

SERS Signal Discrimination of Synthetic Pigment Based on PCA-SVM

Lingfeng Duan, Zhaocong Wang, Bingxin Shou, Lidong Wang*

Qianjiang College, Hangzhou Normal University, Hangzhou Zhejiang
Email: *violet_wld@163.com

Received: Feb. 16th, 2020; accepted: Mar. 2nd, 2020; published: Mar. 9th, 2020

Abstract

Surface enhanced Raman spectroscopy (SERS) is a novel substance detection technology, which has the advantages of fast, high efficiency, and low loss rate. In the analysis of components, it is difficult to distinguish them by conventional means because of the high overlapping of SERS spectra of analytes with similar structures. In this paper, based on the combination of principal component analysis (PCA) and support vector machine (SVM) model to classify and predict the spectrum, taking the edible synthetic pigment of the same type as an example, using SERS and PCA-SVM model to verify the effectiveness of the classification model, the results show that the accuracy of this method is as high as 98%, and the results are basically the same as expected. It provides a basis for SERS signal processing of similar structure substances.

Keywords

SERS, PCA, SVM, Pigment

基于PCA-SVM的合成色素SERS信号判别

段凌风, 王兆聪, 寿彬鑫, 王李冬*

杭州师范大学钱江学院, 浙江 杭州
Email: *violet_wld@163.com

收稿日期: 2020年2月16日; 录用日期: 2020年3月2日; 发布日期: 2020年3月9日

摘 要

表面增强拉曼光谱(SERS)是一种新型的物质检测技术, 有快速、高效、低损耗率等优点。在进行成分分

*通讯作者。

文章引用: 段凌风, 王兆聪, 寿彬鑫, 王李冬. 基于 PCA-SVM 的合成色素 SERS 信号判别[J]. 计算机科学与应用, 2020, 10(3): 437-444. DOI: 10.12677/csa.2020.103045

析时,具有相似结构的待分析物的SERS光谱会出现堆叠的情况,难以采用常规方法区分这类物质。本文基于主成分分析法(PCA)和支持向量机(SVM)相结合的模型对光谱进行分类预测,以同类型的可食用人工合成色素为例,运用SERS和PCA-SVM模型,验证了该分类模型的有效性。结果表明,该方法对不同色素预测的准确度高达98%,且所呈现的结果基本与预期结果相同,具有良好的分类效果。为相似结构物质的SERS信号处理提供依据。

关键词

SERS, 主成分分析, 支持向量机, 色素

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

表面增强拉曼光谱(Surface-Enhanced Raman Spectroscopy, SERS)是一种新型的物质检测技术。由于具有对物质进行定性定量的检测能力,在环境监测、生物化学等领域得到了广泛应用[1] [2] [3]。在 SERS 检测中,研究对象的特征峰是“指纹区”,所以在各个拉曼光谱中进行峰强、峰位的比较,可以很好地区分出各种物质。但 SERS 光谱的形成是由于物质分子的振动,而同类物质的分子结构有一定的相似性,所以产生的特征峰的位置会非常接近,峰强近似相同,难以用常规手段进行区分[4] [5]。如果这些噪声不减少,就会极大程度影响分析判别模型的质量和检测的有效率及准确率。

目前,常见的光谱分析判别方法,如小波变换[6] [7]、偏最小二乘法[8] [9]、人工神经网络[10] [11]等都是基于数据挖掘的经典模型。其中小波变换是基于小波基函数的平移和展宽来获得时域和频域特征的方法。而小波基函数的选取、分解尺度和阈值对去噪效果有很大的影响。偏最小二乘法是一种在理论值和观察值之差即残差的平方和达到最小的前提下求解未知参数的方法,对于线性相关的数据有着较强的回归分析能力。人工神经网络是一种模拟动物神经网络进行机器学习的一种方法,数据量充足的情况下在进行自学习和联想存储记忆方面有明显的适应力,寻找优化解的速度快。

由于 SERS 信号维数过高,直接对信号进行分析判别容易造成维数灾难,对分析产生不良影响[12]。同时噪声也是光谱采集部分不可避免的,噪声的加入会使光谱的特征峰变得模糊,如果这些噪声不减少,就会极大程度影响分析判别模型的质量和检测的准确率。因此必须对原始数据进行降维、降噪的预处理,应用一定的数据处理方法,将数据维数降低的同时最大化体现其特征,提取出各物质主要信息,从而实现物质的区分。主成分分析法(Principal Components Analysis, PCA)是一种特征预处理的分析技术,可以提取一些对于结果有帮助的特征,实现需要分析判别的数据降维和特征位提取。在判别预测模型的选取上,支持向量机(Support Vector Machine, SVM)是一种可以有效区分物质光谱数据信息的模型[13] [14],这种模型可以实现小样本学习,其计算复杂性取决于支持向量的数目,可以在一定程度上避免“维数灾难”,具有较好的鲁棒性。基于主成分分析与支持向量机的判别模型在光谱分析和图像识别中有一定的应用[15] [16] [17],可以将光谱的特征提取出来,提高光谱的判别效率。本论文首先使用主成分分析提取 SERS 信号中的重要特征,使用遗传算法对支持向量机的参数进行优化,再使用 SVM 支持向量机进行分类预测。将输出的值和预定值进行对比,得到模型的预测准确率,如图 1 所示。

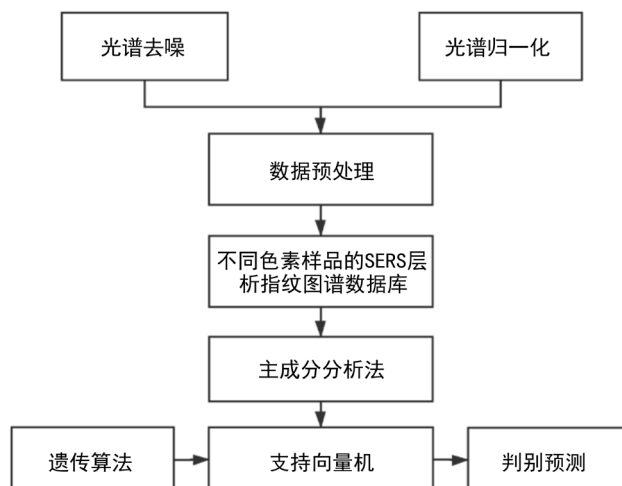


Figure 1. Flow chart of SERS signal feature extraction

图 1. SERS 信号特征提取流程图

2. 实验方法与结果

在前期的研究中, 我们获得了纯苋菜红和诱惑红色素的 SERS 数据, 并采用萃取法得到了香肠中的苋菜红和诱惑红的 SERS 信号[18]。本研究采取相似的办法获得了纯苋菜红色素和蜜饯中苋菜红溶液的 SERS 光谱信号, 以及纯亮蓝色素和蜜饯中亮蓝溶液的 SERS 光谱信号, 如图 2 和图 3 所示。可以看出, 数据特征峰明显, 但是仍然有很多杂峰会影像数据分析, 且数据的维度很高, 需要进行降维处理。

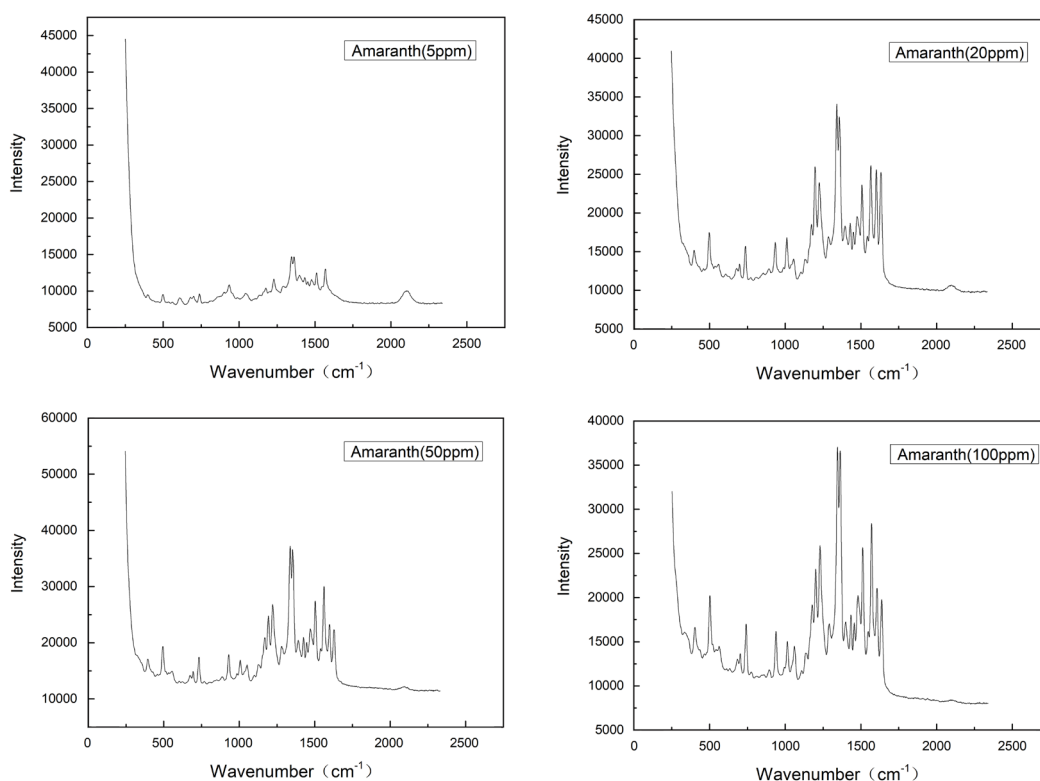


Figure 2. SERS signal of amaranth in different concentrations

图 2. 不同浓度苋菜红 SERS 信号

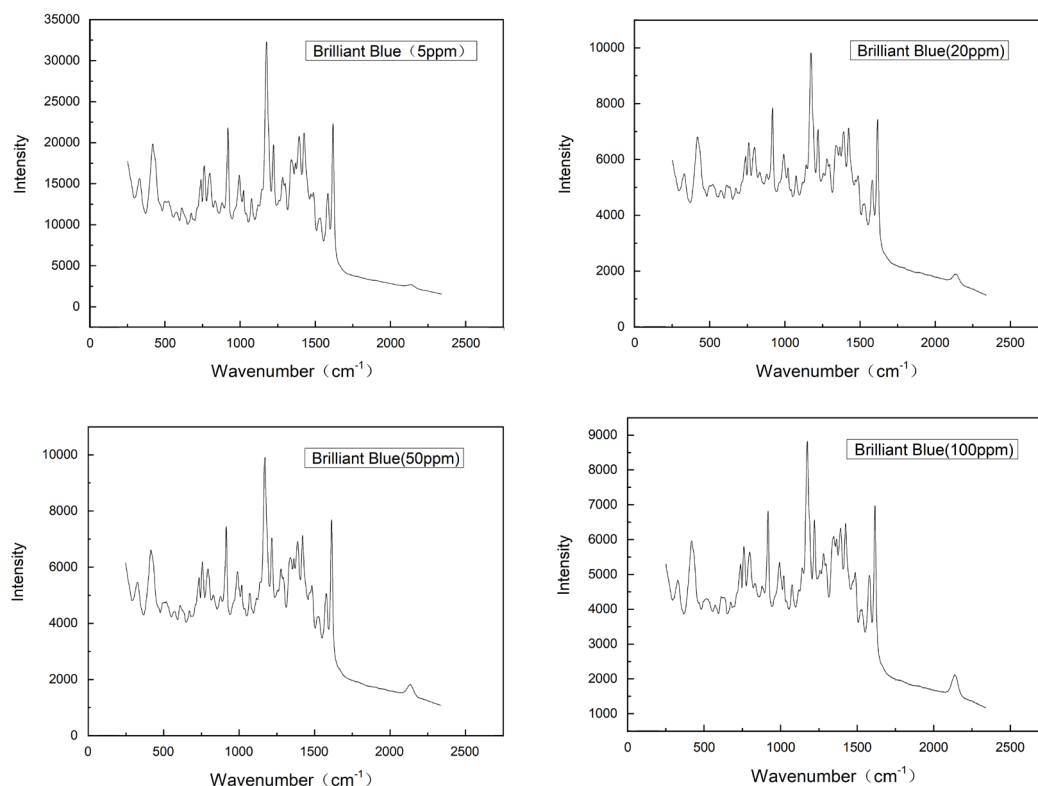


Figure 3. SERS signal with different concentration of brilliant blue

图 3. 不同浓度亮蓝 SERS 信号

3. 数据处理方法

3.1. 主成分分析法[19]

PCA 是一种经典的数学降维模型，它通过一个正交变换，将和其原数据分量相关的原随机样本投影到一个新的空间中去，转化成其分量不相关的新随机向量，然后对多维变量进行降维处理，把携带原有数据信息较多的变量保存下来，实现对原数据的降维的目的。

假设训练集为

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

计算出矩阵协方差为

$$\text{var}(x_j) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \quad (i=1, 2, \dots, n; j=1, 2, \dots, p)$$

进行标准化处理

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_{ij}}{\sqrt{\text{var}(x_j)}} \quad (i=1, 2, \dots, n; j=1, 2, \dots, p) \quad (1)$$

初始化后的矩阵仍用 X 来表示，经处理后数据的相关系数矩阵 R 为

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{bmatrix} \quad (2)$$

其中, $r_{ij} = \text{cov}(x_i, x_j) = \frac{\sum_{k=1}^{k=n} (x_i - \bar{x}_i)(x_j - \bar{x}_j)}{n-1}, n > 1$

贡献率为

$$C = \frac{\lambda_i}{\sum_{i=1}^P \lambda_i} \quad (3)$$

具体实施方法是把检测到的光谱数据作为训练集 X , 取累计共享率达到 95% 的特征数据来代表原光谱数据。

3.2. 支持向量机[20]

SVM 是一种按监督学习方式对数据进行二元分类的广义线性分类器, 其决策边界是对学习样本求解的最大边距超平面。

把纯色素光谱数据设为训练样本集 X 为 $x_i, i=1, 2, \dots, N$, 光谱样本分属两类, w_1 和 w_2 , 线性判别函数 1 的一般形式为: $g(x) = w \cdot x + b$ 。色素判别的问题就转化成寻找最大间隔的分类超平面问题。在这个过程中, 对距离超平面 l_0 最近的样本进行归一化处理, 则有线性可分类 SVM 的优化函数为

$$\arg \min \frac{1}{2} \|w\|^2, \text{ s.t. } y_i (w \cdot x_i + b) \geq 1, i=1, 2, \dots, n \quad (4)$$

解决上述函数的优化问题也就能得出不同物质的区分结果, 由于 SVM 需要数据线性可分, 在遇到线性不可分的情况下一般采用添加核函数的方法, 将原始数据通过核函数映射到更高维平面 v_1 , 在对映射后的数据进行降维至低维平面进行求解。对于多种类数据, 采取“一分为二”的思想, 对每种数据进行逐步区分。

4. 模型应用

4.1. PCA 处理

因为得到的苋菜红和亮蓝 SERS 光谱的维数较高, 因此采取 PCA 进行降维和去噪处理。将用检测存储的拉曼光谱数据文件导入 MATLAB 中, 运行 PCA 程序。经过 PCA 处理后, 前两个特征可以代表 96.7% 的主要成分, 满足 95% 主要成分的需求, 所以取前二特征, 即将原始 1450 维的数据降低至 2 维, 以便投入 SVM 支持向量机中进行处理。将得出的数据用 Excel 表格保存部分数据如表 1。

Table 1. Main feature results after dimension reduction based on PCA

表 1. 基于 PCA 降维后主要特征结果

样品号	特征一峰位	特征一峰强	特征二峰位	特征二峰强
01	1173	32276.4180	1616	22324.7734
02	1172	9821.87087	1616	7433.05144
03	1174	9917.17784	1616	7694.50465
04	1173	9020.20897	1615	6859.59249
05	1344	14717.9075	1361	14687.0474
06	1343	34062.9349	1361	32433.1216
07	1344	37233.3106	1361	36620.9769

4.2. SVM 判别

将经过 PCA 降维处理后的数据进行 SVM 判别。

- 1) 选取经过 PCA 处理过后的光谱数据样本作为训练样本, 经过基线校正、去噪后得到训练样本数据;
- 2) 将训练样本分为正样本和负样本, 其中正样本即为亮蓝的光谱信息, 负样本为苋菜红的光谱信息作图表示验证结果信息;
- 3) 通过前文主成分分析法对光谱进行降维处理并得到代表光谱的特征值, 采用 SVM 进行模型建立, 得到分类器。

由于温度、机械噪声、采集位点等非人为因素, 使得同一物质谱图信息存在一定的噪声。因此, 降低物质与样本相似度, 在增大阈值时, 检验的准确率下降。首先, 通过对光谱信息的降维处理, 实现光谱信息概括较完整的主要特征; 其次, 运用 SVM 寻找物质间的最大几何间隔, 在检验过程中, 保证经验风险最小, 降低置信风险, 实现了在检验中的真实风险最小化。

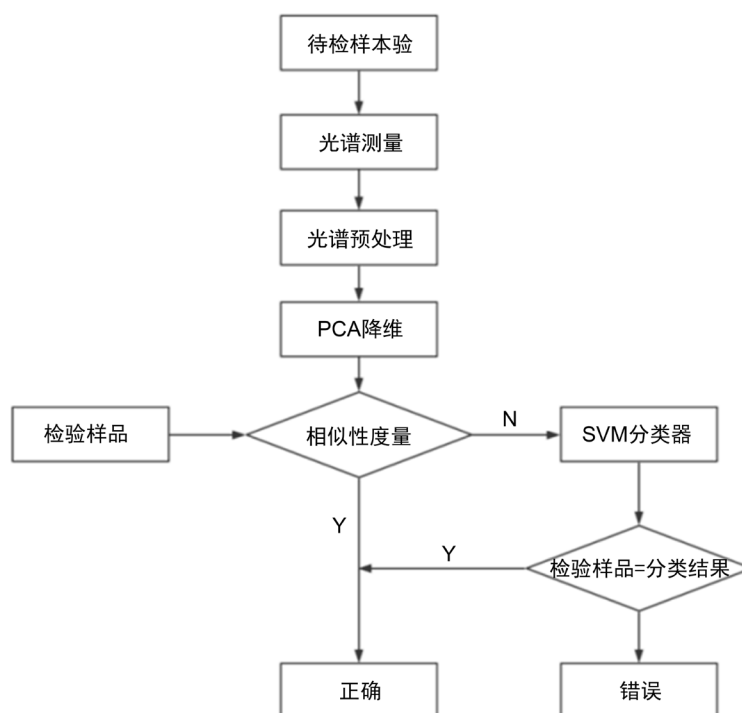


Figure 4. Flow chart of SERS spectral processing

图 4. SERS 光谱处理流程图

SVM 的准确性验证, 识别结果与真实值相同的样本输出正确, 与真实值不同的样本输出错误。应用测试样本对模型的准确性进行验证并输出模型的准确率。具体流程如图 4 所示。

4.3. 结果与分析

将获得的纯品色素数据作为输出变量, 将所有 450 个样本采用选一留一法, 分为训练集(400 个样本)和预测集(50 个样本)。SVM 的参数采用遗传算法进行优化。SVM 最佳参数 c 、 g 分别为 21.69 和 0.08, 交叉验证正确率约为 90.38%。实测集与预测集分类图(图 5)显示 50 个预测样本中预测准确率达 98%, 有一个亮蓝错判为苋菜红。上述结果表明, 基于不同色素的 SERS 光谱, 结合 SVM 分析技术能够较好地识别、区分食品中含有的色素。

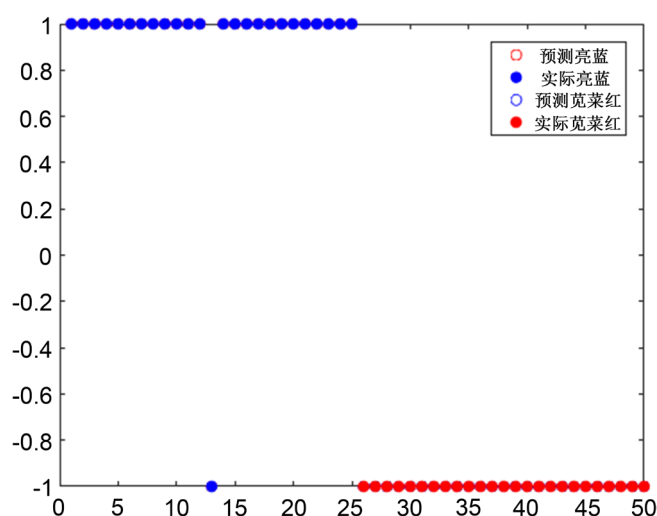


Figure 5. Prediction test chart

图 5. 预测检验图

由图 6 可知, 样品浓度变化时的光谱图会出现明显移动, 本文所选用的 PCA-SVM 模型的平均预测准确率为 98%。其中 50 个预测样本仅一个出错, 和预测准确率以及预期结果一致。SERS 的信号相对于传统的检测手段更为具体, 因为 SERS 是物质“指纹”区域, 本身具有相对较少的杂峰, 特征峰也相对明显, 所以特征的提取和识别度较高。在多组检测中, 模型无法达到 100 % 的准确率, 其主要原因为: (1) 显微镜的焦距误差对数据的影响; (2) 器材的噪声对数据检出的影响; (3) PCA 降维产生噪声; (4) 溶液色素分子沉降但是分布的不均匀导致 9 点检测的光谱信号有所不同; (5) 基底采用银质纳米棒阵列, 纯银在空气中存放时间过长发生氧化在基底的附着面形成了氧化银, 产生了背景信号, 影响 SERS 信号检测。

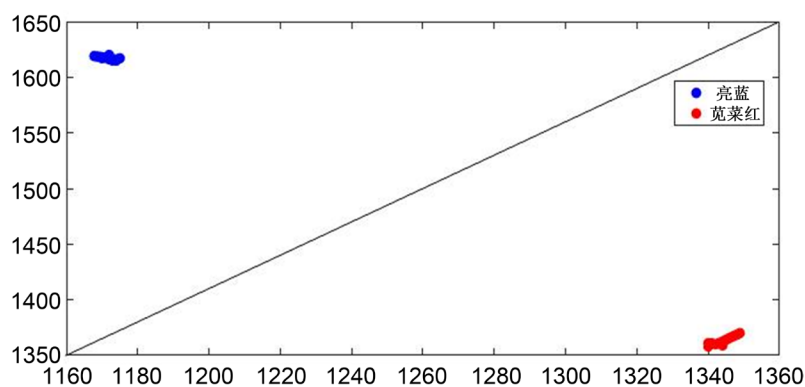


Figure 6. Two classification diagrams of SVM

图 6. SVM 二分类图

5. 总结

本文采用了 PCA-SVM 模型结合 SERS 光谱区分技术对两种常见的人工合成色素苋菜红和亮蓝进行区分, 建立了一个有关苋菜红和亮蓝的区判别模型。利用该模型进行区分的同时, 对从蜜饯中提取的色素溶液成分进行了预测, 实验结果与预测结果基本一致。PCA-SVM 支持向量机模型应用于 SERS 信号的数据挖掘还可以通过改善实验条件、优化模型数据等方法, 减少人为和非人为因素给判别分析带来的不必要误差。在遇到多种物质的情况下, 采用分步处理的方法, 先一分为二再进行区分, 使训练和测

试速度更快, 所需的存储空间更少, 极大程度减少训练学习的时间。在遇到线性不可分的数据的时候可以通过使用一个核函数将原有数据投影到一个线性可分的更高维平面, 再进行分类。这种数据挖掘模型还可应用于其他相似结构物质的 SERS 光谱数据分析, 后续可写入芯片, 应用于小型手持式拉曼光谱仪和微型拉曼检测系统, 从而实现食品添加剂和农药残留等问题的现场快速检测。

基金项目

浙江省新苗人才计划项目(基于多种数据挖掘方法的 SERS 鉴别研究 2019R426046); 浙江省自然科学基金项目(LY19F020022); 2017 年杭州市社会发展科研项目(20170533B05)。

参考文献

- [1] Han, C.Q., Yao, Y. and Wang, W. (2017) Rapid and Sensitive Detection of Sodium Saccharin in Soft Drinks by Silver Nanorod Array SERS Substrates. *Sensors and Actuators B-Chemical*, **251**, 272-279. <https://doi.org/10.1016/j.snb.2017.05.051>
- [2] Yan, B., Li, B. and Wen, Z.N. (2015) Label-Free Blood Serum Detection by Using Surface-Enhanced Raman Spectroscopy and Support Vector Machine for the Preoperative Diagnosis of Parotid Gland Tumors. *BMC Cancer*, **15**, 650. <https://doi.org/10.1186/s12885-015-1653-7>
- [3] 王昕, 何坚, 范贤光. 基于表面增强拉曼光谱的合成色素专利蓝 V 的快速检测[J]. 光谱学与光谱分析, 2019, 39(10): 3205-3209.
- [4] Ma, C.-H., Zhang, J. and Hong, Y.-C. (2015) Determination of Carbendazim in Tea Using Surface Enhanced Raman Spectroscopy. *Chinese Chemical Letters*, **26**, 1455-1459. <https://doi.org/10.1016/j.cclet.2015.10.015>
- [5] Aditya, K. and Venugopal, S. (2019) Paper Swab Based SERS Detection of Non-Permitted Colourants from Dals and Vegetables Using a Portable Spectrometer. *Analytica Chimica Acta*, **1090**, 106-113. <https://doi.org/10.1016/j.aca.2019.08.073>
- [6] 牛宏侠, 张肇鑫, 宁正. 基于小波变换的阈值自适应寻优去噪方法[J/OL]. 传感器与微系统, 2020(2): 33-36.
- [7] 毛焯, 陈亮. 强噪声条件下激光雷达强度图像降噪方法研究[J]. 激光杂志, 2019, 40(11): 80-83.
- [8] 杜馨, 孙晓荣, 刘翠玲. 拉曼光谱结合偏最小二乘法对食用油品质快速检测研究[J]. 中国酿造, 2019, 38(12): 171-174.
- [9] 滕宇. 偏最小二乘法在光谱分析中的应用[J]. 集成电路应用, 2020, 37(1): 16-17.
- [10] 公滨南. BP 神经网络算法在卫星遥感影像地物分类中的应用[D]: [硕士学位论文]. 哈尔滨: 东北农业大学, 2019.
- [11] 全青青. 基于 BP 神经网络的落叶松人工林树种分类[J]. 信息技术与信息化, 2019(9): 43-45.
- [12] Lin, Z.S. and He, L. (2019) Recent Advance in SERS Techniques for Food Safety and Quality Analysis: A Brief Review. *Current Opinion in Food Science*, **28**. <https://doi.org/10.1016/j.cofs.2019.10.001>
- [13] 刘志强, 吕学, 张利. 基于多分类 GA-SVM 的高速公路 AID 模型[J]. 系统工程理论与实践, 2013, 33(8): 2110-2115.
- [14] 高金贺, 黄伟玲, 蒋浩鹏. 城市交通碳排放预测的多模型对比分析[J/OL]. 重庆交通大学学报(自然科学版): 1-7[2019-11-14].
- [15] 郭云香, 陈龙, 李晓瑾. 基于 NIRS 技术和 PCA-SVM 算法快速鉴别国产和进口啤酒花[J]. 药学实践杂志, 2019, 37(4): 322-331.
- [16] 陈勇, 李鹏, 张忠军. 基于 PCA-GA-LSSVM 的输电线路覆冰负荷在线预测模型[J]. 电力系统保护与控制, 2019, 47(10): 110-119.
- [17] 刘翠翠, 杨涛, 马京晶. 基于 PCA-SVM 的麦冬叶部病害识别系统[J]. 中国农机化学报, 2019, 40(8): 132-136.
- [18] 段凌风, 徐璐, 王李冬. 基于 PCA-BP 神经网络的合成色素 SERS 信号判别[J]. 电脑知识与技术, 2017, 13(31): 196-198+204.
- [19] 李海洋, 刘胜. 纺织品近红外光谱定性分析的一种新方法[J]. 光谱学与光谱分析, 2019, 39(7): 2142-2146.
- [20] Hou, X.W., Wang, G.L. and Su, G.Q. (2019) Rapid Identification of Edible Oil Species Using Supervised Support Vector Machine Based on Low-Field Nuclear Magnetic Resonance Relaxation Features. *Food Chemistry*, **280**, 139-145.