基于近红外光谱结合数据增强 CNN 算法的白芷产地溯源方法

郭兆华[1]，文师召[2]，李思凡[3]，王琪[3]，王颖鑫[4]，王鑫国[4]，牛丽颖[4]，李亚薇[5*]，冯薇[4*]（1. 中国电子科技集团公司网络通信研究院微波散射通信专业部，石家庄 050050；2. 南开大学统计与数据科学学院，天津 300192；3. 东北大学理学院，沈阳 110167；4. 河北中医药大学药学院，中药材品质评价与标准化河北省工程研究中心，石家庄 050091；5. 辽宁省检验检测认证中心，辽宁省分析科学研究院，沈阳 110032）

摘要：目的 在中药产地溯源领域，基于近红外光谱结合数据增强卷积神经网络（CNN）算法建立样本量不均衡的白芷产地分类模型具有很大的理论研究价值与实际应用价值。**方法** 研究采集 95 份白芷样本，采用 12 500～4 000 cm^{-1} 波段对不同白芷样品进行近红外光谱采集。本研究所使用的白芷近红外光谱数据集，存在样本量小、样本产地类别分布不均衡等问题。本研究提出了 3 种数据增强算法，包含光谱平移、光谱增噪和光谱组合来提升模型泛化能力，并使用 Focal Loss 作为损失函数来训练 CNN 模型解决样本不平衡的问题。**结果** 将 3 种数据增强算法应用于支持向量机（SVM）模型，对光谱数据添加信噪比为 20 的高斯噪声效果最好，能够将模型正确率提高至 84.2%；在样本不平衡的情况下，通过应用 Focal Loss 作为损失函数来训练 CNN 模型，实现了高达 94.7% 的正确率。**结论** 通过红外光谱技术结合数据增强的 CNN 算法为白芷产地溯源提供了快速、无损的检测手段及可靠的数据分析方法，为中药材产地溯源提供新的方法参考。

关键词：近红外光谱；白芷；产地溯源；数据增强；卷积神经网络

doi：10.11669/cpj.2024.21.005　　**中图分类号：**R284　　**文献标志码：**A　　**文章编号：**1001－2494（2024）21－2022－08

Based on Near Infrared Spectroscopy Combined with Data Enhancement CNN Algorithm Origin Traceability Method of *Angelica Dahurica*

GUO Zhaohua[1], WEN Shizhao[2], LI Sifan[3], WANG Qi[3], WANG Yingxin[4], WANG Xinguo[4], NIU Liying[4], LI Yawei[5*], FENG Wei[4*] (1. *China Electronics Technology Group Corporation Network Communication Research Institute*, *Shijiazhuang* 050050, *China*; 2. *School of Statistics and Data Science*, *Nankai University*, *Tianjin* 300192, *China*; 3. *Northeastern University*, *Shenyang* 110167, *China*; 4. *Quality Evaluation & Standardization Hebei Province Engineering Research Center of Traditional Chinese Medicine*, *School of Pharmaceutical Sciences*, *Hebei University of Chinese Medicine*, *Shijiazhuang* 050091, *China*; 5. *Liaoning Academy of Analytical Sciences*, *Liaoning Inspection*, *Examination and Certification Centrer*, *Shenyang* 110032, *China*)

ABSTRACT: OBJECTIVE To establish an origin classification model of *Angelica dahurica* with unbalanced sample size based on near-infrared spectroscopy combined with data-enhanced convolutional neural network (CNN) algorithm. **METHODS** In this study, 95 samples of *Angelica dahurica* were collected, and near-infrared spectroscopy was performed on different samples within the wavelength range of 12 500 to 4 000 cm^{-1}. The near-infrared spectroscopy dataset of *Angelica dahurica* used in this study faces issues such as small sample size and uneven distribution of sample origins. To enhance the generalizability of the model, three data augmentation algorithms were proposed, including spectral shifting, spectral noise addition, and spectral combination. Additionally, to address the problem of sample imbalance, Focal Loss was used as the loss function for training the CNN model. **RESULTS** The three data enhancement algorithms were applied to the SVM model. Adding Gaussian noise with a signal-to-noise ratio of 20 to the spectral data had the best effect, which could increase the accuracy of the model to 84.2%. Aiming at the problem of sample imbalance, Focal Loss is used as the loss function to train the CNN model, and the accuracy rate can reach 94.7%. **CONCLUSION** The infrared spectroscopy combined with data-enhanced CNN algorithm provides a rapid and non-destructive detection method and reliable data analysis method for the origin traceability of Radix Angelicae Dahuricae, and provides a new method reference for the origin traceability of Chinese medicinal materials.

KEY WORDS: near infrared spectroscopy; *Angelica dahurica*; origin traceability; data enhancement; convolutional neural network

---

基金项目：河北省省级科技计划项目资助（21372503D）；大学生创新创业训练计划项目资助（202414432007）

作者简介：郭兆华，男，硕士，高级工程师　　研究方向：算法设计与模型分析　　***通讯作者：**冯薇，女，博士，教授，博士生导师　研究方向：中药材品质评价　　Tel：（0311）85216828；李亚薇，女，学士，副研究员　　研究方向：工业与信息技术　Tel：（024）24822089

白芷为伞形科植物白芷[*Angelica dahurica* (Fisch. ex Hoffm.) Benth. et Hook. f.]或杭白芷[*Angelica dahurica* (Fisch. ex Hoffm.) Benth. et Hook. f. var. *formosana* (Boiss.) Shan et Yuan]的干燥根[1],具有解表散寒,祛风止痛,宣通鼻窍,燥湿止带,消肿排脓的功效,为我国四十种常用大宗中药材之一,药用历史悠久[2-3]。白芷化学成分复杂,主要包括香豆素类、挥发油类、多糖类、氨基酸类、生物碱类、苷类等化学成分,现代药理研究表明,其具有镇痛、抗炎、舒张血管、抗肿瘤、抗过敏、美白、抑菌等作用[4-6]。据本草文献记载,宋代以前,白芷药材主要来源于山西、吴地(江浙皖赣交界一带)的野生品,宋代以后浙江白芷从野生转为栽培品,于明代形成人工栽培主产区,之后被四川等地引种栽培,中华人民共和国成立后又有河北安国这一新兴产区。由于城市建设,杭州市已无栽种的白芷,川白芷和杭白芷的基原植物相同,但因气候不同性状有所变化。目前,市场上主流白芷药材主要为川白芷、禹白芷、祁白芷、亳白芷,由于各产区地理环境、气候条件、栽培方式、加工采收的方式时间等差异,造成不同产地的药材质量相差较大[7-8],故白芷的产地鉴别对于白芷药材在临床选用具有重要意义。但加工成饮片及粉末的白芷较难通过外观性状进行产地归属。

传统性状鉴别法受检验者主观因素的影响,导致对药材气味、色泽的鉴别差异较大。目前可用于产地鉴别的技术中,气相色谱法、高效液相色谱法等存在耗时、样本前处理复杂、检测对样本有损、检测成本高等缺点。矿物元素指纹图谱技术具有高选择性和灵敏度等优点,可以在宽线性范围内测定多个微量元素,但操作过程复杂繁琐,对环境和人员的要求较高,造成了实际应用中的局限性[9]。核磁共振技术复杂,调试参数较多,对专业操作人员要求高且设备昂贵,限制了其在行业中的应用[10]。近红外光谱技术近年来发展迅速,且广泛应用于各类中药材的产地溯源与质量监控领域中,如山药[11]、茯苓[12]、贝母[13]、天麻[14]等药材。近红外光谱分析技术结合了光谱学、计算机技术与化学计量学方法,具有操作简便、无损检测、分析成本低、分析速度快、样本一般无需预处理、不用化学试剂、不污染环境等优点[15-16]。其能够对中药材从生产运输、采收加工和临床应用的整个过程中进行快速简易的分析,此技术尤其适用于开发对市售药材准确、快速、低成本产地溯源的方法。Liu 等[17]尝试采集自河南、河北、四川、浙江 4 个产地的白芷样本进行近红外漫反射光谱测量,建立了白芷样品产地判别模型。近年来以卷积神经网络(convolutional neural networks,CNN)为代表的深度学习算法开始用于近红外光谱定性分析模型的建立[18-20]。与传统机器学习方法相比,卷积神经网络可以减少光谱分析中对先验知识的依赖,从原始光谱数据中分层次提取微观特征和宏观特征,提高模型预测精度的同时减少建模的工作量。

不同产地的气候及土壤条件等均是药材有效成分积累的重要影响因素,本实验将近红外光谱和数据增强 CNN 算法结合,建立白芷产地溯源模型的快速无损鉴别方法,不仅在保证白芷品质及疗效、保护消费者权益等方面具有重要的意义[21],对其他中药的产地溯源方法研究也具有借鉴意义。

1 仪器与材料

1.1 实验仪器

MPA 型傅里叶变换近红外光谱仪(德国布鲁克光学仪器公司),配备 OPUS 光谱采集软件,测量方式选用固体积分球漫反射方式;YB-150 型多功能粉碎机(永康市速锋工贸有限公司);DHG-9123A 型电热恒温鼓风干燥箱(上海一恒科技有限公司)。

1.2 样品

所用药材从不同产地(安徽亳州、河南禹州、四川遂宁、河北安国)采集,经河北中医药大学侯芳洁副教授鉴定为伞形科植物白芷[*Angelica dahurica* (Fisch. ex Hoffm.) Benth. et Hook. f. 或杭白芷[*Angelica dahurica* (Fisch. ex Hoffm.) Benth. et Hook. f. var. *formosana* (Boiss.) Shan et Yuan]的干燥根,粉碎后过 80 目筛。根据白芷的产地将 95 个样本分为 4 类,分别为安徽亳州 8 份(编号:S1 ~ S4、S27 ~ S29、S71)、河南禹州 9 份(编号:S5 ~ S12、S64)、四川遂宁 23 份(编号:S13 ~ S26、S62 ~ S63、S72 ~ S78)、河北安国 55 份(编号:S30 ~ S61、S65 ~ S70、S79 ~ S95),采集时间为 2021 年 6 ~ 10 月。

2 方 法

2.1 NIRS 信息采集

取适量的白芷药材粉末,放入石英样品杯中至三分之二处,均匀铺平,45 ℃烘干至恒重。采用近红外光谱仪进行数据采集,在室温条件下将仪器预热 30 min,以空气为参比扣除背景采集光谱图,采用积分球漫反射采集光谱,扫描条件为:分辨率

8 cm^{-1}，扫描波段范围：12 500 ~ 4 000 cm^{-1}，样品背景和样品扫描时间：32 s，每次扫描前保持样品处于夯实、均匀、平整的状态，每批样品重复扫描 3 次，取平均值作为样品近红外光谱。

白芷近红外光谱数据集共有 95 个样本，共 2 203个波长点，见图 1。

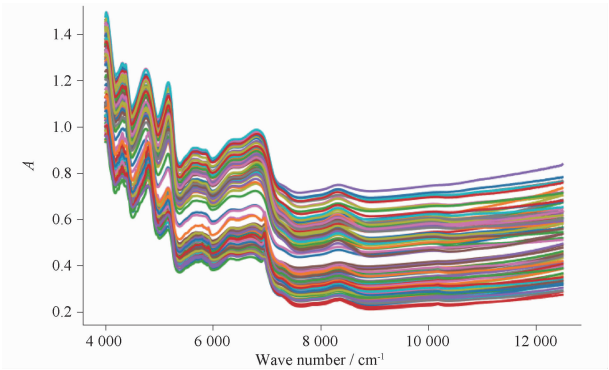


图 1 白芷近红外光谱数据采集图像

Fig. 1 Image of near infrared spectral data set of *Angelica dahurica*

2.2 样本集划分

本实验中采用 SPXY 算法[22-23]进行数据采样，SPXY 算法不仅考虑了样本之间的距离关系，还考虑了样本所属的类别信息。具体而言，SPXY 算法首先计算每个样本与其他样本的距离，然后根据距离大小和样本所属的类别信息，选取一定数量的样本。这种样本选择策略可以保证采样结果具有更好的分类特性。在 SPXY 算法中，所使用的距离度见公式 1 ~ 3。

$$d_x(P,Q)=\sqrt{\sum_{j=1}^{m}(x_{p,j}-x_{q,j})^2} \qquad \text{公式(1)}$$

$$d_y(P,Q)=\sqrt{(y_p-y_q)^2}=|y_p-y_q| \qquad \text{公式(2)}$$

$$d_{SPXY}(P,Q)=\frac{d_x(P,Q)}{\max[d_x(P,Q)]}+\frac{d_y(P,Q)}{\max[d_y(P,Q)]} \qquad \text{公式(3)}$$

其中，$d_x(P,Q)$衡量的是两个样本 P，Q 在特征空间 X 的欧氏距离，$x_{p,j}$，$x_{q,j}$分别是样本 P，Q 在第 j 维上的特征值；$d_y(P,Q)$衡量的是两个样本 P，Q 的绝对差距离；$P,Q\in[1,n]$，$P\neq Q$，划分结果见表 1。

表 1 白芷近红外光谱数据集划分结果

Tab. 1 Division results of near infrared spectral data set of *Angelica dahurica*

Dataset indicators	Value	Dataset indicators	Value
Total number of samples	95	Category distribution	[8,9,23,55]
Proportion of test set	0.2	Training set sample distribution	[5,7,18,46]
Number of samples in training set	76	Test set sample distribution	[3,2,5,9]
Number of samples in test set	19	Sample dimension	2 203

2.3 光谱预处理

光谱数据预处理是近红外光谱分析中的一项重要步骤，其主要作用是消除干扰噪声、背景扰动和仪器波动等因素对光谱数据的影响，提高数据质量和可靠性，从而增强模型的预测能力和稳定性[24]。

在介绍具体算法之前，首先对原始光谱信息矩阵进行定义，对于某一样品，经过红外光谱仪扫描后获取的数据为 $\vec{x}=[x_1,x_2,\ldots,x_m]_{1\times m}$，其中 m 表示仪器扫描时的波长点数。对于 n 个样品，可以构建光谱信息矩阵 X 见公式 4。

$$X=[\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_i,\ldots,\vec{x}_n]^T=\begin{bmatrix}x_{1,1}&x_{1,2}&\cdots&x_{1,m}\\x_{2,1}&x_{2,2}&\cdots&x_{2,m}\\\vdots&\vdots&\ddots&\vdots\\x_{n,1}&x_{n,2}&\cdots&x_{n,m}\end{bmatrix} \qquad \text{公式(4)}$$

其中，$x_{i,k}$表示第 i 个样品的第 k 个波长处的吸光度，$i=1,2,\cdots,n$；$k=1,2,\cdots,m$。

标准正态变量变换（standard normal variate transformation，SNV）被广泛用于消除固体表面散射、样本颗粒大小以及光程变化对红外光谱数据的影响[25]。相对于前文的标准化方法，SNV 的不同之处在于它基于光谱矩阵的行来对某一条光谱进行预处理，计算方式见公式(5)。

$$\mathrm{SNV}(\vec{x}_i)=\frac{\vec{x}_i-\bar{x}_i}{\sqrt{\frac{\sum_{k=1}^{m}(x_{i,k}-\bar{x}_i)^2}{m-1}}} \qquad \text{公式(5)}$$

其中，$\bar{x}_i=\frac{\sum_{k=1}^{m}x_{i,k}}{m}$，$m$ 为波长点数，$x_{i,k}$表示第 i 个样品的第 k 个波长处的吸光度。

2.4 数据增强算法

为增加数据的多样性，提高模型的鲁棒性和泛化能力，本实验提出 3 种针对于近红外光谱的数据增强算法[26]，以改善近红外光谱数据集普遍存在的样本量较小的问题。

2.4.1 光谱平移　光谱平移是一种常见的光谱数

据增强算法，其基本思想是将原始光谱数据沿着波长轴平移一定距离，生成新的数据。通过平移，可以产生一系列新的光谱数据，这些数据具有与原始数据相同的特征和分布，但在数据分布上略有变化。在本实验中，光谱数据水平平移的方向和平移距离 d 是可以提前设定的超参数。

对于采取的白芷样品中某一样本的光谱数据，$\vec{x}_i=[x_{i,1},x_{i,2},\cdots,x_{i,m}]_{1\times m}$，假设向左平移 2 个波长点，则最终其对应的增强样本光谱数据如下为 $\vec{x}_i^{\ \Delta}=[x_{i,3},x_{i,4},\cdots,x_{i,m-1},x_{i,m},x_{i,m}]$。

2.4.2 光谱增噪 光谱增加噪声是一种常见的光谱数据增强算法，其基本思想是向原始光谱数据中添加一定强度和类型的噪声，生成新的数据，这些数据具有与原始数据不同的特征和分布，可以使模型更好地学习到数据的特征和规律。本实验通过添加指定信噪比的高斯噪声进行数据集扩充，以下详述该方法的实施步骤：

输入：光谱向量$\vec{x}_i$，信噪比（SNR）；

步骤 1：计算原始光谱向量 $\vec{x}_i$ 的功率 $P_{\vec{x}_i}=\frac{1}{m}\sum_{j=0}^{m-1}|x_{i,j}|^2$，其中，$x_{i,j}$表示光谱向量 $\vec{x}_i$ 的第 j 个元素；

步骤 2：计算信噪比 $\gamma=10^{\frac{\mathrm{SNR}}{10}}$；

步骤 3：计算噪声功率 $P_{\vec{n}_i}=\frac{P_{\vec{x}_i}}{\gamma}$；

步骤 4：生成噪声 $\vec{n}_i\sim\mathrm{N}(0,P_{\vec{n}_i})$，$\mathrm{N}(0,P_{\vec{n}_i})$是均值为 0，方差为 $P_{\vec{n}_i}$的高斯分布；

步骤 5：生成加噪声的光谱向量 $\vec{x}_i^{\ \Delta}=\vec{x}_i+\vec{n}_i$；

输出：增噪光谱向量 $\vec{x}_i^{\ \Delta}$。

SNR 的大小决定了噪声功率和信号功率之间的比例关系，因此可以影响噪声的大小。如果 SNR 越低，噪声的功率就会越高，这会使得噪声信号更强，对信号的影响也更大。反之，如果 SNR 越高，噪声功率就会越低，噪声信号也会更弱，对信号的影响就会更小。

2.4.3 光谱组合 对于同一类别的光谱数据，进行线性组合也可以作为一种光谱数据增强的算法。线性组合可以将多个相似的光谱数据合并成一个新的光谱数据，扩充数据集，提高模型的性能和泛化能力。线性组合的优点是可以保持数据的特征和分布，不会产生新的噪声和误差。

在本实验中，具体实现方式如下。假设 $\vec{x}_{(1)}$，$\vec{x}_{(2)}$，$\cdots\vec{x}_{(s)}$为某一类样品的所有样本光谱信息，则生成的增强样本为 $\vec{x}^{\Delta}=\sum_{j=1}^{S}\varepsilon_j\cdot\vec{x}_{(j)}$，其中 ε_j表示线性组合系数，在实验中随机生成，以确保生成样本的多样性，其中 $\sum_{j=1}^{S}\varepsilon_j=1$。

2.5 Focal Loss 损失函数

Focal Loss（FL）是一种用于解决类别不平衡和难样本学习问题的损失函数[27]。FL 通过对易分类样本的减弱权重，更加关注困难样本，使得模型更加专注于困难样本的学习，从而提高了模型在类别不平衡数据集上的性能。见公式 6。

$$\mathrm{FL}(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \qquad \text{公式(6)}$$

其中，α_t表示样本的类别权重，用于解决类别不平衡问题，p_t为模型预测样本属于正类的概率，γ 为可调的超参数，控制易分类样本的惩罚程度。在本实验中取 γ 为 2。

该公式中$(1-p_t)^{\gamma}$项表示减弱易分类样本的权重，因为易分类样本的预测概率 p_t会趋近于 1，从而使得该项的值趋近于 0，使得易分类样本的损失函数值权重更小。而对于困难样本，由于其预测概率 p_t相对较低，因此$(1-p_t)^{\gamma}$的值更大，从而使得困难样本的损失函数值权重更大，更加关注困难样本的学习。

2.6 卷积神经网络

一维卷积神经网络（1D Convolutional Neural Network，1D-CNN）是一种针对序列数据进行特征提取和分类的神经网络模型[28-30]。1D-CNN 可以看作是对序列数据进行卷积运算和池化运算的组合。

在 1D-CNN 中，输入数据集合被视为一维向量，其每个元素代表一个波长点的吸光度。与标准卷积神经网络不同，1D-CNN 只在一个方向的滑动窗口中对连续的近红外光谱数据进行卷积运算。

1D-CNN 的典型组成部分包括卷积层、激活函数层、池化层、全连接层和输出层。在卷积层中，通过对输入光谱数据的每个波长点进行滑动卷积运算，提取出数据中连续的模式。激活函数层增强网络的非线性表达能力。池化层通过选择最大值或平均值来减小卷积层输出的数据集大小。全连接层通过连接提取的特征向量和输出层来决定输入数据的类别。输出层通常使用 softmax 函数来计算各个类别的概率，并得出最终的分类结果。

本实验所建立的 1D-CNN 参数见表 2，网络结构见图 2。

2.7 模型建立

首先将扫描获得的白芷近红外光谱数据集进行 SNV 预处理，再按 SPXY 方法进行划分，将训练集

表 2 一维卷积神经网络结构(1D-CNN)网络结构

Tab. 2 One-dimensional convolutional neural network structure

Number of layers	Structure
1	1D convolutional kernel(1,16,21)
2	Batch normalization
3	Relu activation function
4	1D convolutional kernel(16,32,19)
5	Batch normalization
6	Relu activation function
7	1D convolutional kernel(21,64,17)
8	Batch normalization
9	Relu activation function
10	Fully connected(137 536,512)
11	Fully connected(512,4)
12	Softmax

注:一维卷积核(1,16,21)表示输入维度为1,输出维度16,卷积核大小为21,其他参数如padding取为0,步长为1,其余卷积核格式相同。

Note:1D convolutional kernel of (1,16,21) indicates an input dimension of 1, an output dimension of 16, and a kernel size of 21. Other parameters such as padding are set to 0, and stride is set to 1. Other convolutional kernels follow the same format.

作为卷积神经网络的输入,以 FL 作为损失函数训练网络,训练 200 轮,优化器使用 Adam 方法,初始学习率 0.001,学习率衰减率 0.000 1。训练完成后,将测试集作为卷积神经网络模型的输入,预测样品的产地类别,并与实际产地类别进行对比,分析模型的性能优劣。

2.8 评价指标

混淆矩阵(confusion matrix,CM)是分类任务中常用的评估模型性能的工具[31]。CM 将模型对数据集中每个样本的分类结果进行汇总,在多分类任务中,CM 的表现形式是一个 N × N 的矩阵,其中 N 表示分类的类别数目。CM 中的每一个元素 $C_{i,j}$表示实际标签为第 i 类的样本被预测为第 j 类的样本数目。公式 7 是多分类任务的 CM 示例:

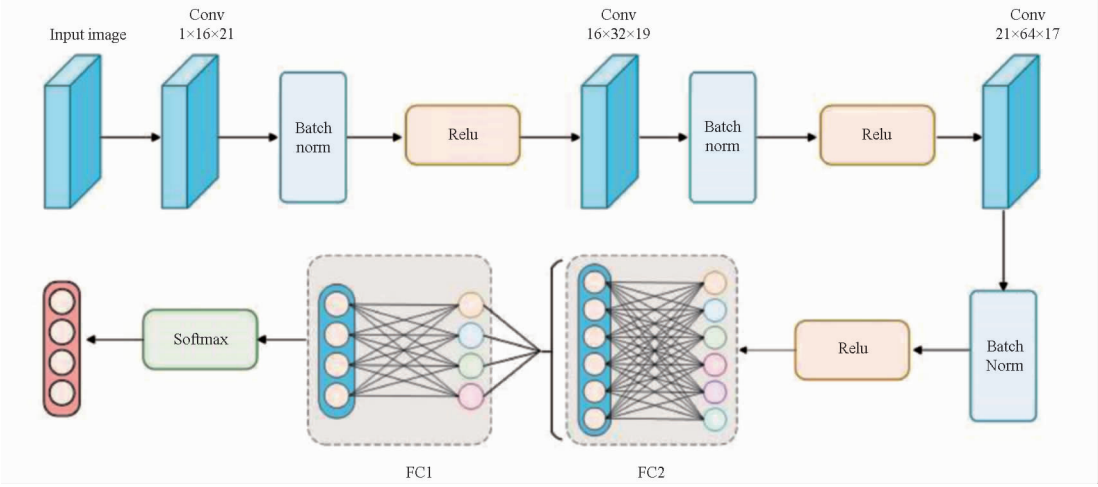


图 2 1D-CNN 网络结构示意图

Fig. 2 1D-CNN network structure diagram

$$\begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,N} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ C_{N,1} & C_{N,2} & \cdots & C_{N,N} \end{bmatrix}_{N\times N} \quad \text{公式(7)}$$

在 CM 的基础上,本实验以总分类正确率 OA、F_1值和 Kappa 系数[32]作为模型性能评价指标。其中 Kappa 系数 κ 相对于总体正确率 OA 而言更加准确地反映了模型的分类性能,其越接近于 1,说明分类器预测结果与真实情况的一致性越好。

3 结果与讨论

对白芷数据集进行数据增强后的样本产地类别分布信息见表3。

表 3 数据增强后白芷产地分类数据集信息

Tab. 3 Data set information of *Angelica dahurica* origin classification after data enhancement

Dataset indicators	Value	Dataset indicators	Value
Total number of samples	171	Category distribution	[13,16,41,101]
Number of samples in training set	152	Training set sample distribution	[10,14,36,92]
Number of samples in test set	19	Test set sample distribution	[3,2,5,9]

首先验证数据增强算法的有效性，以常见的支持向量机（support vector machine，SVM）算法建立分类器，其余预处理操作和样本划分与“2.7”项下描述保持一致，SVM 核函数采用高斯核函数，惩罚参数设置为 30，实验结果见表 4、图 3。2～5 号实验进行了光谱平移，6～8 号实验进行了光谱增噪，9 号实验进行了光谱数据的线性组合，不同的数据增强算法相对于对照组（1 号）而言，在总正确率 OA 和 Kappa 系数均有不同幅度的提升，其中对光谱数据添加信噪比为 20 的高斯噪声效果最好，将模型正确率提高至 84.2%，同时 Kappa 系数也达到 0.776。相比较而言，对光谱数据进行线性组合的方式对于模型提升的性能并不大，可能是由于简单的线性组合产生了部分偏离类别群体的样本，导致模型准确率下降。

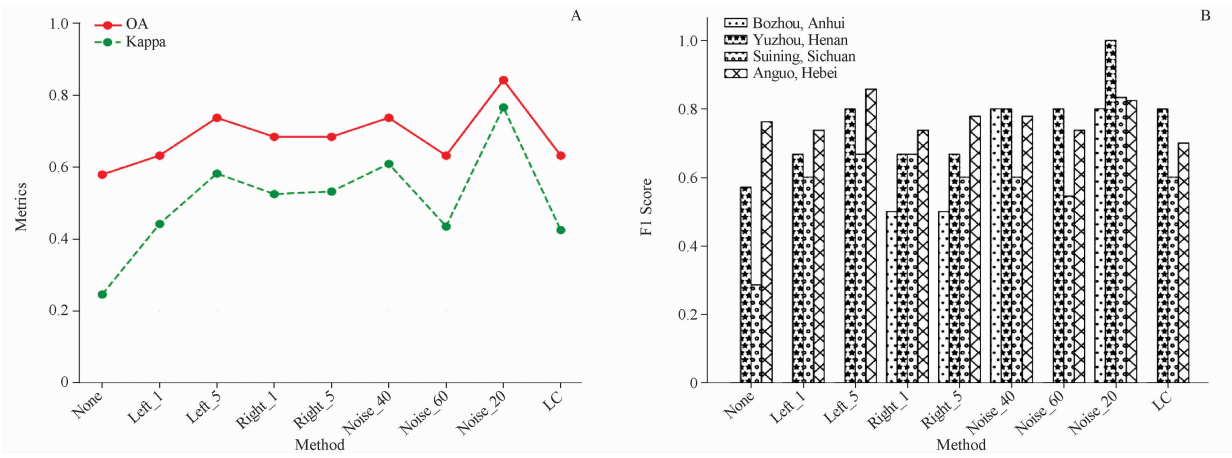
表 4 基于样本增强算法进行白芷产地分类实验结果

Tab. 4 Experimental results of *Angelica dahurica* origin classification based on sample enhancement algorithm

Serial number	Classifier	Sample augmentation algorithm	OA[4)]	κ[4)]	$F_{1,1}$[4)]	$F_{1,2}$[4)]	$F_{1,3}$[4)]	$F_{1,4}$[4)]
1	SVM	None	0.579	0.348	Nan[5)]	0.571	0.286	0.762
2	SVM	Left_1[1)]	0.632	0.442	Nan[5)]	0.667	0.600	0.737
3	SVM	Left_5	0.737	0.582	Nan[5)]	0.800	0.667	0.857
4	SVM	Right_1	0.684	0.525	0.500	0.667	0.667	0.737
5	SVM	Right_5[1)]	0.684	0.532	0.500	0.667	0.600	0.778
6	SVM	Noise_40[2)]	0.737	0.609	0.800	0.800	0.600	0.778
7	SVM	Noise_60	0.632	0.435	Nan[5)]	0.800	0.545	0.737
8	SVM	Noise_20	0.842	0.766	0.800	1.000	0.833	0.824
9	SVM	Linear combination[3)]	0.632	0.425	Nan[5)]	0.800	0.600	0.700

注：[1)]Left_1 表示将光谱向量向左平移 1 个波长点，Right_5 表示将光谱向量向右平移 5 个波长点；[2)]Noise_40 表示向光谱向量中添加信噪比为 SNR = 40 的高斯噪声；[3)]Linear combination 对应“2.2.3”项下提到的同类别光谱向量线性组合的方法；[4)]首行中，OA 表示总体分类正确率，κ 表示 Kappa 系数，$F_{1,i}$（i = 1，2，3，4）分别表示安徽亳州、河南禹州、四川遂宁、河北安国 4 个产地类别的 F_1 值；[5)]Nan 表示由于某一类别的精确率与召回率均为 0 而导致 F_1 值无法计算。

Note: [1)]Left_1 indicates shifting the spectral vector one wavelength point to the left, while Right_5 means shifting the spectral vector five wavelength points to the right; [2)]Noise_40 refers to adding Gaussian noise with a signal-to-noise ratio of 40 to the spectral vector; [3)]Linear combination corresponds to the method of linear combination of spectral vectors from the same category mentioned in section “2.2.3”; [4)]In the first line, OA represents the overall classification accuracy, Kappa coefficient is a statistical measure of prediction accuracy, representing the values for the four geographical origins: Bozhou in Anhui, Yuzhou in Henan, Suining in Sichuan, and Anguo in Hebei; [5)]Nan indicates that the value cannot be calculated due to both precision and recall being zero for a particular category.



A－正确率和 Kappa 系数对比；B－F1 系数对比。

A－comparison of accuracy and Kappa coefficient; B－comparison of F1 coefficients.

图 3 支持向量机（SVM）模型-不同数据增强方法的分类结果对比

Fig. 3 SVM model-comparison of classification results of different data enhancement methods

通过将不同的数据增强算法应用于 SVM 模型中，验证了数据增强算法的有效性。观察表 5 发现，虽然整体正确率有所提升，但是在部分组别实验中，对于少样本类别的第一类和第二类白芷的学习效果并不理想，这是由于样本不平衡所导致模型更倾向于将未知样本预测为多样本类别。因此，下文使用 Focal Loss 作为损失函数训练 CNN 模型，使模型更注重于学习少样本的困难类别样本。首先通过 SPXY 样本划分法将 SNV 预处理后的样本划分为训练集与测试集，其次通过“2.4”项下提出的数据增

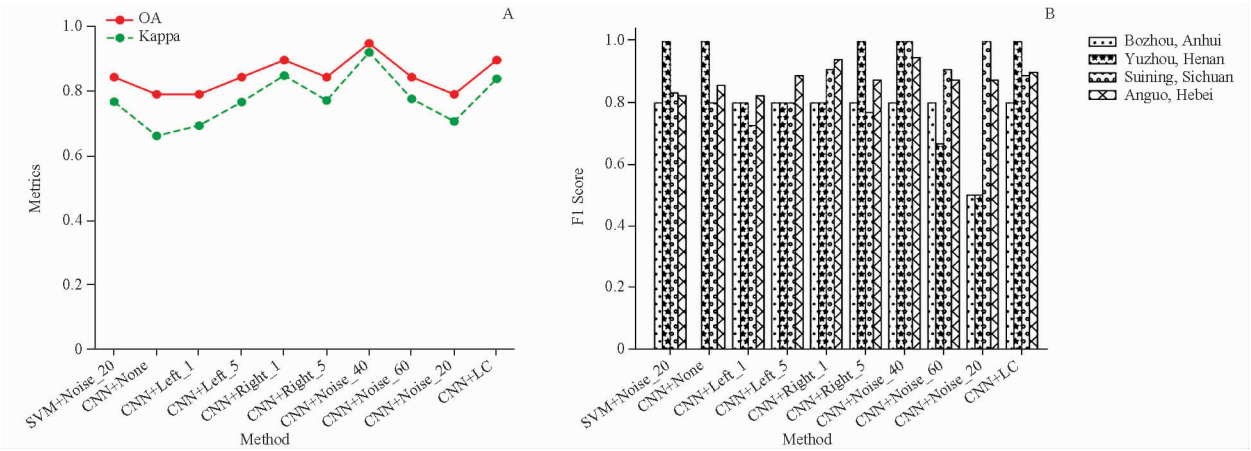
强算法向训练集光谱进行样本增强，然后利用增强训练集光谱建立以 Focal Loss 作为损失函数的一维卷积神经网络模型，参数见“2.6”和“2.7”，接着将样本中的测试集作为 1D-CNN 的输入，从而预测出样品的产地类别，最后与实际产地类别对比，分析 CNN 模型的优劣。

通过表 5 和图 4 可以发现，传统的 1D-CNN 模型（2 号）不使用数据增强算法，分类正确率为 78.9%，Kappa 系数为 0.66。3～6 号实验进行了光谱平移，7～9 号实验进行了光谱增噪，10 号实验进行了光谱数据的线性组合，使用不同数据增强后的白芷近红外光谱数据集的 1D-CNN 模型（3～10 号）与传统的 1D-CNN 模型（2 号）相比，分类正确率与 Kappa 系数均有不同幅度的提升。利用 Focal Loss 作为损失函数去训练 CNN 模型的策略能在保证高准确率和 Kappa 系数的前提下，显著提高第一类和第二类白芷样本的 F_1 值，这是因为 Focal Loss 的加入使得模型更加注重于学习少样本的特征，从而更进一步地提升总体的正确率。其中，向光谱数据中添加信噪比为 40 的方法得到的模型正确率为 94.7%，Kappa 系数也说明分类器预测结果的一致性非常好。

表 5 基于样本增强算法和 Focal Loss 进行白芷产地分类实验结果

Tab. 5 Experimental results of *Angelica dahurica* origin classification based on sample enhancement algorithm and Focal Loss

Serial number	Classifier	Sample augmentation algorithm none	OA	κ	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$
1	SVM	Noise_20	0.842	0.766	0.800	1.000	0.833	0.824
2	CNN	None	0.789	0.660	Nan	1.000	0.800	0.857
3	CNN	Left_1	0.789	0.692	0.800	0.800	0.727	0.824
4	CNN	Left_5	0.842	0.765	0.800	0.800	0.800	0.889
5	CNN	Right_1	0.895	0.847	0.800	0.800	0.909	0.941
6	CNN	Right_5	0.842	0.770	0.800	1.000	0.769	0.875
7	CNN	Noise_40	0.947	0.919	0.800	1.000	1.000	0.947
8	CNN	Noise_60	0.842	0.775	0.800	0.667	0.909	0.875
9	CNN	Noise_20	0.789	0.705	0.500	0.500	1.000	0.875
10	CNN	Linear Combination	0.895	0.837	0.800	1.000	0.889	0.900



A－正确率和 Kappa 系数对比；B－F1 系数对比。

A－comparison of accuracy and Kappa coefficient; B－comparison of F1 coefficients.

图 4 CNN 模型-不同数据增强方法的分类结果对比

Fig. 4 CNN model-comparison of classification results of different data enhancement methods

4 结 论

本实验围绕白芷数据集的样本量小且样本分布不均衡两个问题，提出 3 种数据增强算法，应用于 SVM 模型，模型性能有一定提高。在使用数据增强算法的基础上，又使用 Focal Loss 作为损失函数训练 CNN 模型，发现不仅能够提升模型的正确率和 Kappa 系数，同时对于少样本的学习也更加充分有效，显著提高了前两类样本的准确率与召回率。本实验提出的模型不仅能够提高白芷产地分类的准确性，还特别优化了对小样本、分布不均

衡数据的处理能力，从而能够在实际生产和应用中迅速、精准地进行白芷产地的溯源。目前使用的数据集可能在地理和环境条件上存在局限，未来需要扩展数据集的多样性，以确保模型的泛化能力。还可以探索光谱数据以外的其他信息，如气候、土壤和种植技术数据，以增强产地溯源的准确度和可靠性。

REFERENCES

[1] *Ch. P*(2020) Vol Ⅰ(中国药典 2020 年版. 一部)[S]. 2020:109.

[2] ZHU Y X, LI B L, MA H S, *et al*. Research progress of *Angelica dahurica* for extracting effective components, pharmacological action and clinical application [J]. *China Med Her*(中国医药导报), 2014, 11(31): 159-162,166.

[3] ZHANG J, DENG R, FAN G, *et al*. Quality control and discrimination of angelicae dahurica radix by RRLC fingerprints combined with chemomtrics method [J]. *Chin Pharm J*(中国药学杂志), 2011, 46(6): 418-421.

[4] JI Q, MA Y H, ZHANG Y. Research progress on chemical constituents and pharmacological effects of Angelicae Dahuricae Radix [J]. *Food Drug*(食品与药品), 2020, 22(6): 509-514.

[5] ZOU J Y, SU W, PAN Y, *et al*. Chemical components and pharmacological action for *Angelica dahurica* sinensis and predictive analysis on its Q-marker [J]. *World Sci Technol Mod Tradit Chin Med*(世界科学技术-中医药现代化), 2023, 25(7): 2535-2548.

[6] WANG R, LIU J, YANG D Y, *et al*. Research progress in chemical constituents and pharmacological action of *Angelica dahurica* [J]. *Chin J Inf Tradit Chin Med*(中国中医药信息), 2020, 37(2): 123-128.

[7] ZHANG S Y, WANG J, SHENG Y C, *et al*. Comparative study on the volatile constituents of Baizhi from six different producing regions [J]. *Storage Process*(保鲜与加工), 2019, 19(4): 176-183.

[8] ZHOU B, LIU P, CHEN J, *et al*. Analysis and evaluation of chemical composition of coumarins and polysaccharides in Angelica Dahuricae Radix from different areas [J]. *J Nanjing Univ Tradit Chin Med*(南京中医药大学学报), 2015, 31(1): 68-73.

[9] HU H W, ZHU L, CHEN S Y, *et al*. Research progress on traceability technologies of *Zanthoxylum bungeanum* maxim. origins [J]. *J Food Saf Qual*(食品安全质量检测学报), 2023, 14(13): 110-116.

[10] AN L, WANG H, MA J W, *et al*. Analysis of differences in chemical composition of different strawberry(Fragaria × ananassa Duch.) cultivars based on nuclear magnetic resonance metabolomics [J]. *J Food Saf Qual*(食品安全质量检测学报), 2020, 11(14): 4750-4756.

[11] LI C B, NIU C W, SU L, *et al*. Identification and variance analysis of chinese yam from different origins by near infrared spectroscopy [J]. *Food Res Dev*(食品研究与开发), 2022, 43(15): 175-181.

[12] LI J Y, YU M, ZHENG Y, *et al*. Nondestructive identification of *Poria cocos* blocks from different origins based on near infrared spectroscopy [J]. *Chin J Anal Lab*(分析试验室), 2021, 40(12): 1381-1386.

[13] ZHOU T, FU S B, XIE H M, *et al*. Identification and validation of bulbs of Fritillariae species using near-infrared spectroscopy data [J]. *West China J Pharm Sci*(华西药学杂志), 2021, 36(2): 193-197.

[14] BAI Q X, HOU Y, YANG P P, *et al*. Identification method of the producetion site of *Gastrodia elata* blume based on near infrared spectroscopy [J]. *J West China For Sci*(西部林业科学), 2021, 50(3): 124-130.

[15] XIE H J, GAN Y, CHEN Q H, *et al*. Application of near infrared spectroscopy analysis technology in the field of preparation [J]. *Chin Pharm J*(中国药学杂志), 2009, 44(2): 87-91.

[16] HAN S H, GUO Y S, LI X, *et al*. Determination of ethanol in base liquor of Baijiu based on near infrared spectroscopy technology [J]. *China Brew*(中国酿造), 2018, 37(9): 158-161.

[17] LIU M H, ZHANG X G, ZHOU Q, *et al*. Determination of geographical origins of Chinese medical herbs by nir and pattern recognition [J]. *Spectrosc Spect Anal*(光谱学与光谱分析), 2006, 26(4): 629-632.

[18] ZHENG Z S, LIU B, LU P, *et al*. Spectral classification and characteristic spectral analysis of nearshore aquatic plants based on AlexNet [J]. *Chin J Lasers*(中国激光), 2023, 50(2): 129-138.

[19] YUAN W, JIANG H, SUN M, *et al*. Geographical origin identification of Chinese tomatoes using long-wave Fourier-transform near-infrared spectroscopy combined with deep learning methods [J]. *Food Anal Method*, 2023: 1-13.

[20] ZHANG X L. Convolutional neural network-based spectral analysis and its application in quality evaluation of agro-products [D]. Hangzhou: Zhejiang University, 2021.

[21] LI J H, ZHANG P J, REN S L, *et al*. Classification and origin identification of Chinese medicinal materials based on infrared spectroscopy analysis [J]. *Digit Technol Appl*(数字技术与应用), 2022, 40(6): 53-55.

[22] GALVAO R K H, ARAUJO M C U, JOSÉ G E, *et al*. A method for calibration and validation subset partitioning [J]. *Talanta*, 2005, 67(4): 736-740.

[23] ZHAN X R, ZHU X R, SHI X Y, *et al*. Determination of hesperidin in tangerine leaf by near-infrared spectroscopy with SPXY algorithm for sample subset partitioning and Monte Carlo cross validation [J]. *Spectrosc Spect Anal*(光谱学与光谱分析), 2009, 29(4): 964-968.

[24] RINNAN Å. Pre-processing in vibrational spectroscopy-when, why and how [J]. *Anal Method*, 2014, 6(18): 7124-7129.

[25] BARNES R J, DHANOA M S, LISTER S J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra [J]. *Appl Spectrosc*, 1989, 43(5): 772-777.

[26] LIU J, OSADCHY M, ASHTON L, *et al*. Deep convolutional neural networks for Raman spectrum recognition: a unified solution [J]. *Analyst*, 2017, 142(21): 4067-4074.

[27] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal loss for dense object detection [J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, 42(2): 318-327.

[28] XU L, ZHU D, CHEN X, *et al*. Combination of one-dimensional convolutional neural network and negative correlation learning on spectral calibration [J]. *Chemom Intell Lab Sys*, 2020, 199: 103954.

[29] WANG B, DENG J, JIANG H. Markov transition field combined with convolutional neural network improved the predictive performance of near-infrared spectroscopy models for determination of aflatoxin B1 in maize [J]. *Foods*, 2022, 11(15): 2210.

[30] MA D, SHANG L, TANG J, *et al*. Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network [J]. *Spectrochim Acta Part A: Mol Biomol Spectrosc*, 2021, 256: 119732.

[31] SUN D W. *Infrared Spectroscopy for Food Quality Analysis and Control* [M]. Cambridge: Academic press, 2009.

[32] KRAEMER H C. Extension of the kappa coefficient [J]. *Biometrics*, 1980, 36(2): 207-216.

(收稿日期:2024-01-20)