

基于卷积神经网络和近红外光谱的太平猴魁茶产地鉴别分析

陈琦^{1,3}, 潘天红^{2,4*}, 李鱼强⁴, 林鸿⁴

1. 合肥工业大学食品与生物工程学院, 安徽 合肥 230601
2. 安徽大学电气工程与自动化学院, 安徽 合肥 230601
3. 黄山海关茶叶质量安全研究中心, 安徽 黄山 245000
4. 江苏大学电气信息工程学院, 江苏 镇江 212013

摘要 太平猴魁茶因其特有的“喉韵”深受广大消费者喜爱,不同产地太平猴魁茶市场价格相差较大,如何实现产地精准鉴别是目前促进绿茶产业发展的关键因素。依赖于人工经验的感官评审方法主观性强、稳定性差,无法应用于实际生产检测过程。作为目前主要的检测分析方法,化学分析方法周期长、检测成本高,而且目前没有用于茶叶产地鉴别的统一标准。近红外光谱(NIR)作为一种无损检测分析方法,具有快速、非破坏性、无污染等特点,但是不同产地太平猴魁茶主要内含成分及其含量基本相同,不同产地样本光谱特征峰分布相似,导致常规分析方法无法有效选择特征变量。卷积神经网络(CNN)作为经典深度学习网络模型之一,具有强特征提取和模型表达能力。采用太平猴魁茶产地光谱特征分析,利用一维卷积神经网络模型(1-D CNN)提取太平猴魁茶 NIR 特征,提出一种基于 1-D CNN 和 NIR 的太平猴魁茶产地鉴别分析方法。试验以 6 个不同产地共 120 个样本为研究对象,分析 10 000~4 000 cm^{-1} 范围内的光谱信息;将样本随机划分为训练集(84,占 70%)和测试集(36,占 30%),分别讨论不同间隔采样、网络结构、卷积核大小及激活函数对产地鉴别结果的影响,并引入 Dropout 方法对比分析模型过拟合现象;最终建立一个具有 9 层结构的 1-D CNN 模型。蒙特卡罗试验结果表明,相比于基于原始光谱数据(40.57%, 7.06)和 PCA 方法(31.93%, 6.96)的太平猴魁茶产地预测模型准确率和标准差,基于 1-D CNN 的太平猴魁茶产地鉴别模型预测精度和稳定性更高,其测试集预测准确率平均值和标准差分别为 97.73%和 3.47。因此,1-D CNN 可有效提取太平猴魁茶不同产地 NIR 特征,提高太平猴魁茶产地鉴别精度,为太平猴魁茶精准产地鉴别及溯源分析提供参考。

关键词 产地鉴别;卷积神经网络;近红外光谱;特征提取;太平猴魁茶

中图分类号: Q657.3 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2021)09-2776-06

引言

太平猴魁茶属于绿茶尖茶,主要产于安徽省黄山市黄山区三合村猴坑、猴岗和严家三个村落^[1]。太平猴魁茶成品具有“两叶一芽、扁平挺直、魁伟重实、色泽苍绿、兰香高爽、滋味甘醇”等特点,深受广大消费者喜爱^[1-2]。随着市场的不断扩大,茶叶掺假现象的不断发生损害了太平猴魁茶的市场形象,实现精准产地鉴别分析对促进太平猴魁茶发展具有重要意义。

传统感官分析方法主要通过外观、汤色、滋味及香气等感官指标实现产地鉴别分析,但手工业工艺制备的茶叶成品外观差异较小,导致基于人工经验的感官评审方法无法实现快速、精准的产地鉴别分析^[3]。目前主要通过化学分析方法实现茶叶高精度产地鉴别,然而化学分析检测繁琐、成本高,并且目前没有统一的化学检测标准可用于太平猴魁茶产地鉴别分析。

近红外光谱(near infrared spectroscopy, NIR)作为一种无损分析方法,具有快速、非破坏性、无污染等特点,已在茶叶生产过程得到广泛应用^[4-5]。然而不同产地茶叶内含成

收稿日期: 2020-08-04, 修订日期: 2020-12-10

基金项目: 国家重点研发计划项目(2017YFF0211302, 2017YFF0211301); 安徽省高校协同创新项目(GXXT-2019-012), 安徽省科技重大专项项目(202003a06020001)资助

作者简介: 陈琦, 1986年生, 合肥工业大学食品与生物工程学院高级工程师 e-mail: chenqiyu@126.com

* 通讯作者 e-mail: thpan@live.com

分种类及其含量基本相同,不同产地样本光谱特征峰分布基本相同,导致常规分析方法无法有效选择光谱特征。卷积神经网络(convolutional neural networks, CNN)作为典型的深度学习网络模型之一,具有很强的特征提取和模型表达能力,已被广泛应用于 NIR 特征提取分析^[6-9]。

本研究以不同产地光谱特征以及 CNN 模型,以太平猴魁茶为研究对象,利用 1-D CNN 模型提取 NIR 特征,建立基于 1-D CNN 和 NIR 的太平猴魁茶产地鉴别模型,并采用蒙特卡罗方法进行产地鉴别模型稳定性分析,为太平猴魁茶产地鉴别及溯源分析提供新方法。

1 实验部分

1.1 材料与仪器

材料:选取太平猴魁茶地理保护范围黄山市黄山区的猴坑、猴岗、颜家、三合、石河坑、汪王岭共 6 个不同产地样本,每个产地各 20 个样品。按照采摘时间从指定地点采集样品,委托黄山市猴坑茶业有限公司按照太平猴魁茶传统工艺进行制备,样本信息如表 1 所示。

表 1 样品信息
Table 1 Sample information

产地	样品规格数量	采摘时间	样品编号
猴坑	50 g×20 个	2018.4.18	101-120
猴岗	50 g×20 个	2018.4.17	201-220
颜家	50 g×20 个	2018.4.17	301-320
三合	50 g×20 个	2018.4.16	401-420
石河坑	50 g×20 个	2018.4.16	501-520
汪王岭	50 g×20 个	2018.4.16	601-620

试验仪器:IRTracer-100NIR 仪(日本岛津)、KN 295 Knifetec 样品磨(瑞典 FOSS)、石英样品杯(日本岛津)。

1.2 方法

1.2.1 样品制备

将茶叶用粉碎机粉碎 50 s,将粉碎的试样过 0.154 mm 的筛,密封放于 4 °C 冰箱中备用。

1.2.2 仪器条件

光阑:自动,延迟 10 s;增益:1;模式:能量扫描;零填充:4 倍;动镜速度:2.8 mm·s⁻¹;测定模式:吸光度;扫描波数范围:10 000~4 000 cm⁻¹;扫描次数:45 次;光束:内部;检测器:MCT;光源:钨灯光源;变迹函数:SqrTriangle;分辨率:4 cm⁻¹。

1.2.3 光谱采集

取适量样品于石英样品杯中,高度约为 5 mm,保证样品杯底部样品均匀无缝隙、厚度均匀,放在漫反射扫描装置的放样口处,盖好样品室盖板,扫描光谱。

1.3 光谱数据预处理

因背景噪声、杂散光和人工操作等因素影响,导致所采集数据包含噪声信号^[10]。在分析之前对所采集的原始光谱数据进行预处理。选取标准正态变换(standard normal variants, SNV)作为预处理方法^[11]。

1.4 卷积神经网络

卷积神经网络是一种多层非全连接的高性能非线性深度学习学习方法,其基本结构由卷积层、池化层和全连接层组成,在正向传播过程中利用卷积层和池化层相互交替学习实现原始数据特征提取;在反向传播过程中则利用梯度下降法最小化误差函数实现参数调整,以此完成权值更新^[12]。

对于图像等二维形式的的数据,一般采用大小的卷积核和池化核进行特征提取,但本试验采集的 NIR 是一维数据,因此采用 1-D CNN 进行特征分析,其一维卷积和池化分析过程如图 1 所示。

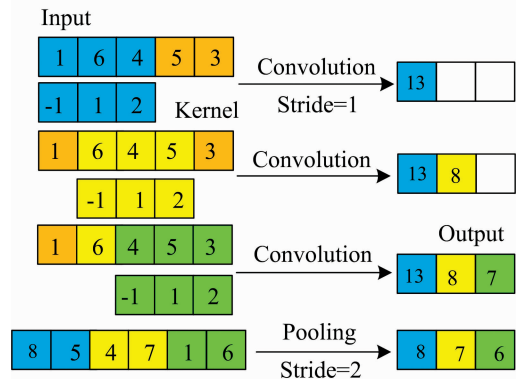


图 1 1-D CNN 提取特征过程

Fig. 1 Feature selection process of 1-D CNN

1.5 模型评价方法

为确定最佳网络结构和系统参数,选择预测准确率(correct identification rate, CIR)和运行时间作为模型性能评估指标,其中 CIR 定义为

$$CIR = \frac{N_c}{N} \times 100\% \quad (1)$$

式中, N 表示全部样本, N_c 表示对应分析数据集中预测准确样本数。

2 结果与讨论

2.1 近红外光谱及预处理

选取黄山区太平猴魁茶共 120 个样本(表 1)进行光谱采集,原始光谱如图 2(a)所示。经 SNV 预处理后的光谱如图 2(b)所示,相比于图 2(a),预处理光谱特征峰更加明显。

2.2 采样间隔对模型的影响

太平猴魁茶原始光谱波长点(12 446)较多,容易导致网络训练时间过长。为降低模型计算成本,需要对原始数据通过间隔采样实现降维处理。为选择最佳采样间隔,将样本随机划分为训练集(84, 占 70%)和测试集(36, 占 30%),分析对比不同采样间隔数据预测精度和计算时间(表 2)。

分析结果表明,基于原始光谱数据的 1-D CNN 模型计算时间过长且模型预测准确率较低,当采样间隔为 6 时, CNN 模型测试集预测精度达到最大(96.67%),当采样间隔继续增大时,1-D CNN 模型计算时间减少,但预测性能快速下降,综合考虑模型预测精度和计算时间,选择 6 作为最佳

采样间隔。

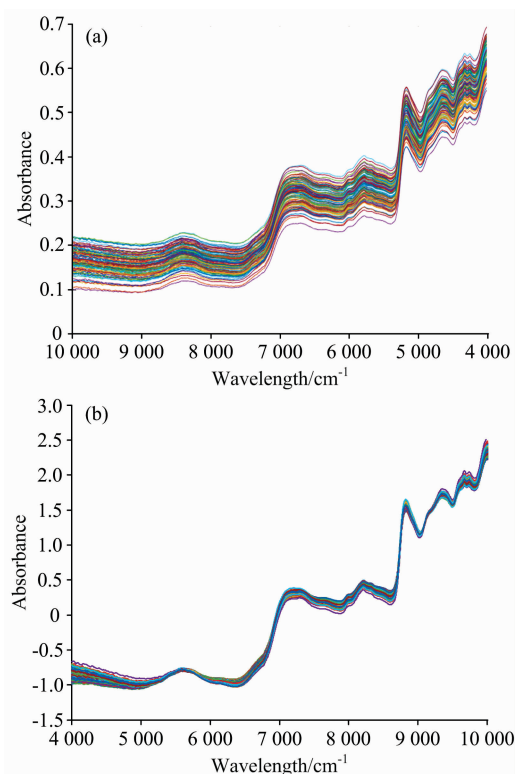


图 2 太平猴魁茶光谱数据

(a): 原始光谱; (b): 预处理光谱

Fig. 2 Spectra of Taiping Houkui tea

(a): Original data; (b): Preprocessed data

表 2 不同采样间隔预测结果

Table 2 Prediction results with different sampling intervals

采样	训练集/%	测试集/%	时间/s
0	100	75.83	217.83
2	100	87.50	112.74
4	100	88.33	60.40
6	100	96.67	43.02
8	100	91.67	35.27
10	100	87.50	29.30

2.3 网络结构设置

为确定最佳网络结构,分别计算 2—5 层网络结果模型损失函数(图 3)。分析结果表明,3 层网络结构模型目标函数可在最短(约 85 步,如图 3 红色实线所示),迭代步长内实现收敛,因此选择具有 3 层网络的 1-D CNN 模型。

2.4 卷积核大小及数目

对于 NIR 数据,重叠特征峰和独立特征峰的存在导致模型结果对于卷积核大小及数目更加敏感,卷积核过小导致模型计算过程复杂;卷积核过大则容易造成特征丢失。为确定最佳卷积核大小,分别讨论 7—39 不同大小卷积核分析结果(图 4)。

对于卷积核数目,按照 $2n$ 进行设置,并分别测试了卷积核数目为 $n=3, 4, 5, 6, 7$ 的分析结果。考虑到池化操作会

减少卷积特征维度,以 2 为梯度递减相邻层卷积核数目,具体分析结果如表 3 所示。

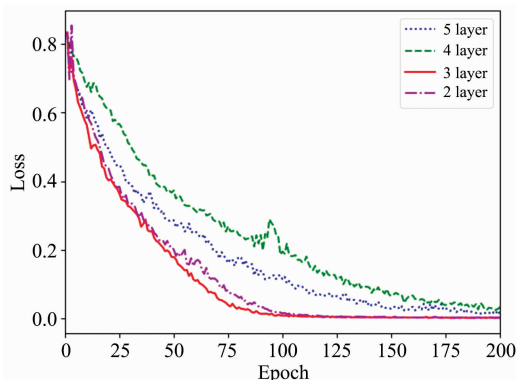


图 3 不同 1-D CNN 结构训练集损失函数

Fig. 3 Loss function values of training set for different 1-D CNN structure

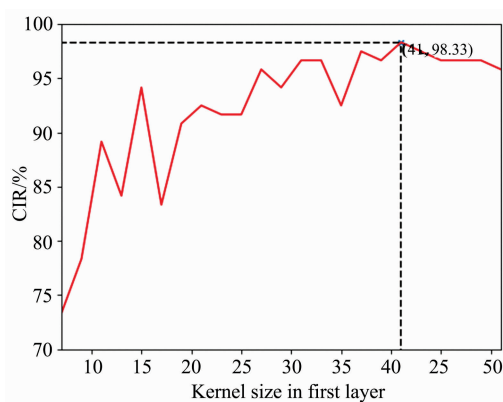


图 4 不同大小卷积核 CIR

Fig. 4 CIR of different convolution kernel sizes

表 3 不同数目卷积核 CIR

Table 3 CIR of different convolution kernel number

	训练集/%	测试集/%	时间/s
16	100	82.49	55.77
32	100	96.16	46.80
64	100	98.33	96.84
128	100	96.67	218.82

由图 4 可知,当卷积核大小为 41 时, CIR 达到最大值,由表 3 可知,当卷积核数目大于 32 时,运行时间显著增大。结合图 4 和表 3,确定最佳卷积核大小和数目分别为 41 和 32。

2.5 Dropout 的影响

为防止因数据集参数过大而引起过拟合现象,采用 Dropout 方法进行网络结构优化。对比网络结构优化前后模型的预测结果发现,在引入 Dropout 后模型性能有明显提升,其预测集准确率从 91.67% 提高到 100%。

2.6 激活函数的比较

为确定最佳激活函数,分别对比含有 4 种不同激活函数(ReLU, Sigmoid, PReLU 和 ELU)网络模型预测结果。其

中, Sigmoid 激活函数预测准确率最低 (75%), PReLU (95.83%) 和 ReLU (93.33%) 次之, ELU 激活函数预测准确率最高 (98.33%)。因此, 选择 ELU 作为 1-D CNN 模型激活函数。

2.7 产地鉴别结果

根据图 5 模型结构和系统参数所建 1-D CNN 模型如图 5

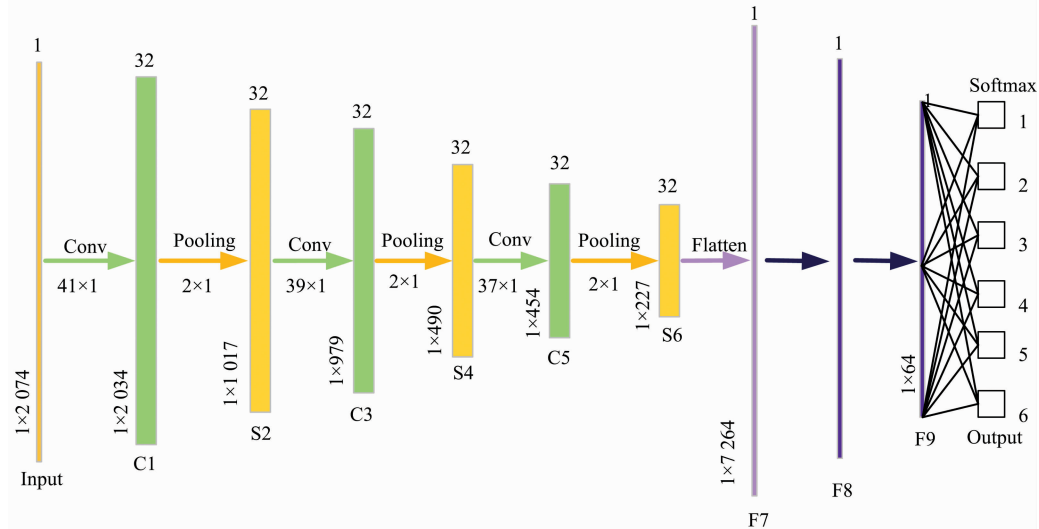


图 5 1-D CNN 模型结构

Fig. 5 Model structure of 1-D CNN model

根据图 5 模型结构所建 1-D CNN 模型的预测结果如图 6 所示, 图中横纵轴数值表示不同产地属性, 对角线变量大小表示正确预测数, 可知基于 CNN 和 NIR 的太平猴魁茶产地鉴别模型预测准确率为 100%。

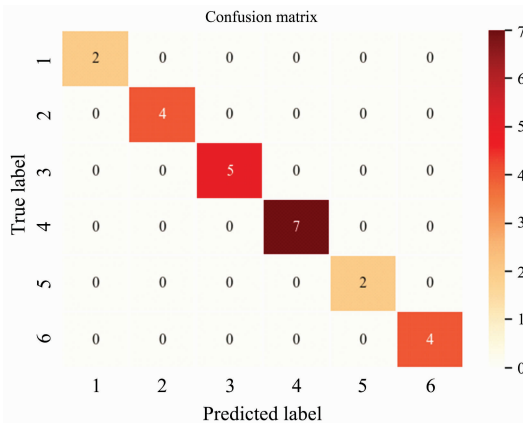


图 6 1-D CNN 模型预测结果

Fig. 6 Prediction results of 1-D CNN model

2.8 结果分析

为验证 1-D CNN 模型筛选 NIR 特征变量的有效性, 分别选取不同产地样本, 对比原始数据和不同卷积层筛选数据的分布情况[见图 7(a)–(d)]。由图 7(a)可知, 原始光谱数据主要分布在 8 700~5 600 和 5 300~4 000 cm^{-1} 两个区间, 经第一层卷积网络获得特征峰主要分布在 7 800~6 700, 6 400~5 200 和 5 200~4 200 cm^{-1} , 第二层筛选的特征峰主

所示, 图中所示结构含有 3 个卷积层(C1, C3, C5), 其卷积核大小分别为 41, 39 和 37, 由 MSRA 初始化卷积核权重。3 个池化层(S2, S4, S6)按照设定步长和窗口大小降低数据维度。在分析过程中, 为防止梯度消失, 在每一次卷积后进行批处理化(batch normalization, BN)。通过 Softmax 分类器实现对特定分析对象的分类预测。

要分布在 10 000~7 100, 6 300~5 400 和 5 400~4 200 cm^{-1} , 第三层筛选特征峰主要分布在 10 000~7 300, 6 300~5 800 和 5 500~4 400 cm^{-1} 。结合太平猴魁茶主要品质成分官能团的分布, 可以发现 1-D CNN 能够有效提取太平猴魁茶内含成分粗纤维 (7 502~5 446 cm^{-1})、茶多酚 (6 101~4 246 cm^{-1})、咖啡碱 (7 502~5 446 和 4 424~4 246 cm^{-1}) 和游离氨基酸 (7 502~5 800 和 5 450~4 246 cm^{-1}) 的 NIR 特征^[13]。对比分析图 7(a)和(d)可知, 1-D CNN 模型在有效选择光谱特征的同时能够实现不同产地特征变量差异化表示, 有助于提高太平猴魁茶产地鉴别模型分析能力。

2.9 模型对比

为验证 1-D CNN 模型有效性, 采用相同的训练集和测试集样本, 对主成分分析 (principal component analysis, PCA) 和 1-D CNN 模型分别进行 100 次蒙特卡罗试验, 不同方法所建模型的训练集和测试集分析结果如表 5 所示。分析结果表明, 相比于原始数据 (41.29%, 40.57%), 基于 PCA 特征筛选的产地鉴别模型准确率有所下降 (30.96%, 31.93%), 其主要原因是不同产地光谱变量分布相似, PCA 无法有效选择特征光谱; 1-D CNN 模型能够显著提高太平猴魁茶产地鉴别精度 (98.48%~97.73%), 可实现太平猴魁茶高精度产地鉴别分析。此外, 测试结果标准差表明基于 1-D CNN 的产地鉴别模型稳定性更高。

3 结论

基于太平猴魁茶不同产地 NIR 信息, 结合 CNN 特征表

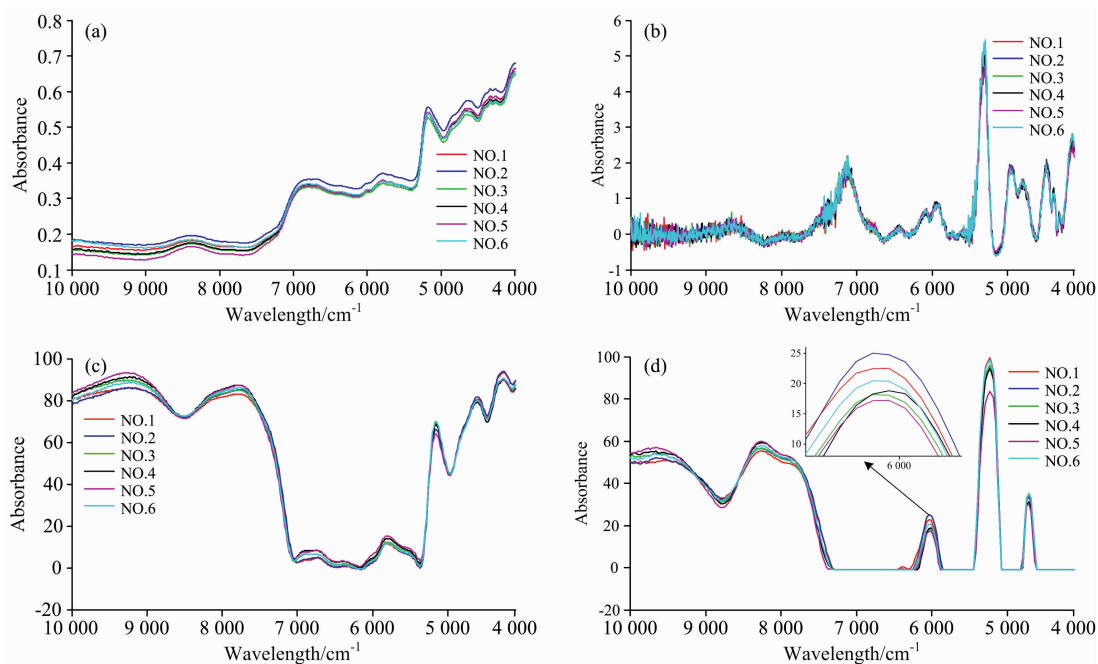


图 7 光谱特征分布

(a): 原始光谱; (b): 第一卷积层; (c): 第二卷积层; (d): 第三卷积层

Fig. 7 Spectral feature distribution

(a): Original spectrum; (b): First convolutional layer; (c): Second convolutional layer; (d): Third convolutional layer

表 5 蒙特卡罗试验结果对比

Table 5 Comparison of Monte Carlo experimental results

方法	变量数	累计贡献率/%	训练集精度平均值/%	预测集精度	
				平均值/%	标准差
原始光谱数据	2 074	100	41.29	40.57	7.06
PCA	3	99.99	30.96	31.93	6.96
1-D CNN	64	—	98.48	97.73	3.47

示方法, 提出基于 1-D CNN 的太平猴魁茶产地鉴别模型。试验结果表明 CNN 能够有效筛选不同产地样本 NIR 特征, 实现太平猴魁茶高精度产地鉴别分析(100%); 100 次蒙特卡罗试验表明, 相比于原始数据(40.57%, 7.06)和 PCA 模型(31.93%, 6/96), 基于 1-D CNN 的产地鉴别模型具有更高的预测精度(97.73%)和更小的标准差(3.47), 该方法为太平猴魁茶等名贵茶叶产地鉴别及溯源分析提供新思路。

References

- [1] ZHAN Hui-bin(占辉斌). Journal of Northeast Agricultural University(东北农业大学学报), 2018, 16(1): 34.
- [2] LEI Pan-deng, HUANG Jian-qin, WU Qiong, et al(雷攀登, 黄建琴, 吴琼, 等). China Tea Processing(中国茶叶加工), 2016, 1(1): 33.
- [3] PAN Tian-hong, LI Yu-qiang, CHEN Qi, et al(潘天红, 李鱼强, 陈琦, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2020, 36(13): 264.
- [4] Li Chunlin, Guo HaoWei, Zong Bangzheng, et al. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc., 2019, 206: 254.
- [5] Li Xiaoli, Jin Juanjuan, Sun Cahnjun, et al. Food Chem., 2019, 270: 236.
- [6] Chen Xiaoyi, Chai Qinqin, Lin Ni, et al. Anal. Methods, 2019, 40(11): 5118.
- [7] Satoru Hiwa, Kenya Hanawa, Ryota Tamura, et al. Comput. Intell. Neurosci., 2016, 2016(3): 1841945.
- [8] Thanawin Trakoolwilaiwan, Bahareh Behboodi, Jaeseok Lee, et al. Neurophotonics, 2017, 5(1): 011008.
- [9] MENG Shi-yu, HUANG Ying-lai, ZHAO Peng, et al(孟诗语, 黄英来, 赵鹏, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(1): 284.
- [10] Zou Xiaobo, Zhao Jiewen, Povey M J W, et al. Anal. Chim. Acta, 2010, 667(1): 14.
- [11] Li Yuqiang, Pan Tianhong, Li Haoran, et al. J. Food Process Eng., 2020, 43(8): 13445.
- [12] Ren Shaoqing, He Kaiming, Girshick Ross, et al. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39(6): 1137.
- [13] Chen Quansheng, Zhao Jiewen, Sumpun Chaitep, et al. Food Chem., 2009, 113(4): 1272.

Geographical Origin Discrimination of Taiping Houkui Tea Using Convolutional Neural Network and Near-Infrared Spectroscopy

CHEN Qi^{1,3}, PAN Tian-hong^{2,4*}, LI Yu-qiang⁴, LIN Hong⁴

1. School of Food and Biological Engineering, Hefei University of Technology, Hefei 230601, China

2. School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China

3. Huangshan Customs Research Center for Tea Quality and Safety, Huangshan 245000, China

4. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China

Abstract Taiping Houkui Tea, one of China's precious tea series, occupies an important position in public consumption and the tea market. The price of Taiping Houkui tea varies greatly from different geographical origins, and accurate geographical origin discrimination is currently an important factor in promoting the green tea industry. Sensory evaluation methods that rely on the manual experience are highly subjective and poor instability and cannot be applied to the actual analysis process. As the main detection and analysis method at present, the chemical analysis method is time-consuming and laborious. More importantly, there is currently no uniform standard for the geographical origin discrimination of tea. Near-infrared spectroscopy (NIR), as non-destructive testing and analysis method, has the characteristics of fast, non-destructive, and non-polluting. However, the types and contents of main components of Taiping Houkui tea from different origins are similar, which results in the same spectral peak distributions of various samples, and conventional analysis methods are limited for selecting feature variables. As one of the typical deep learning network models, convolutional neural network (CNN) has strong feature extraction and model expression capabilities. Based on the analysis of the spectral characteristics of Taiping Houkui tea from different geographical origins, the 1-dimension CNN (1-D CNN) is used to extract the NIR features, and a discriminant method combing NIR with 1-D CNN is explored to identify the geographical origin of Taiping Houkui tea in this work. In this paper, 120 samples were collected from 6 different geographical origins. The NIR were sampled from $10\ 000\sim 4\ 000\ \text{cm}^{-1}$ and preprocessed by standard normal variate (SNV). The sample is randomly divided into a training set (84, 70%) and test set (36, 30%), and the effects of CNN structure, convolution kernel size, activation function and other parameters on the analysis results were discussed separately. As a result, a 1-D CNN model with 9-layer was constructed for the geographical origin discrimination of Taiping Houkui tea. The principal component analysis (PCA) was compared, and the Monte-Carlo method was used to evaluate the stability and robustness of the proposed method. Compared with the prediction accuracy and standard deviation of the models based on original spectral data (40.57%, 7.06) and the PCA method (31.93%, 6.96), the prediction accuracy and stability of the 1-D CNN-based geographical origin discrimination model are higher, and the average prediction accuracy and standard deviation of the testing set are 97.73% and 3.47, respectively. The comparison results demonstrate the proposed 1-D CNN model can effectively extract NIR features and has the ability to identify the geographical origins of Taiping Houkui tea, which provides an effective method for the identification and traceability analysis of the geographical origin and production of valuable tea products such as Taiping Houkui tea.

Keywords Geographical origin discrimination; Convolutional neural network; Near-infrared spectroscopy; Feature selection; Taiping Houkui tea

(Received Aug. 4, 2020; accepted Dec. 10, 2020)

* Corresponding author