DOI: 10.20103/j.stxb.202404240918

陈冬英,翁伟雄,陈培亮,魏建崇.基于一维卷积神经网络与自编码算法的松属物种鉴别机制.生态学报,2025,45(5):2401-2411.

Chen D Y, Weng W X, Chen P L, Wei J C.Identification mechanism of *Pinus* species based on one dimensional convolutional neural network model and self-coding algorithm.Acta Ecologica Sinica,2025,45(5):2401-2411.

基于一维卷积神经网络与自编码算法的松属物种鉴别机制

陈冬英[1,2,*],翁伟雄[1],陈培亮[1],魏建崇[1,2]

1 福建江夏学院电子信息科学学院,福州 350108

2 数字福建智能家居信息采集及处理物联网实验室,福州 350108

摘要:松属植物具有重要的生态和经济价值。但松属植物的基因组庞大、分子进化慢,物种的特征相似性极高,辨别难度大。为解决传统松属物种鉴别方法存在的成本高、耗时长、准确率低、操作复杂等问题,提出了一种基于松属近红外光谱数据(NIRS)并结合一维连续型卷积神经网络(1D-CS-CNN)与自编码技术的松属物种检测机制。使用更高效率的连续型结构替代传统 1D-CNN 模型中隐含层结构,并针对松属 NIRS 数据适应性改进为 1D-CS-CNN 模型,使其可直接应用于一维 NIRS 数据。结合自编码器的重构误差设计一种考虑未知类别的松属物种鉴别方法,通过待测样本的自编码重构误差来解决卷积神经网络置信度过高的问题,将修正的置信度与预先设定的阈值进行比较,判断该样本是否为未知品种。实验结果表明,1D-CS-CNN 训练集与测试集准确率均达到近 100%,损失值收敛为 0.015,改进后的 1D-CS-CNN 模型识别速度更快;同时,自编码模型对未知类别松属检测机制识别率为 99%。实验结果证明,该模型可快速高效分类出不同松属物种,同时检测出松属新物种。

关键词:松属物种;近红外光谱(NIRS);自编码器;一维连续卷积神经网络(1D-CS-CNN);鉴别

Identification mechanism of *Pinus* species based on one dimensional convolutional neural network model and self-coding algorithm

CHEN Dongying[1,2,*], WENG Weixiong[1], CHEN Peiliang[1], WEI Jianchong[1,2]

1 *College of Electronic Information Science, Fujian Jiangxia University, Fuzhou* 350108, *China*

2 *Smart Home Information Collection and Processing on Internet of Things Laboratory of Digital Fujian, Fuzhou* 350108, *China*

Abstract: *Pinus* species hold important ecological and economic value. However, the large genomes and slow molecular evolution of Pinus species result in highly similar morphological traits, complicating inter-specific identification. In order to solve the problems of high cost, long time, low accuracy and complex operation of traditional identification methods, this paper designed a detection method based on near-infrared spectral data (NIRS), one-dimensional continuous convolutional neural network (1D-CS-CNN) and Auto-encoder technology for *Pinus* species. Initially, the 1D-CS-CNN model uses the efficient Continuous Structure (CS) to replace the hidden layer structure in the traditional 1D-CNN model. The model can be directly applied to analyze one-dimensional near-infrared spectral data (NIRS). Next, combining the reconstruction error of the auto-encoder, a identification method considering an unknown origin is designed for *Pinus* species, which can solve the problem of high confidence in convolution neural networks by using an auto-encoder and reconstruction errors of the samples to be tested. Whether the sample is an unknown variety can be determined by comparing the corrected confidence

基金项目:福建省自然科学基金项目(2023J011094);福建省高校产学合作项目(2021H6003);全国大学生创新创业国家级项目(202413763002)

收稿日期:2024-04-24; **网络出版日期**:2024-11-28

* 通讯作者 Corresponding author.E-mail: cdy@ fjjxu.edu.cn

level with the preset threshold value. The results show that the 1D-CS-CNN training set achieves 100% accuracy, with the loss value stabilizing at 0.015. Compared with the traditional 1D-CNN model, the improved 1D-CS-CNN model has faster recognition speed. Meanwhile, the accuracy rate of the auto-encoder for the category detection mechanism of *Pinus* species from an unknown origin is 99%. The experimental results show that the model can quickly and efficiently classify different species of *Pinus* and detect the new species in *Pinus*.

Key Words: *Pinus* species; near-infrared spectral data (NIRS); auto-encoder; one-dimensional continuous convolutional neural network (1D-CS-CNN); identification

松属植物在我国大量分布[1],是木材、纸浆、松香和松节油的主要来源,同时也是现存最古老的针叶树之一,具有极其重要的生态和经济价值[2]。近年来,随着生态保护意识的增强和生物资源利用的深入,松属植物的研究逐渐受到国内外学者的高度关注[3—4]。但是,松属植物的基因组十分庞大、分子进化速率慢,造成不同松属物种的特征相似性极高,辨别其物种归属难度很大[5—6]。为了更有效的辨别松属物种的种类,建立一种方便快捷的鉴别方法是十分急迫的。

目前已现存的松属物种类别鉴定的研究技术主要有:植物形态学物种鉴定方法、物种的 DNA 分子鉴定技术、显微鉴定法等。其中,植物形态学物种鉴定,是通过采集样本的根部、茎部、叶部等性状特征,根据性状展现的颜色,纹理特征等性状特点来对松属物种进行鉴别。由于植物形态学物种鉴定方法不需要特定的实验设备和实验试剂,随时可以通过观察即可进行鉴定,因此是目前在松属物种鉴别中应用最广的鉴定方法[7]。但是植物形态学物种鉴定法依赖于已有的经验和现有的知识,存在积累时长和主观性的限制。同时显微鉴定法[8]、DNA 分子鉴定法[9—10]等传统分析方法虽然分析精度较高,但是分析过程时间长、效率低,需要大量的专业知识,并且会破坏样本。因此不适用于松属物种的快速鉴别。

近红外技术具备简单、快速、无损等优点,其光谱主要是由含氢基团 X—H(X=C, N, O)振动的合频和倍频组成,所含信息量丰富,被广泛应用于各类植物识别。然而不同的松属物种内含成分种类及其含量基本相同,不同类别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[11]。近年来,综合近红外光谱数据(Near-infrared spectral data, NIRS)与合适的模型进行分类识别,现有研究已取得良好的效果[12]。尤其是机器学习中的卷积神经网络(Convolutional neural network, CNN)算法应用在数据特征提取[13],目标检测[14]以及图像分类识别[15]等各方面均获得很好的效果。如:文献[16]设计了一种基于一维卷积神经网络(1D Convolutional neural network, 1D-CNN)算法识别马兜铃酸和其相似物,其准确率可以达到约100%;文献[17]对 LeNet-5 网络结构进行改进,应用在不同产地的烟叶识别,训练集与测试集分别可以达到98.2%和95.0%的准确率;文献[18]分别建立三种模型:1D-CNN,2D-CNN 及 PLS-DA 对烟叶的不同产地进行溯源,三种方法比较表明 1D-CNN 的效果最佳。近年来,近红外数据和机器学习的传统模型已广泛应用在植物物种鉴别分析[19]。通过查阅相关研究文献,将 1D-CNN 应用在基于 NIRS 数据的松属物种研究仍处于初步探索阶段。

虽然在 NIRS 数据定性分析方面,模式识别方法已从传统的浅层学习发展到深度学习方法并取得了有效的成果,但上述文献工作所使用的深度模型,仅限于识别训练集中存在的已知类别识别,无法应用于新的未知类别识别[20]。现有的分类器往往将未知类别区分成已知类别,使得在松属物种检测过程带来了极大的风险[21]。因此,如何提高定性分析模型识别未知类别的能力也是亟待解决的关键问题。

本文以不同松属物种的近红外光谱数据为研究对象,提出一种基于一维连续卷积神经网络(One Dimensional Continuous Structure Convolutional Neural Network,1D-CS-CNN)的鉴别模型。同时,采用自编码器的重构误差来修正未知类别在 CNN 模型中置信度过高的问题,使 1D-CS-CNN 模型能够使用阈值法来判断当前数据是否为未知类别。实验证明,本设计模型实现快速无损的鉴别出不同松属物种,同时准确的判断出松属新物种,并且具有效率高、适应性高、识别精度高等优点。

1　样本数据获取与预处理

1.1　松属近红外光谱数据的采集

采集的样本是各类松属,该属分布于全球各地[22]。在我国,当属云南、东北等多个省份松属物种生长旺盛、特征明显。所以最终从云南获取的松属物种的 NIRS 数据共 1700 个。数据来源于 CSDN 网站的开源数据[23]。本设计对所获得的原始数据进行样本分配,具体情况如表 1 所示。

表 1　实验数据分配情况

Table 1　Distribution of experimental data

样本集 Samples	样本名称 Sample name	样本数量 Numbers of sample	样本编号 Sample number
已知类别	乔松(代号 1)	300	1—300
Known class	云南松(代号 2)	300	301—600
	华山松(代号 3)	300	601—900
	思茅松(代号 4)	300	901—1200
	樟子松(代号 5)	300	1201—1500
未知类别	油松	100	1501—1600
Unknown class	红松	100	1601—1700

1.2　样本预处理及划分

松属物种的进化过程极其缓慢,因此,不同松属物种的特征有很大的相似之处。为了更好的提取不同松属物种独有的特征,需要对原始数据进行预处理。本设计预处理采用均值标准差标准化法[24]。该方法通过将某个特征的值映射到某一区间之内,消除量纲对最终结果的影响,将原值减去均值后除以标准差,使得得到的特征满足均值为 0,标准差为 1 的正态分布。其转化函数如公式(1)所示。

$$x^{*}=\frac{x-\mu}{\sigma} \tag{1}$$

式中,μ 表示所有样本数据平均值,σ 表示所有样本数据标准差。预处理前后的结果如图 1 所示。

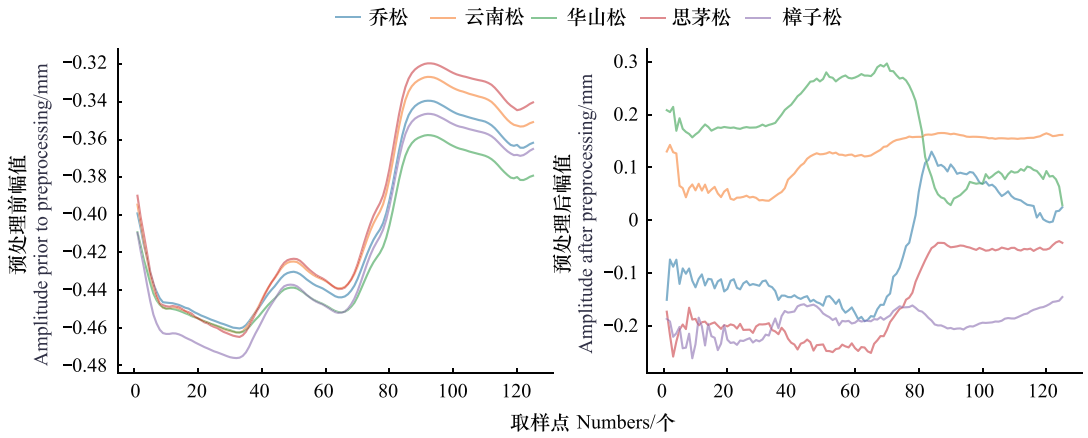

图 1　样本预处理前后

Fig.1　Pre-processing data and post-processing data of samples

为进一步实验,本设计对预处理后的 1500 个已知松属物种样本进行分层划分。划分方法如下:将总样本随机按比例 8:2 分为模型的训练集与模型的测试集。其中,训练集再随机按比例 8:2 切分为模型的训练数据与模型的训验证练数据。所以,样本数据包括三个类别:训练集(64%)、验证集(16%)和测试集(20%)。根

据训练集数据完成模型建立,采用验证集数据进行模型选择,测试集进行最终模型的识别性能验证。不同松属物种样本的数据划分如表 2 所示。

表 2 样本数据集的分层划分

Table 2 Hierarchical division of the sample data set

样本 Samples	样本总数 Numbers of sample	训练集 Training set	验证集 Validating set	测试集 Testing set
乔松 *Joss pine*	300	192	48	60
云南松 *Pinus yunnanensis*	300	192	48	60
华山松 *Pinus armandii*	300	192	48	60
思茅松 *Pinus cohinchinensis*	300	192	48	60
樟子松 *Mongolian pine*	300	192	48	60
总数 Total	1500	960	240	300

2 分析方法

2.1 1D-CS-CNN 模型设计

一般卷积神经网络包括输入层、隐含层和输出层,如图 2 所示[25]。图中的隐含层由卷积层、池化层以及全连接层组成。首先,通过卷积层自动提取数据特征;接着,采用最大池化层操作实现降采样;而后,设计激活层来计算一个输出张量;最后,设置一层或多层全连接层堆叠,起到分类器的作用。同时,为了避免过度拟合并加快收敛速度,通常使用随机失活机制和批归一化(Batch normalization,BN)机制。

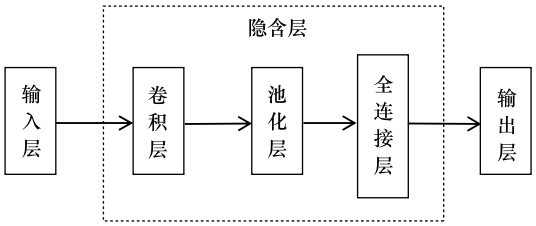

图 2 卷积神经网络结构图

Fig.2 Structure diagram of convolutional neural network

在 CNN 基础上,本设计通过对 LeNet-5 进行了改进得到一个适应于松属物种 NIRS 数据分类效果良好的 1D-CNN 模型。本设计结构在 1D-CNN 的隐含层采用交替连续递减模式,构建 3-2-1 的卷积层与最大池化迭代,命名为一维递进卷积神经网络(One dimensional progressive structure convolutional neural network,1D-PS-CNN)。其核心结构如图 3 所示。

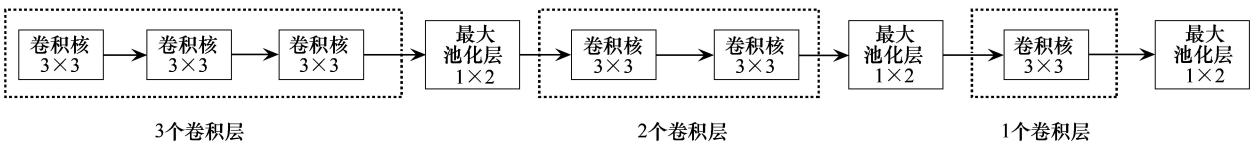

图 3 模型改进前的核心结构

Fig.3 The core structure before the model improvement

为了降低参数量、提高准确率,使用连续型结构替代原有的递进型结构,主要有两个方面的改进:

1、若采用递减连续多次卷积,易导致计算量大且运行时间长、CPU 运行内存增大等问题,为了克服以上不足,本设计改进层数为平均值,即改原来层数为 3-2-1 的递减卷积数为平均 2-2-2 卷积层数;

2、在特征图缩小相同维度的前提条件下,卷积核越小,占用的内存越小,同时该数据为一维数据。为了更高效且降低计算量,本设计把原来卷积层大小为 3×3 用大小为 1×3 卷积层与一个带有 Dropout 的池化层来替代。把数据大小降采样到只有一个矢量的向量大小。改进后的核心结构如图 4 所示。

改进后的结构命名为 1D-CS-CNN,其网络模型如图 5 所示。图中的 128@ 1×3 表示通道数是 128 个,且每个通道的数据大小是 1×3。具体实现的步骤如下:首先,输入大小为 1×25 的数据源,经过 2 层卷积后,输出 128 个特征图,且每个特征图的大小为 1×119。其中,每层卷积核为 128 个,每个卷积核大小为 1×3;其次,对

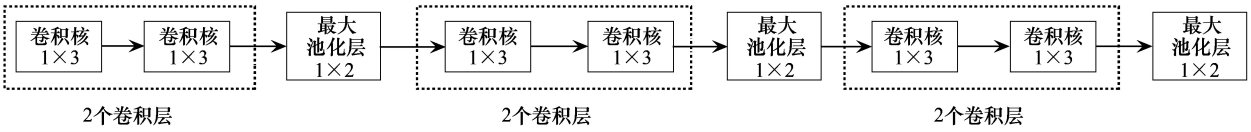

图 4　模型改进后的核心结构

Fig.4　The improved core structure of the model

128 个特征图进行最大池化操作,设置步长为 2,得到特征图 128 个,且各个特征图大小均为 1×59;而后,重复上述操作,在相同卷积核的大小的前提下经过 2 层卷积后,再对卷积后的特征图步长为 2 的最大池化操作,得到 128 个特征图,大小变成 1×27;最后,重复第一个和第二个步骤,最终得到的 128 个特征图的大小均为 1×12。利用展平层将数据大小平铺为 1×1536,由此可以使得数据更好的经过全连接层。最终,通过 2 个全连接层输出分布概率得到分类结果。

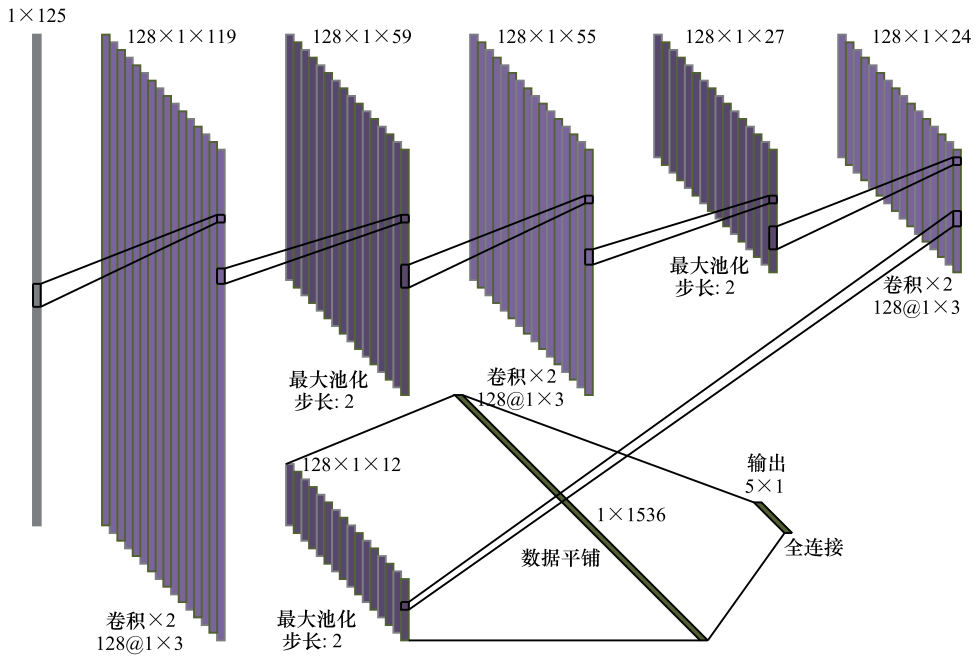

图 5　1D-CS-CNN 网络结构

Fig.5　Network structure of 1D-CS-CNN

128@ 1×3:通道数为 128 个,且每个通道大小为 1×3

2.2　基于自编码的未知产地检测模型建立

鉴别模型中训练集已知的类别称为样本内分布数据(In-distributions data,*ID* data),鉴别模型中训练集未知的类别称之为样本外分布数据(Out-of-distributions data,OOD data)。本设计采用阈值法来预测 OOD 数据。主要思路是利用 *ID* 数据训练好的鉴别模型,将鉴别模型最后输出的 MAE 函数值定义为置信度,设置合适的阈值,通过以下公式判断检测数据是否为 OOD 样本。若置信度大于阈值 σ,则该样本为 *ID* 样本,否则为 OOD 样本。设计公式如式(2)。

$$m(x;\sigma)=\begin{cases} IN & \text{if} \quad M_{\text{loss}}(x)\leqslant\sigma \\ \text{OUT} & \text{if} \quad M_{\text{loss}}(x)>\sigma \end{cases} \tag{2}$$

其中,$M_{\text{loss}}(x)$ 为模型输出的样本置信度,*IN* 表示输出为 ID 样本,OUT 表示输出为 OOD 样本。

2.2.1 自编码器

自编码器(Auto-encoder,AE)结构图如图 6 所示。该结构由编码器、压缩编码、解码器组成。编码器通常是一个前馈、密集连接的人工神经网络,目的是获取输入数据并将其压缩成一个潜在的空间表征,从而生成一个具有降维特性的压缩数据[26—27]。解码器负责获取压缩数据并将其转换回与原始数据相近的空间表征。

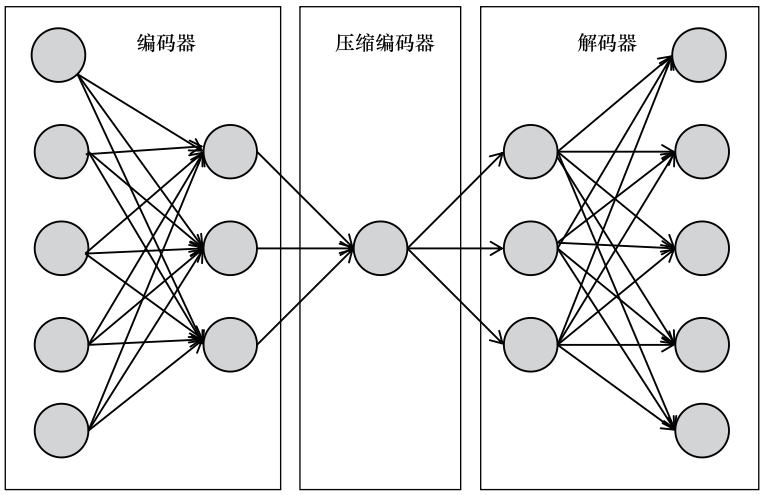

图 6 自编码器结构

Fig.6 The structure of Auto-Encoder(AE)

2.2.2 松属物种检测机制设计

本文使用自编码器的重构误差来修正过高的 OOD 样本的概率值。思路如下:本设计使用 ID 近红外光谱数据样本训练一个自编码器。当使用 ID 光谱数据输入自编码器重建光谱数据时,重建后的光谱数据较为平滑,而 OOD 样本重构后的光谱数据较为粗糙。相应地 ID 重构数据与原始数据之间的误差相对较小,而 OOD 重构误差则较大。因此,通过比较重构前后的误差,可实现对 OOD 样本的检测。

结合一维卷积神经网络与自编码技术,本设计的松属物种检测机制如图 7 所示。具体步骤如下:输入一个待测样本的 NIR 数据,分别经过 1D-CS-CNN 鉴别模型和自编码器;通过重构误差修正后的 MAE 值与设置好的阈值对比,从而判断待测样本是 OOD 样本,还是 ID 样本;若为样本内数据,将其输入至 1D-CS-CNN 模型,并判断输出松属类别。

由图 7 可得,本文所设计的松属物种检测机制并不影响最初的鉴别模型分类任务,可以将其看成是外置模块,该机制可在不影响原有鉴别模型的基础上将其移植到鉴别模型中。

2.3 模型评价方法

本设计采用浮点运算 FLOPs(Floating-point Operations)衡量模型的性能复杂度。根据模型中的乘加数来计算 FLOPs[28]。本模型中卷积层的 FLOPs 采用公式(3)进行计算。

$$\mathrm{FLOPs}_{\mathrm{conv}} = (k_w \times k_h \times c_{in} \times c_{\mathrm{out}} + c_{\mathrm{out}}) \times H \times W \tag{3}$$

式中,k_w、k_h、c_{in}、c_{out}分别表示卷积核的宽度、卷积核的高度、输入通道数以及输出通道数(卷积核的数量),H 以及 W 分别表示为特征向量的高与宽。

本设计以准确率(Accuracy, ACC)为模型性能评价的第二个指标。通过将每一类别的 TP(即模型预测为正样本的数量)与 TN(模型预测负样本的数量)相加求和,而后与预测总样本数相除获得模型多标签任务的准确度。准确率采用公式(4)进行计算。

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n}(TP_i + TN_i)}{\sum_{i=1}^{n}(TP_i + TN_i + FP_i + FN_i)} \tag{4}$$

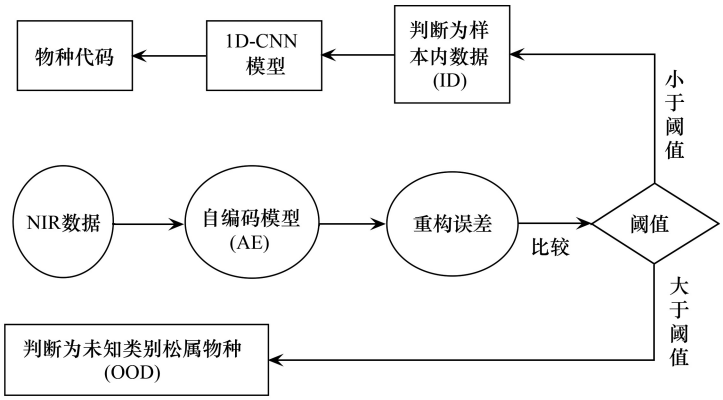

图 7　松属物种检测机制结构框图

Fig.7　The structural diagram of the detection of pine species

NIR: 近红外光谱 Near-infrared spectral;1D-CNN: 一维卷积神经网络 One dimensional convolutional neural network

式中,*FP* 与 *FN* 分别表示预测错误的负样本数量和预测错误的正样本数量。ACC 的值越大,则预测准确率越高。

1D-CNN 模型另一个性能指标为交叉熵损失函数(Categorical Crossentropy,Loss)[29],计算公式如下:

$$\text{Loss} = -\sum_{i=1}^{\text{OutputSize}} (y_i \times \log \hat{y}_i) \tag{5}$$

式中,OutputSize 指样本的种类,在本设计中样本的种类为 5。y_i 代表种类代号(即松属代号),$\hat{y}_i$代表模型输出的某个种类的 SoftMax 值(即模型预测某个样本属于某个产地的置信度)。交叉熵损失函数越小,代表模型的分类性能越好。

Auto-Encoder 模型的主要性能指标是平均绝对误差(Mean absolute error, MAE)。通过对输入与输出每一点的误差进行求和,而后将总和除以总数,即为 MAE 值。MAE 的计算公式如下:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{6}$$

式中,$\hat{y}_i$代表自编码器输出数据某个点的大小,y_i 代表输入数据某个点的大小。MAE 的值越低,说明输入与输出的误差越小,模型性能越好。

3　结果与讨论

3.1　1D-CS-CNN 模型的性能分析

为了测试 1D -CS- CNN 对松属物种 NIRS 数据集的识别性能,分别建立了基于标准差标准化预处理的 1D-PS-CNN 和改进 1D-CS-CNN 进行对比。数据集中的[训练集(训练集:验证集):测试集]按比例[8(8:2):2]完成划分。本系统所涉及的参数设置如表 3。

表 3　超参数的设置

Table 3　Parameter settings

批尺寸 Batch size	激活函数 Activate function	训练次数 Epoch	学习率 Learning rate	优化器 Optimizer
32	Tanh	300	0.0003	Adam

根据 2.3 节的公式实现各个性能指标的计算。图 8 与图 9 分别为模型训练过程中的损失值与准确率。根据图 8,1D-CS-CNN 相较 1D-PS-CNN 的损失值下降速度更快。当训练次数达到 250 左右时,1D-CS-CNN 的训练集及验证集的损失值,分别为 0.05 和 0.15,与 1D-PS-CNN 模型相比分别降低 0.1188 和 0.0617。结果表

明 1D-CS-CNN 比 1D-PS-CNN 具有更好的学习能力,能够提取更多的数据特征,具有更强的识别能力。由图 9 可知,1D-CS-CNN 的测试集与训练集上升速度均更快,其准确率分别为 100%和 98.33%,相比与 1D-PS-CNN 模型,该模型训练集和验证集准确率更高。说明该模型可以更准确、无损地实现不同松属 NIRS 数据物种的识别。

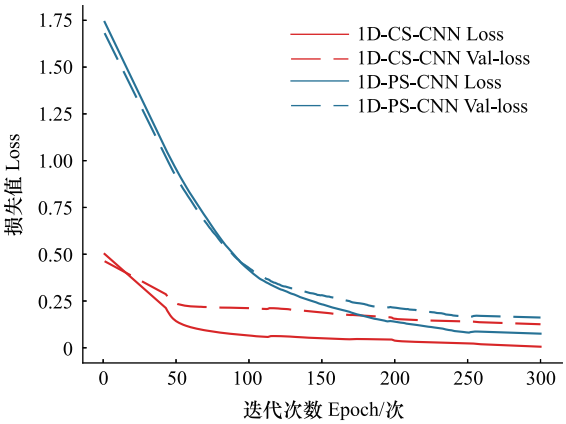

图 8 改进的 1D-CS-CNN 和 1D-PS-CNN 的损失值

Fig.8 The Loss of improved 1D-CS-CNN and 1D-PS-CNN

1D-CS-CNN Loss:一维连续卷积神经网络训练集损失值;1D-CS-CNN Val-loss:一维连续卷积神经网络验证集损失值;1D-PS-CNN Loss:一维递进卷积神经网络训练集损失值;1D-PS-CNN Val-loss:一维递进卷积神经网络验证集损失值

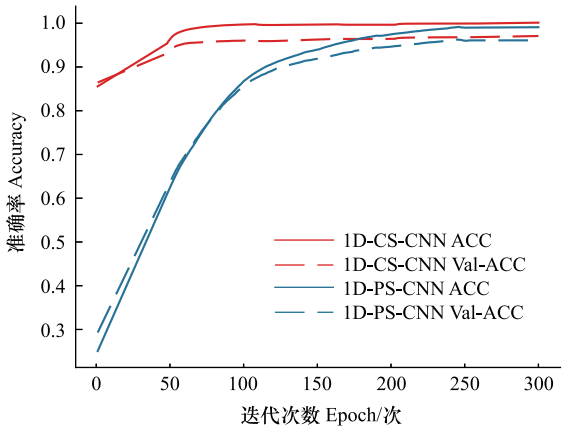

图 9 改进的 1D-CS-CNN 和 1D-PS-CNN 的准确值

Fig.9 The Accurate of improved 1D-CS-CNN and 1D-PS-CNN

为了保证模型训练占有尽量小的内存,同时达到尽可能好的效果,本设计中的两个模型训练次数均设为 150。表 4 为改进前后模型所得的性能指标对比结果。根据表 4 可得,1D-CS-CNN 模型相比于 1D-PS-CNN 模型,训练集与测试集的准确率(ACC),分别提升了 3.22%和 2.77%,FLOPs 分别减少近 10.6%,且识别速率加快了 4.75s。由此表明,在数据有限的前提下,若仅通过模型容量的增加来提高模型性能,代价相对比较昂贵,但适当改变卷积核结构可以大大提高 CNN 模型的效率。以上实验结果的对比和分析表明,改进的 1D-CS-CNN 模型对松属物种近红外光谱分类具有更高的效率和更好的识别性能。

表 4 两种模型的评价指标对比

Table 4 Comparison of evaluation indicators of the two models

性能 Performance	一维递进卷积神经网络 1D-PS-CNN		一维连续卷积神经网络 1D-CS-CNN	
	训练集	验证集	训练集	验证集
准确率 ACC/%	96.78	95.56	100	98.33
损失值 Loss	0.1351	0.1905	0.0163	0.1288
浮点运算 FLOPs	178.1M		159.2M	
运行时间 Runtime/s	43.01		38.26	

3.2 1D-CS-CNN 与传统方法的比较

为验证本设计方案的可行性与高效性,分别与原始光谱数据分析法和植物形态学物种鉴定法比较。选取 30 个样本及其光谱数据,按比例 8:2 分为训练集数据与测试集数据,分别采用三种方法进行判断。其对比实验结果如表 5。由表 5 可知,三种方法的训练与测试结果,以 1D-CS-CNN 的方法识别准确率最高,性能最优。因此,在有限的高维数据前提下,本设计模型与不采用特征提取算法的传统方法拥有更好的分类结果。

表 5　1D-CS-CNN 与传统方法对比结果

Table 5　Results of comparison between 1D-CS-CNN and traditional methods

方法 Method	训练集精度平均值/% The precision average of the training set	测试集精度平均值/% The accuracy average of the testing set
原始光谱数据 Raw spectral data	62.50(15/24)	50(3/6)
植物形态学 Plant morphology	91.67(22/24)	83.33(5/6)
一维连续卷积神经网络 1D-CS-CNN	100	100

(前数据/后数据)如(15/24)表示前数据为实验识别准确的样本数是 15 个,后数据为实验总样本数是 24 个

3.3　未知松属物种检测机制的结果分析

3.3.1　自编码器模型的建立

自编码器结构如图 10 所示。本设计的编码器和解码器,采用的全连接层构成分别为 32 个神经元和 16 个神经元,压缩编码层均使用包含 16 个神经元的全连接层,迭代次数 Epoch 为 100,批大小 Batch-size 为 32。

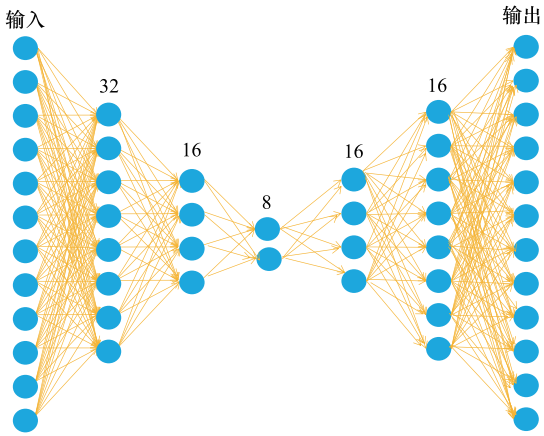

图 10　自编码器网络结构

Fig.10　The network structure of the Auto-encoder

将不同松属物种的 NIRS 数据按照 8(8:2):2 的比例划分数据集、验证集与测试集,利用训练集建立自编码器模型。自编码器的训练过程与 1D-CS-CNN 模型相似,通过不断调整神经元的权重和差值,来降低理想输出和实际输出的误差(通过 MAE 函数值体现),从而获得一个理想的模型。本设计训练过程的 Loss 函数值变化如图 12 所示。与 1D-CS-CNN 模型不同,1D-CS-CNN 模型输出的是一个值(代表产地代号),而自编码器的输出是 1D-CS-CNN 的输入。从图 12 可知,模型的训练集和测试集损失值最终分别收敛在 0.000001。

3.3.2　未知类别检测机制结果分析

根据 3.3.1 完成自编码器的建立。图 12 为自编码器模型对所有样本数据的输出 MAE 函数值的统计。其中,纵坐标代表样本的数量,横坐标代表 MAE 值的区间。因大部分样本内数据 MAE 值均小于 0.05,因此,本文的阈值设置为 0.05。若某物种数据经过自编码器重构后的误差大于 0.05,则该样本判断为样本外数据,否则判断其为样本内数据。

输入样本内数据思茅松(4)和样本外数据红松至系统。自编码器的输出结果如图 13 所示。图中的蓝色线代表自编码器的输入,红色线代表自编码器重构误差后的输出,红色面积区域代表两者之间的误差。若误差大于本设计的阈值 0.05,代表样本属于样本外数据。图 13 中左图为输入思茅松的重构误差结果图,图 13 中右图为输入红松的重构误差图。在可视化界面中运行日志输出结果,得该样本内数据的重构误差值为 0.009,小于 0.05,输出预测物种代码为样本内种类 4(思茅松),与输入结果一致。对样本外数据的重构误差为 0.0618,大于阈值 0.05,输出判断为未知类别。因此,该系统对样本内外数据具有很好的识别能力。

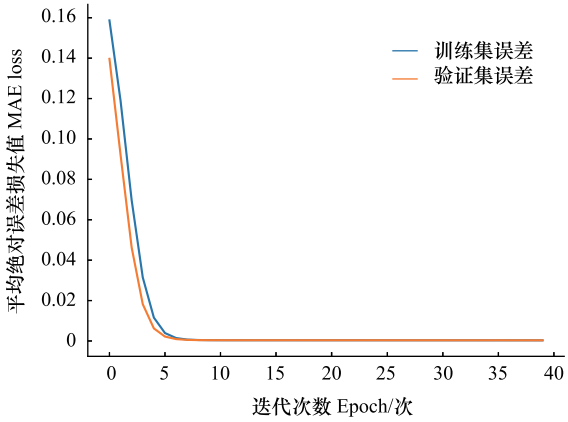

图 11 自编码器模型训练的 MAE 值

Fig.11 The MAE of the Auto-encoder

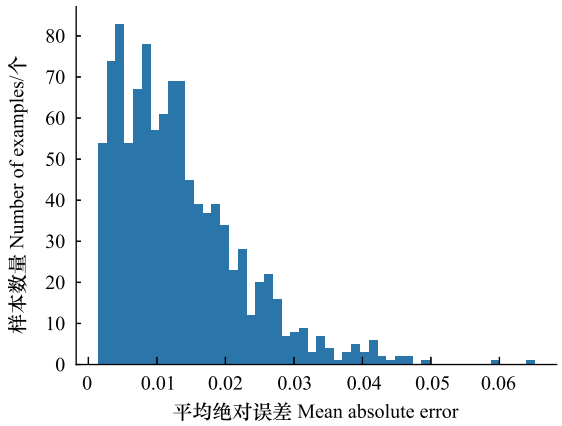

图 12 自编码器输出的 MAE 函数值

Fig.12 MAE function value output from the encoder output

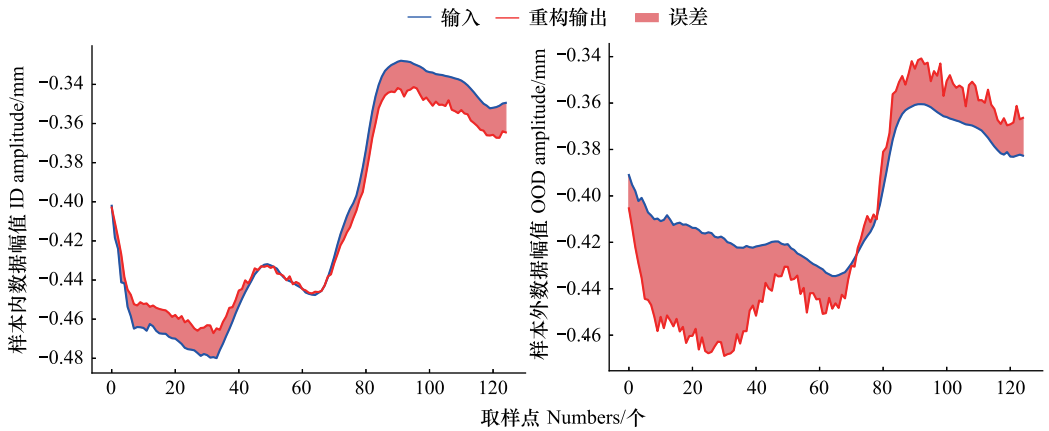

图 13 样本内数据样本外数据自编码器重构误差结果

Fig.13 Auto-encoder Reconstruction error results of reconstruction of data in the sample and reconstruction of out-of-sample data

为了更好评价系统预测的性能,本设计采用混淆矩阵[30]对 OOD 检测机制预测性能进行评估,具体结果如图 14 所示。其中,混淆矩阵横向代表样本的实际类别,纵向代表样本的预测类别。从图中可得,模型将 12 个 ID 数据错误分类成 OOD 数据,将 3 个 OOD 数据误判成 ID 数据。OOD 检测机制鉴别准确率为 99%,可以有效地鉴别未知伪品。

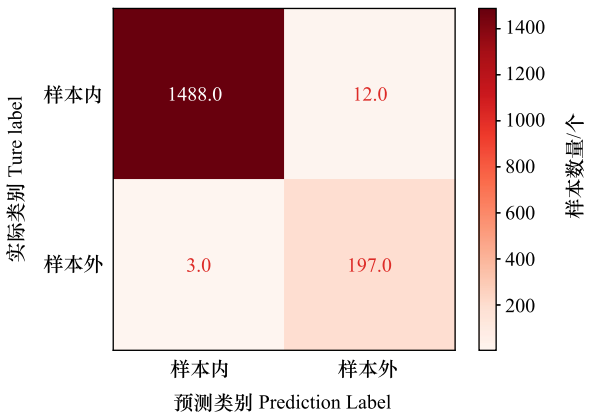

图 14 松属物种检测机制混淆矩阵

Fig.14 The confusion matrix of the *Pinus* species detection mechanism

4 结论

针对不同松属物种鉴别问题,本研究提出一种基于一维卷积神经网络与自编码算法的松属物种 NIRS 分类模型。根据 NIRS 数据特点,替换递进结构为连续结构,改进 1D-PS-CNN 为 1D-CS-CNN 模型。将该模型输出的 MAE 概率值,定义为样本的置信度,设置阈值为 0.05,通过重构误差设计自编码模型。测试结果表明,改进后的 1D-CS-CNN 模型可实现高效准确的松属物种

识别,加入的自编码模型对未知类别的识别准确率达 99%。综上,本研究对植物分类鉴定研究具有重要的意义。

参考文献(References):

[1] 马庆,马庆辉. 分子标记技术在松属植物遗传改良中的应用. 辽宁林业科技,2023,(2): 40-42, 69.

[2] 杨海涵, 王治炜, 杨静莉. 松科树种体胚发生研究进展. 温带林业研究, 2021, 4(2): 1-7, 12.

[3] Falcon-Lang H J, Mages V, Collinson M. The oldest *Pinus* and its preservation by fire. Geology, 2016, 44(4): 303-306.

[4] 李卫英, 章正仁, 辛雅萱, 王飞, 辛培尧, 高洁. 云南松、思茅松和卡西亚松天然种群间的针叶表型变异. 植物生态学报, 2023, 47(6): 833-846.

[5] 洑香香, 赵虎, 王玉. 松属近缘种形态和分子鉴定及其亲缘关系探讨. 林业科学, 2011, 47(10): 51-58.

[6] 李晓辰, 贡璐, 魏博, 丁肇龙, 朱海强, 李岳峰, 张涵, 马勇刚. 气候变化对新疆雪岭云杉潜在适宜分布及生态位分化的影响. 生态学报, 2022, 42(10): 4091-4100.

[7] Rodríguez S M, Ordás R J, Alvarez J M. Conifer Biotechnology: An Overview. Forests, 2022, 13(7): 1061.

[8] 李俊, 段雅萍, 蔡秀珍, 王婷, 潘柏含. 松属针叶角质层微形态特征在分类学中的应用. 植物研究, 2022, 42(3): 341-351.

[9] 杨辰. 松属五针松组物种的分子鉴定研究[D]. 兰州: 兰州大学, 2014.

[10] 于巧宁, 仲米存, 王锐莹, 王桐, 王洪涛. 马尾松和云南松花粉的特异性分子鉴定. 食品科技, 2020, 45(8): 56-60.

[11] Zhu L H, Chu X F, Sun T Y, Ye J R, Wu X Q. Micropropagation of Pinus densiflora and the evaluation of nematode resistance of regenerated microshoots in vitro. Journal of Forestry Research, 2019, 30(2): 519-528.

[12] Zeng J, Chai Q Q, Peng X, Li S M. Geographical Origin Identification for Tetrastigma Hemsleyanum Based on High Performance Liquid Chromatographic Fingerprint. 2019 Chinese Automation Congress (CAC). November 22-24, 2019, Hangzhou, China. IEEE, 2019: 1816-1820.

[13] 拱健婷, 李莉, 邹慧琴, 徐东, 王大仟, 丛悦, 刘长利. 基于近红外光谱和梯度提升决策树建立当归药材及伪品的定性判别模型. 世界科学技术-中医药现代化, 2019, 21(10): 2237-2243.

[14] Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[15] Mishkin D, Sergievskiy N, Matas J. Systematic evaluation of convolution neural network advances on the Imagenet. Computer Vision and Image Understanding, 2017, 161: 11-19.

[16] Zhang L, Ding X Q, Hou R C. Classification Modeling Method for Near-Infrared Spectroscopy of Tobacco Based on Multimodal Convolution Neural Networks. Journal of Analytical Methods in Chemistry, 2020, 2020: 9652470.

[17] 鲁梦瑶, 杨凯, 宋鹏飞, 束茹欣, 王萝萍, 杨玉清, 刘慧, 李军会, 赵龙莲, 张晔晖. 基于卷积神经网络的烟叶近红外光谱分类建模方法研究. 光谱学与光谱分析, 2018, 38(12): 3724-3728.

[18] Hein M, Andriushchenko M, Bitterwolf J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 41-50.

[19] 柴琴琴, 曾建, 张勋. 基于贝叶斯优化卷积神经网络的金线莲伪品鉴别. 浙江农业学报, 2022, 34(2): 391-396.

[20] 王选齐, 杨锋, 曹斌, 刘静, 魏德健, 曹慧. 卷积神经网络在甲状腺结节诊断中的应用. 激光与光电子学进展, 2022, 59(8): 0800002.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 2017, 60(6): 84-90.

[22] 刘艳艳, 刘畅, 魏晓新. 我国及周边地区松属白松亚组系统学研究进展和保护现状. 生物多样性, 2022, 30(2): 136-147.

[23] wood-nirs: 四种松属物种的近红外光谱(NIRS)数据, 2021 年 4 月 8 号. 网址: https://download.csdn.net/download/weixin_42127937/16557677.

[24] Chen D Y, Zhang H, Xiao Y Y, Zhang Z L, Chen W J, Huang S Y, Chen H X. Fiber grating loop ring-down strain sensors using overlap spectrum demodulation and machine learning algorithm. Optical Fiber Technology, 2023, 76: 103248.

[25] Chen X Y, Chai Q Q, Lin N, Li X H, Wang W. 1D convolutional neural network for the discrimination of aristolochic acids and their analogues based on near-infrared spectroscopy. Analytical Methods, 2019, 11(40): 5118-5125.

[26] Chai Q Q, Zeng J, Lin D H, Li X H, Huang J, Wang W. Improved 1D convolutional neural network adapted to near-infrared spectroscopy for rapid discrimination of Anoectochilus roxburghii and its counterfeits. Journal of Pharmaceutical and Biomedical Analysis, 2021, 199: 114035.

[27] 袁非牛, 章琳, 史劲亭, 夏雪, 李钢. 自编码神经网络理论及应用综述. 计算机学报, 2019, 42(1): 203-230.

[28] 陈冬英, 张昊, 张子龙, 余沐昕, 陈璐. 基于改进 1D-VD-CNN 与近红外光谱数据的金银花产地溯源研究. 光谱学与光谱分析, 2023, 43(5): 1471-1477.

[29] Chen D Y, Zhang H, Lin L Y, Zhang Z L, Zeng J, Chen L, Chen X G. Auto-encoder design based on the 1D-VD-CNN model for the detection of honeysuckle from unknown origin. Journal of Pharmaceutical and Biomedical Analysis, 2023, 234: 115572.

[30] Lipton Z C, Elkan C, Naryanaswamy B. Optimal Thresholding of Classifiers to Maximize F1 Measure. Machine Learning and Knowledge Discovery in Databases, 2014, 8725: 225-239.