第 58 卷 第 1 期/2021 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展

基于PLS-DA拉曼光谱特征提取的中性笔油墨MLP模式识别

王晓宾[1,2]，马枭[1]，杨蕾[1,2]，李春宇[1,2*]

[1]中国人民公安大学刑事科学技术学院，北京 100038；

[2]中国人民公安大学刑事科学技术实验教学中心，北京 100038

摘要 中性笔油墨是司法鉴定中同一认定的重要物证。为提高油墨检验的准确性，本文利用拉曼光谱法对油墨样本进行无损检测。首先对预处理后的光谱数据进行降维处理，构建偏最小二乘判别分析模型；然后采用受试者工作特征曲线线下面积对预测效果进行验证，提取出36个变量投影重要性最高的特征变量；接着将特征变量作为数据输入到隐藏层神经元数目为13的多层感知器中，最终的训练正确率为87%且无过拟合现象。将变量投影重要性的特征提取与有监督的多层感知器训练相结合，可以有效压缩数据，缩短分析时间。感知器层间的连接权重可通过自主学习进行调节，提高了中性笔油墨分类结果的可信度与正确率。

关键词 光谱学；拉曼光谱法；偏最小二乘判别分析；变量投影重要性；多层感知器

中图分类号 O657.37 **文献标志码** A **doi**：10.3788/LOP202158.0130002

Multi-Layer Perceptron Pattern Recognition of Handwriting Ink Based on PLS-DA Raman Spectral Feature Extraction

Wang Xiaobin[1,2], Ma Xiao[1], Yang Lei[1,2], Li Chunyu[1,2*]

[1]*School of Forensic Science, People's Public Security University of China, Beijing* 100038, *China*;

[2]*Forensic Science Experimental Teaching Center, People's Public Security University of China, Beijing* 100038, *China*

Abstract Handwriting ink is an important physical evidence of the identification in judicial appraisal. In order to improve the accuracy of ink inspection, we employed Raman spectroscopy for the non-destructive inspection of ink samples. First, the pre-processed spectral data were dimensionally reduced to construct a model of partial least squares discrimination analysis. Then, after the prediction effect was verified by the area under the receiver operating characteristic curve, 36 feature variables with the highest variable importance for the projection were extracted. Furthermore, the feature variables were input as data into a multi-layer perceptron with 13 neurons in the hidden layer, and the final training accuracy rate was 87%, without overfitting. We also found that combining the feature extraction of variable importance for the projection with supervised multi-layer perceptron training could effectively compress the data and shorten the analysis time. Besides, the connection weight between perceptron layers could be adjusted through autonomous learning, which improved the credibility and accuracy of handwriting ink classification results.

收稿日期：2020-04-29；**修回日期**：2020-05-28；**录用日期**：2020-06-15

基金项目：国家重点研发计划(2019YFF0303405)、公安部技术研究计划(2019JSYJC21)、中国人民公安大学基本科研业务费项目(2019JKF109)

*E-mail：lichunyu@ppsuc.edu.cn

Key words spectroscopy; Raman spectroscopy; partial least squares discrimination analysis; variable importance for the projection; multi-layer perceptron

OCIS codes 300.6450; 100.4996

1 引 言

在文书司法鉴定领域，书写笔油墨的种类认定及溯源一直以来都是法庭科学工作者研究的热点之一。通过对油墨进行识别可以准确地进行伪造文件、变造文件的鉴定，从而识别文件的真伪。随着科技的发展，越来越多的现代分析仪器被用于书写笔油墨的分析检验中，检验手段不断丰富，色谱法、质谱法等半破坏性方法已在油墨检验领域得到了广泛应用。如：Jones 等[1]利用实时质谱法对纸张上书写笔迹的油墨种类进行了检验，Djozan 等[2]通过薄层色谱法对从文档中提取出的墨迹进行了检验。在法庭科学中，光谱法等非破坏性检验方法凭借其无损检验的特性，可对油墨物证的损伤程度降至最低。因此，光谱法可为油墨的种类鉴定和溯源提供新的思路和方法。Da Silva 等[3]利用可见光谱法对 25 个品牌的蓝色油墨进行了无损检验，然后利用光谱数据建立了偏最小二乘判别分析模型，最终实现了快速识别文件中油墨品牌的目的。Teixeira 等[4]借助拉曼光谱成像及均场独立成分分析对使用 7 种不同的圆珠笔伪造的笔迹进行检验，准确地区分出了不同油墨成分的添改笔迹。

相对于检验方法的不断发展，数据处理方面的发展较为缓慢，特别是在进行溯源鉴别时，由于书写笔油墨成分基本相同，所以识别区分的难度极大。近年来，模式识别（pattern recognition）逐渐成为一种解决法庭科学物证溯源实际问题的主要手段，将模式识别技术用于法庭科学油墨的识别，可以根据化学实验测得的数据揭示物质的潜在性质，从而获得数据中的有用信息。但是目前关于书写笔油墨光谱的研究报道主要是通过构建不同的算法模型对光谱整体进行识别[5]，在光谱数据特征提取以及计算机深度学习领域的研究还有待进一步深入。因此，本文选取使用得最为广泛的签字笔油墨作为研究对象，采集其拉曼光谱数据，利用计算机多层感知器对特征变量进行学习，构建了一种特征提取结合分类预测的中性笔油墨拉曼光谱分析方法。

2 实 验

2.1 拉曼光谱的测定

实验所用仪器为 InVia 激光显微共聚焦拉曼光谱仪，选择 532 nm 激光器，设置激光功率为 1%（0.2 mW），设定扫描范围为 100～2000 cm^{-1}，扫描时间为 20 s，积分次数为 2。

为贴近实际情况，提高方法的实用性，本次实验收集了市面上不同品牌的 100 支中性笔样本（样品表略），利用中性笔在同一规格的 A4 纸上划出一条横线，截取墨迹均匀、笔画厚重部分作为实验样本。同时，为了避免污染样本，在样本制作过程中要注意避免手与纸张直接接触，可采用镊子夹取等方式进行操作。将待测样本放置于采集台上，待显微镜调焦清晰后选择采集点进行测量，重复该操作直至完成对所有样本的检测。虽然纸张本底等条件不会影响光谱结果，但为了保证测试结果的准确性，本实验对每个样本均进行重复测试，每个样本测试三次，三次测量结果一致才可将其作为该样本的谱图。

2.2 光谱的预处理

根据预处理的目的，光谱预处理方法大致可以分为基线校正、归一化校正、散射校正和平滑校正这四类[6]。基线校正是为了消除测量时背景产生的干扰，例如对原谱图进行一阶求导、二阶求导[7]等。归一化校正的目的是消除测量过程中产生的数据之间的数量级差异，其中，Z-score 法是一种高效的归一化校正方法[8]，它可将光谱数据集中到一定的区间内，同时还可保持原始数据自身的差异性。在光谱采集过程中，因纸张表面颗粒尺寸不均匀，测得的拉曼光谱中可能伴随有散射噪声，这些散射噪声甚至会覆盖原始光谱[9]。包括多元散射校正（MSC）在内的散射校正可以通过构造理想光谱来消除这种影响[10]。平滑校正作为一种最常见的光谱预处理方法，其目的是消除或限制样品制备、测量以及测量参数的设置过程中带来的不可避免的噪声。Savitzky-Golay（S-G）平滑法作为一种较为成熟且有效的平滑方法最早由 Savitzky[11]等提出，已在光谱平滑预处理领域得到了广泛应用[12]。本研究团队对测量所得的原始拉曼光谱图进行观察，未发现谱图存在基线

漂移等误差，但存在数量级尺度差异、噪声较大和荧光强度不均等问题[13]，因此对其进行S-G平滑、MSC和Z-score归一化预处理，处理后的结果如图1所示。

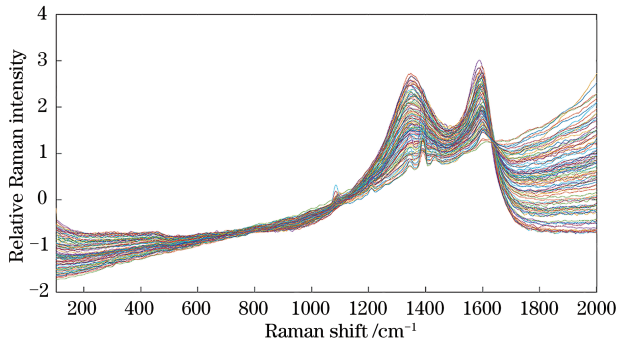

图1　预处理后样本的拉曼光谱图

Fig. 1　Raman spectra of samples after pre-processing

3　结果与分析

3.1　PLS-DA模型的建立

由于拉曼光谱所含的数据信息较大，每一拉曼位移对应着特定的光强度，数据总量高达数千个，因此在分析时进行降维处理是十分重要的[14]。主成分分析(PCA)作为一种常见的降维分析方法，在光谱数据分析领域得到了广泛应用，其通过计算出少量主成分变量来解释较多的原始变量，达到了降低数据维度的目的[15]。但由于PCA仅简单地利用变量数据进行分析，属于无监督的分类方法，在分析结果的准确性方面仍然有待加强。与无监督算法相对应的是有监督算法。偏最小二乘判别分析(PLS-DA)作为一种有监督的降维、判别分析方法，近年来被广泛应用于理化检验领域[16-17]。PLS-DA将变量数据与分类信息划分为两组数据集，将降维分析与组类别相结合，更能凸显组间差异，从而将每一类样本区分开来[18]。

首先，按照光谱特征峰与背景荧光强度将100个样本分为5个类别，将类别信息与已经过预处理的拉曼光谱数据一同导入模型中。通过构建彼此完全独立的正交新变量t_1、t_2来解释各样本之间的差异。PLS-DA得分图如图2所示。

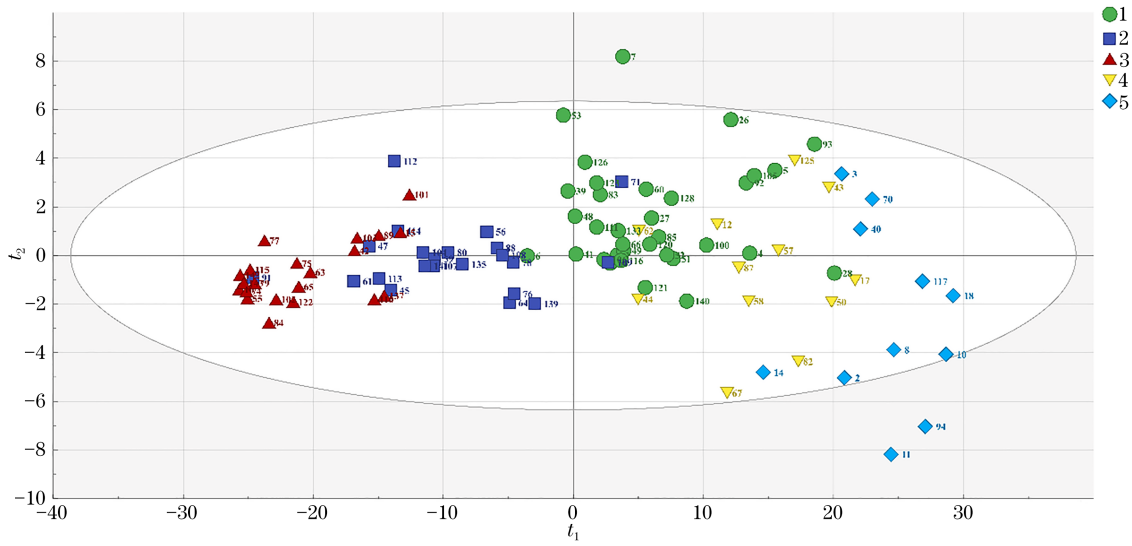

图2　PLS-DA得分图

Fig. 2　PLS-DA score graph

由图2可知，采用PLS-DA对光谱高维数据进行降维，投影产生对光谱数据解释能力累计为97.59%的两个正交变量t_1、t_2，其中解释能力最强的t_1变量能够解释光谱数据中94.7%的变化。各类样本在投影变量二维平面上分散开来，总体呈类间分离、类内收缩的特点。但值得注意的是，PLS-DA作为一种有监督的判别分析方法，在建立模型时可能会出现拟合效果不佳的情况。因此，对已构建的PLS-DA模型进行诊断是十分有必要的。因此，绘制受试者工作特征曲线(ROC)，如图3所示，以检验模型的预测效果[19]。

ROC作为一种反映预测准确率的曲线，以伪阳性率(R_{FP})为x轴，真阳性率(R_{TP})为y轴[20]。ROC下方的面积A_{UC}是一种综合评价模型预测准确值的参数，A_{UC}值越高代表该模型对某一类样本的分类预测效果越好[21]。由图3可知，5类样本的A_{UC}值分

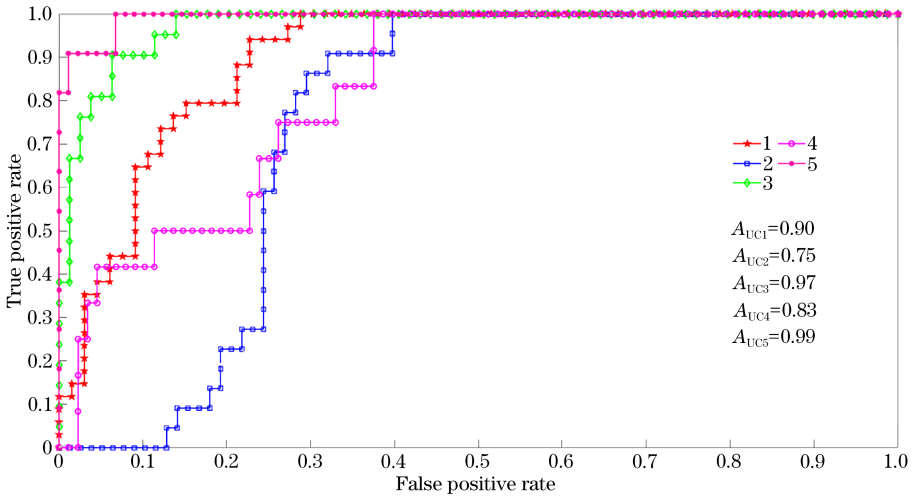

图 3　5类样本的ROC曲线及曲线下方的面积A_{UC}

Fig. 3　ROC curve and area under the curve A_{UC} of 5 types of samples

别为 0.90、0.75、0.97、0.83 和 0.99(保留两位小数),所有样本的A_{UC}均大于 0.75 且曲线向坐标轴左上方靠拢,表明模型的预测能力较好。

3.2　V_{IP}值的提取

前文已采用构建的 PLS-DA 模型对样本进行了初步分类,但根据得分图中样本的散点分布以及A_{UC}值的大小可知,预测分类效果有待进一步加强。因此,基于已构建的 PLS-DA 模型,通过提取变量投影重要性(V_{IP})对复杂变量进行筛选,以达到提取特征变量的目的。

V_{IP}值是一种基于 PLS-DA 的特征变量筛选指标,表示每一独立自变量对因变量的解释能力[22]:若所有自变量对因变量的解释能力都相同,则所有自变量的 V_{IP} 值均为 1。因此,若某一变量的 V_{IP} 值越大,就表明该变量对因变量的解释能力越强,反之,则解释能力越弱。通过筛选出 V_{IP} 值较大的一系列自变量,既能很好地解释样本类别这一因变量,又能降低数据复杂程度和提取特征变量的目的,如图 4 所示。

由图 4 可知,在 100～2000 cm^{-1}范围内,1076～1095 cm^{-1}、1382～1398 cm^{-1}、1972～2000 cm^{-1}拉曼位移波段内的 V_{IP} 值均大于 1.5,会对因变量的预测产生较大影响。结合谱图进行分析:在 1076～1095 cm^{-1}和 1382～1398 cm^{-1}波段内,部分类别样本出现了明显的特征峰,这对类别的判断具有重要帮助;而在 1972～2000 cm^{-1}波段内,由于不同类别样本所含荧光成分不同,在拉曼光谱图中出现了不同强度的背景荧光,该特征也可用于类别的判断。

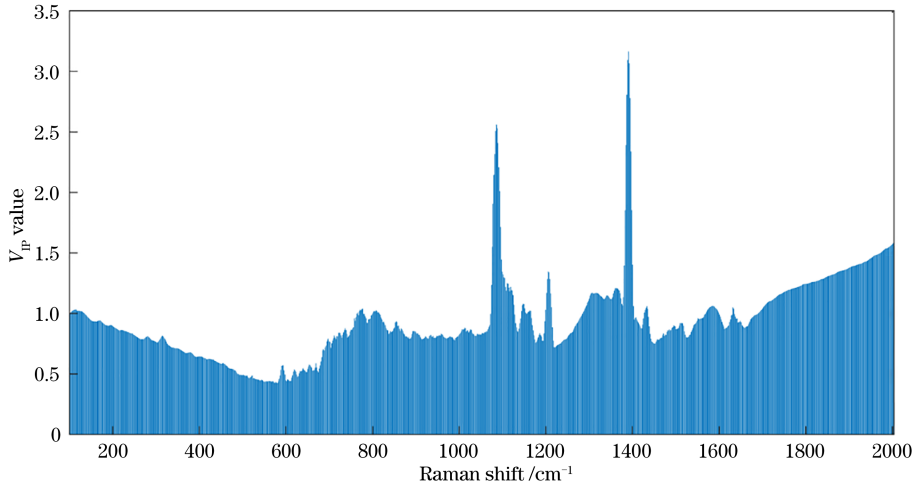

图 4　自变量的 V_{IP} 值

Fig. 4　V_{IP} value of independent variables

最终提取了上述3个波段内拉曼位移自变量特征共36个。

3.3　多层感知器模式识别

多层感知器(MLP)被又称为人工神经网络(ANN),主要由输入层、隐藏层、输出层三层结构组成。输入层与输出层只有一层结构,而隐藏层有一层或多层结构,如图5所示。

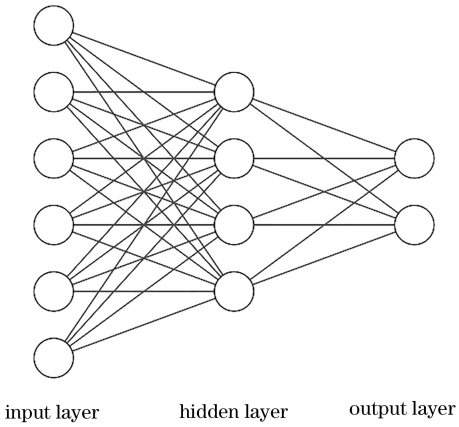

图5　多层感知器的结构

Fig. 5　Structure of multi-layer perceptron

通过输入层将样本数据传递至与其相连接的隐藏层,再传递至输出层就可得到识别结果[23]。在这一过程中,层与层之间的神经元全连接,但层内的神经元之间无连接[24]。因此,数据在两个神经元之间的连接上传递时,便会进行权重计算,直至最终输出结果。本文所采用的后向传播(BP)神经网络算法最早由Rumelhart等[25]提出。作为一种具有自主学习能力的反向传播算法,BP神经网络算法通过计算网络输出值与实际值的拟合程度,不断地从后向前对网络的连接权重进行优化,直至网络趋于稳定[26]。

在一个完整的MLP结构中,神经元的数量可以影响整个模型的识别效果,而输入层和输出层神经元数目是固定的(根据数据量和输出类别而定),所以隐藏层神经元数目的选择便成为了对模型结果影响较大的重要因素。若隐藏层神经元数量较少,就会导致分类效果较差,而神经元数量过多,则会导致分类效率较低。因此,隐藏层神经元数量 m 可以通过(1)式进行计算[27]。

$$m=\sqrt{a\times b}\ , \tag{1}$$

式中:a 为输入层的神经元数目;b 为输出层的神经元数目。在本实验中,输入层数据为通过PLS-DA模型 V_{IP} 值提取出的36个拉曼特征光谱数据,输出层数据为5类样本的分类结果。综合考虑多层感知器的分类效果与效率,设置隐藏层神经元数目为13,分类结果如表1所示。

表1　多层感知器的分类结果

Table 1　Classification results of multi-layer perceptron

Actual class	Predicted class				
	1	2	3	4	5
1	32	2	0	4	0
2	1	17	1	0	0
3	0	3	20	0	0
4	1	0	0	7	0
5	0	0	0	1	11

由表1可知,MLP最终的分类正确率为87%,分类效果良好。在训练过程中,已将所有样本数据不重叠地分为训练集、验证集和测试集。训练集不断地将训练结果与真实值拟合,从后向前调整各连接上的权重等普通参数;验证集通过验证神经元数目等超参数检验网络的泛化能力;测试集利用已训练的网络对测试数据进行拟合,检验网络是否过拟合。为了直观地判断训练集、验证集和测试集在神经网络上的拟合情况,引入可衡量多层感知器中期望值与实际值接近程度的交叉熵(cross-entropy),并将其作为损失函数来绘制损失函数的变化,如图6所示。

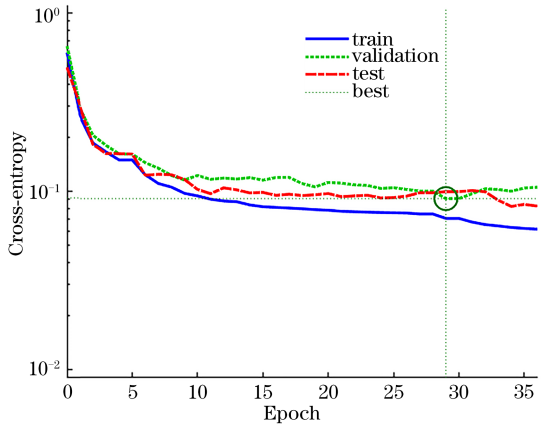

图6　交叉熵的时期变化

Fig. 6　Changes of cross-entropy in epoch

交叉熵作为损失函数可用于判断实际输出值与期望输出值的拟合程度。交叉熵越小,则期望值与实际值的概率分布越接近,拟合效果越好。由图6可知,训练集、验证集和测试集的交叉熵随着时期(epoch)的增加而逐渐变小并最终趋于平缓。训练集、验证集和测试集的损失函数曲线比较接近,无论是在权重设置还是泛化能力检验方面都表现良

好，对于未参与训练的测试集数据都能很好地拟合，且未出现过拟合的情况。

4 结 论

本文构建了一种特征提取结合分类预测的笔迹油墨拉曼光谱分析方法，该方法将基于PLS-DA模型的V_{IP}值提取的特征变量作为多层感知器输入层数据进行自主学习，网络的最终分类正确率为87%，达到了较好的分类效果。本文在训练的同时划分了验证集与测试集，观察损失函数可知模型参数的拟合效果较好，对于未知类别的笔迹油墨样本也可进行分类预测。本文的研究思路与方法可对司法鉴定中关于笔迹油墨同一认定的问题起到一定的启示作用，同时本文所构建的笔迹油墨检验方法将光谱法与数据分析方法相结合，使得学科之间相互交叉并且相互促进。但值得注意的是，本文仅使用了100个笔迹油墨样本进行分析检验，若要实现对市面上任意样本的模式识别，则需要不断地补充全新数据并优化模型参数。此外，多层感知器作为一种普遍使用的模式识别方法在本文中得到了很好的应用，接下来可以进一步研究其他适用于模式识别的分类算法模型。

参 考 文 献

[1] Jones R W, McClelland J F. Analysis of writing inks on paper using direct analysis in real time mass spectrometry [J]. Forensic Science International, 2013, 231(1/2/3): 73-81.

[2] Djozan D, Baheri T, Karimian G, et al. Forensic discrimination of blue ballpoint pen inks based on thin layer chromatography and image analysis [J]. Forensic Science International, 2008, 179(2/3): 199-205.

[3] da Silva V A G, Talhavini M, Peixoto I C F, et al. Non-destructive identification of different types and brands of blue pen inks in cursive handwriting by visible spectroscopy and PLS-DA for forensic analysis [J]. Microchemical Journal, 2014, 116: 235-243.

[4] Teixeira C A, Poppi R J. Discriminating blue ballpoint pens inks in questioned documents by Raman imaging and mean-field approach independent component analysis (MF-ICA) [J]. Microchemical Journal, 2019, 144: 411-418.

[5] He X L, Chen L B, Wang J F, et al. Raman spectroscopy analysis of plastic steel window based on K nearest neighbors algorithm [J]. Laser & Optoelectronics Progress, 2018, 55(5): 053001.
何欣龙, 陈利波, 王继芬, 等. 基于K近邻算法的塑钢窗拉曼光谱分析[J]. 激光与光电子学进展, 2018, 55(5): 053001.

[6] Diwu P Y, Bian X H, Wang Z F, et al. Study on the selection of spectral preprocessing methods [J]. Spectroscopy and Spectral Analysis, 2019, 39(9): 2800-2806.
第五鹏瑶, 卞希慧, 王姿方, 等. 光谱预处理方法选择研究[J]. 光谱学与光谱分析, 2019, 39(9): 2800-2806.

[7] He X L, Wang J F, Wang F, et al. Rapid identification of rubber particles based on second derivative infrared spectra [J]. China Measurement & Test, 2019, 45(9): 60-64, 83.
何欣龙, 王继芬, 王飞, 等. 二阶导数红外光谱快速鉴别轮胎橡胶颗粒[J]. 中国测试, 2019, 45(9): 60-64, 83.

[8] Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems [J]. Pattern Recognition, 2005, 38(12): 2270-2285.

[9] Chen H Z, Song Q Q, Tang G Q, et al. The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models [J]. ISRN Spectroscopy, 2013, 2013: 642190.

[10] Romero-Torres S, Pérez-Ramos J D, Morris K R, et al. Raman spectroscopic measurement of tablet-to-tablet coating variability [J]. Journal of Pharmaceutical and Biomedical Analysis, 2005, 38(2): 270-274.

[11] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures [J]. Analytical Chemistry, 1964, 36(8): 1627-1639.

[12] Xie J, Pan T, Chen J M, et al. Joint optimization of Savitzky-Golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose [J]. Chinese Journal of Analytical Chemistry, 2010, 38(3): 342-346.
谢军, 潘涛, 陈洁梅, 等. 血糖近红外光谱分析的Savitzky-Golay平滑模式与偏最小二乘法因子数的联合优选[J]. 分析化学, 2010, 38(3): 342-346.

[13] Zhu L L, Feng A M, Jin S Z, et al. Fluorescence suppression methods in Raman spectroscopy detection and their application analysis [J]. Laser & Optoelectronics Progress, 2018, 55(9): 090005.
朱磊磊, 冯爱明, 金尚忠, 等. 拉曼光谱检测中荧光抑制方法及其应用分析[J]. 激光与光电子学进展,

2018, 55(9): 090005.

[14] Tian G Y, Yuan H F, Liu H Y, et al. Application of wavelet transform to compressing near infrared spectra data[J]. Journal of Instrumental Analysis, 2005, 24(1): 17-20, 24.
田高友, 袁洪福, 刘慧颖, 等. 小波变换用于近红外光谱数据压缩[J]. 分析测试学报, 2005, 24(1): 17-20, 24.

[15] Shi R J, Xia F Z, Zeng W D, et al. Raman spectroscopic classification of foodborne pathogenic bacteria based on PCA-Stacking model[J]. Laser & Optoelectronics Progress, 2019, 56(4): 043003.
史如晋, 夏钒曾, 曾万聃, 等. 基于PCA-Stacking模型的食源性致病菌拉曼光谱识别[J]. 激光与光电子学进展, 2019, 56(4): 043003.

[16] Almeida M R, Fidelis C H V, Barata L E S, et al. Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation[J]. Talanta, 2013, 117: 305-311.

[17] de Almeida M R, Correa D N, Rocha W F C, et al. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation [J]. Microchemical Journal, 2013, 109: 170-177.

[18] Aa J Y. Analysis of metabolomic data: principal component analysis[J]. Chinese Journal of Clinical Pharmacology and Therapeutics, 2010, 15(5): 481-489.
阿基业. 代谢组学数据处理方法: 主成分分析[J]. 中国临床药理学与治疗学, 2010, 15(5): 481-489.

[19] Zhu L, Chen P J. Determination of best cut off value of activity count in diagnosis exercise intensity of adolescents by receiver operating characteristic (ROC) curve analysis[J]. China Sport Science, 2012, 32(11): 70-75.
朱琳, 陈佩杰. 应用ROC曲线确定活动计数在青春期少年运动强度诊断中的最佳临界值[J]. 体育科学, 2012, 32(11): 70-75.

[20] Wang X B, Ma X, Wang X C. Infrared spectral pattern recognition of watercolor pen ink based on artificial neural network[J]. Laser & Optoelectronics Progress, 2020, 57(15): 153005.
王晓宾, 马枭, 王新承. 基于人工神经网络的水彩笔油墨红外光谱模式识别[J]. 激光与光电子学进展, 2020, 57(15): 153005.

[21] Ke C F, Wu X Y, Li K. A comparative analysis of four PLS-DA diagnostic statistics in the application of metabolomics[J]. Chinese Journal of Health Statistics, 2014, 31(3): 403-406.
柯朝甫, 武晓岩, 李康. PLS-DA模型四种诊断统计量在代谢组学应用中的比较[J]. 中国卫生统计, 2014, 31(3): 403-406.

[22] Zhang Z, Feng G S. Application of variable importance for projection in the variables selection [J]. Modern Preventive Medicine, 2012, 39(22): 5813-5815.
张政, 冯国双. 变量投影重要性分析在自变量筛选中的应用[J]. 现代预防医学, 2012, 39(22): 5813-5815.

[23] Wang Q Q, Tang J T, Zhang L, et al. Seismic data denoising based on multi-layer perceptron[J]. Oil Geophysical Prospecting, 2020, 55(2): 272-281, 228.
王琪琪, 汤井田, 张良, 等. 利用多层感知机的地震数据去噪[J]. 石油地球物理勘探, 2020, 55(2): 272-281, 228.

[24] Shen H Y, Wang Z X, Gao C Y, et al. Determining the number of BP neural network hidden layer units [J]. Journal of Tianjin University of Technology, 2008, 24(5): 13-15.
沈花玉, 王兆霞, 高成耀, 等. BP神经网络隐含层单元数的确定[J]. 天津理工大学学报, 2008, 24(5): 13-15.

[25] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323(6088): 533-536.

[26] Zhang Y H. The beauty of deep learning: data processing and best practices in the AI era[M]. Beijing: Publishing House of Electronics Industry, 2018: 214-215.
张玉宏. 深度学习之美: AI时代的数据处理与最佳实践[M]. 北京: 电子工业出版社, 2018: 214-215.

[27] Quan Y, Wang Z Q, He M. Application of neural network based on cross-entropy method in pathological image analysis[J]. Journal of China Medical University, 2009, 38(6): 446-448.
全宇, 王忠庆, 何苗. 基于交叉熵的神经网络在病理图像分析中的应用[J]. 中国医科大学学报, 2009, 38(6): 446-448.