流形学习方法及近红外透射光谱的新疆冰糖心红富士水心鉴别

郭俊先[1]，马永杰[1]，郭志明[2]，黄 华[3]，史 勇[1]，周 军[1]

1. 新疆农业大学机电工程学院，新疆 乌鲁木齐 830052
2. 江苏大学食品与生物工程学院，江苏 镇江 212013
3. 新疆农业大学数理学院，新疆 乌鲁木齐 830052

摘 要 苹果水心病在众多苹果主产区都有发生，现阶段没有合适的方法实现快速鉴别和分类。为了探索苹果水心鉴别新方法，采用近红外透射光谱与化学计量学方法结合非线性流形学习数据降维技术，逐个采集好果与疑似水心病果样本590～1 250 nm的近红外透射光谱，将经光谱校正后的原始光谱做多元散射校正(multivariate scattering correction，MSC)、标准正态变量变换(standard normal variate transformation，SNVT)、二阶求导(2nd derivative)、一阶求导(1st derivative)、归一化(normalization)、卷积平滑法(savitzky-golay smoothing，SG)、均值中心化(mean centering，MC)、移动平均平滑(moving average，MA)、直接差分二阶求导(direct differential second derivative，DDSD)以及直接差分一阶求导(direct differential first derivative，DDFD)等10余种光谱预处理；先对预处理后的光谱数据建立全波长模式识别模型从而找出多元散射校正是最优预处理方法，而后再分别用多维尺度分析(multidimensional scaling，MDS)、分布邻域嵌入(stochastic neighbor embedding，SNE)、对称分布邻域嵌入(symmetric stochastic neighbor embedding，Sym-SNE)、t分布邻域嵌入(t-distributed stochastic neighbor embedding，t-SNE)、拉普拉斯映射(laplacian eigenmaps，LE)、等距特征映射(isomap)、地标等距映射(landmark isomap)、局部线性嵌入(locally linear embedding，LLE)、扩散映射(diffusion maps，DM)等多种流形学习方法对经多元散射校正预处理后的光谱数据做降维处理，并结合马氏距离判别(mahalanobis distance discrimination，MD)、二次判别分析(quadratic discriminant analysis，QDA)、贝叶斯判别(Bayesian discrimination，BD)、K最近邻法(K nearest neighbor，KNN)识别其水心存在与否。结果表明，提取前12主成分，采用多元散射校正-地标等距映射-K最近邻法(MSC-landmark isomap-KNN)模型识别效果最优，校正集和预测集识别率分别为97.5%和96.3%。故，流形学习方法结合近红外透射光谱可成功、高效地实现苹果水心鉴别，为进一步研发水心鉴别设备提供新的理论指导。

关键词 苹果水心病；近红外光谱；化学计量法；流形学习；模式识别
中图分类号：S661.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)08-2415-06

引 言

苹果水心病，又叫苹果蜜病、糖化病，在众多苹果产区是一种常见的果实生理病害[1]。水心病的产生与钙含量的缺失有关，高钙果实水心病发生率要低于低钙果实[2]。由于苹果水心病主要产生于果心部位，故，水心病果在染病情况下一般无法直接用肉眼察觉，除非染病情况非常严重，已经在苹果表面形成水渍状白斑[3]。正是因为钙含量的缺失导致水心病果由于山梨糖醇的累积使其拥有得天独厚的味觉优势，这已经日趋成为商家提高苹果商品价格的一个卖点，新疆阿克苏冰糖心红富士苹果更是以此闻名于国内市场，但这也仅限于水心含量较低的病果；严重的水心病果会随着储藏时间的增加从而霉变、褐变、甚至黑心，这在一定程度上影响到苹果的质量和销售链各环节人员的经济收入。

截止目前，苹果水心病的无损检测已经吸引了一大批国内外学者对其进行相关研究，并且都取得了一定的成果。包括计算机视觉、电子鼻、电学特征法、密度法、核磁共振、色彩像素、X光、热成像在内的多种方法[4]。但是受限于成本、反应时间、甚至识别率不可靠的问题，并没有得到广泛的应

收稿日期：2019-07-05，**修订日期**：2019-11-18
基金项目：国家自然科学基金项目(61367001)资助
作者简介：郭俊先，1975年生，新疆农业大学机电工程学院教授 e-mail：junxianguo@163.com

用。国内目前的水心病检测还是以人力经验分类为主，准确率及可靠性都不能得到确切保证。因此，急需无损、可靠、快捷的方法对摘后苹果进行水心鉴别、分类储存、销售，进而提高苹果的经济价值。

近红外光谱技术是近年来无损检测的热门技术。广泛应用于农产品、化工、药材、木材、医学领域的定性、定量分析，由于分析速度快、不破坏样本等优点受到了众多专家学者的青睐[5]。但是近红外光谱波段数量较大、波段间的相关性较高，一定程度上存在数据冗余等问题。因此数据降维也一直是近红外光谱数据分析的必经流程之一，常用的线性降维方法有主成分分析、独立成分分析等。但由于线性降维算法在降维后不能很好保持复杂结构高维数据的完整信息，所以产生了对非线性降维的需求，也就有了从数学拓扑出发的流形映射，这是因为大部分现实中非线性结构都可以看做是流形结构[6]；流形学习是近年来发展起来的一种非线性数据降维方式，其主要思想是将高维的数据映射到低维，并且在数据降维的同时能够表征近红外光谱数据的某些本质结构特征[7]。目前主要的流形学习算法有多维尺度分析(MDS)[8]、分布邻域嵌入(SNE)[9]、对称分布邻域嵌入(SymSNE)[10]、t分布邻域嵌入(t-SNE)[11]、拉普拉斯映射(LE)[12]、等距特征映射(isomap)[13]、地标等距映射(landmark isomap)[14]、局部线性嵌入(LLE)[15]、扩散映射(diffusion maps)[16]等。以上研究结果虽然预示了利用近红外光谱技术结合流形学习方法应用于检测苹果内部水心存在与否的可能性，但是利用近红外透射光谱结合多种流形学习方法对苹果水心的鉴别研究还未有相关报道。

以新疆冰糖心红富士好果与水心病疑似病果为研究对象，采集其近红外透射光谱，结合化学计量学建模方法及流形学习方法。第一，探究在多元散射校正、标准正态变量变换、二阶求导、一阶求导、归一化、卷积平滑法、均值中心化、移动平均平滑、直接差分二阶求导以及直接差分一阶求导等 10 种光谱预处理方法下全波长建模的性能，从中筛选最优预处理方法；第二、以最优预处理方法下的光谱数据集作为自变量，结合多维尺度分析(MDS)、分布邻域嵌入(SNE)等 9 种流形学习方法进行非线性数据降维，分别对比其在马氏距离判别(MD)、K 最近邻法(KNN)、二次判别分析(QDA)、贝叶斯判别(BD)模式识别模型下水心鉴别的性能，从中挑选最优模型建立近红外透射光谱结合非线性流形学习方法的苹果水心病预测模型，为开发水心快速无损检测设备提供新思路。

1 实验部分

1.1 材料

样本“新疆冰糖心红富士”好果与疑似水心病果于 2018 年 3 月 6 日购买于新疆乌鲁木齐市北园春水果批发市场。由经验丰富的果商对同批次同品牌苹果拆箱挑选大小尺寸均匀、无明显损伤的正常果及疑似病果共计 230 个，装箱转运至实验室后开箱平铺、室温 20 ℃静置 24 h，擦净苹果表面浮土并检查损伤情况并逐个编号、剔除受损的苹果共计 13 个，其余 217 枚参与数据采集。

1.2 设备

近红外透射光谱采集系统来自江苏大学食品与生物工程学院无损检测重点实验室，如图 1 所示，采集系统由苹果托架、配套小型风扇的光源套件(JCR12V 100 W 卤钨灯)、近红外光谱仪(USB2000＋，Ocean Optics，USA)、大芯径双包层石英光纤(标准 SMA905 接口)、机架、暗箱与计算机等组成。光纤探头一端连接近红外透射光谱仪，另一端安装于果托圆心正下方，实现对透射光谱的高效采集。

图 1 近红外透射光谱采集系统

Fig. 1 Near-infrared transmission spectrum acquisition system

1.3 方法

1.3.1 近红外透射光谱的采集

近红外透射光谱数据采集由配套软件 SpectraSuite (Ocean Optics，USA)实现，后续数据处理使用 Matlab 2014b (MathWorks，USA)，绘图使用 OriginPro 8 (OriginLab，USA)。

USB2000＋光谱仪使用前预热约 1 h，之后通过测试采样设置 SpectraSuite 软件界面参数，确定样品光谱的采集参数为：平均次数 3；平滑度 5；积分时间 120 ms；波段数 512。采集光谱时，将苹果按图示置于光谱采集仪器的果托上，注意苹果与底部果托之间不能留有光缝，确保光纤接收光信号的点完全屏蔽光源，使其只能接收到透过苹果的光。等待软件界面显示的光谱稳定且在微小时间内不再发生变化时，保存光谱；然后将苹果顺时针旋转 120°，保存光谱之后再次将苹果顺时针旋转 120°并保存光谱，最终将三次获得光谱的平均值作为该样本的光谱数据。共计采集 217 个苹果近红外透射光谱信息。每测量 10 个样本就需要保存一次该时刻的暗光谱用于后续光谱校正，这是为了避免由于 USB2000＋光纤光谱仪预热不充分导致暗光谱所产生的试验误差。

1.3.2 光谱校正处理

光谱采集过程中，由于摄像头中的暗电流噪声以及苹果表面不均匀性会对光谱数据产生一定的噪声影响，因此需要对获得的光谱数据按照式(1)进行校正

$$R = \frac{I_0 - I_D}{I_W - I_D} \tag{1}$$

式(1)中：R 为校正后的光谱；I_0 为原始采集光谱；I_D 为拧上镜头盖采集的全黑暗光谱；I_W 为全反射光谱。

1.3.3 苹果分类

光谱采集后，沿赤道面横向剖切并记录水心状况，经统

计共有水心病果 97 个，正常苹果 120 个。将所有采集后的近红外透射光谱数据用 SPXY 方法将样本按照 3∶1 的比例分为校正集与预测集。其中校正集为 72 个水心苹果与 90 个正常苹果，预测集为 25 个水心苹果与 30 个正常苹果。

1.4　数据处理

①将 SpectraSuite 软件获得的原始近红外光谱信息导入 matlab 2014b 软件进行黑白校正；②随后将校正过的光谱信息作为输入变量对其进行数据中心化、移动平均平滑等多种光谱预处理；③再将经过预处理后的光谱建立全波长模式识别模型，从中筛选最优光谱预处理方法；④将最优光谱预处理后提取的光谱数据应用多种流形学习方法进行数据降维；⑤将数据降维后的结果分别用于模式识别。

2　结果与讨论

2.1　光谱数据预处理与筛选

正常新疆冰糖心红富士苹果与水心病果经黑白校正的所有原始光谱见图 2，其形状、趋势相似，肉眼几乎无法辨别。并且原始光谱数据的获取除了包含样本自身有用的化学信息外，常常伴随着其他无关信息，如样本背景、杂散光、以及设备自身的干扰。因此，旨在消除光谱无关信息的光谱预处理方法变得尤为重要。为了消除以上干扰，同时也为排除其他不确定干扰及筛选最优预处理方式；分别采用 10 种预处理方法对原始光谱进行预处理，随后将预处理后的数据作为建模集进行全光谱模式识别建模研究。结果如表 1。

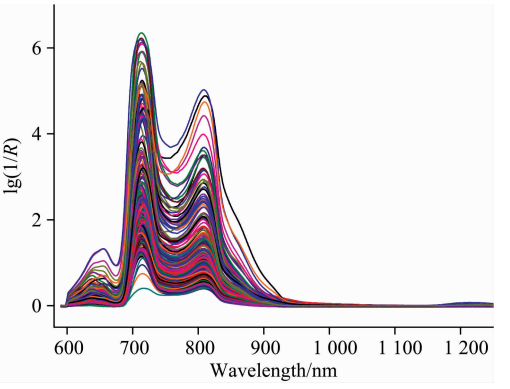

图 2　正常苹果与水心病果的近红外透射光谱

Fig. 2　Near-infrared transmission spectra of healthy and watercore apples

表 1　全光谱下多种预处理方法结合多种模式识别方法建模的苹果类型识别率

Table 1　Recognition rate of apple type with multiple preprocessing and multiple pattern recognition under full spectra

	BD		MD		QDA		KNN	
	Calibration set accuracy /%	Prediction set accuracy /%	Calibration set accuracy /%	Prediction set accuracy /%	Calibration set accuracy /%	Prediction set accuracy /%	Calibration set accuracy /%	Prediction set accuracy /%
SNVT	83.9	72.7	75.9	70.9	76.5	69.0	87.0	72.7
MSC	85.1	80.0	82.7	87.2	83.3	87.2	93.2	87.2
2nd derivative	84.5	76.3	82.7	85.4	81.4	81.8	84.5	85.4
Normalization	83.9	72.7	79.6	81.8	81.4	81.8	88.8	85.4
SG	83.9	72.7	82.0	83.6	82.0	83.6	86.4	85.4
MC	84.5	85.4	82.0	83.6	82.0	83.6	85.1	83.6
1st derivative	85.1	74.5	82.7	87.2	83.9	85.4	85.1	81.8
MA	85.1	72.7	82.0	83.6	82.0	83.6	86.4	87.2
Original	83.9	72.7	82.0	83.6	82.0	83.6	83.9	85.4
DDFD	81.4	74.5	82.7	81.8	85.1	85.4	83.9	83.6
DDSD	85.1	76.3	82.7	83.6	83.3	85.4	85.8	83.6

由表 1 可见，预处理在一些情况下会起到“促进”建模的过程，有时也会“抑制”。模式识别方法亦是如此，不同的方法也会因为原理不同得到相左的效果。预处理方法最优的是多元散射校正，模式识别方法中较好的是马氏距离判别、K 最近邻法及二次判别分析。图中 MSC-KNN 模型的校正集与预测集的成功率为 93.2%与 87.2%，达到了较优的识别效果，但是由于全光谱建模耗时长、变量复杂并且不稳定等缺点，所以并不能直接应用于生产实际中，于是需要将经过最优预处理方法，即多元散射校正处理后的光谱经流形学习方法数据降维后再结合本试验匹配较优的模式识别方法(即 MD，KNN，QDA)进行建模分析。

2.2　基于流形学习方法的数据降维

为了对比各种流形学习方法对数据降维的效果。将经过多元散射校正预处理后的光谱数据集结合流形学习方法进行数据降维，输出维度确定为 3，绘制多种流形学习方法下提取的三维特征散点图如图 3 所示。

图 3 可见：绿色为水心病果、蓝色为正常苹果。流形学习方法在前三个主成分下的降维效果各有特点；图 3(a)，(b)，(i)，(l)将正常苹果与水心病果几乎完全分开，只是在空间上看起来有重叠；并且图 3(a)与(l)有相似的降维效果，这与成超[17]所得结果一致；图 3(c)和(g)在空间中产生了流形状曲线，并且在水心分类上产生了较好的结果；图 3(d)与(e)不仅没有将正常苹果与水心病果很好分开，相比常规 PCA 方法使数据变得更加混淆，对于本试验来说并不是合

适的数据降维方法；图 3(f)很好地将水心病果与正常苹果分开，说明相对于 SNE 方法，t-SNE 既能获得原始高维数据的重要信息，同时也能很好地揭示全局簇结构；图 3(h)虽然也将两类样本成功分开，但是水心病果在图中重合太多，并不能直观看出聚类分布；图 3(j)与(k)也能很好地将两类苹果分开。纵观图 3，发现大部分小图中的绿色区域(即水心病果)都呈散开趋势，推测这或许与水心病果的水心大小存在差异有一定联系。

上述经流形学习方法数据降维的结果都被用来结合经前段测试后与本试验匹配较优的模式识别方法(即 MD，KNN，QDA)进行建模分析；在分析中 0 与 1 被分别用来代表水心病果与正常苹果。本次分析的目的在于从中找出最高效、最简单、最快捷的基于流形学习方法与近红外透射光谱的水心鉴别模型，结果如表 2；本次分析数据只考虑主成分数目与不同流形学习方法对苹果识别率的影响，故为了消除其他干扰对建模的影响，其他参数均不做任何改变。(经优选后确定 KNN 中 K 全部取 5)

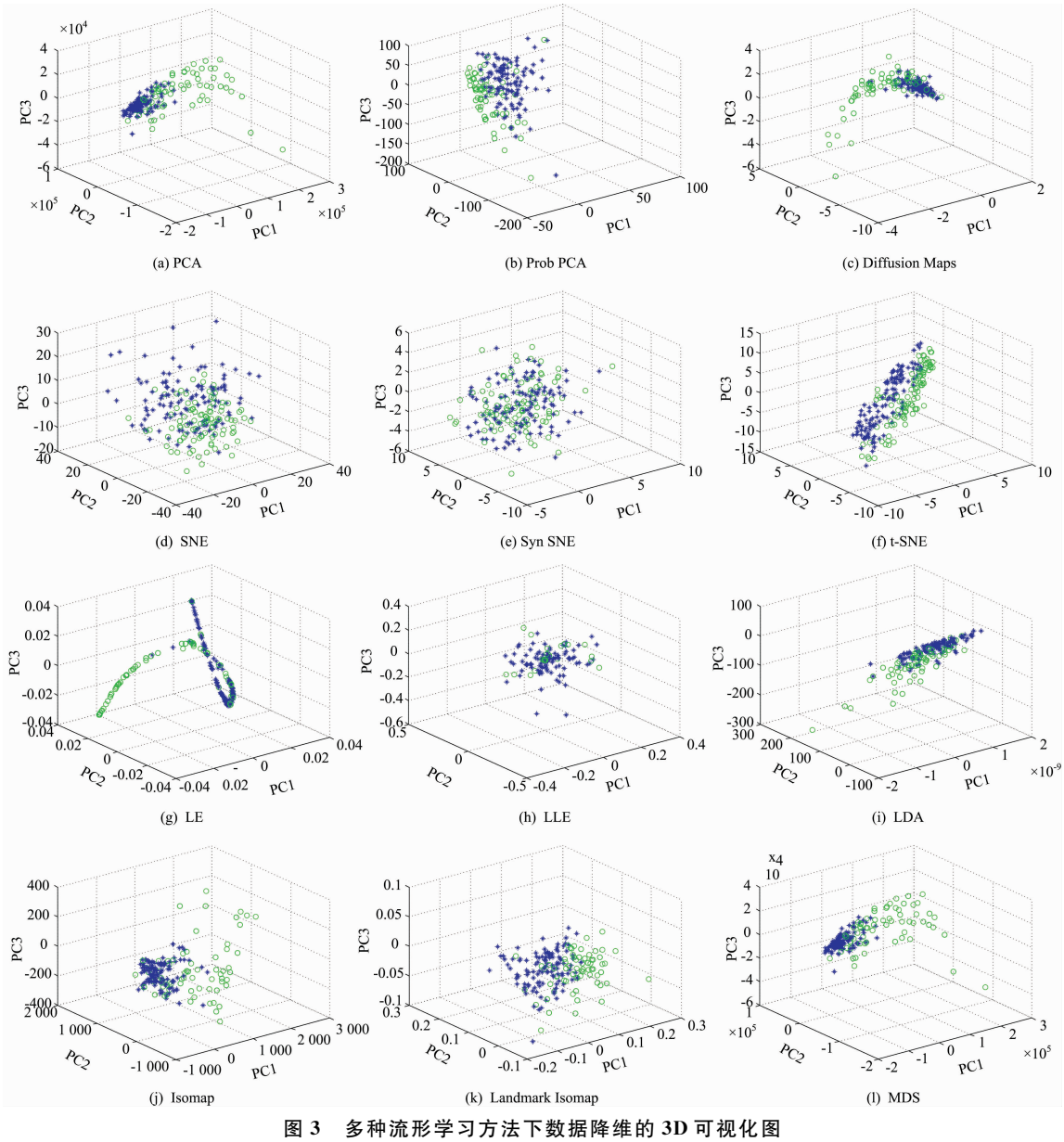

图 3 多种流形学习方法下数据降维的 3D 可视化图

Fig. 3 3D visualization of dimension reduction under various manifold learning methods

参照图 3 分析结果，马氏距离判别分析下最优模型为 MSC-Diffusion Maps-MD，校正集与预测集识别率分别为 91.3%与 90.9%。这与图 3 经多种流形学习方法数据降维后的 3D 可视化图结果吻合。K 最近邻法方法下最优模型为 MSC-Landmark Isomap-KNN，校正集识别率为 97.5%，预测集为 96.3%。在二次线性判别分析方法下，最优模型为 MSC-t-SNE-QDA，校正集与预测集苹果类型的识别率分别为 95%与 90.9%。由表 2 同时获悉，SNE，SymSNE 并不适

表 2　多种流形学习数据降维方法结合多种模式识别方法建模的苹果类型识别率

Table 2　Recognition rate of apple type with multiple manifold learning dimension reducing algorithm and multiple pattern recognition

	MD			KNN(K=5)			QDA		
	number of principal component	Calibration set accuracy /%	Prediction set accuracy /%	number of principal component	Calibration set accuracy /%	Prediction set accuracy /%	number of principal component	Calibration set accuracy /%	Prediction set accuracy /%
PCA	10	93.8	85.4	16	94.4	92.7	17	93.2	85.4
Prob PCA	14	83.3	85.4	13	93.2	92.7	19	90.7	83.6
Diffusion Maps	18	91.3	90.9	19	97.5	90.9	15	90.1	89.0
SNE	13	83.3	74.5	19	90.1	83.6	14	82.7	81.8
SymSNE	7	68.5	65.4	12	73.4	65.4	8	71.6	65.4
t-SNE	13	81.4	80.0	19	97.5	90.9	18	95.0	90.9
LE	16	88.2	89.0	13	94.4	92.7	13	88.2	89.0
LLE	16	82.0	87.2	17	90.7	87.2	16	88.8	85.4
LDA	15	83.3	85.4	14	94.4	92.7	13	90.7	83.6
Isomap	17	90.1	90.9	11	97.5	92.7	13	93.2	85.4
Landmark Isomap	19	91.3	87.2	12	97.5	96.3	19	91.3	87.2
MDS	10	93.8	85.4	16	94.4	92.7	17	93.2	85.4

合应用于本试验的数据降维，这也与本文 2.2 节“基于流形学习方法的数据降维”的内容相符。

研究以新疆冰糖心红富士为试材，数据采集后切开检验水心病果占比仅为 44.7%，证明人工对水心筛选并不可行；经过本试验的分析，结合流形学习方法能够很好地提高水心识别率，正确率达到了 96.3%。从表 2 获悉，MSC-Diffusion Maps-KNN 与 MSC-t-SNE-KNN 虽然都能成功将水心病果与正常苹果分开，预测集识别率达到 90.9%，但是需要 19 个主成分的共同参与，推测正是因水心病果光谱与正常苹果光谱之间存在难以区分的差异性，所以导致需要多个主成分参与建模分析才能获取较优识别率。虽然基于流形学习方法与近红外透射光谱的新疆冰糖心红富士水心鉴别达到了成功率 96.3%的鉴别效果，但是在图 3(a)，(c)，(j)，(k)与(l)中发现了正常苹果成功聚类、水心病果呈现局部分散的问题，推测这可能与试验阶段没有考虑苹果水心大小对透射光谱的影响有关，在后续的试验与研究中，有望开展关于水心病果中水心大小预测的定量分析研究。

3　结　论

利用近红外透射光谱结合流形学习非线性数据降维方法与多种光谱预处理及多种模式识别方法同时对新疆冰糖心红富士进行水心分鉴，得到如下结论。

(1)对采集后经黑白校正后的光谱数据用多元散射校正、标准正态变量变换、二阶求导、一阶求导、归一化、卷积平滑法、均值中心化、移动平均平滑、直接差分二阶求导以及直接差分一阶求导进行光谱预处理；随后对所有光谱预处理后的数据做全波长建模分析，比较得出多元散射校正是最优预处理方法；其很好地消除了颗粒分布不均匀及散射对重要信息的影响，提高光谱信噪比。

(2)结合多维尺度分析(MDS)、分布邻域嵌入(SNE)、对称分布邻域嵌入(SymSNE)、t 分布邻域嵌入(t-SNE)、拉普拉斯映射(LE)、等距特征映射(Isomap)、地标等距映射(Landmark Isomap)、局部线性嵌入(LLE)、扩散映射(DM)等流形学习算法对经多元散射校正预处理后的数据集做数据降维处理，比较得出地标等距映射是最优数据降维方法，其在数据降维的同时很好地保留了数据原始的本构特征。

(3)对所有经过流形学习方法数据降维后的自变量分别建立马氏距离判别、K 最近邻法、二次判别分析的模式识别模型，比较得出 MSC-Landmark Isomap-KNN 是最优模型，校正集识别率为 97.5%，预测集识别率为 96.3%。

本方法为苹果水心鉴别提供技术支持，可以有效地对苹果进行水心鉴别、分类储存并销售，进而提高苹果的经济价值，为开发新型苹果水心病检测设备提供新思路。

References

[1] LIU Xiao-yong, ZHANG Hui-yuan, DONG Tie, et al(刘小勇，张辉元，董　铁，等). Journal of Fruit Science(果树学报), 2008, 25(5): 721.

[2] DU Yan-min, WANG Wen-hui, HANG Bo, et al(杜艳民，王文辉，杭　博，等). Acta Horticulturae Sinica(园艺学报), 2015, 42(10): 2023.

[3] ZHANG Xin-sheng, XIONG Xue-lin, ZHOU Wei, et al(张新生，熊学林，周　卫，等). Soil and Fertilizer Sciences in China(土壤肥料), 1999, (4): 3.

[4] YUAN Hong-fei, HU Xin-mu, YANG Jun-lin, et al(袁鸿飞, 胡馨木, 杨军林, 等). Food Science(食品科学), 2018, 39(16): 306.
[5] WANG Xue, MA Tie-min, YANG Tao, et al(王 雪, 马铁民, 杨 涛, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2018, 34(13): 203.
[6] SUN Jin-guang, DING Sheng-feng(孙劲光, 丁胜锋). Journal of China University of Mining & Technology(中国矿业大学学报), 2017, 46(4): 932.
[7] YUE Xue-jun, QUAN Dong-ping, HONG Tian-sheng, et al(岳学军, 全东平, 洪添胜, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2015, 46(6): 244.
[8] Chen L, Buja A. Publications of the American Statistical Association, 2009, 104(485): 209.
[9] KE Jia-jia, HU Jian-zhong(柯佳佳, 胡建中). Application Research of Computers(计算机应用研究), 2015, 32(10): 2992.
[10] Kumar S M, Balakrishnan G. Indian Journal of Science & Technology, 2016, 9(47): 12.
[11] MENG Xiao-chen, WANG Yue, ZHU Lian-qing(孟晓辰, 王 玥, 祝连庆). Journal of Biomedical Engineering(生物医学工程学杂志), 2018, 35(5): 697.
[12] SUN Wei-wei, LIU Chun, LI Wei-yue(孙伟伟, 刘 春, 李巍岳). Geomatics and Information Science of Wuhan University(武汉大学学报·信息科学版), 2015, 40(9): 1151.
[13] DING Ling, TANG Ping, LI Hong-yi(丁 玲, 唐 娉, 李宏益). Infrared and Laser Engineering(红外与激光工程), 2013, 42(10): 2707.
[14] Orsenigo C. Pattern Recognition Letters, 2014, 49: 131.
[15] Liu X, Tosun D, Weiner M W, et al. Neuro Image, 2013, 83: 148.
[16] NI Jia-peng, SHEN Tao, ZHU Yan, et al(倪家鹏, 沈 韬, 朱 艳, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(8): 2360.
[17] CHENG Chao, YANG Chen-hui(成 超, 杨晨晖). Journal of Xiamen University·Natural Science(厦门大学学报·自然科学版), 2017, 56(1): 123.

Watercore Identification of Xinjiang Fuji Apple Based on Manifold Learning Algorithm and Near Infrared Transmission Spectroscopy

GUO Jun-xian[1], MA Yong-jie[1], GUO Zhi-ming[2], HUANG Hua[3], SHI Yong[1], ZHOU Jun[1]
1. College of Mechanical and Electronic Engineering, Xinjiang Agricultural University, Urumqi 830052, China
2. College of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China
3. College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi 830052, China

Abstract Apple watercore occurs in many major apple producing areas, while there is no suitable way to sort apple type with watercore quickly. This research applies near infrared transmission spectroscopy, chemometric methods and manifold learning algorithm, selecting Xinjiang Red Fuji apple and watercore disease ones as samples, collecting near infrared transmission spectrum within 590 to 1 250 nm, spectroscopically corrected spectrum is used to do ten more speciesof spectral pretreatment. Firstly, full-wavelength pattern recognition is performed on the pre-processed spectral data to find out that multivariate scattering correction is the best pretreatment method. Then dataset preprocessed by multivariate scattering correction is used to make dimension reduction by using many other manifold learning algorithms such as Multidimensional Scaling, Stochastic Neighbor Embedding, Symmetric Stochastic Neighbor Embedding, t-Distributed Stochastic Neighbor Embedding, Laplacian Eigenmaps, Isomap, Landmark Isomap, Locally Linear Embedding, Diffusion Maps, combining Mahalanobis distance discrimination, quadratic discriminant analysis, K-nearest neighbor method to identify if watercore exist or not. Results indicate that an optimal identification model is obtained by using MSC-Landmark Isomap-KNN when principal components equal to twelve, and the identification rates for the calibration set and prediction set are 97.5% and 96.3% respectively. Hence, manifold learning algorithm and near infrared transmission spectroscopy technology can successfully realize the watercore identification of Xinjiang Red Fuji apple, which provides a theoretical basis for developing identification device in further research.

Keywords Apple watercore disease; Near infrared spectroscopy; Chemometrics; Manifold learning; Pattern recognition

(Received Jul. 5, 2019; accepted Nov. 18, 2019)