doi:10.6041/j.issn.1000-1298.2023.01.019

基于植物电子病历多类型数据融合的作物病害诊断方法

丁俊琦[1]　李　博[2]　乔　岩[3]　张领先[1]

(1.中国农业大学信息与电气工程学院,北京 100083;2.北京信息科技大学经济管理学院,北京 100083;
3.北京市植物保护站,北京 100029)

摘要:植物电子病历(EMR)以结构化和非结构化的形式记录了大量关于疾病症状、环境特征以及诊断开方的信息,为病害的智能诊断提供了优质知识来源,但是其样本量少、公开数据集缺乏和多种类型数据并存的特点给相关研究带来困难。根据植物 EMR 多类型数据混合的特点,提出了一种基于 BERT-MPL 数据融合与注意力机制优化的作物病害诊断模型(BERT-MPL data fusion model based on attention mechanism,BM-Att)。首先采用 BERT 预训练语言模型抽取电子病历中非结构化部分的文本语义特征;其次通过 one-hot 编码和多层感知机(MLP)对结构化数据进行编码和向量维度的扩增;最后在特征融合阶段采用注意力机制强调关键特征,利用多层全连接层实现病害诊断。构建了番茄、黄瓜、生菜和西瓜 4 种作物的 15 种病害数据集验证模型的效果并进行消融实验,并且对比了 CNN、RCNN、AttRNN、FastText、Transformer、BERT 和 ERNIE 等处理文本数据的常见模型,以及 BERT-ALEX、BERT-1dCNN、BERT-1dLSTM、BERT-1dAttLSTM、BERT-MLP、ERNIE-ALEX、ERNIE-1dCNN、ERNIE-1dLSTM、ERNIE-1dAttLSTM、ERNIE-MLP 等不同数据融合策略。结果表明,BM-Att 取得最优结果,在测试集的准确率、精确率、召回率和 F1 值宏平均值分别达到 95.82%、96.38%、95.48% 和 95.85%,能够实现作物病害的有效诊断。在特征融合阶段添加注意力机制的策略将模型 F1 值宏平均值提高 1.47 个百分点,显著提升了模型对生菜霜霉病、西瓜线虫等小样本病害的分类效果。该研究可为电子病历数据挖掘及实现智能辅助病害诊断提供参考。

关键词:病害诊断;数据融合;电子病历;BERT;MLP;注意力机制

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1000-1298(2023)01-0196-09　　**OSID**:

Crop Disease Diagnosis Method Based on Fusion of Multiple Types of Data from Plant EMRs

DING Junqi[1]　LI Bo[2]　QIAO Yan[3]　ZHANG Lingxian[1]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing* 100083, *China*
2. *School of Economics and Management, Beijing Information Science and Technology University, Beijing* 100083, *China*
3. *Beijing Plant Protection Station, Beijing* 100029, *China*)

Abstract: The rapid diagnosis of crop diseases is crucial for agricultural production. A large amount of information on disease symptoms, drug prescriptions and environmental characteristics is recorded in the plant electronic medical record (EMR) in both structured and unstructured forms. Plant EMRs can provide a high-quality source of knowledge for intelligent diagnosis of diseases. However, their small sample size, the lack of publicly available datasets and the co-existence of multiple types of data posed difficulties for related research. A crop disease diagnosis model based on BERT-MPL data fusion and attention mechanism (BM-Att) was proposed for the characteristics of multiple types of data mixing in plant EMR. Firstly, BERT pre-trained language model was used to extract text semantic features from the unstructured part of the electronic medical record. Secondly, one-hot coding and multi-layer perceptron (MLP) was used to encode the structured data and augment the vector dimension. Finally, an attention mechanism was used to selectively highlight key features in the feature fusion phase and multiple fully connected layers were used to enable disease diagnosis. To verify the validity of the model, a dataset of 15 diseases of four crops, namely tomato, cucumber, lettuce and watermelon, was constructed and the

收稿日期:2022-03-23　修回日期:2022-06-01

基金项目:国家自然科学基金项目(62176261)

作者简介:丁俊琦(1996—),女,博士生,主要从事农业信息技术和自然语言处理研究,E-mail:1163577256@qq.com

通信作者:张领先(1970—),男,教授,博士生导师,主要从事农业信息技术研究,E-mail:zhanglx@cau.edu.cn

following experiments were carried out. Ablation experiments were conducted; representative deep learning models for text classification were compared, such as CNN, RCNN, AttRNN, FastText, Transformer, BERT and ERNIE; representative models with different approaches to structured data processing were compared, such as BERT - ALEX, BERT - 1dCNN, BERT - 1dLSTM, BERT - 1dAttLSTM, BERT - MLP, ERNIE - ALEX, ERNIE - 1dCNN, ERNIE - 1dLSTM, ERNIE - 1dAttLSTM, ERNIE - MLP, etc. The results showed that BM - Att achieved optimal results with accuracy, precision, recall and F1-score of 95.82%, 96.38%, 95.48% and 95.85%, respectively in the test set, indicating that effective diagnosis of crop diseases can be achieved. The strategy of adding an attention mechanism to the feature fusion stage improved the F1 macro mean of the model by 1.47 percentage points, significantly improving the model's classification of small sample diseases such as lettuce downy mildew and watermelon nematode. The research result can provide a reference for data mining of electronic medical records and the implementation of intelligent diagnosis of diseases.

Key words: disease diagnosis; data fusion; electronic medical records; BERT; MLP; attention mechanism

0　引言

为了帮助农户减少作物病害损失,植物诊所(由国际农业和生物科学中心(CABI)领导的全球性计划 Plantwise 在全球 30 多个国家开设植物诊所,建立了由专业植物医生管理的全球植物诊所网络)通过为农民提供植物健康咨询服务,对亚洲、非洲和美洲 3 100 多万用户的生活产生了积极影响。经过专门培训的植物医生对作物疾病进行诊断,并为前来咨询的农民提出建议。这些诊疗活动细节被记录为植物电子病历(Electronic medical record, EMR),输入植物智慧(Plantwise)知识库中。植物电子病历记录了农户在农作物种植过程中遇到的病虫害问题及植物医生提供的诊疗建议。通过制定结构化病历模板,植物诊所实现病历数据的规范化书写和数据采集。这些数据以结构化和非结构化文本的形式储存了大量病害症状、作物情况和环境特征的信息,并且带有植物医生的诊断标签,为病害的智能辅助诊断研究提供了完整样本[1],如何对其进行有效挖掘以实现智能辅助诊断是一个亟待解决的问题。

在农业领域,信息管理系统在植物病害防治机构中还未普及,植物电子病历还没有得到足够重视。而在医学领域,已有大量利用电子病历实现计算机辅助疾病诊断的研究[2-4]。在不同的疾病诊断上,将疾病诊断问题转换为不同疾病病历的分类问题,这些研究证明了电子病历在辅助诊断方面的价值,为探索基于植物电子病历的病害智能诊断提供了方法参考。先前工作利用决策树[5-6]、支持向量机[7]、贝叶斯网络[8]、词嵌入和各种神经网络模型[9]来挖掘病历信息,进而实现分类。但这些方法存在难以区分词语多义性或者需要学习大量参数的问题,基于预训练的深度学习模型可以有效地解决这一问题[10],如语言模型词向量(Embeddings from language models, ELMo)[11]、生成式预训练(Generative pre-training, GPT)[12]、基于 Transformer 的双向编码器表示(Bidirectional encoder representations from Transformer, BERT)[13]等。另一方面,植物电子病历需要由专业植物医生标记和录入,样本量较少且缺乏公开数据集。针对此问题,预训练语言模型无需人工标签,可以从海量的语料中学习到通用的语言表示,在公开数据集较少的农业病害领域非常适用。如,杨国峰等[14]将 BERT 用于作物病害问答系统,李林等[15]将 BERT 与堆叠式长短期记忆网络(Long short term memory networks, LSTM)模型结合用于解决农业病虫害问句分类的语料稀缺问题。

但植物电子病历不是简单的文本描述,而是经过科学设计的、符合植物病理学中病害诊断基本原理的规范结构,包含结构化的地理、时间、环境、分布等特征。研究证明,病害发生的环境特征、时空分布等信息对病害的准确识别意义重大,但是这些信息在病害智能诊断的研究中尚未得到有效利用[16-17]。如果仅聚焦于植物电子病历中的单一类型数据,仅对问诊记录文本进行特征抽取,将会造成大量多类型结构化数据的信息损失。

综上,植物电子病历缺乏公开数据集,样本量较少,且包含多种数据类型。针对这些特点,本文提出一种基于多类型数据融合的病害诊断模型(BERT-MPL data fusion model based on attention mechanism, BM - Att),通过预训练模型和不同数据分支的特征融合,多维度、全方位地对植物电子病历中的信息进行抽取,进而实现病害准确诊断。

1　数据描述与处理

1.1　数据来源

本文植物电子病历样本来源于北京市植保站所

建立的 115 家植物诊所线上诊断系统。植物诊所依托合作社、绿控基地、专业化防治组织等社会化主体,覆盖北京市昌平、延庆、顺义、平谷、密云、房山、通州、大兴、怀柔、海淀等 10 个区。

植物诊所采用公益性的农作物病虫诊断与咨询服务模式,由具备资质的植物医生为农户提供病虫害诊断和防治技术服务并记录开方。所有植物电子病历经过“植物医生填写录入——区级数据管理员初步协调验证——市级数据管理员二次协调验证”的三级数据验证程序导入数据库(图 1)。

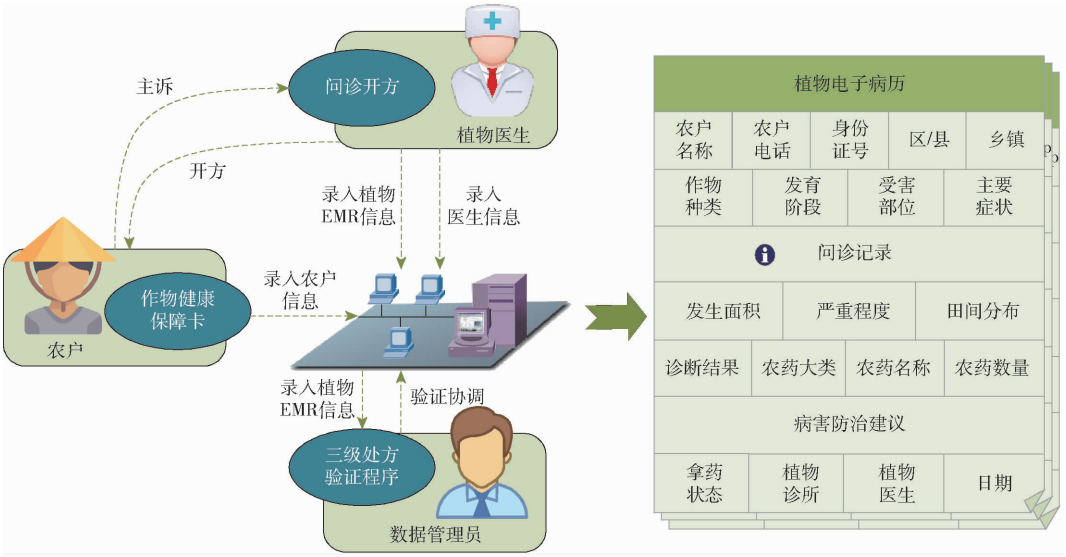

图 1 植物诊所病历填写流程

Fig. 1 Flow of filling out medical records in plant clinic

1.2 数据分析

植物诊所开具的电子病历记录了农户在农作物种植过程中遇到的病虫害问题及植物医生提供的诊疗建议,包括问诊农户、植物诊所、作物和症状的详细信息,以及诊断结果和病虫害防治建议等内容。其中不仅有结构化的信息,如就诊时间、诊所位置、发病时间、发病面积等,还包括非结构化的自由文本信息,如主要症状、检查所见、诊断结果、治疗建议等,如图 1 所示。本文收集了 2017 年 11 月到 2021 年 7 月期间约 44 个月的数据,共 44 000 余条作物病害电子病历。本文获取的样本共包含 89 种病害,涉及 46 种植物。初步分析数据发现,北京市种植植物种类和病害种类繁多,但是大多数作物及相关病害的数量占总样本量比例在 1% 以下,样本量仅几十条,属于冷门作物或不常见病害。从过少的样本中无法提取足够的有效信息以支持智能诊断,因此本文选择常见作物的常见病害为研究对象。常见的种植作物包括番茄、黄瓜、生菜和西瓜,共占全部样本的 95.68%。其次,在这 4 种作物的相关病害中选择样本量大于 400 条的病害,共 15 种,占全部样本的 74.69%。具体包括:番茄晚疫病、番茄病毒病、黄瓜霜霉病、番茄灰霉病、黄瓜角斑病、黄瓜白粉病、番茄线虫、番茄早疫病、番茄叶霉病、西瓜炭疽病、生菜线虫、黄瓜灰霉病、生菜霜霉病、黄瓜线虫和西瓜线虫。其中番茄病害 6 种、黄瓜病害 5 种、生菜病害和西瓜病害各 2 种。相比原始数据,这样的选择尽可能地覆盖了北京市常见作物病害,同时极大减缓了样本过少导致的误诊问题。

1.3 数据预处理

研究中未获取原始植物电子病历中涉及个人隐私的信息,例如农户姓名、手机号、身份证号码等。此外,由于数据存在冗余、缺失等问题,剔除了部分缺失率过高的记录和字段,对冗余记录进行了去重处理。

对于电子病历的非结构化数据部分,采用正则表达式、字符匹配等方法提取了问诊记录中的主诉部分,并且对空格、换行等字符进行清理。对于电子病历的结构化数据部分,进行删除重复值、缺失值处理、一致化处理和异常值处理等操作。

经过上述清洗和筛选,最终获得 22 842 条数据。每种病害的具体数据统计结果如图 2 所示。其中,番茄晚疫病和番茄病毒病为较大样本病害,数量超过 4 000 条;黄瓜霜霉病、番茄灰霉病、黄瓜角斑病和黄瓜白粉病次之,数量在 1 000 条以上;番茄线虫、番茄早疫病、番茄叶霉病、西瓜炭疽病、生菜线虫、黄瓜灰霉病、生菜霜霉病、黄瓜线虫和西瓜线虫为小样本病害,数量在 1 000 条以下。

利用分层抽样的方法,选取 60% 的数据作为训练集来训练模型,20% 作为测试集,20% 作为验证集。数据处理流程如图 3 所示。

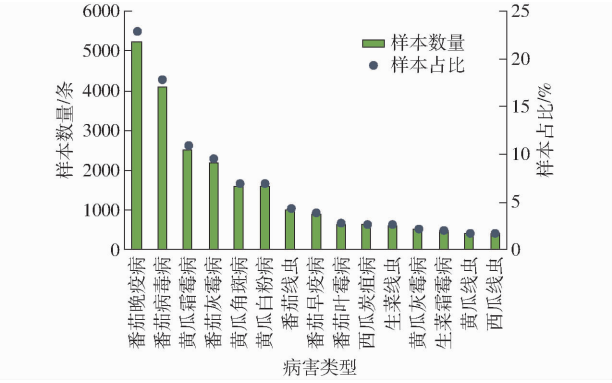

图 2　数据统计结果

Fig. 2　Statistical results

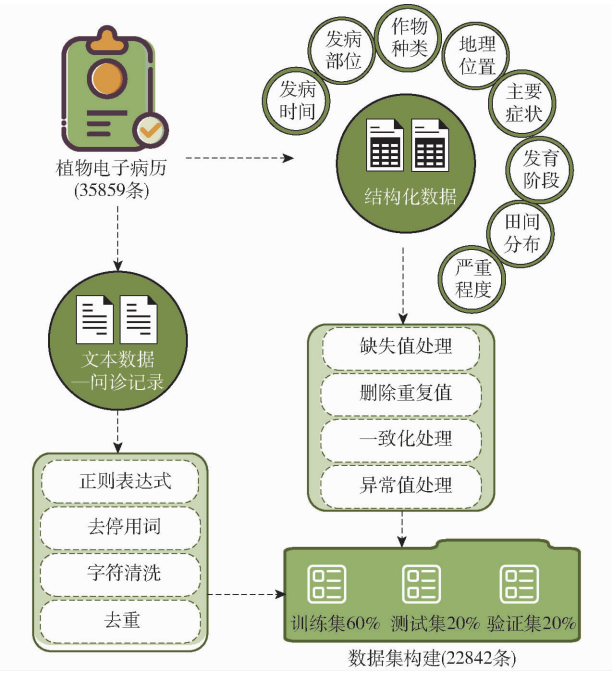

图 3　数据处理流程图

Fig. 3　Data pre-processing flowchart

2　病害诊断方法

2.1　模型结构

为了融合不同类型结构的数据，提出了一种用于病害诊断的多类型数据融合深度学习模型 BM-Att。通过结构化和非结构化数据之间优势互补，可以使融合后的效果超过任意单一类型数据能够达到的效果。其主要包含：文本数据处理模块和结构化数据处理模块。模型整体架构如图 4 所示，文本数据处理模块采用预训练语言模型对电子病历中的问诊记录进行建模。结构化数据处理模块基于电子病历中的作物种类、主要症状、生长阶段、发病部位、发病时间、地理位置、田间分布和严重程度等结构化数据进行建模，鉴于结构化数据的维度远低于文本数据，模型中加入多层感知器(Multi-layer perceptron，MLP)对结构化信息进行抽取，并将其扩增到适当的维度。最后进行基于注意力机制的特征融合，综合了结构化数据分支和文本分支获得的关键特征进行预测。

2.2　文本数据处理分支

植物电子病历问诊记录描述了病害的主要症状和外界环境，其中包含了大量信息，采用合适的算法对这些记录关键信息进行提取有利于更好地诊断作物病害。但是受限于植物电子病历数据数量的不足，难以获得足够信息实现有效的语义信息抽取。预训练语言模型的发展将自然语言处理(Natural language processing，NLP)领域的研究提升到了一个新的阶段，即无需人工标签，可以从海量的语料中学习到通用的语言表示，并显著提升下游的任务[18]。BERT 是 2018 年谷歌人工智能(Google AI)团队提出的在大规模通用领域语料库预先训练的语言模型，通过对下游任务进行微调的方式，在各种文本分类任务上有着明显的改进[19-20]。

BERT 是由双向变换器(Transformer)模型组成的编码结构。Transformer 用作特征提取器，它比 LSTM 特征提取功能更强大。同时，多层双向 Transformer 可以提取上下文的文本特征[21]。相比使用单向 Transformer 的 GPT 和双向 LSTM 编码的 ELMO，BERT 通过使用双向变压器编码器提取上下文特征，该编码器具有更深的层次和更好的并行性[13]。多头注意力机制是 Transformer 层的核心，其主要思想是通过计算词与词之间的关联度来调整词的权重，反映了该词与这句话中所有词之间的相互关系以及每个词的重要程度。首先，输入序列 $\boldsymbol{T}$，通过线性变换得到目标矩阵 $\boldsymbol{Q}$、上下文矩阵 $\boldsymbol{K}$ 以及原始矩阵 $\boldsymbol{V}$。然后通过计算放缩点积求得自注意力(Self-attention)的分数值，该分数值决定了当模型对一个词进行编码时，对输入句子的其他词的关注程度，具体计算公式为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中　d_k——输入维度

最后，将经过 i 次计算的 Self-attention 分数值进行拼接和线性变换，最终获得一个与原始字向量长度相同的增强语义向量，作为多头注意力层的输出，具体计算公式为

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(h_1,h_2,\cdots,h_i)\boldsymbol{W}^0 \tag{2}$$

其中　$$h_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^q,\boldsymbol{K}\boldsymbol{W}_i^k,\boldsymbol{V}\boldsymbol{W}_i^v) \tag{3}$$

式中　h_i——多头注意力层的输出

$\boldsymbol{W}_i^q$——$\boldsymbol{Q}$ 对应的 h_i 的权重矩阵

$\boldsymbol{W}_i^k$——$\boldsymbol{K}$ 对应的 h_i 的权重矩阵

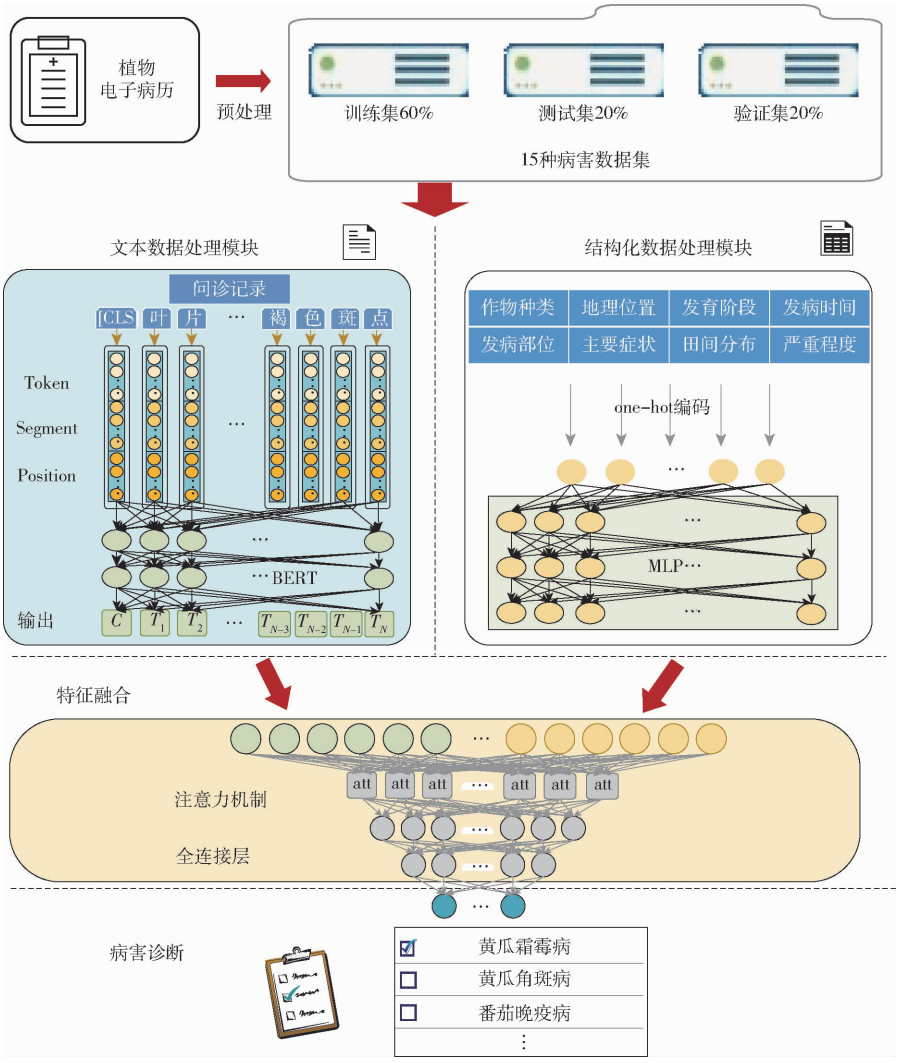

图 4 基于多类型数据融合的病害诊断模型框架

Fig. 4 Framework of a disease diagnosis model based on fusion of multiple types of data

$\boldsymbol{W}_i^v$——$\boldsymbol{V}$ 对应的 h_i 的权重矩阵

$\boldsymbol{W}^0$——附加的权重矩阵

BERT 已在多个下游 NLP 任务(例如机器翻译,命名实体识别、文本分类、阅读理解和问题解答)中取得了显著的效果。在本文的实验中,使用了谷歌发布的中文预训练模型“BERT - Base, Chinese”。该模型采用了 12 层的 Transformer,输出为 768 维度的向量,多头注意力的参数为 12,模型总参数量为 110 MB,共包含约 2 万个中文简体字和繁体字,含有部分英文单词和数字。将模型载入后,可以直接输出训练好的字向量或句向量。本文使用该模型获取句向量并将其作为后续网络模型的部分输入。

2.3 结构化数据处理分支

针对结构化的数据,首先对其进行独热编码,即 1 位有效编码。独热编码使用 N 位状态寄存器来对 N 个状态进行编码,每个状态都有它独立的寄存器位,并且其中只有 1 位有效。本文对电子病历中的作物种类、主要症状、生长阶段、发病部位、发病时间、地理位置、田间分布和严重程度等特征完成独热编码。例如,将生长阶段属性对应的苗期、生长期、开花期、结果期、成熟期、收获期共 6 个阶段转换为 6 位的标签值。具体编码内容如图 5 所示。

对植物电子病历中的结构化数据完成编码后,得到 107 维的向量。与文本分支得到的 768 维向量相比,维度差异较大。为了避免低维度的信息被高维度的信息淹没,减少数据融合前的信息损失,需要在数据融合之前增加低维模态[22-23]。因此,本文进一步采用 MLP 来扩增结构化数据向量表示的维度。

本文所使用 MLP 网络共有 107 个输入节点,512 个输出节点,以及 3 层包括 512 个隐藏节点的隐藏层。通过输入层、隐藏层、输出层进行前向传播得到预测值,利用均方误差(MSE)损失函数来衡量真实值和预测值之间的误差。MSE 损失函数利用梯度下降的方法来不断更新权重参数和偏置参数,不断缩小误差,最终获取和真实值最接近的预测结果。

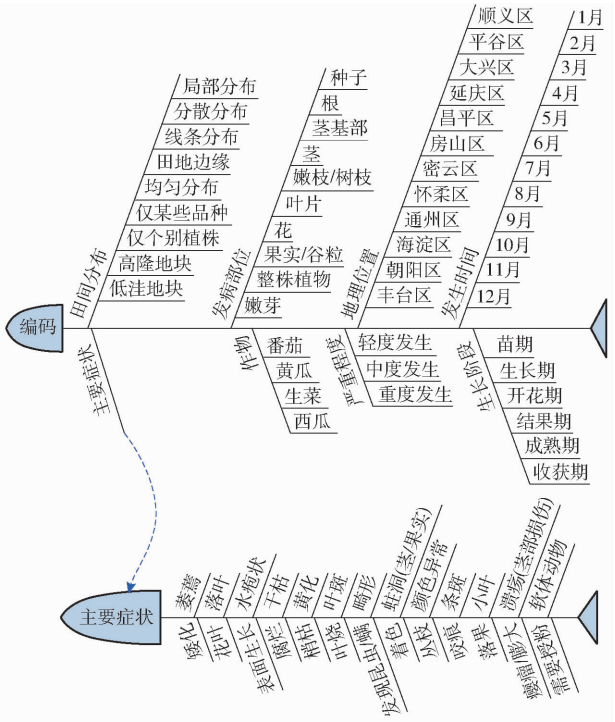

图 5　编码鱼骨图

Fig. 5　Fishbone diagram of data coding

2.4　特征融合

对于文本数据，BERT 将每条问诊记录转换为 768 维向量作为特征表示。同时，为了确保不同数据可以在相同尺度上被融合，结构化数据处理模块将 107 维的结构化数据扩增到 512 维进行特征表示。得到每种类型数据的向量表示之后，一般而言，有多种进行向量融合的方法，例如相加、相乘、拼接等。由于向量拼接能够最大程度地保留不同向量中的信息，因此选择拼接的方式。拼接后形成 1 280 维向量。然后，使用注意力机制和 3 层全连接层抽取重要的特征，同时对向量进行适当压缩。最后使用一个带有 Sigmoid 激活函数的神经元对作物病害进行判断。本文使用的注意力机制是在拼接向量 $\boldsymbol{f}_{ij}$ 的基础上，计算每一个元素的概率权重 α_{ij}，即 $\boldsymbol{u}_{ij}$ 和 $\boldsymbol{f}_{ij}$ 之间的映射。

$$\boldsymbol{u}_{ij} = \tanh(\boldsymbol{W}\boldsymbol{f}_{ij}) \tag{4}$$

$$\alpha_{ij} = \frac{\exp(\boldsymbol{u}_{ij}^{\mathrm{T}}\boldsymbol{u}_{w})}{\sum_{j}\exp(\boldsymbol{u}_{ij}^{\mathrm{T}}\boldsymbol{u}_{w})} \tag{5}$$

$$\boldsymbol{c} = \sum_{j}\alpha_{ij}\boldsymbol{f}_{ij} \tag{6}$$

式中　$\boldsymbol{W}$——参数矩阵，需要从训练数据里学习

$\boldsymbol{c}$——最后注意力，由权重和相应的键值进行加权求和得到

$\boldsymbol{u}_{ij}$——隐藏表示

$\boldsymbol{u}_{w}$——上下文向量，需从训练数据里学习

3　实验设计与结果分析

3.1　实验环境

本研究分类模型实验是基于 Pytorch 的深度学习库实现的。所有模型都采用 NVIDIA Quadro P2000 GPU(5GB)训练。

文本数据处理分支里对输入文本进行长截短填，将输入数据的最大文本长度(Pad-size)设定为 80。若文本真实长度小于 Pad-size，则用 0 填充；若文本真实长度超出 Pad-size，则截断至 Pad-size。为了防止过度拟合，加入了 0.1 的 Dropout。

训练过程中使用随机梯度下降法，经过多次参数调整，最终将学习率设置为 3×10^{-5}。训练共进行 4 个 Epoch(每个 Epoch 约 900 次迭代)，批量大小设置为 16，并对所有的参数应用权重衰减为 0.01 的 L2 正则化。

3.2　评价指标

由于作物病害诊断数据集的不平衡性，准确率作为评价指标容易产生误导，F1 值(F1-score)更适用作为不平衡数据集分类任务的模型评价指标。因此，本文采用精确率(Precision)、召回率(Recall)、F1 值和准确率(ACC)来评价模型对不同病害的分类效果，并使用宏平均法来评估模型的整体性能。此外，还使用混淆矩阵来检查每种作物病害诊断正确率。

3.3　实验结果

3.3.1　消融实验

本研究构建的 BM－Att 模型核心在于文本数据和结构化数据的特征提取和融合，通过添加处理结构化数据的 MLP 模块和特征融合阶段的注意力模块改进基于 BERT 的原始诊断模型。为了验证提出的 BM－Att 病害诊断模型的有效性，在 15 种病害数据集上进行了消融实验研究，并使用测试集评估模型性能。在模型训练中，使用仅处理文本数据的 BERT 作为基线模型，分别通过添加结构化数据处理分支和注意力机制方式优化 BERT 模型，得到 BERT－MLP 和 BERT－Att 2 个优化模型。最后，通过在处理文本数据的 BERT 模型之外，同时添加处理结构化数据的 MPL 模块和数据融合后的注意力机制获得 BM－Att(BERT＋MLP＋Att)。表 1 为不同模型诊断结果的准确率、精确率、召回率和 F1 值。

结果表明，消融实验中的各个模型结果的 F1 值都优于仅处理文本数据的原始 BERT，并且本文提出的 BM－Att 模型准确率最高。

相对于 BERT 模型，BERT－MLP 模型的准确率、精确率、召回率和 F1 值分别提高 1.16、1.27、

表 1 各模型消融实验结果对比

Tab. 1 Comparison of each model in ablation studies

%

模型	BERT	MLP	Attention	准确率	精确率	召回率	F1 值
BERT	√			94.23	93.58	91.98	92.47
BERT - MLP	√	√		95.39	94.85	94.26	94.38
BERT - Att	√		√	94.14	94.81	92.20	92.69
BM - Att	√	√	√	95.82	96.38	95.48	95.85

注:√表示该模型含有该模块。

2.28、1.91 个百分点,主要由于结构化数据带来的不同于文本数据的信息,为病害诊断特征融合模型提供了有效基础。

相对于 BERT 模型,BERT - Att 模型的精确率、召回率和 F1 值分别提高 1.23、0.22、0.22 个百分点,但是准确率略有下降,整体变化不大。主要由于 BERT 预训练语言模型的 Transformer 层包含多头注意力机制,能够有效提取文本语义特征,并在下游分类任务上实现很好的效果,额外添加注意力机制没有太大变化。

相对于 BERT 模型,BM - Att 模型的准确率、精确率、召回率和 F1 值分别提高 1.59、2.8、3.5、3.38 个百分点,F1 值宏平均值达到 95.85%,主要由于添加了 MLP 结构化数据处理模块和在特征融合阶段添加注意力机制能够有效改进融合效果。

3.3.2 单文本分支影响

植物电子病历中的文本数据包含了大量重要信息,能否对其进行有效提取是实现病害智能诊断的关键。为了验证提出的 BM - Att 模型在文本信息提取方面的有效性,比较多种具有代表性的文本数据处理模型。在 15 种常见作物病害数据集上,对比了 BERT 语言模型同其它处理文本数据的常见模型:CNN、RCNN、带有注意力机制的 RNN(AttRNN)、FastText、Transformer 和 ERNIE。表 2 为不同模型诊断结果的精确率、召回率、F1 值和准确率。

表 2 文本分支模型结果对比

Tab. 2 Comparison of model results of text branches

%

模型	准确率	精确率	召回率	F1 值
CNN	94.01	92.97	91.73	92.02
RCNN	94.17	91.93	92.04	91.55
AttRNN	94.04	92.16	92.22	91.91
FastText	89.60	88.69	84.54	84.52
Transformer	91.73	90.39	89.28	89.23
ERNIE	94.33	92.91	92.44	92.38
BERT	94.23	93.58	91.98	92.47

结果显示,本文提出的模型在所有指标上均取得了最好的结果,在测试集的准确率、精确率、召回率和 F1 值的宏平均值分别达到 95.82%、96.38%、95.48% 和 95.85%,准确率比 ERNIE 高出 1.49 个百分点,F1 值比 BERT 高出 3.38 个百分点,这证明了结构化数据的加入有助于诊断准确率的提升。

在其它模型中,基于预训练的深度学习模型 BERT 和 ERNIE 取得了最好的效果。具体来说,BERT 取得了更高的精确率和 F1 值,分别达到 93.58% 和 92.47%,而 ERNIE 取得了更好的召回率和准确率,分别达到 92.44% 和 94.33%。预训练语言模型可以从海量的语料中学习到通用的语言表示,并显著提升深度学习在下游 NLP 任务中的表现[18]。

3.3.3 结构化数据影响

经过 one-hot 编码,植物电子病历中的结构化数据转换为一维数据。为了验证所提出的 BM - Att 模型对一维数据信息提取的有效性,比较了结构化数据的不同处理方式对模型效果的影响。所提出的 BM - Att 模型在结构化数据处理分支采用的是 MLP 模型,将其与具有代表性的一维数据分类模型进行比较,包括一维 ALEX、一维卷积神经网络(1dCNN)、一维长短期记忆网络(1dLSTM)和基于注意力机制的一维长短期记忆网络(1dAttLSTM)。在实验中,选择文本分支中效果较好的 2 种预训练模型 BERT 和 ERNIE 作为基础模型,添加上述 5 种处理方式的结构化数据分支,对比不同数据融合模型的诊断效果。表 3 为 BERT - ALEX、BERT - 1dCNN、BERT - 1dLSTM、BERT - 1dAttLSTM、BERT - MLP、ERNIE - ALEX、ERNIE - 1dCNN、ERNIE - 1dLSTM、ERNIE - 1dAttLSTM、ERNIE - MLP 等模型在测试集上的准确率、精确率、召回率和 F1 值。

表 3 结构化数据分支模型结果对比

Tab. 3 Comparison of model results for structured data branches

%

文本分支	结构化数据分支	准确率	精确率	召回率	F1 值
BERT	ALEX	93.88	92.12	91.45	91.39
	1dCNN	94.53	94.69	92.02	93.05
	MLP	95.39	94.85	94.26	94.38
	1dLSTM	94.39	93.67	92.24	92.62
	1dAttLSTM	94.02	93.56	91.65	92.21
ERNIE	ALEX	94.11	93.35	92.20	92.44
	1dCNN	94.24	93.89	91.84	92.46
	MLP	94.62	93.69	93.37	93.29
	1dLSTM	94.52	93.87	92.33	92.80
	1dAttLSTM	94.04	92.96	91.88	92.12

结果显示,不管文本处理分支采用的是 BERT

还是 ERNIE，相比 ALEX、1dCNN、1dLSTM 以及基于注意力机制的 1dLSTM（1dAttLSTM），MLP 均取得了最好的结果。尤其是 BERT－MLP，F1 值和准确率分别为 94.38% 和 95.39%。对于植物电子病历，结构化数据分支处理的是低维度数据。一般来说，MPL 简明的结构适用于一维数据并且在高噪声条件下表现良好，在医学诊断领域有着广泛应用[24－26]。

3.3.4　注意力机制影响

表 4 为注意力机制对模型的影响。在数据融合阶段，BM－Att 模型引入注意力机制提取重要特征，以优化特征融合效果。结果显示，注意力机制在各项指标上均提升了模型的分类效果。相比没有添加注意力机制的模型，在数据融合阶段增加注意力层的模型在大多数病害类别上达到了更好的分类效果。尤其是对于生菜、西瓜等小样本病害，诊断效果有了明显的改进，将西瓜线虫诊断的 F1 值从 90.64% 提高到 100%。从整体上看，F1 值的宏平均值提高 1.47 个百分点，F1 加权平均值和准确率分别提高 0.42、0.43 个百分点。相比加权平均值，宏平均值更加注重模型对不平衡数据集中较小样本的分类效果。注意力机制通过权重分配融合全局信息，选择性地强调关键特征，发掘模型对不同类别数据特征的关注点，改进了模型的重要特征提取能力[27－28]。在数据融合阶段添加的注意力机制让模型对于小样本病害投入了更多的关注，这对病害智能诊断的实际应用具有重要意义。

表 4　注意力机制影响对比

Tab.4　Comparison of effects of attentional mechanisms

%

模型	病害种类	精确率	召回率	F1 值
BERT－MLP	番茄病毒病	98.94	99.11	99.03
	番茄灰霉病	97.19	89.77	93.33
	番茄晚疫病	94.19	98.03	96.07
	番茄线虫	98.62	95.65	97.11
	番茄叶霉病	89.37	95.36	92.27
	番茄早疫病	91.57	84.13	87.69
	黄瓜白粉病	98.63	90.36	94.31
	黄瓜灰霉病	96.64	92.9	94.74
	黄瓜角斑病	97.33	90.48	93.78
	黄瓜霜霉病	91.06	98.15	94.48
	黄瓜线虫	97.06	80.49	88.00
	生菜霜霉病	96.77	100	98.36
	生菜线虫	100	100	100
	西瓜炭疽病	92.50	99.46	95.85
	西瓜线虫	82.88	100	90.64
	准确率		95.39	
	宏平均值	94.85	94.26	94.38
	加权平均值	95.54	95.39	95.37
BM－Att（本文模型）	番茄病毒病	99.43	99.43	99.43
	番茄灰霉病	98.01	90.23	93.96
	番茄晚疫病	93.39	98.03	95.65
	番茄线虫	99.30	95.32	97.27
	番茄叶霉病	90.59	94.33	92.42
	番茄早疫病	90.73	83.03	86.71
	黄瓜白粉病	98.86	91.19	94.87
	黄瓜灰霉病	97.30	92.90	95.05
	黄瓜角斑病	97.52	89.65	93.42
	黄瓜霜霉病	90.78	98.68	94.56
	黄瓜线虫	99.19	100	99.60
	生菜霜霉病	98.04	100	99.01
	生菜线虫	100	100	100
	西瓜炭疽病	92.50	99.46	95.85
	西瓜线虫	100	100	100
	准确率		95.82	
	宏平均值	96.38	95.48	95.85
	加权平均值	95.94	95.82	95.79

4　结论

（1）针对电子病历中的文本数据部分，采用预训练 BERT 语言模型提取深层语义特征。对比 CNN、RCNN、带有注意力机制的 RNN、FastText、Transformer、ERNIE 和 BERT 模型，发现与传统深度学习模型相比，基于预训练的 BERT 语言模型取得了好的效果，F1 值达到 92.47%。原因是预训练语言模型可以从海量的语料中学习到通用的语言表示，能更好地提取文本中的重要信息。

（2）针对电子病历中的结构化数据部分，对比不同处理方式对模型效果的影响，结果显示，不管文本处理分支采用的是 BERT 还是 ERNIE，在结构化数据分支采用 MLP 均为最佳选择。相比 ALEX、1dCNN、1dLSTM 以及基于注意力机制的 1dAttLSTM 等复杂网络，MPL 的多层全连接层结构对于编码形成的低维度数据更加简明有效。

（3）在数据融合阶段添加注意力机制，通过权重分配强调关键信息特征，发掘模型对不同类别数据特征的关注点，改进了模型对重要特征的提取能力，将 F1 值宏平均值提高 1.47 个百分点。更重要的是，注意力机制能显著改进模型对小样本类别的分类效果，将西瓜线虫诊断的 F1 值从 90.64% 提高到 100%，对于病害智能诊断的实际应用具有重要意义。

（4）综合来看，本文提出的方法在基于植物电子病历的 15 种常见作物病害的诊断任务上结果最佳，在所有指标上均取得了最好的结果，F1 值宏平均值和准确率分别达到 95.85% 和 95.82%，表明本文模型能够实现作物病害的有效诊断。

参 考 文 献

[1] 徐畅,丁俊琦,赵聃桐,等. 基于 LightGBM 和处方数据的番茄病害诊断方法[J]. 农业机械学报,2022,53(9):286 - 294.
XU Chang, DING Junqi, ZHAO Dantong, et al. Tomato disease diagnosis method based on LightGBM and prescription data [J]. Transactions of the Chinese Society for Agricultural Machinery, 2022,53(9):286 - 294. (in Chinese)

[2] NI P, LI Y, ZHU J, et al. Disease diagnosis prediction of EMR based on BiGRU - ATT - Capsnetwork model[C]//2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019: 6166 - 6168.

[3] MILED Z B, HAAS K, BLACK C M, et al. Predicting dementia with routine care EMR data[J]. Artificial Intelligence in Medicine, 2020, 102: 101771.

[4] SONG J, WANG Y, TANG S, et al. Local-global memory neural network for medication prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(4): 1723 - 1736.

[5] AZAR A T, EL-METWALLY S M. Decision tree classifiers for automated medical diagnosis[J]. Neural Computing and Applications, 2013, 23(7 - 8): 2387 - 2403.

[6] HAMOUD A. Applying association rules and decision tree algorithms with tumor diagnosis data[J]. International Research Journal of Engineering and Technology, 2017, 3(8): 27 - 31.

[7] GOKULNATH C B, SHANTHARAJAH S P. An optimized feature selection based on genetic approach and support vector machine for heart disease[J]. Cluster Computing, 2019, 22(6): 14777 - 14787.

[8] SHEN Y, ZHANG L, ZHANG J, et al. CBN: constructing a clinical bayesian network based on data from the electronic medical record[J]. Journal of Biomedical Informatics, 2018, 88: 1 - 10.

[9] SEGURA-BEDMAR I, COLÓN-RUÍZ C, TEJEDOR-ALONSO M Á, et al. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches[J]. Journal of Biomedical Informatics, 2018, 87: 50 - 59.

[10] ZHANG K, LIU C, DUAN X, et al. Bert with enhanced layer for assistant diagnosis based on chinese obstetric EMRS[C]// 2019 International Conference on Asian Language Processing (IALP). IEEE, 2019: 384 - 389.

[11] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. ArXiv Preprint ArXiv:1802.05365, 2018.

[12] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. [2022 - 04 - 13]. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. ArXiv Preprint ArXiv:1810.04805, 2018.

[14] 杨国峰, 杨勇. 基于 BERT 的常见作物病害问答系统问句分类[J]. 计算机应用, 2020, 40(6): 1580 - 1586.
YANG Guofeng, YANG Yong. Question classification of common crop disease question answering system based on BERT[J]. Computer Applications, 2020, 40(6): 1580 - 1586. (in Chinese)

[15] 李林, 刁磊, 唐詹,等. 基于 BERT_Stacked LSTM 的农业病虫害问句分类方法[J]. 农业机械学报, 2021, 52(增刊): 172 - 177.
LI Lin, DIAO Lei, TANG Zhan, et al. Question classification method of agricultural diseases and pests based on BERT_Stacked LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(Supp.): 172 - 177. (in Chinese)

[16] 杨森, 冯全, 张建华, 等. 基于深度学习与复合字典的马铃薯病害识别方法[J]. 农业机械学报, 2020, 51(7): 22 - 29.
YANG Sen, FENG Quan, ZHANG Jianhua, et al. Identification method for potato disease based on deep learning and composite dictionary[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(7): 22 - 29. (in Chinese)

[17] 王春山, 赵春江, 吴华瑞, 等. 采用双模态联合表征学习方法识别作物病害[J]. 农业工程学报, 2021, 37(11): 180 - 188.
WANG Chunshan, ZHAO Chunjiang, WU Huarui, et al. Recognizing crop diseases using bimodal joint representation learning [J]. Transactions of the CSAE, 2021, 37(11): 180 - 188. (in Chinese)

[18] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: a survey[J]. Science China Technological Sciences, 2020, 63(10): 1872 - 1897.

[19] GONZÁLEZ-CARVAJAL S, GARRIDO-MERCHÁN E C. Comparing BERT against traditional machine learning text classification[J]. ArXiv Preprint ArXiv:2005.13012, 2020.

[20] AL-GARADI M A, YANG Y C, CAI H, et al. Text classification models for the automatic detection of nonmedical prescription medication use from social media[J]. BMC Medical Informatics and Decision Making, 2021, 21(1): 1 - 13.

[21] SONG Z, XIE Y, HUANG W, et al. Classification of traditional Chinese medicine cases based on character-level bert and deep learning[C]//2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2019: 1383 - 1387.

[22] 陈先来, 贾一珍, 安莹, 等. 基于多类型数据的胃癌风险预测融合模型[J]. 计算机工程, 2022, 48(9): 254 - 261.
CHEN Xianlai, JIA Yizhen, AN Ying, et al. Fusion model of gastric cancer risk prediction based on multi-type data[J]. Computer Engineering, 2022, 48(9): 254 - 261. (in Chinese)

[23] YAN R, ZHANG F, RAO X, et al. Richer fusion network for breast cancer classification based on multimodal data[J]. BMC Medical Informatics and Decision Making, 2021, 21(1): 1 - 15.

(下转第 223 页)

[13] KAMARUDIN J A M, ABDULLAH A, SALLEHUDDIN R. A review of deep learning architectures and their application[C]// 17th Asia Simulation Conference, 2017: 7522.
[14] 刘媛媛,张凡,师琪,等.基于高光谱和集成学习的库尔勒香梨黑斑病潜育期诊断[J]. 农业机械学报,2022,53(6):295－303.
LIU Yuanyuan, ZHANG Fan, SHI Qi, et al. Diagnosis of Korla pear black spot disease in incubation period based on hyperspectral imaging and ensemble learning algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(6): 295－303. (in Chinese)
[15] 孙红,李松,李民赞,等. 农业信息成像感知与深度学习应用研究进展[J]. 农业机械学报,2020,51(5):1－17.
SUN Hong, LI Song, LI Minzan, et al. Research progress of image sensing and deep learning in agriculture[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(5): 1－17. (in Chinese)
[16] 郭祥云,台海江. 深度学习在大田种植中的应用及展望[J]. 中国农业大学学报,2019,24(1):119－129.
GUO Xiangyun, TAI Haijiang. Current situation and prospect of deep learning application in field planting[J]. Journal of China Agricultural University, 2019, 24(1): 119－129. (in Chinese)
[17] 任志尚,彭慧慧,贺壮壮,等. 基于高光谱成像技术的面条中马铃薯全粉含量检测[J]. 农业机械学报,2020,51(增刊2):466－470,506.
REN Zhishang, PENG Huihui, HE Zhuangzhuang, et al. Detection of potato powder addition in noodles based on hyperspectral imaging[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp. 2): 466－470, 506. (in Chinese)
[18] ABDULRIDHA J, BATUMAN O, AMPATZIDIS Y. UAV-based remote sensing technique to detect citrus canker disease utilizing hyperspectral imaging and machine learning[J]. Remote Sensing, 2019, 11(11): 1373.
[19] 苏鸿,温国泉,谢玮,等. 基于区域卷积神经网络模型的广西柑橘病虫害识别方法研究[J]. 西南农业学报,2020,33(4):805－810.
SU Hong, WEN Guoquan, XIE Wei, et al. Research on citrus pest and disease recognition method in Guangxi based on regional convolutional neural network model[J]. Southwest China Journal of Agricultural Sciences, 2020, 33(4): 805－810. (in Chinese)
[20] BARI B S, ISLAM N, RASHID M, et al. A real-time approach of diagnosing rice leaf disease using deep learning-based Faster R－CNN framework[J]. PeerJ Computer Science, 2021, 432(7): 1－27.
[21] 黄敏,夏超,朱启兵,等. 融合高光谱图像技术与 MS－3DCNN 的小麦种子品种识别模型[J]. 农业工程学报,2021,37(18):153－160.
HUANG Min, XIA Chao, ZHU Qibing, et al. Recognizing wheat seed varieties using hyperspectral imaging technology combined with multi-scale 3D convolution neural network[J]. Transactions of the CSAE, 2021, 37(18): 153－160. (in Chinese)
[22] 张号逵,李映,姜晔楠. 深度学习在高光谱图像分类领域的研究现状与展望[J]. 自动化学报,2018,44(6):961－977.
ZHANG Haokui, LI Ying, JIANG Yenan. Deep learning for hyperspectral imagery classification: the state of the art and prospects[J]. Acta Automatica Sinica, 2018, 44(6): 961－977. (in Chinese)
[23] HACHAJ T, STOLINSKA A, ANDRZEJEWSKA M, et al. Deep convolutional symmetric encoder-decoder neural networks to predict students' visual attention[J]. Symmetry, 2021, 13(12): 13122246.
[24] TREISMAN A M, GELADE G. A feature-integration theory of attention[J]. Cognitive Psychology, 1980, 12(1): 97－136.
[25] TANG Y, HUANG Y, WU Z, et al. Question detection from acoustic features using recurrent neural network with gated recurrent unit[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016: 6125－6129.

(上接第 204 页)

[24] DESAI M, SHAH M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and convolutional neural network (CNN)[J]. Clinical EHealth, 2021, 4: 1－11.
[25] 尹柯, 巴文娟, 汤敏, 等. 多层感知器模型预测纯磨玻璃结节肺腺癌浸润程度[J]. 中国医学影像技术, 2020, 36(11): 1652－1656.
YIN Ke, BA Wenjuan, TANG Min, et al. Multilayer perceptron model in predicting infiltration degree of pure ground glass opacity lung adenocarcinoma[J]. Chinese Journal of Medical Imaging Technology, 2020, 36(11): 1652－1656. (in Chinese)
[26] 王德广, 黄盈朵. 基于 SVM－MLP 的乳腺癌预测[J]. 微型电脑应用, 2022, 38(1): 130－133,138.
WANG Deguang, HUANG Yingduo. Prediction of breast cancer based on SVM－MLP[J]. Microcomputer Applications, 2022, 38(1): 130－133,138. (in Chinese)
[27] 牛迪. 基于自注意力机制的多特征融合槽抽取模型[J]. 南京理工大学学报, 2022, 46(1): 69－75.
NIU Di. Multidimensional features slot filling model based on self-attention mechanism[J]. Journal of Nanjing University of Science and Technology, 2022, 46(1): 69－75. (in Chinese)
[28] 韩旭, 赵春江, 吴华瑞, 等. 基于注意力机制及多尺度特征融合的番茄叶片缺素图像分类方法[J]. 农业工程学报, 2021, 37(17): 177－188.
HAN Xu, ZHAO Chunjiang, WU Huarui, et al. Image classification method for tomato leaf deficient nutrient elements based on attention mechanism and multi-scale feature fusion[J]. Transactions of the CSAE, 2021, 37(17): 177－188. (in Chinese)