对比主成分分析的近红外光谱测量及其在水果农药残留识别中的应用

陈淑一[1]，赵全明[1]，董大明[2*]

1. 河北工业大学电子信息工程学院，天津 300401
2. 北京农业智能装备技术研究中心，北京市农林科学院，北京 100097

摘 要 近红外光谱(NIR)分析具有测试方便、不破坏样本、响应快速等优势，但是，由于在谱带分布和结构分析中存在着许多复杂因素，使得在提取特征光谱信息时存在许多困难。现阶段，虽然已经有多种光谱数据降维方式被广泛使用，但是这些传统的数据降维方式都有一个局限性，就是数据的降维仅仅针对于一个数据集，当数据集中有多个关键因素形成干扰时，数据降维和分类的结果往往不是很理想，得不到想要分析的信息。这一问题造成了在分析近红外光谱时建立的数据降维模型极差，无法正确的对样品进行预测分类。对比主成分分析(contrastive principle component analysis，cPCA)是一种基于主成分分析(PCA)的改进算法，起源于对比学习，并应用于基因组信息解析。cPCA算法的优势就是能够将一个数据集中的降维推广到两个相关联数据集之间的降维，从而能够得到数据集中的关键信息。将cPCA算法应用于近红外光谱处理中，建立了准确的近红外光谱数据降维模型。在实验验证中，使用cPCA算法对不同类型水果(苹果和梨)表面农药残留进行分析。结果表明，在对不同类型的水果进行农药残留分析时，使用PCA算法进行数据降维只能区分出不同的水果类型，而水果表面是否喷洒农药这一关键的特征信息并不能分析出来；而使用cPCA算法进行数据降维分析时，由于对背景光谱的约束作用，能够清晰的将有无喷洒农药的样本分类。这说明了，cPCA在近红外光谱数据降维中有着明显的优势，解决了近红外光谱数据降维模型中数据集受限和特征信息的提取问题，进而建立准确的近红外光谱数据降维模型。

关键词 近红外；cPCA；数据降维；模型建立

中图分类号：X56 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)03-0917-05

引 言

近红外光谱测量中高维数据集的无关因素干扰给光谱数据的分析带来了许多困难，数据降维[1]和特征提取是解决这一问题的重要手段。比较常见的数据降维方式包括：线性降维方式中的主成分分析[2-3](PCA)和线性判别分析[4](LDA)；非线性降维方式中的局部线性嵌入算法[5](LLE)和T分布随机近邻算法[6](t-SNE)等。其中，主成分分析(PCA)是使用最为广泛的数据降维方式。然而，这些常见的数据降维方式都是针对于一个数据集，当我们要研究的信息涉及到两个数据集或者一个数据集存在研究者不感兴趣的干扰信息时，传统的数据降维方法就不再准确，而对比主成分分析算法[7-8](cPCA)就有效地解决了这一问题。

对比主成分分析(cPCA)算法是Abubakar Abid等2018年提出的一种新的算法，是我们所熟知的主成分分析算法的改进，属于无监督学习。cPCA通过引入背景数据集(background dataset)有效的将我们研究的目标数据集(target dataset)中不感兴趣的干扰信息消除，从而更好的实现数据的降维和分类。cPCA算法主要应用于基因组的数据降维，并且已经在不同类型的正常小鼠和白化病小鼠的分类、不同白血病人细胞移植前后的分类中得到了成功的实验。我们将cPCA算法应用到不同类型水果表面农药残留分析[9]中，对测量的近红外光谱进行数据降维，实现了该算法在近红外光谱模型建立中的首次应用。

收稿日期：2019-03-02，**修订日期**：2019-07-16

基金项目：国家自然科学基金优秀青年科学基金项目(31622040)资助

作者简介：陈淑一，女，1994年生，河北工业大学硕士研究生 e-mail：806733129@qq.com

*通讯联系人 e-mail：damingdong@hotmail.com

1 实验部分

1.1 材料

所用的水果包括新鲜的红富士苹果 30 个，皇冠梨 30 个，总计 60 个两种不同类型的水果，均购买于北京市果香四溢水果超市。首先将水果清洗干净，沿着水果的赤道部分均匀采样，间隔角度约为 70°左右，一个水果样本共计 5 个采样点。配置好 1∶1 500 毒死蜱农药，取 10 个富士苹果，均匀涂抹到苹果表面后晾干。再取另外 10 个富士苹果表面均匀涂抹上水，晾干后进行采样，梨的采样方法类似。剩余的 10 个富士苹果和 10 个皇冠梨洗净后不做任何处理进行采样。共采得涂抹毒死蜱的苹果和梨的样本 100 个；涂抹水的苹果和梨的样本 100 个，不做处理的苹果和梨的样本 100 个，共计 300 个样本。

1.2 仪器

实验用的 DLP NIRscan Nano(v2.1.0)近红外光谱仪，光谱的测量范围为 950～1 700 nm，每条光谱共计 228 个数据点，每个样本测量前都使用标准白板为背景进行背景光谱采集。Unscrambler 9.7(CASMO 公司)光谱分析软件，主要用于光谱数据的预处理和分析。

1.3 光谱采集

将上述的 300 个样本使用 DLP 仪器进行近红外光谱扫描，为了区分的更加清楚，分别选取四种不同类型样本中的一条原始光谱图，如图 1 所示。

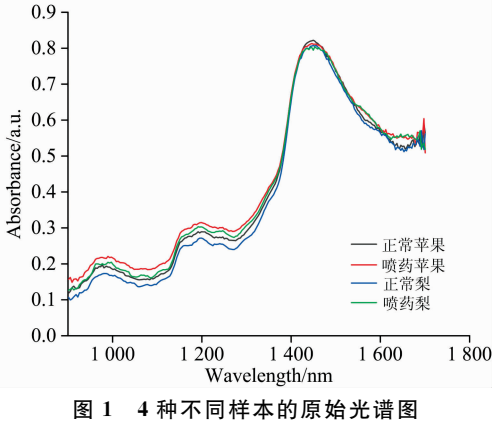

图 1 4 种不同样本的原始光谱图

Fig. 1 Original spectra of four different samples

由图 1 可以看出，四种不同样本的原始光谱图略有差异，但是大致特征相似。分别在 950～1 250 nm 处和1 400～1 600 nm 处光谱吸光度较强，并且有明显变化，说明这一部分包含的信息量较多。同时，原始光谱图中存在噪声，需要对光谱数据进行预处理。

2 结果与讨论

2.1 算法描述

为了保证该算法的有效性，更加有效地利用光谱信息，首先要对 300 个样本光谱数据进行预处理。主要使用的光谱预处理手段包括均值中心化(mean centering)、基线校正(baseline)、一阶求导(S-G)和标准正态变换(SNV)，目的是消除光谱数据中的基线漂移和无关噪声信息，如样品背景和杂散光等。经过预处理后的光谱重复性更好。

将处理过后的光谱数据进行 cPCA 算法分析，具体流程图如图 2 所示。

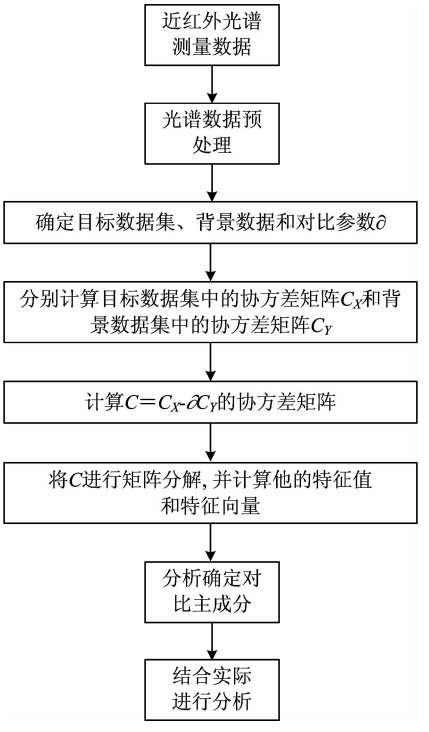

图 2 cPCA 算法流程分析图

Fig. 2 Algorithm flow chart of cPCA

具体算法数学过程描述如下：

(1)使用 DLP 近红外光谱仪进行光谱数据采集，并对光谱数据进行预处理。

(2)确定 d 维目标数据集$\{x_i \in R^d\}$和 d 维背景数据集$\{y_i \in R^d\}$，分别计算两者的协方差矩阵 $\boldsymbol{C}_X$，$\boldsymbol{C}_Y$；

(3)将目标数据集和背景数据集的方差分别用单位向量表示为

Target dataset variance

$$\lambda_X(v) = v^{\mathrm{T}}\boldsymbol{C}_X v$$

Background dataset variance

$$\lambda_Y(v) = v^{\mathrm{T}}\boldsymbol{C}_Y v$$

(4)设对比强度为∂，表示背景数据集在目标数据集中的对比消除强度，计算后的单位化向量 C 表示为

$$C = \mathrm{v}^T(\boldsymbol{C}_X - \partial\boldsymbol{C}_Y)v$$

(5)计算协方差矩阵 $\boldsymbol{C}$ 并进行矩阵分解，求出相应的特征值和特征向量

$$v^* = \mathrm{argmax}_{v\in R^d_{\mathrm{unit}}} v^{\mathrm{T}}(\boldsymbol{C}_X - \partial\boldsymbol{C}_Y)v$$

(6)将得到的特征值和特征向量 v^* 进行由高到低排序，保留贡献率较高的对比主成分，分别命名为 cPC1，cPC2，…，cPCn；

由基本原理可以看出，使用 cPCA 算法的关键是背景数

据集的选择和对比参数∂的选择。

2.2　不同类型水果和有无喷农药水果的 PCA 分析

实验的主要目的是为了在不同水果类型中区分出喷洒农药的水果和未喷洒农药的水果。将 200 个喷洒农药的苹果、未喷洒农药的苹果、喷洒农药的梨和未喷洒农药的梨进行混合后使用 PCA 方法进行降维分类，结果发现 PCA 只能大致区分出不同水果类型(苹果和梨)这一我们不感兴趣的无关信息，如图 3 所示。其中，黑色和红色的点集分别代表没有喷洒农药的苹果和喷洒农药的苹果(fruit 0 和 fruit 1)；蓝色和绿色的点集分别代表没有喷洒农药的梨和喷洒农药的梨(fruit 2 和 fruit 3)，并且散点图的区分度也不是很好，总体效果较差。

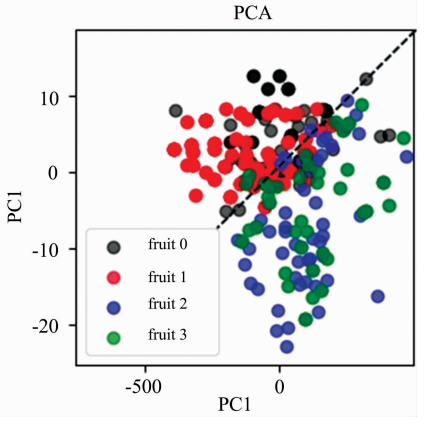

图 3　PCA 得分结果图

Fig. 3　Score plot of PCA model

为了更加清晰的分析出影响 PCA 得分结果的影响因素，我们对 PCA 模型的主成分进行分析，如图 4 所示。PCA 中前两个主成分得分分别为 79%和 6%，其方差的累计贡献率达到 85%，表明 PCA 分析结果对原始光谱有比较好的代表性。

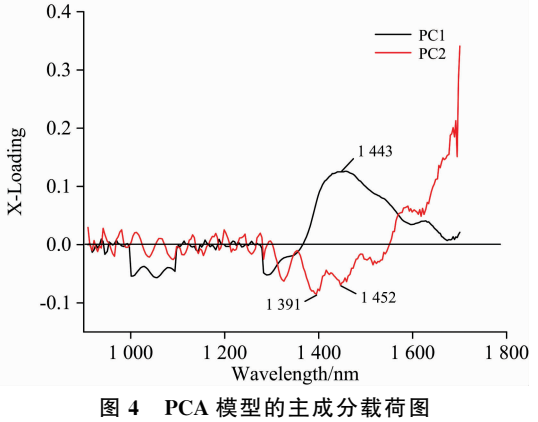

图 4　PCA 模型的主成分载荷图

Fig. 4　Principal component loading of PCA model

PC1 和 PC2 仅在 1 350～1 500 nm 波段处有明显的特征峰，经分析可知，该波段主要是区分不同水果类型的特征波段。而在其他波段处没有明显的特征峰且载荷值较低，其特征载荷向量代表的主要是其他干扰信息和噪声。所以 PCA 模型仅仅能够区分不同的水果类型，而不能够给出水果表面有无喷农药这一特征信息。

2.3　不同类型水果和有无喷农药水果的 cPCA 分析

上述结果的原因是由于不同水果类型背景信息所占有的方差比例较大，我们通过优化模型算法来解决这一问题。根据 cPCA 的核心思想，引入背景数据集消除目标数据集中占有较大方差的干扰信息。将剩余的 100 个健康苹果和健康梨作为背景约束，设置最佳对比参数∂(最佳对比参数为∂＝8.89)，运行 cPCA，并和 PCA 结果进行比较以说明该算法的优越性。

运行结果如图 5 所示，cPCA 算法能够清晰的将喷洒农药的水果和未喷洒农药的水果区分开。其中，黑色点集和蓝色点集聚成一类，分别代表没有喷洒农药的苹果和梨(fruit 0 和 fruit 2)；而红色点集和绿色点集聚成一类，分别代表喷洒农药的苹果和梨(fruit 1 和 fruit 3)。红色点集和绿色点集样本间距离略小于黑色点集和蓝色点集的样本间距离，这是由于喷洒农药后的苹果和梨的表面光谱特征比没有喷洒农药的苹果和梨的表面光谱特征更为相似造成的。交界处的个别点存在偏差，可能是由于在实验过程中存在测量误差造成的。

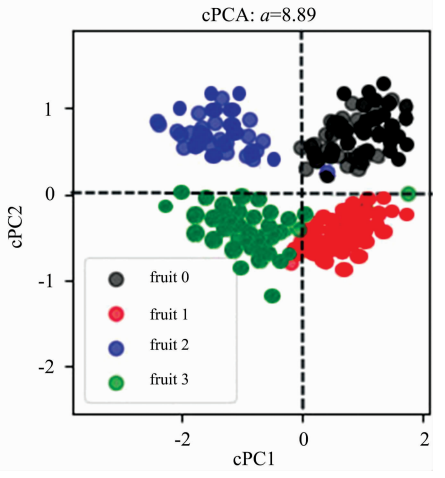

图 5　cPCA 得分结果图

Fig. 5　Score plot of cPCA model

图 6 显示的是 cPCA 模型的对比主成分载荷图，cPCA 中前两个对比主成分 cPC1 和 cPC2 的得分分别为 85%和 5%，其方差的累计贡献率达到 90%，说明对比主成分的分析结果能够很好的代表原始光谱信息。从图中可以看出，cPC1 和 cPC2 的有效特征峰集中在两个波段，分别为 1 000～1 150 和 1 400～1 550 nm 处。这两个波段主要反映的是不同水果类型和水果表面有无喷农药的差异。并且，相比于 PCA 来说，在其他波段处的干扰信息较少。所以，cPCA 模型能够对样本进行正确的分类，优于 PCA 模型结果。

2.3　cPCA 算法不同∂值下的分类效果

cPCA 算法中最重要的参数就是对比强度∂，它表示背景数据集于目标数据集中的消除强度，∂值越大代表背景数据集在目标数据集中的消除强度越强。选取了不同∂值下 cPCA 算法运行结果，如图 7(a)和(b)所示。

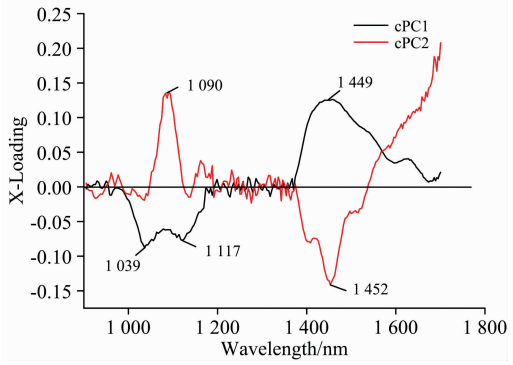

图 6 cPCA 模型的主成分载荷图

Fig. 6 Principal component loading of cPCA model

图 7(a)显示的是当对比强度∂的值过小时(∂=2.03)cPCA 得分结果图。从该得分图我们可以看出，样本之间的区分度更加明显，并且喷洒农药的水果(fruit 0 和 fruit 2)和未喷洒农药的水果(fruit 1 和 fruit 3)也有聚成一类的趋势。这说明了，cPCA 引入的背景数据集起到了作用，消除了目标数据集中的一部分方差，但由于对比强度∂过小，约束条件不足，并没有达到理想的分类效果。

图 7(b)显示的是当对比强度∂的值过大时(∂=228.5)cPCA 得分结果图。可以明显的看出，得分散点图中丢失了大部分有用信息，并且在第一对比主成分中的得分为零。当对比强度∂过大时，背景数据集于目标数据集中的约束条件过强，在目标数据集中的消除方差过大，导致一部分有用信息的丢失。所以，对比强度∂不是越大越好，而是要根据具体的实验结果来选择。

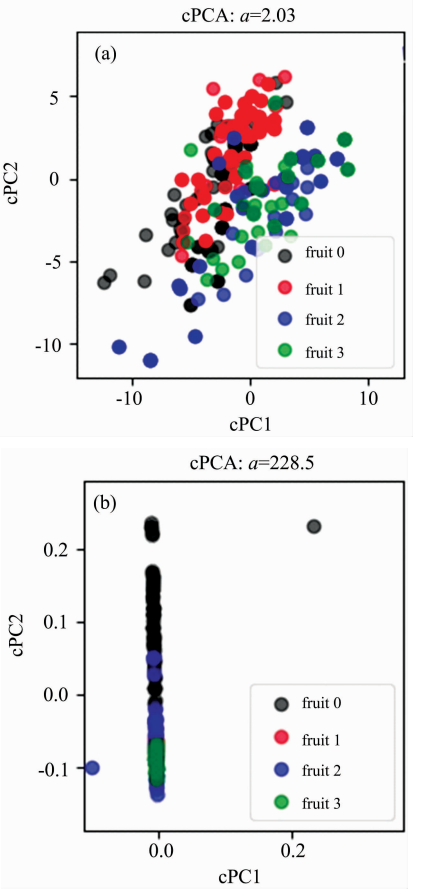

图 7 (a) ∂=2.03 时 cPCA 得分结果图；(b) ∂=228.5 时 cPCA 得分结果图

Fig. 7 (a) cPCA score when ∂=2.03; (b) cPCA score when ∂=228.5

3 结 论

对比主成分分析算法(cPCA)是一种新兴的数据降维方式，通过引入背景数据集作为约束，消除背景干扰信息，从而得到数据集中的关键信息。在对两种不同类型水果(苹果和梨)进行农药残留分析时，使用 PCA 算法只能区分出不同的水果类型这一背景信息，而使用 cPCA 算法能够将不同类型水果表面是否喷洒农药的信息特征正确的展示出来，说明 cPCA 算法能够有效地建立数据降维模型，在近红外光谱分析中有着广阔的应用前景。

References

[1] Niu D, Dy J, Jordan M. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, 15: 552.
[2] Jolliffe I T, Cadima J. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, 374 (2065): 20150202. doi: org/10.1098/rsta.2015.0202.
[3] Wetzel S J. Physical Review E, 2017, 96(2): 022140.
[4] Wu L, Shen C, van den Hengel A. Pattern Recognition, 2017, 65(0031-3203): 238.
[5] Wu Y C, Hwang H T, Hsu C C, et al. INTERSPEECH, 2016, 567: 1652.
[6] Gisbrecht A, Schulz A, Hammer B. Neurocomputing, 2015, 147(0925-2312): 71.
[7] Abid A, Zhang M J, Bagaria V K, et al. Nature Communications, 2018, 9(1): 2134.
[8] Severson K, Ghosh S, Ng K. arXiv preprint arXiv: 1811.06094, 2018.
[9] SUN Jun, ZHOU Xin, MAO Han-ping, et al(孙 俊，周 鑫，毛罕平，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 32(19): 302.

Application of Near Infrared Spectroscopy Combined with Comparative Principal Component Analysis for Pesticide Residue Detection in Fruit

CHEN Shu-yi[1], ZHAO Quan-ming[1], DONG Da-ming[2*]

1. Hebei University of Technology, Tianjin 300401, China
2. Beijing Research Center for Intelligent Equipment for Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

Abstract Near-infrared spectroscopy (NIR) analysis is considered as a promising chemical analysis technique because its advantages of convenient-testing, no damaging and fast response. However, due to the many unknown factors in the band distribution and structural analysis of the near-infrared spectrum, there are many difficulties in extracting the characteristic spectral information. Nowadays, although a variety of spectral data dimensionality reduction methods have been widely used, the traditional data dimensionality reduction methods have a limitation that the dimensionality reduction is restricted in one dataset. The results of data dimensionality reduction are often not ideal when there are many factors in dataset . This problem makes the data establish dimensionality reduction model extremely hard in near-infrared spectrum. Comparative Principal Component Analysis (cPCA) is an improved algorithm based on principal component analysis (PCA), which originated from Comparative Learning and applied to genomic information analysis. The advantage of the cPCA algorithm is that it can realize the dimensionality reduction between two related data sets. In this paper, the cPCA algorithm is applied to near-infrared spectroscopy for the first time and establish an accurate spectral dimensionality reduction model. In the experimental, we used the cPCA algorithm to analyze the surface of different types of fruits (apples and pears) with pesticide residues and without pesticide residues . The result showed that the PCA algorithm just distinguishes different fruit types, while the cPCA algorithm classifies the fruits with or without pesticides due to the constraint of the background dataset. This showed that cPCA outperforms in data dimensionality reduction of near-infrared spectra. It solves the problem of dataset limitation and feature information extraction in the near-infrared spectral data dimensionality, and cPCA could establish an accurate spectral data dimensionality reduction model.

Keywords Near-infrared spectroscopy; cPCA; Data dimensionality reduction; Model establishment

(Received Mar. 2, 2019; accepted Jul. 16, 2019)

* Corresponding author