第42卷，第10期
2022年10月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42，No. 10，pp3079-3083
October，2022

主成分分析排序和模糊线性判别分析的生菜近红外光谱分类

武 斌[1]，沈嘉棋[2]，汪 鑫[2]，武小红[3]，侯晓蕾[2]

1. 滁州职业技术学院信息工程学院，安徽 滁州 239000
2. 江苏大学卓越学院，江苏 镇江 212013
3. 江苏大学电气信息工程学院，江苏 镇江 212013

摘 要 贮存时间是影响生菜品质的一项重要因素，传统的贮存时间鉴别方法主要依靠人工经验，但是这种方法的准确率和可信度并不高。研究的目标是建立一种基于模糊识别的模型进行生菜光谱分析以实现生菜贮存时间的鉴别，并与其他鉴别方法作比较。为此，在当地超市购买60份新鲜生菜样品，存放于冰箱中待用。首先，通过Antaris Ⅱ近红外光谱检测仪采集生菜样品的近红外光谱数据，每隔12小时检测一次，每个样本检测重复三次，并取三次平均值作为实验数据。其次，利用多元散射校正(MSC)减少近红外光谱中的冗余信息。为了进一步去除近红外光谱中的无用信息以及简化随后的数据分类过程，分别运用主成分分析(PCA)和排序主成分分析(PCA Sort)。其中，PCA Sort通过改进对主成分的排序方法能提高分类准确率，同时便于模糊线性鉴别分析(FLDA)进一步提取特征。PCA和PCA Sort的计算仅运用了前15个主成分(能充分反映光谱的主要信息)。最后，利用模糊线性鉴别分析算法(FLDA)和K近邻算法(KNN)进一步分类所得的低维数据。基于PCA和KNN算法的模型鉴别准确率达到43%，而基于PCA，FLDA和KNN算法的模型鉴别准确率可达83%。上述结果说明基于PCA，FLDA和KNN算法的模型鉴别准确率已经得到较大程度提高。当用PCA Sort替代了模型中的PCA算法后，结合FLDA和KNN算法则鉴别准确率达到98.33%。实验结果表明PCA Sort结合FLDA和KNN所建立的模型是有效的生菜贮存时间鉴别模型。

关键词 近红外光谱；主成分分析；生菜；模糊鉴别线性分析；K近邻算法

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)10-3079-05

引 言

目前，冰箱可以用来保鲜生菜，但随着贮藏时间的延长，生菜中亚硝酸盐含量在不断增加而损害人体健康。更重要的是，长期贮存会导致其中的水和大多数营养物质的含量下降[1]。例如，董伟等人选取华南4种高叶酸含量作物作为实验材料，实验表明，因长期贮存平均叶酸损失达到23%。因此，确定食物的贮存时间具有重要意义[2]。本工作以生菜为例，探讨一种有效的蔬菜新鲜度检测方法。

传统的人工筛选是评估食物新鲜度最常见的方法。有经验的人可通过观察食物的外部特征(比如颜色、形状和味道)快速做出判断。然而，受一些内部和外部因素的影响，人工筛选是主观的，缺乏准确性。因此，研究人员通过多种方法进行实验以检测食品的质量和贮存时间。高婷婷等[3]结合时间-温度指标(time-temperature indicators，TTIS)，借助一些常用的建模方法，如测定反应速率常数(k)和活化能(E_a)以及E_a匹配来监测新鲜食品质量。为了快速准确地评价罗非鱼鱼片的新鲜度，Han等[4]利用电子舌结合线性和非线性多元算法检测鱼的新鲜度。

近红外光谱技术具有快速、无损、操作简单、精度高、成本低等优点。目前，环境分析、食品工程[5-6]、食品新鲜度检测[7]等不同领域的许多研究人员都应用了近红外反射光谱(near infrared reflectance spectroscopy，NIRS)技术。

模糊线性判别分析(fuzzy linear discriminant analysis，FLDA)是一种有监督的特征提取和降维方法，该算法也被广泛应用于分类及其他领域。例如Guidea等[8]借助FLDA算法对矿泉水中的矿物成分进行分析分类，有效区分了来自罗

收稿日期：2021-07-19，**修订日期**：2022-03-17

基金项目：国家自然科学基金项目(31471413)，滁州职业技术学院校级自科重点项目(YJZ-2020-12)，滁州职业技术学院院级人才项目“优秀骨干教师”(YG2019026，YG2019024)，安徽省质量工程项目(2020SJJXSFK1864，2020kfkc370)，江苏大学大学生创新训练计划项目(202010299246Y)资助

作者简介：武 斌，1978年生，滁州职业技术学院信息工程学院副教授 e-mail：wubind2003@163.com

马尼亚和德国的矿泉水，正确分类率达到 88%。Shen 等[9]应用 FLDA 对白菜的中红外光谱进行特征提取，并使用 K-最近邻法(K-nearest neighbor，KNN)进行样本分类，实现了无损检测白菜是否有 λ-三氯氟氰菊酯农药残留。

K-最近邻法(KNN)是一种有监督的分类方法[10]。Chen 等[11]为快速无损地检测猪肉的储存时间，分别使用线性判别分析(linear discriminant analysis，LDA)、K-最近邻(KNN)、反向传播人工神经网络(back propagation artificial neural network，BP-ANN)等算法建立了猪肉储存时间判别模型，结果表明 BP-ANN 模型在训练集和预测集中的判别率分别为 99.26%和 96.21%。

在主成分分析算法(principal component analysis，PCA)的基础上，采用新排序原则对特征向量进行重组的 principal component analysis sort (PCA Sort)算法，并建立生菜贮藏时间的判别模型。首先，利用 Antatis Ⅱ型近红外光谱仪采集生菜的近红外光谱数据，并利用多元散射校正(multiple scatter correction，MSC)消除光散射的影响，对预处理后的数据分别采用 PCA，PCA+FLDA 和 PCA Sort+FLDA 等方法进行分析。最后利用 KNN 进行分类，确定各组生菜的贮存时间，计算并比较这三种方法的鉴别结果。

1 实验部分

1.1 样本

从镇江一家超市购买生菜。为了减小误差，实验材料应符合一定的标准。所有的生菜样品(60 个样品)保证是在同一时间(新鲜和成熟)采摘的，大小、颜色、重量和叶子的完整性没有太大的差异。用水清洗和晾干后，生菜样品被放入有标签的塑料袋中，并放入 4 ℃保鲜柜中备用。

1.2 实验环境与计算运行环境

采用美国 Thermo Antaris Ⅱ型近红外光谱仪获取生菜的近红外反射光谱。在整个实验过程中，由于近红外光谱对外界环境敏感，实验室保持温度在 20～25 ℃，空气相对湿度在 50%～60%。

所有计算均在 Windows 10 的 MATLABR2020a(Math Works，Natick，MA，USA)运行。

1.3 近红外光谱数据采集

光谱仪需要提前开机预热 1 h。采用反射积分球模式采集样品的近红外光谱，对每个样品扫描 32 次，得到漫反射光谱的平均值。光谱扫描的波数范围为 10 000～4 000 cm^{-1}，扫描间隔为 3.856 cm^{-1}。实验开始后，每隔 12 h 取出生菜样品进行近红外光谱检测，共检测三次，取其平均值，每个样品采集的近红外光谱为 1557 维数据。

在生菜原始近红外光谱中，受环境影响，易发生噪声、样本异质性、基线漂移和偏移[12]。多元散射校正(MSC)可有效消除不同散射水平引起的光谱差异。故采用 MSC 对初始近红外光谱进行预处理。图 1 为 MSC 预处理后的光谱图。

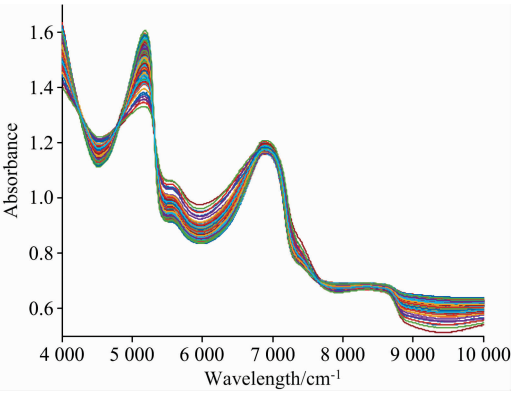

图 1 MSC 处理后的生菜样本近红外光谱

Fig. 1 NIR spectral data of lettuce samples treated by MSC

1.4 PCA Sort

采集的生菜样品近红外光谱有 1557 维，属于高维数据，同时光谱中含有大量无用信息和噪声数据，增加了分析、建模和计算的难度，故需对近红外光谱进行降维以提取生菜近红外光谱的主要特征信息。主成分分析(PCA)可对生菜近红外光谱数据进行降维，同时较好地保留主要特征信息。然而，PCA 在降维过程中会丢失一些鉴别信息而导致分类准确率降低。为提高分类的准确率，对 PCA 算法进行了改进，按照一定的规则改变其特征向量的顺序。具体算法如下：

(1)设训练样本组成的矩阵为 $\boldsymbol{A}$，$\boldsymbol{A}\in R^{n\times d}$($n$ 为训练样本数，d 为训练样本维数)。

(2) 用训练样本矩阵 $\boldsymbol{A}$ 组成协方差矩阵 $\boldsymbol{S}$

$$\boldsymbol{S}=\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})^{\mathrm{T}} \tag{1}$$

式(1)中，x_k 为第 k 个训练样本，$\bar{x}$ 为训练样本的平均值，$\bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j$。

(3) 根据式 $\boldsymbol{S}v=\lambda v$，对矩阵 $\boldsymbol{S}$ 进行特征分解，得到一组特征向量 v_1，v_2，…，v_n，λ 和 v 分别是特征值和对应的特征向量。

(4) 计算类内散射矩阵 $\boldsymbol{S}_w$ 与类间散射矩阵 $\boldsymbol{S}_b$

$$\boldsymbol{S}_w=\sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_i^{(j)}-m_i)(x_i^{(j)}-m_i)^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{S}_b=\sum_{i=1}^{c}n_i(m_i-\bar{m})(m_i-\bar{m})^{\mathrm{T}} \tag{3}$$

式(2)中，c 为样本总类别数；n_i 为第 i 类样本个数，$i=1,2,\cdots,c$；$x_i^{(j)}$ 为第 i 类中的第 j 个训练样本，$j=1,2,\cdots,n_i$，m_i 为第 i 类样本均值，$\bar{m}$ 为总体样本均值，T 为矩阵转置运算。

(5) 根据式(4)计算特征向量 v_1，v_2，…，v_n 的 $J(v_k)$值，并按 $J(v_k)$从大到小的规则将 v_1，v_2，…，$v_n(n=\sum_{i=1}^{c}n_i)$ 进行排序，得到一组新的特征向量 $w=[w_1,w_2,\cdots,w_n]^{\mathrm{T}}$，其中 w_1 为式(4)最大值时对应的特征向量。

$$J(v_k)=\frac{|v_k^{\mathrm{T}}\boldsymbol{S}_Bv_k|}{|v_k^{\mathrm{T}}\boldsymbol{S}_wv_k|} \tag{4}$$

(6) 将第 k 个训练样本 x_k 和测试样本 y_k 根据式(5)和式(6)投影到特征向量 w 上，其中 z_k 是训练样本 x_k 在特征向量 w 上的投影，t_k 是测试样本 y_k 在特征向量 w 上的投影。

$$z_k = w^{\mathrm{T}} x_k \tag{5}$$

$$t_k = w^{\mathrm{T}} y_k \tag{6}$$

1.5　模糊线性判别分析(FLDA)

模糊线性判别分析(FLDA)用式(5)中训练样本 z_k 的每类均值作为聚类中心，计算出类内离散度矩阵 $\boldsymbol{S}_{fw}$ 和模糊类间离散度矩阵 $\boldsymbol{S}_{fb}$。计算矩阵$[\boldsymbol{S}_{fw}]^{-1}\boldsymbol{S}_{fb}$的特征值和特征向量及投影空间，并将训练样本和测试样本投影到得到的特征向量上。FLDA 具体算法描述见文献[13]，m 为 FLDA 中的权重指数。

2　结果与讨论

2.1　PCA 和 PCA Sort 的特征向量数对鉴别准确率的影响

分别使用 PCA 和 PCA Sort 对生菜近红外光谱降维。计算结果表明，前 15 个主成分充分反映了生菜近红外光谱的大部分信息。分别使用 PCA 和 PCA Sort 算法得到它们的前 6～15 个特征向量，并用 FLDA 和 KNN 进一步处理，得到各自的准确率如表 1 所示。由表 1 可知 PCA Sort 的准确率要高于 PCA，当取 $m=2$，$K=3$ 时，PCA Sort 的准确率达到 98.33%，高于 PCA 算法(83.33%)。

表 1　PCA 和 PCA Sort 的前 6～15 个特征向量及其准确率

Table 1　The first 6～15 eigenvectors and accuracies of PCA and PCA Sort

The number of eigenvectors	PCA	PCA Sort
	accuracy/%	accuracy/%
6	48.33	78.33
7	61.67	88.33
8	83.33	86.67
9	75.00	85.00
10	81.67	88.33
11	85.00	95.00
12	78.33	93.33
13	76.67	90.00
14	76.67	96.67
15	83.33	98.33

2.2　参数 m 和 K 对鉴别准确率的影响

为确定恰当的权重系数 m，首先计算出 m 在 1～15 范围内取值时的准确率，当 m 增加时，总体准确率下降。结果如图 2(a)所示，发现权重系数较小时，PCA+FLDA+KNN 和 PCA Sort+FLDA+KNN 的分类结果相对准确。因此，将计算范围缩小到 m 为 1～5，寻找更为精确的 m。缩小范围[如图 2(b)所示]，当权重系数取 $m=2$ 时，分类的准确率最高。

KNN 参数 K 的取值也是影响分类结果的一个因素。计算 K 取 1～20 内的奇数时对应的各项准确率，结果如图 3 所示，由此可知，$K=3$ 时 PCASort+FLDA+KNN 的分类准确率最高，此时 PCA+KNN 及 PCA+FLDA+KNN 两种方法的分类准确率也近似最优结果，故 K 取 3。

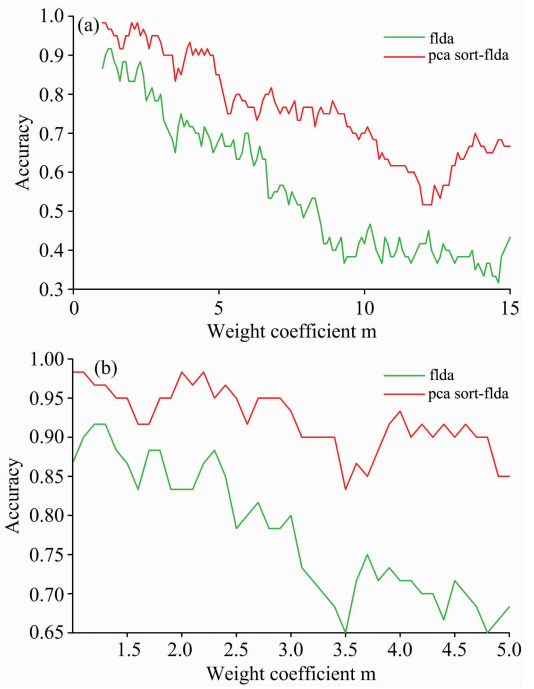

图 2　改变权重系数时准确率的变化

(a)：权重系数 m 为 1～15；(b)：权重系数 m 为 1～5

Fig. 2　The accuracy changing with the weight coefficient (m)

(a)：m in the range of 1～15；(b)：m in the range of 1～5

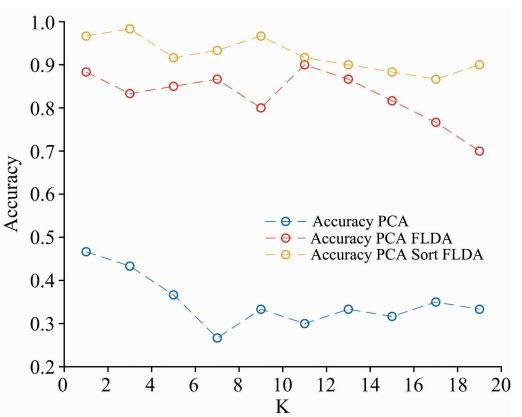

图 3　K 改变时分类准确率的变化

Fig. 3　The accuracy changing with the number of cluster center (K)

三种方法的最高准确率如表 2 所示。分析表 2 可以看出，采用 FLDA 算法并结合 PCA 和 KNN，准确率近似为 PCA 和 KNN 算法的两倍，达到 83%。而用 PCA Sort 代替 PCA，准确率则进一步提高，达到 98%。因此，使用 PCA Sort，并用 FLDA 进行特征提取，再用 KNN 进行分类，具有更好的优越性。

表 2 三种方法的最高准确率

Table 2 The highest accuracy of three methods

Algorithm	Accuracy/%
PCA+KNN	43
PCA+FLDA+KNN	83
PCA Sort+FLDA+KNN	98

3 结 论

提出一种新的特征提取方法，即 PCA Sort+FLDA，以降低数据的维数，提取生菜近红外光谱的特征信息，基于该方法及近红外光谱技术，建立了一种比传统人工筛选方法具有更多优势的分类模型。通过比较 PCA+KNN，PCA+FLDA+KNN 和 PCA Sort+FLDA+KNN 三种方法的分类准确率，发现当使用 PCA Sort 得到新的特征向量空间，并用 FLDA 进行特征提取，可以提高 KNN 分类的准确率(达到最高的 98%)。综上所述，近红外光谱结合 PCA Sort，FLDA 和 KNN 可大幅提高生菜贮藏时间的识别准确率，也为其他食品贮藏时间的测定提供了可行的参考方法。

References

[1] Bie T Y, Xu J S, Yang L L, et al. Food Research and Development, 2012, 33(12): 205.
[2] Dong W, Cheng Z J, Wang X L, et al. International Journal of Food Sciences and Nutrition, 2011, 62(5): 537.
[3] Gao T T, Tian Y, Zhu Z W, et al. Trends in Food Science & Technology, 2020, 99: 311.
[4] Sun L, Yuan L M, Cai J R, et al. Food Analytical Methods, 2015, 8(4): 922.
[5] Sun J, Ge X, Wu X H, et al. Journal of Food Processing and Preservation, 2018, 41(6): e12816.
[6] Jiang S Y, Sun J, Zhou X, et al. Journal of Food Processing and Preservation, 2017, 40(4): e12510.
[7] Yu H D, Zuo S M, Xia G H, et al. Food Analytical Methods, 2020, 6: 1.
[8] Guidea A, Gaceanu R D, Pop H F. Studia Universitatis Babes-Bolyai Chemia, 2020, 65: 45.
[9] Shen Y J, Wu X H, Wu B, et al. Agriculture, 2021, 11: 275.
[10] Nayak S K, Panda M, Palai G. Optik, 2020, 212: 164675.
[11] Chen Q S, Cai J R, Wang X M, et al. LWT—Food Science and Technology, 2011, 2053: 2058.
[12] Wu X H, Zhou H X, Wu B, et al. Journal of Food Processing and Preservation, 2020, 44(8): e14561.
[13] Wu X H, Wu B, Sun J, et al. Journal of Food Process Engineering 2017, 40(2): e12355.

NIR Spectral Classification of Lettuce Using Principal Component Analysis Sort and Fuzzy Linear Discriminant Analysis

WU Bin[1], SHEN Jia-qi[2], WANG Xin[2], WU Xiao-hong[3], HOU Xiao-lei[2]

1. Department of Information Engineering, Chuzhou Polytechnic, Chuzhou 239000, China
2. Institute of Talented Engineering Students, Jiangsu University, Zhenjiang 212013, China
3. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China

Abstract The storage time of lettuce is an important factor affecting the quality. The traditional way of detecting lettuce storage time mostly depends on artificial experience, so it lacks accuracy and reliability. This study aims to provide a fuzzy recognition model for spectral analysis of lettuce to identify the storage time of lettuce compared with other discriminant methods. For this objective, sixty samples of fresh lettuce bought in the local supermarket were prepared and stored in a refrigerator for later detection. These samples were detected by near-infrared spectroscopy (NIR). Firstly, the Antaris II NIR spectrometer (the wave number range: 10 000～4 000 cm^{-1}) was utilized to collect the near-infrared spectral data of lettuce samples every 12 hours, and every sample detection was repeated three times, taking the average value as experiment data. Secondly, NIR spectra were preprocessed with multiple scatter correction (MSC) for decreasing reductant information. PCA and PCA Sort were used to further clear the useless data of NIR spectra and simplify the following classification of data. PCA Sort was based on PCA with sorting principal components and could improve the classification accuracy and help the FLDA extract features effectively. In this step, only the first fifteen components of PCA and PCA Sort were used to compress NIR spectra. Finally, fuzzy linear discriminant analysis (FLDA) algorithm and k-nearest neighbor (KNN) were performed to classify the previous low-dimensional data. The classification accuracy of the model based on PCA coupled with KNN was 43%, and that based on PCA as well as FLDA and KNN was 83%. The classification results in experiments showed that the discriminant of the model based on PCA,

FLDA and KNN was significantly improved. Replacing PCA in the model with PCA Sort, the recognition accuracy of this new model based on the algorithm PCA Sortcoupled with FLDA and KNN was better and achieved 98.33%, which was higher than other classification algorithms. The classification results in experiments showed that PCA Sort plus FLDA and KNN could build an efficient discrimination model for the identification of the storage time of lettuce.

Keywords　NIR spectra; Principal component analysis; Lettuce; Fuzzy linear discriminant analysis; K-nearest neighbor

(Received Jul. 19, 2021; accepted Mar. 17, 2022)