两个特征进行分类，从而达到对茶叶进行定性分类的目的。实验结果表明，该算法能有效地对 3 种茶叶进行分类，正确识别率达到 100%。本研究为不同品种茶叶的分类识别提供了一种新思路。

关键词：茶叶；近红外光谱；主成分分析；线性判别分析； K_ 近邻分类

中图分类号： TH744.1 **文献标志码**： A **DOI**: 10.3969/j.issn.1672-8785.2015.05.009

Classification and Identification of Leshan Tea Using Near Infrared Spectroscopy

LI Min

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

Abstract: Taking the real Zu Yeqing tea produced in Leshan, the inferior Zu Yeqing tea and the Maofeng tea produced in Emei Mountain as the research objects, a classification algorithm for different kinds of tea based on near infrared spectroscopy is put forward. The algorithm uses Multiplicative Scatter Correction (MSC) to preprocess the near infrared spectral data of the above three kinds of tea for removing the random variation in the spectral data maximally. Then, it uses Principal Component Analysis (PCA) to reduce the dimensionality of the spectral data for removing redundant. Next, it carries out Linear Discriminant Analysis (LDA) for further feature extraction. Finally, it uses the K_Nearest Neighbor algorithm to classify the first features in the LDA result so as to realize the qualitative tea classification. The experimental results show that this algorithm can classify the above three kinds of tea effectively. Its correct recognition rate is up to 100%. This study provides a new idea for the classification of tea.

Key words: tea; near infrared spectroscopy; principal component analysis; linear discriminant analysis; K_nearest neighbor classification

收稿日期：2015–03–20

基金项目：乐山市科技局项目 (14NZD017)

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

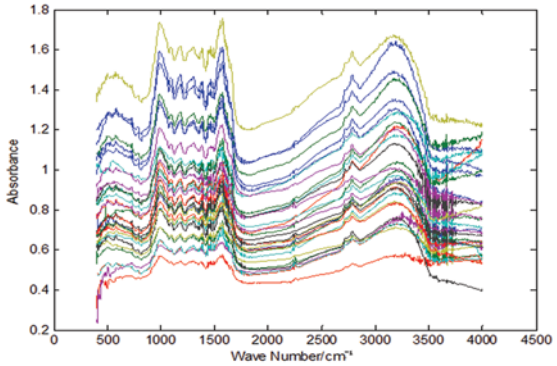
0 引言

乐山市是中国西部盛产名优绿茶的区域之一，具有悠久的产茶历史。据初步统计，全市目前有茶园基地约 60 万亩，茶叶年生产总量达 2.5 万吨，茶叶年产值达 5.6 亿。现在已经形成“竹叶青”、“仙芝竹尖”、“森林雪”、“一枝春”和“峨眉雪芽”等品牌。目前中国的茶叶市场由于缺乏有效的茶叶品种鉴别方法比较混乱，尤其是名优绿茶市场，贴牌现象、以次充好和以假乱真现象比较严重。这样既损害了消费者的利益，又严重损害了乐山名茶的市场声誉。为了保护乐山名茶品牌，抢占茶叶市场，急需研究出一些高效的茶叶品种分类识别方法。

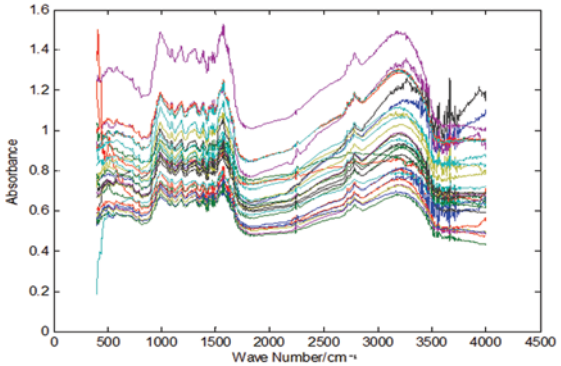
传统的茶叶鉴别方法主要有化学方法和感官评价方法两种 [1]。化学方法虽然能正确鉴别茶叶的品种，但步骤繁琐，价格昂贵；感官鉴别法受外部环境和人为因素等干扰，正确鉴别率不高。近红外光谱 (near-infrared spectroscopy, NIRS) 分析是近年来兴起的一种分析与研究手段，具有分析速度快、产出多、不破坏样品和适于产品在线分析等优点，已在食品、医药、农业和石化等领域得到广泛的应用 [2]。目前已涌现出一些基于近红外光谱分析技术的茶叶分类算法，但这些算法一般都没对茶叶光谱进行预处理，而且正确分类的鉴别率偏低，普遍在 80% 左右。本研究以乐山茶叶为对象，采用多元散射校正法对茶叶的近红外光谱数据进行预处理，再采用主成分分析对光谱数据进行降维，去除冗余；用线性判别分析进一步提取特征；最后采用 K- 近邻算法对 LDA 结果的前两个特征进行分类。该算法对茶叶定性分类的正确识别率达到 100%，明显优于其它算法。

1 茶叶光谱的采集

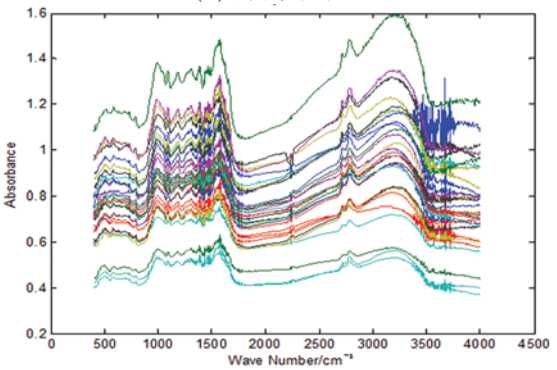
实验仪器为 FTIR–7600 型傅里叶红外光谱仪，其波数范围为 7800 cm^{-1} ~350 cm^{-1}，分辨率为 4 cm^{-1}，扫描次数为 32，数据点的间隔为 1.928 cm^{-1}。选取乐山市峨眉山竹叶青公司生产的正品竹叶青和从乐山市场上购买的散装劣质竹叶青和峨眉山产毛峰为实验对象。三种茶叶经研磨粉粹，再用 40 目筛进行过滤后，各取 0.5 g 分别与溴化钾按 1:100 均匀混合。每个样本取混合物 1 g 进行压膜，然后用光谱仪扫描 3 次，取 3 次的平均值作为样本光谱数据。采集环境温度为 25.2 °C, 相对湿度为 49%，电压为 220 V。每种茶叶采集 32 个样本，共获得 96 个样本。每个样本为一个 1868 维的数据，波数范围为 4001.569 cm^{-1} ~401.1211 cm^{-1}。三种茶叶的近红外光谱如图 1 所示。样本情况见表 1。



(a) 正品竹叶青



(b) 劣质竹叶青



(c) 峨眉山毛峰

图 1 三种茶叶的近红外光谱

表 1　实验样本

茶叶品种	样本数	训练集	测试集
正品竹叶青	32	15	17
劣质竹叶青	32	15	17
峨眉山毛峰	32	15	17

2 分类算法

本文算法流程如图 2 所示。

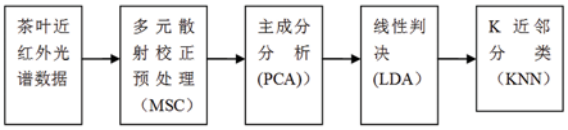


图 2　本文算法的流程

2.1 MSC 预处理

样品的不均匀性 (粒度分布) 常导致样品的光谱存在很大的差异，散射引起的光谱变化往往大于样品成分引起的光谱变化 [3]。从仪器采集的近红外光谱数据若不经过预处理就直接用于分类识别，会导致正确识别率低下。经 MSC 预处理后，茶叶近红外光谱的差异会大大减少，同时随机变异会被最大程度地扣除，为后续的分类识别奠定了良好的基础。

2.2 PCA

主成分分析又称为抽象因子分析，就是利用数据降维的思想把原来多个变量划分为少数几个综合变量。由于综合变量为原变量的线性组合，可以达到消除众多冗余信息的目的 [4]。新变量能最大限度地表征原变量的数据特征，而且没有信息损失 [5]。但主成分分析并不是最终目的，最主要的目的是达到数据降维。经处理后的数据还需要作进一步分析，如判别分析、聚类分析等。本文对 3 种茶叶共 96 个样本数据采用主成分分析，使其降为 20 维数据，即获得 20 个主成分。

2.3 LDA

LDA 也叫做 Fisher 线性判别，是一种模式识别的经典算法 [6]。其定性判别分析的基本思想是将高维的样本矢量投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的目的。投影后，须保证样本矢量在新的子空间有最大的类间距离 (即获得最大的类间散布矩阵) 和最小的类内距离 (即获得最小的类内散布矩阵)。茶叶的近红外光谱数据经 PCA 降维后，再采用 LDA 有效地进行特征抽取。

2.4 KNN 分类识别

KNN 法逐一计算每个待测样本与各训练样本之间的距离 (本文选用欧氏距离)，找出最近的 K 个样本进行判决。KNN 分类结果的准确性与 K 值关系较大， K 值的选择目前尚无规律可循，这里是采用反复实验对比法来确定最佳 K 值的。

3 实验分析

对 3 种茶叶做了如下三种实验分析： (1) 光谱数据不经预处理，直接进行 PCA 分析，然后进行 LDA 判别，最后进行 KNN 分类，即 “ PCA+LDA+KNN ” 法； (2) 光谱数据经 Savitzky — Golay 预处理后，再进行 PCA 分析，然后进行 LDA 判别，最后进行 KNN 分类，即 “ SZG+PCA+LDA+KNN ” 法； (3) 光谱数据经 MSC 预处理，然后采用 PCA 分析降维，再采用 LDA 判别进一步提取特征，最后进行 KNN 分类，即 “ MSC+PCA+LDA+KNN ” 法，即本文提出的算法。

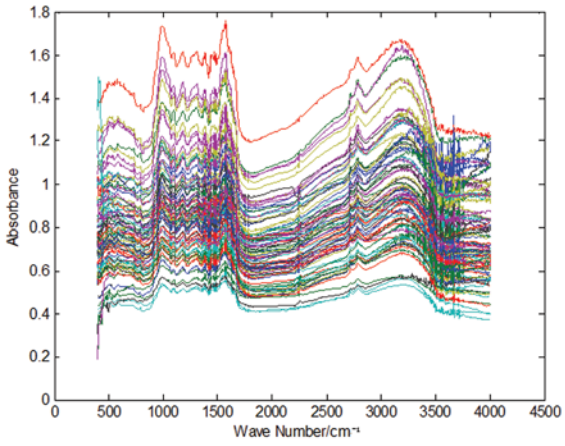


图 3　3 种茶叶的原始近红外光谱

三种茶叶共 96 个样本的原始近红外光谱如图 3 所示，经 MSC 预处理后的光谱如图 4 所示，经 SZG 预处理后的光谱如图 5 所示。各种算法的分类结果如图 6、图 7 和图 8 所示。其中“ tea1 ”代表峨眉山毛峰，“ tea2 ”代表正品竹叶青，“ tea3 ”代表劣质竹叶青。各种算法的最佳 K 值和最高的正确识别率见表 2 。通过比较可知，预处理茶叶的光谱数据可大大提高正确分类识别率，而 MSC 预处理方法的效果优于 SavitZky 一 Golay 预处理法。本文提出的“ MSC+PCA+LDA+KNN ”算法能对 3 种乐山茶叶进行有效的分类识别，正确识别率达 100% 。

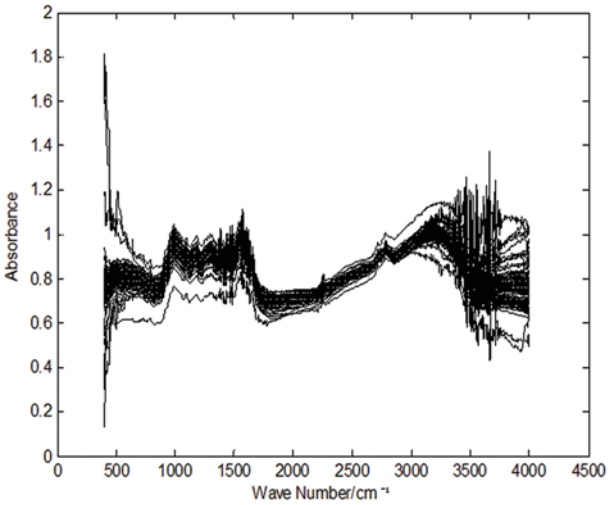


图 4 经 MSC 预处理后的光谱

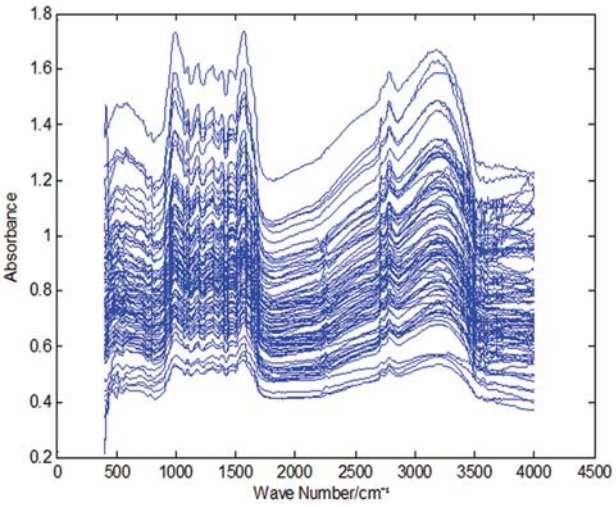


图 5 SZG 预处理后光谱数据

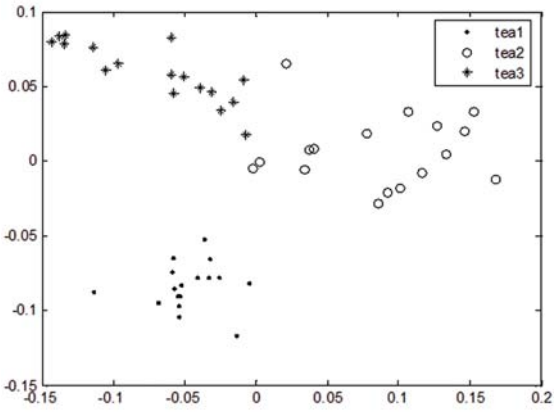


图 6 PCA+LDA+KNN 分类结果

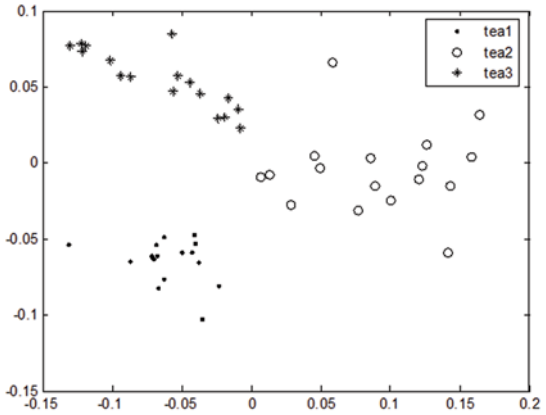


图 7 SZG+PCA+LDA+KNN 分类结果

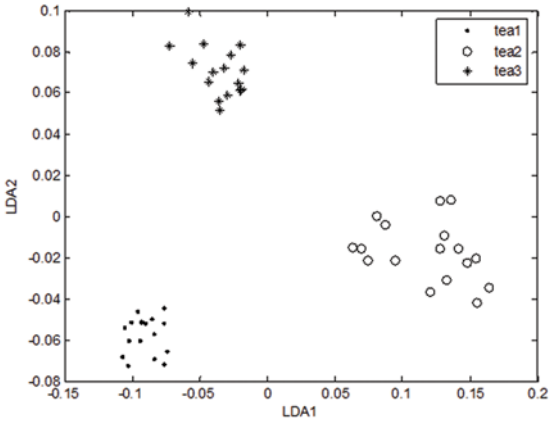


图 8 MSC+PCA+LDA+KNN 分类结果

表 2 各种分类算法比较

算法	最佳 K 值	正确识别率 (%)
PCA+LDA+KNN	4	92.16
SZG+PCA+LDA+KNN	4	94.12
MSC+PCA+LDA+KNN	**4**	**100**

(下转第 48 页)