DOI: 10.12398/j.issn.2096-7217.2023.04.001

基于电子病历的作物病虫害关联挖掘及智能诊断

徐畅[1]，张领先[1*]，乔岩[2]

(1. 中国农业大学信息与电气工程学院，北京市，100083；2. 北京市植物保护站，北京市，100029)

摘要：作物病虫害研究是人工智能技术与智慧农业交叉领域的热点问题。现有的研究受到数据获取困难、技术实施成本高以及作物病虫害发生态势复杂等因素的限制。北京市“植物诊所”形成的植物电子病历(plant electronic medical records，PEMRs)为作物病虫害的诊断与防治提供了新的研究方向。PEMRs以多模态数据的形式存储，包含了丰富的植物信息、病虫害信息和环境信息，如何挖掘PEMRs信息并利用其辅助后续研究是亟待解决的问题。鉴于知识图谱的信息表示能力、机器学习的挖掘能力和深度学习的特征抽取能力，根据电子病历特点，利用结构化数据构建作物病虫害知识图谱，利用非结构化数据和领域知识进行知识增强，进一步利用Neo4j图数据库和图数据科学(graph data science，GDS)结合机器学习算法从“热”点发现、联系链路发现、相似病虫害发现3个维度进行关联挖掘。在此基础上，将基于Transformer的双向编码器(bidirectional encoder representation from transformers，BERT)与卷积神经网络(convolutional neural network，CNN)结合，利用非结构化文本数据实现文本特征抽取和病虫害诊断，模拟植物医生实现智能化服务，在20种常见病虫害上的综合准确率可达到93.13%。本研究可为作物病虫害的及时诊断、对症防治、科学用药和辅助决策提供理论支持，创新了农业科技社会化服务新模式、新业态。

关键词：作物病虫害；知识图谱；关联挖掘；植物电子病历；病虫害诊断；机器学习

中图分类号：S24；TP391 **文献标识码**：A **文章编号**：2096-7217(2023)04-0001-10

徐畅，张领先，乔岩. 基于电子病历的作物病虫害关联挖掘及智能诊断[J]. 智能化农业装备学报(中英文)，2023，4(4)：1－10

XU Chang，ZHANG Lingxian，QIAO Yan. Association mining of crop diseases and intelligent diagnosis based on electronic medical records [J]. Journal of Intelligent Agricultural Mechanization，2023，4(4)：1－10

0 引言

作物病虫害是制约农业可持续发展的主要因素之一，也是影响作物产量和品质的主要原因[1]，借助人工智能技术研究作物病虫害自动检测问题已成为目前的研究热点。作物病虫害研究需要大量的实测数据，目前相关研究主要集中在两个方面，一是基于深度学习的计算机视觉技术[2]，通过显微图像、多光谱图像、高光谱图像和遥感图像等成像技术无损地诊断作物病虫害；二是基于传统机器学习的多源数据融合技术，通过传感器、气象站和物联网等硬件技术监测作物病虫害的影响因子[3]。但是它们均会受到数据获取困难、技术实施成本高以及作物病虫害发生态势复杂等因素的限制。

由北京市“植物诊所”形成的电子病历为数字化赋能作物病虫害挖掘与决策提供了新的思路[4-5]，也是小农户与现代农业有效衔接的重要措施。作物病虫害是病原物、寄主作物和环境条件相互作用而产生的(3个要素的关系被称为作物病虫害三角关系，简称病三角原理)[6]。而电子病历中每条处方代表一个真实病虫害发生“点”，包含“病原物—寄主—环境”三者及其相关关系的重要信息。通过电子病历挖掘，可以多角度揭示作物病虫害关联关系及其变化规律。

目前生物医学研究领域已有较为深入的探索[7]，多项研究证明处方数据具有回溯性和可预测性及辅助构建临床决策支持系统的能力[8]。但是，农业领域电子病历应用还处于起步阶段。PEMRs作为一种未得到充分利用的多模态数据源，可以通过挖掘历史数据的潜在价值，发现新知识，因此对其进行深入挖掘具有一定的研究意义。中国农业大学张领先团队已在基于电子病历的作物病虫害智能诊断与处方推荐进行了初步探索[9]，徐畅等[4, 10]基于处方结构化信息出发，利用

收稿日期：2023-07-20 修回日期：2023-10-20

基金项目：国家自然科学基金(62176261，62376272)

第一作者：徐畅，女，1998年生，山东烟台人，博士研究生；研究方向为机器学习和数据挖掘。E-mail：xc199873@126.com

*通信作者：张领先，男，1971年生，河南信阳人，博士，教授；研究方向为农业信息化技术。E-mail：zhanglx@cau.edu.cn

Stacking集成学习和优化的LightGBM实现了不同场景下番茄作物的病虫害诊断;丁俊琦等[5, 11]和张领先等[12]基于处方的非结构化信息利用深度学习模型实现了多种作物的处方推荐。然而,作物病虫害发生原因复杂多变,各种因子相互关联、相互影响。因此仅从统计数据来确定特征和病虫害间更深层次的关系是一项困难的任务。如何高效地挖掘处方大数据并辅助后续作物病虫害研究,目前还是亟待解决的问题。

知识图谱、机器学习和预训练语言模型等人工智能技术发展为解决上述现实难点和痛点带来了新的契机,利用信息技术手段开展作物处方数据的管理和服务是一项前所未有的尝试。早期作物病虫害关联挖掘最常见的两种研究思路:一是利用Meta等分析方法系统分析公开发表的研究[13];二是利用农业物联网和传感器采集温湿度、风、光照等环境因子,结合机器学习模型探究区域范围影响病虫害的关键因素[14-15]。

知识图谱(knowledge graph, KG)的本质是一个由大量实体及其之间的关系组成的大规模知识库,用于可视化描述知识资源及其载体,挖掘、分析、构建、绘制和展示知识及它们之间的相互联系[16]。知识图谱作为一种语义网络,具有可扩展性强、支持智能应用等优点[17-18]。随着大数据和深度学习的发展,改善了人工开具处方效率低和成本高的问题,以BERT[19]为代表的预训练语言模型可以利用大规模文本数据的自监督性质来构建预训练任务,辅助抽取出良好的文本特征为文本分类提供了有效的方法[20],为基于处方文本数据进行诊断提供了可行性。同时以K-近邻(K-nearest neighbor, KNN)为代表的机器学习算法可以快速从大量数据中挖掘能够代表一类事物的规律,从而对事物进行预测和分类[21]。

针对上述问题,基于电子病历中的结构化数据构建作物病虫害知识图谱,将复杂的数据可视化,利用半结构化数据和领域知识对知识图谱进行增强,并存储于Neo4j图数据库。

进一步知识图谱的挖掘可从以下3个方面展开:高发病虫害、病虫害间的关联关系、相似病虫害发现,转化为计算机领域的查找“热”(popular)的节点,联系链路发现、相似节点等关联挖掘问题,利用GDS与机器学习模型相结合解决。在此基础上基于电子病历中的非结构化文本数据(问诊记录),将病虫害诊断转化为文本分类问题,利用BERT算法挖掘文本特征后利用CNN实现文本分类,模拟植物医生诊断思维模式[22]。

1 作物病虫害知识图谱的构建

1.1 作物病虫害多源数据采集与预处理

本研究数据主要包括3部分,分别是:北京市植物保护站提供的作物病虫害处方数据,Plantwise网站总结归纳的农业技术明白纸,百度百科、中国农业信息网、兴农网以及蔬菜网等权威网站。

1.1.1 处方数据

2012年,北京首次引入了国际应用生物科学中心植物诊所的先进概念。植物医生基于双向信息流的原则和有害生物综合防治(integrated pest management, IPM)原则为农户提供个性化精准的作物病虫害问诊服务,问诊过程形成标准的处方,通过3级数据验证程序后录入数据库。

1.1.2 农业技术明白纸

以动态数据为导向,处方数据库每年更新,明白纸每年进行动态开发和修订,共完成了覆盖全市22种主要作物在内的127种病虫害的明白纸,以农业防治和高效、低毒的化学农药等绿色防控技术为主,形成了一套动态、绿色、高效的病虫害解决方案库,基本实现了对主要作物常见病虫害问题的全覆盖。

1.1.3 权威网站

选择百度百科、中国农业信息网、兴农网以及蔬菜网等权威网站,利用XPath解析网页提取结构化数据,作为知识图谱的补充数据。

针对以上多源数据,将其划分为结构化数据和文本数据两部分,借助深度学习、知识计算和可视化等大数据分析技术进行预处理,并转化为统一表示形式。对结构化数据进行数据清洗、描述统计、数据分析、图表展示、数据编码和特征提取等;对文本数据进行数据清洗、中文分词、去停用词、词频统计、特征处理和特征提取等。将预处理后的多源数据作为后续构建病虫害知识图谱、关联挖掘及智能诊断的数据源,整体流程如图1所示。

1.2 知识图谱构建整体流程

作物病虫害知识图谱的构建流程如图2所示,将作物病虫害领域知识拓扑的构建归纳为数据获取、本体层构建、实体层构建和知识加工。作物病虫害领域知识图谱的构建过程分为4步:首先,根据领域知识、文献阅读和病虫害三角原理初步设计本体层;第2步通过实体识别、关系抽取等技术进行知识提取,从大量半结构化或非结构化的作物病虫害数据中提取知识图谱的元素,如实体和关系等,补充完善本体层;第3步,通过知识融合消除歧义,并链接知识库和作物病虫害领域的实体,以此增强知识的逻辑和表达能力,并通过

不断的调整和融合，完成实体层构建。最后，借助本体推理完成作物病虫害领域知识图谱的构建。

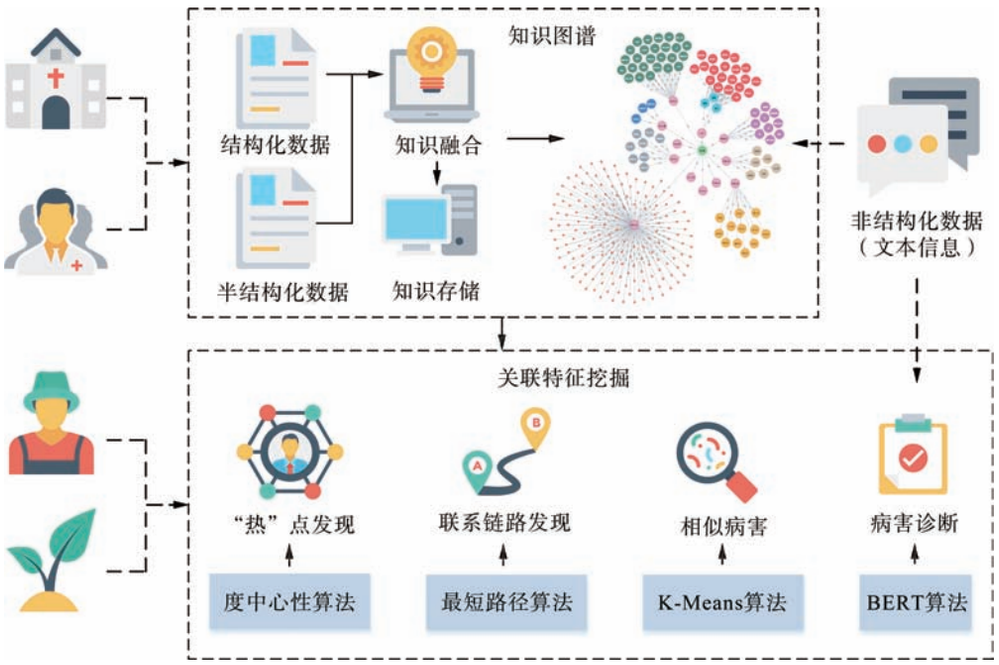


图 1　基于电子病历的作物病虫害知识图谱构建及关联挖掘

Figure 1　Construction of crop pest knowledge graph and correlation mining based on electronic medical records

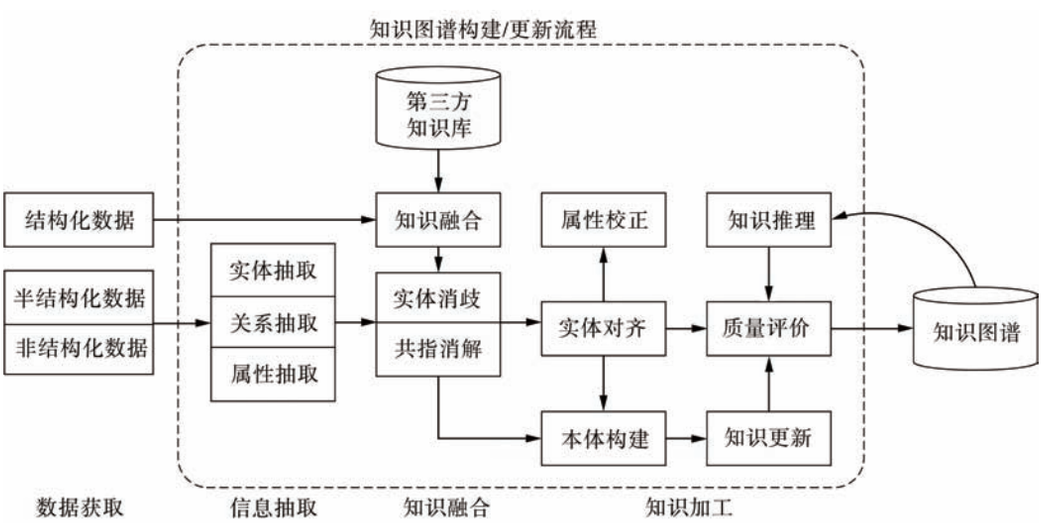


图 2　作物病虫害知识图谱的构建流程

Figure 2　Construction process of knowledge graph of crop pests and diseases

1.3　知识图谱本体层构建

以常见的 20 种作物病虫害为主要研究对象，在充分参考主要数据源的基础上，通过设计全面细粒度揭示病虫害的本体概要模型，为进一步详细设计本体模型奠定了基础。

病虫害领域本体概要模型图如图 3 所示。作物病虫害本体模型的设计与构建应与领域特点紧密结合，且该领域本体模型应具有可复用性。图 3 从核心描述对象——作物病虫害出发，根据病虫害三角原理为核心向外扩展。

1.4　知识图谱实体层构建

将已有实体层数据进行清洗整理，按照 1.3 节中设计的本体层框架导入。但是对于多源异构数据，通常会出现在实际场景中对应多种表达方式，例如缩写简写等，这种现象会导致融合后的数据存在大量冗余数据。

因此本研究进一步进行了实体消歧等数据处理操作。将清理整理后的数据，按照设计好的本体层框架导入，通过 Neo4j 查询语言 Cypher 来检索数据，得到知识图谱部分内容如图 4 所示。

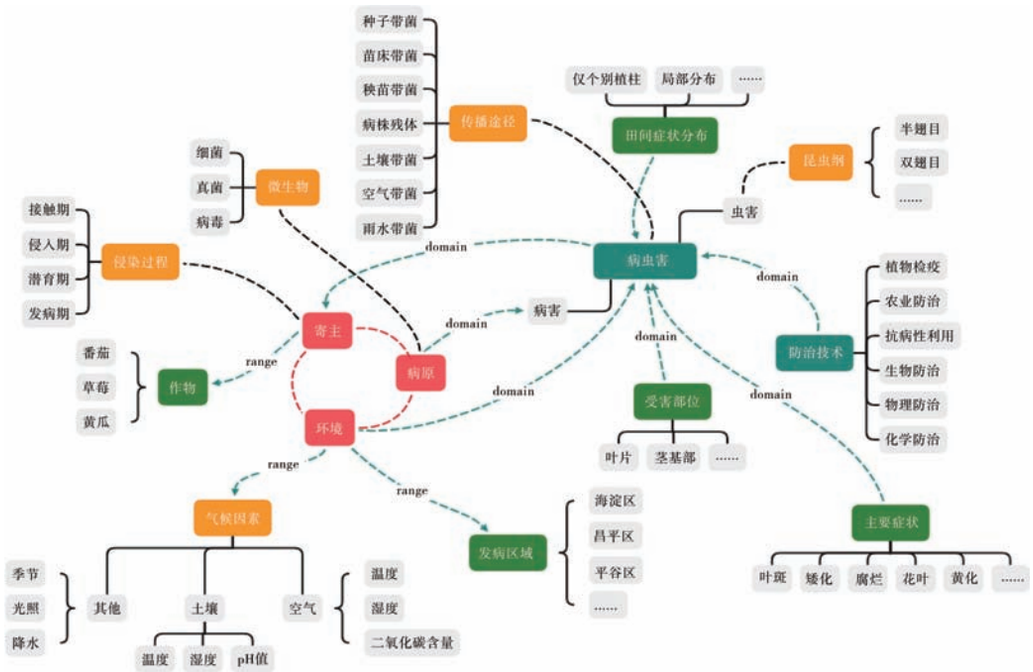


图 3　作物病虫害领域本体模型概要图

Figure 3　Summary of crop pest and disease domain ontology model

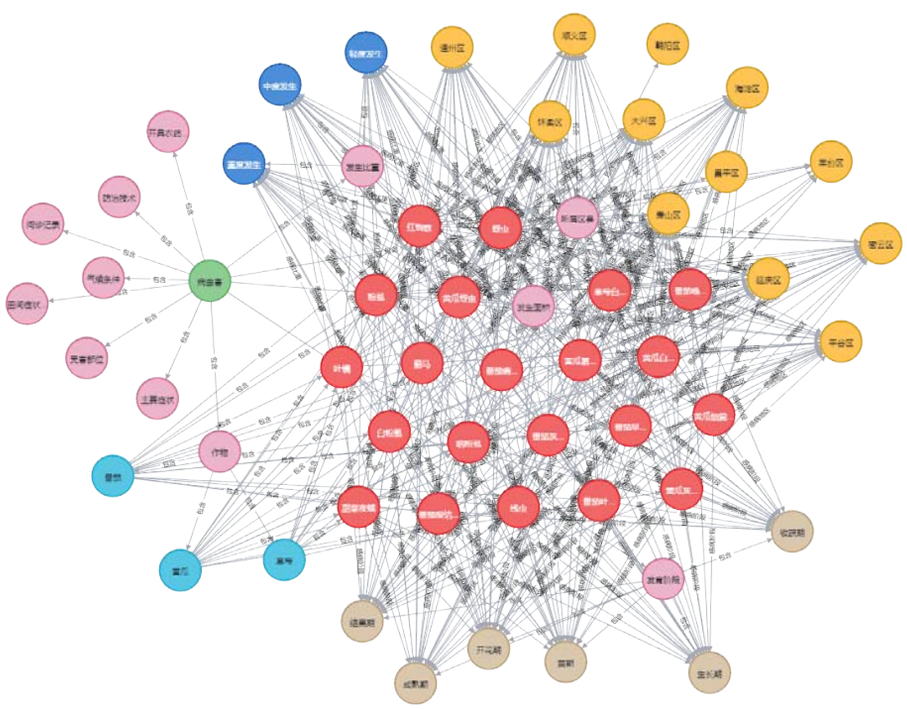


图 4　作物病虫害知识图谱部分展示

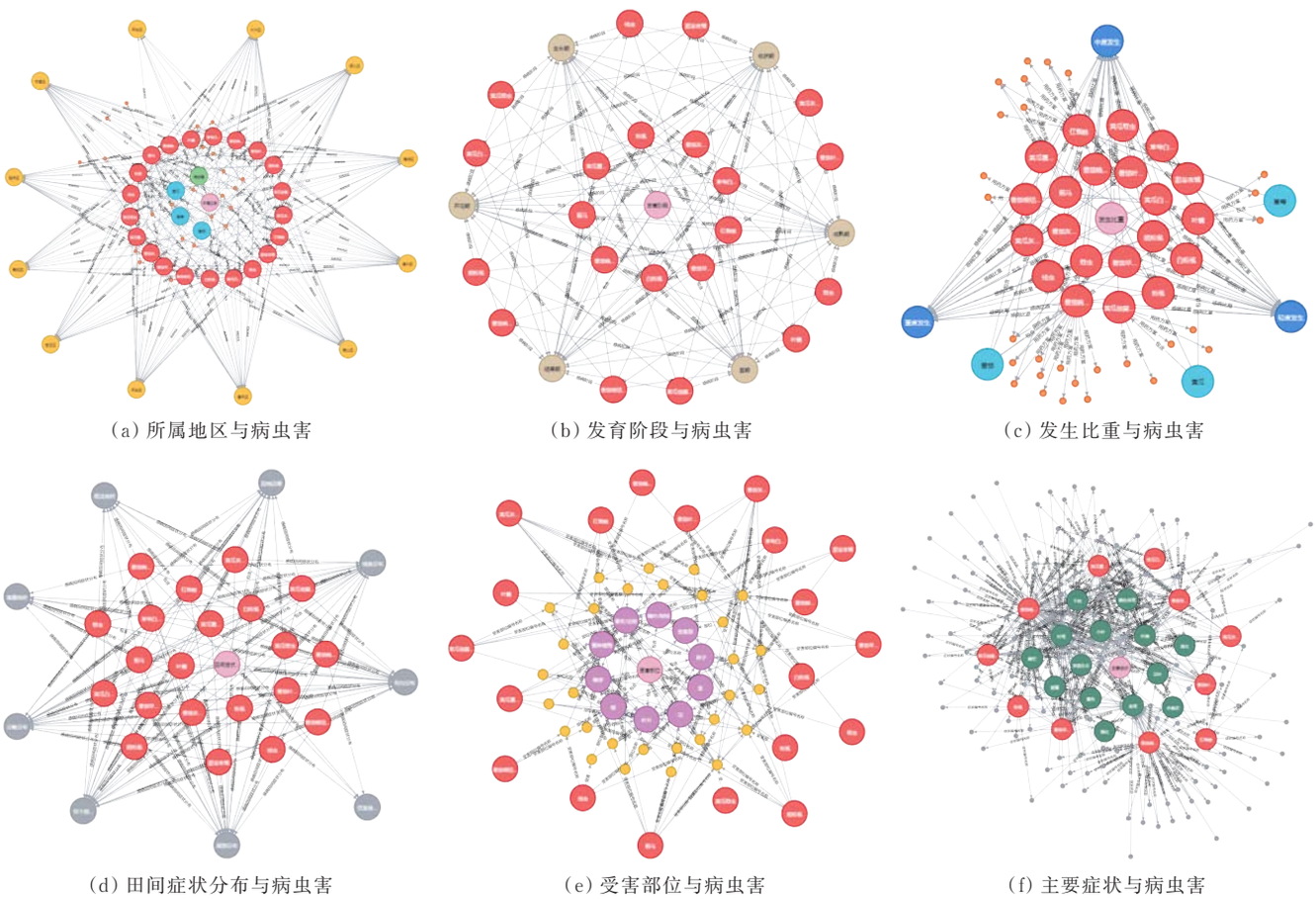
Figure 4　Part of the crop pest knowledge graph

2　基于知识图谱的关联挖掘

2.1　作物病虫害数据分析

基于知识图谱,结合对处方数据的统计分析,探究各种实体(作物病虫害、所属区、作物、发育阶段、受害部位、发生比重、主要症状、田间症状分布、诊断结果)之间的关系,利用 Neo4j 图数据库实现高效的关联关系查询和挖掘,如图 5 所示。根据知识图谱统计分析,管理者可以对北京市整体或区域施行宏观把控,例如:顺义区保护地蔬菜病虫害以番茄晚疫病、灰霉病、黄瓜霜霉病、角斑病为主,发生程度为轻—中等发生,其中春季保护地番茄病毒病轻度发生;白粉病、叶霉病、瓜类角斑病、瓜类霜霉病在部分棚室中等发生;其他病虫害偏轻发生。保护地虫害轻—偏重发生,其中烟粉虱

春季轻发生，夏秋季轻—中等发生；红蜘蛛、蓟马、蚜虫等小型害虫中等至偏重发生。当前生产菜农预防意识增强，物理预防措施越来越完善。通过提前定植，错过害虫发生高峰期，生产上采取阻断传播途径，创造不利于病源、虫源繁殖和传播环境条件，对防止病虫害发生有一定作用。



(a) 所属地区与病虫害　(b) 发育阶段与病虫害　(c) 发生比重与病虫害

(d) 田间症状分布与病虫害　(e) 受害部位与病虫害　(f) 主要症状与病虫害

图 5　作物病虫害知识图谱实体关系展示

Figure 5　Demonstration of entity relationship of knowledge graph of pests and diseases

2.2　基于 BERT-CNN 的文本分类算法

BERT 是采用 Transformer Encoder block 进行连接的预训练语言模型，Transformer 中的多头注意力（Multi-Head）可以让网络捕捉更丰富的信息。BERT－CNN 模型是通过 BERT 语言模型生成词向量，即得到每句的文本表示后（768 维向量），进一步使用 CNN 提取文本特征，最后将每条文本经过全连接层进行分类输出，模型结构如图 6 所示。

输入文本 $X=[X_1, X_2 \cdots X_n]$，经过输入层得到向量形式 $E(X)$，在经过编码层后，得到基于 BERT 模型的动态词向量矩阵 $T=[T_1, T_2 \cdots T_n]$。由于词向量维度过大，需要对其进行降维，以达到提取特征的目的。

将 BERT 的输出 T_i 输入 CNN 的卷积层，卷积核在动态词向量矩阵 $T=[T_1, T_2 \cdots T_n]$ 上通过上下滑动来提取特征，其中卷积核的长度主要用于捕获 d 个词之间的关系，最后得到特征映射矩阵 $C_i=[C_1, C_2 \cdots C_{n-d+1}]$，计算公式如公式(1)所示。

利用 CNN 的最大池化层提取主要的局部特征，得到文本特征 M_i，最后输入全连接层中。

为避免过拟合，通常会在全连接层后接一个 Dropout，其作用是以一定的概率丢弃神经元，将其置为 0。

通过 $softmax(M_i)$ 实现文本分类，如公式(3)所示。

$$C_i=f(W_i \cdot T+b) \tag{1}$$

$$M_i=max(C_i)=max(C_1, C_2 \cdots C_{n-d+1}) \tag{2}$$

$$softmax(M_i)=\frac{1}{1+e^{-M_i}} \tag{3}$$

式中：W_i——卷积核的权重矩阵；

b——偏置项。

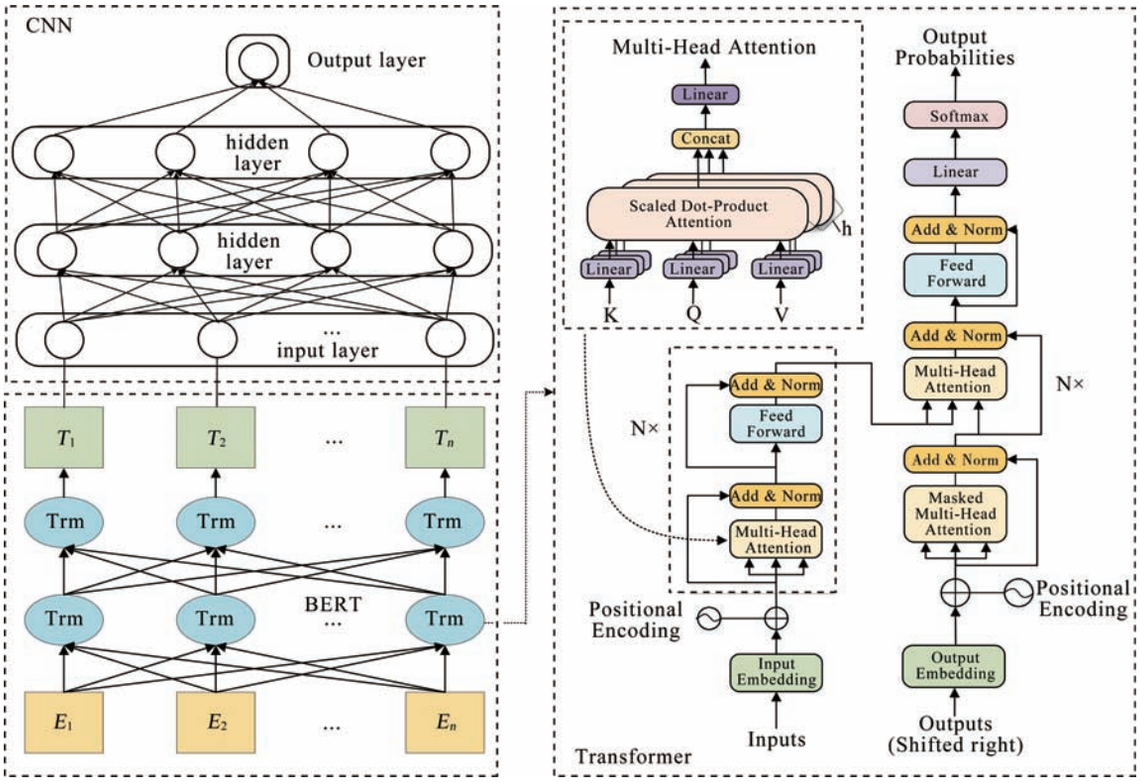


图 6 BERT 算法框架

Figure 6 Framework of BERT algorithm

2.3 基于 Neo4j 和机器学习的关联挖掘

2.3.1 中心性与重要性——度中心性(Degree centrality)算法

度中心性是在网络分析中刻画节点中心性最直接的度量指标[23],一个节点的节点度越大就意味着这个节点的度中心性越高,由此可确定图中的最重要节点。对于一个拥有 g 个节点的无向图,节点 i 的度中心性是 i 与其他 $g-1$ 个节点的直接联系总数,节点 i 的度中心度表示如公式(4)所示。

$$C_{\mathrm{D}}(N_i)=\sum_{j=1}^{g}x_{ij}\quad(i\neq j) \tag{4}$$

2.3.2 目标最短路径——Dijkstra 算法

Dijkstra 算法是从一个顶点到其余各顶点的最短路径算法。它是从起始点开始,采用贪心算法的策略,每次遍历到起始点距离最近且未访问过的顶点的邻接节点,直到扩展到终点为止。输入有向图 $G=(V,E)$,起始顶点 $s\in V$,并且每条边 $e\in E$ 的长度 e 均为非负值,输出从 s 到 v 的最短路径的长度 $dist(s,v)$。

2.3.3 相似性——KNN 算法

GDS 的 KNN 可以用于大型图的全局推断,根据图中的接近程度,节点属性,社区结构,重要性等来确定节点之间的相似性。KNN 的度量选择由节点属性类型驱动,选择为 jaccard 相似度进行相似度的计算,如公式(5)所示。

$$d_j(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}=\frac{A\Delta B}{|A\cup B|} \tag{5}$$

3 试验分析

3.1 试验环境及试验数据

以 Neo4j 社区版 5.5.0 为图数据库构建和存储知识图谱,Neo4j 图数据科学 2.3 为 Neo4j 图形算法版本,以 Cypher 形式进行关联挖掘。BERT 相关模型使用 NVIDIA Quadro P2000 GPU(5 GB)进行训练,并基于 Pytorch 的 Python 深度学习库实现,训练共 4 个 Epoch(每个 Epoch 约 1 800 次迭代),batch_size 设置为 16,学习率设置为 3×10^{-5}。本研究选取 2019 年 3 月 25 日至 2021 年 6 月 13 日的 45 833 条数据作为电子病历多模态数据的来源,并按照 6∶2∶2 的比例划分训练集、验证集和测试集。

3.2 评价指标

基于 GDS 实现的算法采用 AUCPR(area under the precision—recall curve),即“精度—召回率曲线下面积”为评价指标。基于 BERT 的病虫害多分类模型采用准确率(*Accuracy*)以及宏平均(Macro avg)和加权平均(Weighted avg)的精确度(*Precision*)、召回率(*Recall*)和 F_1 值(F_1-score)的作为评价指标。

3.3　试验结果与分析

3.3.1　基于度中心性算法的病虫害“热”点发现

在 GDS 的应用中，需要计算外度（out-degree）中心性，即计算从一个节点发出的关系。表 1 是使用度中心性来计算每种病虫害使用的药物处方数量，黄瓜霜霉病、黄瓜白粉病、蚜虫、番茄晚疫病和粉虱作为病虫害发生的“热”点，可进一步基于此提前实现病虫害预警或处方推荐。

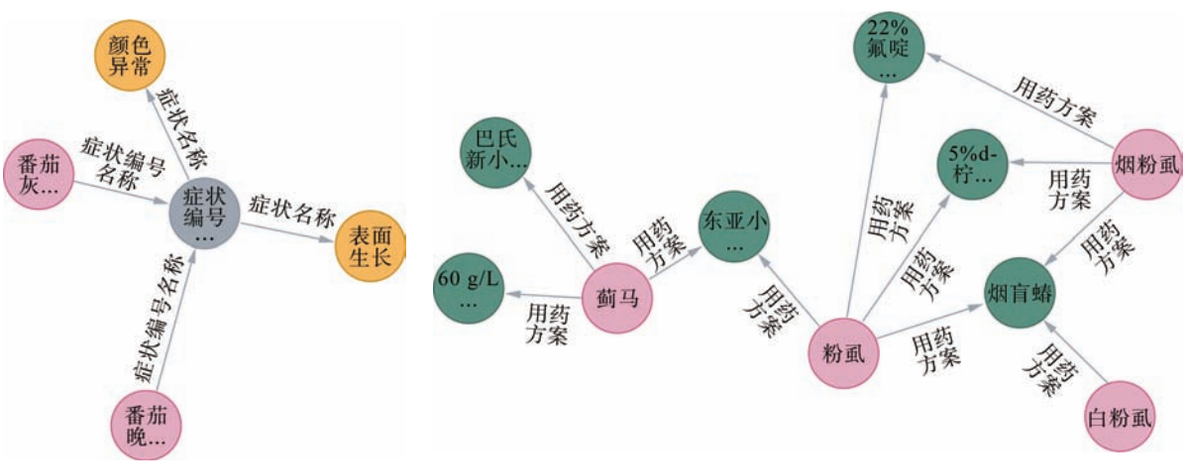
3.3.2　基于 Dijkstra 算法的病虫害间目标最短路径发现

基于构建的作物病虫害知识图谱，利用 Dijkstra 算法可以挖掘存在联系的实体的所有链路。例如，涉及同一病虫害的症状可能不存在直接的关联，但通过最短路径的计算可以发现隐藏的联系链路，相应挖掘出重要的症状，为相似病虫害诊断提供关键路径。由于 Dijkstra 算法支持加权关系，在比较路径时考虑了属性信息。图 7（a）为使用 Dijkstra 目标最短路径算法寻找到的病虫害“番茄灰霉病”和“番茄晚疫病”之间的最短路径，图 7（b）为“蓟马”“粉虱”“白粉虱”和“烟粉虱”之间最短路径。

表 1　病虫害“热”点分析

Table 1　Analysis of “popular” points of pests and diseases

病虫害名称	药物处方数量/条
黄瓜霜霉病	31
黄瓜白粉病	25
蚜虫	20
番茄晚疫病	16
粉虱	15



（a）番茄灰霉病和番茄晚疫病　　（b）蓟马、粉虱、白粉虱和烟粉虱

图 7　病虫害间的最短路径图

Figure 7　A map of the shortest paths between pests and diseases

3.3.3　基于 KNN 算法的相似病虫害发现

在作物病虫害知识图谱应用中，首先利用快速随机投影（fast random projection，FastRP）算法实现图嵌入，进一步利用 KNN 算法通过 jaccard 计算某两个病虫害节点的相似性，将 Top-K 限制为 1，得到每个节点的最相似的配对，试验的结果见表 2。

表 2　病虫害相似度分析

Table 2　Similarity analysis of pests and diseases

病虫害名称 1	病虫害名称 2	相似度
番茄晚疫病	黄瓜霜霉病	0.820 7
黄瓜霜霉病	番茄晚疫病	0.820 7
粉虱	白粉虱	0.949 6
番茄病毒病	黄瓜白粉病	0.700 1
番茄灰霉病	番茄早疫病	0.711 4

其中番茄晚疫病和黄瓜霜霉病以及粉虱和白粉虱都具有较高的相似性，高达 0.8 以上。

3.3.4　基于 BERT－CNN 的病虫害诊断

基于 BERT 挖掘的文本信息作为 CNN 和 RNN 分类模型的输入，构建基于 BERT－CNN 和 BERT－RNN 的多分类模型，实现作物病虫害的精准诊断。

试验结果如表 3 所示，针对常见的 20 种病虫害，BERT－CNN 在我们的特定任务上表现效果最好，其综合诊断准确率可以达到 0.931 3，同时宏平均和加权平均的 F_1-score 达到 0.876 5 和 0.931 4，相比于原始的 BERT 模型准确率、宏平均的 F_1-score 和加权平均的 F_1-score 分别提升 0.48%、3.76% 和 0.57%，进一步说明能有效提取文本信息并实现分类。

表 4 展示了 BERT－CNN 对 20 种病虫害诊断的详细试验结果，其中 75% 的病虫害的诊断 F_1-score 可

以达到80%以上,对于番茄根结线虫的诊断效果最佳,F_1-score可以高达0.984 2。

对于表2中相似度最高的番茄晚疫病和黄瓜霜霉病具有较好的区分能力,可以满足实际的诊断需求。

表3 BERT、BERT-CNN和BERT-RNN的对比结果

Table 3 Comparison of BERT, BERT－CNN and BERT－RNN

模型名称	*Accuracy*	Macro avg 的F_1-score	Weighted avg 的F_1-score
BERT	0.926 8	0.844 7	0.926 1
BERT－CNN	0.931 3	0.876 5	0.931 4
BERT－RNN	0.924 4	0.820 9	0.924 0

表4 基于BERT-CNN的病虫害诊断结果

Table 4 Diagnosis results based on BERT－CNN

病虫害名称	精确度	召回率	F_1值	样本数量
白粉虱	0.538 5	0.700 0	0.608 7	10
草莓白粉病	0.974 8	0.975 9	0.975 3	831
番茄病毒病	0.971 0	0.973 0	0.972 0	481
番茄根结线虫	0.975 9	0.992 6	0.984 2	408
番茄灰霉病	0.961 8	0.916 4	0.938 5	275
番茄晚疫病	0.700 0	0.875 0	0.777 8	8
番茄叶霉病	0.725 7	0.866 2	0.789 7	284
番茄早疫病	0.974 7	0.974 7	0.974 7	316
粉虱	0.976 5	0.965 5	0.971 0	1 680
红蜘蛛	0.824 6	0.921 6	0.870 4	102
黄瓜白粉病	0.907 8	0.885 8	0.896 7	289
黄瓜灰霉病	0.885 3	0.751 1	0.812 7	442
黄瓜霜霉病	0.965 4	0.967 5	0.966 5	1 818
黄瓜细菌性角斑病	0.893 6	0.910 3	0.901 9	1 015
蓟马	0.812 1	0.785 3	0.798 5	545
甜菜夜蛾	0.901 2	0.793 5	0.843 9	92
线虫	0.926 8	0.936 0	0.931 4	203
蚜虫	0.907 7	0.956 8	0.931 6	185
烟粉虱	0.948 6	0.985 4	0.966 7	206
叶螨	0.515 2	0.772 7	0.618 2	22
Accuracy		0.931 3		9 212
Macro avg	0.864 4	0.895 3	0.876 5	9 212
Weighted avg	0.932 9	0.931 3	0.931 4	9 212

4 结论

针对作物品种及病虫害种类繁多、作物病虫害与特征间的关联关系复杂和研究数据获取困难等难点,结合电子病历多模态数据的特点,探究人工智能技术在作物病虫害关联挖掘中的应用。电子病历的数据包括结构化和非结构化数据,根据不同模态数据的优势对电子病历进行深入挖掘,其中利用结构化数据构建知识图谱,并结合非结构化数据和领域知识进行知识增强,可以直观可视化病虫害的概况。

为了挖掘潜在信息,进一步基于知识图谱进行关联挖掘,结合度中心性算法、Dijkstra算法和KNN算法实现“热”点发现、最短路径发现和相似病虫害发现,为后续诊断和推荐应用提供支撑。在此基础上,利用非结构化数据基于BERT－CNN模型提取文本特征并实现病虫害诊断,并与BERT和BERT－RNN进行比较。试验证明BERT－CNN模型的分类效果最优,对于常见的20种病虫害的综合诊断准确率可以达到0.931 3。本文的主要创新点在于对电子病历大数据进行数据挖掘并辅助作物病虫害诊断,可以模拟植物医生的思维模式实现智能化服务,辅助早期智能决策与主动防控。未来将构建多模态数据集开展后续研究。

参 考 文 献

[1] ZHAO S Y, PENG Y, LIU J Z, et al. Tomato leaf disease diagnosis based on improved convolution neural network by attention module [J]. Agriculture, 2021, 11(7): 1－15.

[2] WANI J A, SHARMA S, MUZAMIL M, et al. Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: Methodologies, applications, and challenges [J]. Archives of Computational Methods in Engineering, 2022, 29(1): 641－677.

[3] FAHEY T, PHAM H, GARDI A, et al. Active and passive electro-optical sensors for health assessment in food crops [J]. Sensors, 2021, 21(1): 171.

[4] XU C, DING J Q, QIAO Y, et al. Tomato disease and pest diagnosis method based on the stacking of prescription data [J]. Computers and Electronics in Agriculture, 2022, 197: 106997.

[5] DING J Q, LI B, XU C, et al. Diagnosing crop diseases based on domain-adaptive pre-training BERT of electronic medical records [J]. Applied Intelligence, 2023, 53(12): 15979－15992.

[6] GARCÍA-ESTRADA R S, DIAZ-LARA A, AGUILAR-MOLINA V H, et al. Viruses of economic impact on tomato crops in Mexico: From diagnosis to management—A review [J]. Viruses, 2022, 14(6): 1251.

[7] HUANG S C, PAREEK A, ZAMANIAN R, et al. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: A case-study in pulmonary embolism detection [J]. Scientific Reports, 2020, 10(1): 22147.

[8] LIU Y, YU Z X, YANG Y L, et al. Diabetes risk data mining method based on electronic medical record analysis [J]. Journal of Healthcare Engineering, 2021: 6678511－6678526.

[9] 张领先, 韩梦瑶, 丁俊琦, 等. 作物病害智能诊断与处方推荐技术研究进展[J]. 农业机械学报, 2023, 54(6): 1－18.
ZHANG Lingxian, HAN Mengyao, DING Junqi, et al. Research progress in intelligence diagnosis and prescription recommendation of crop diseases [J]. Transactions of Chinese Society for Agriculture Machinery, 2023, 54(6): 1－18.

[10] 徐畅, 丁俊琦, 赵聘桐, 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 Chinese Society for Agriculture Machinery, 2022, 53(9): 286－294.

[11] 丁俊琦, 李博, 乔岩, 等. 基于植物电子病历多类型数据融合的作物病害诊断方法[J]. 农业机械学报, 2023, 54(1): 196－204, 223.
DING Junqi, LI Bo, QIAO Yan, et al. Crop disease diagnosis method based on fusion multiple types of data from plant EMRs [J]. Transactions of Chinese Society for Agriculture Machinery, 2023, 54(1): 196－204, 223.

[12] 张领先, 赵聘桐, 丁俊琦, 等. 基于CDSSM的作物病害处方推荐方法[J]. 农业机械学报, 2023, 54(3): 308－317.
ZHANG Lingxian, ZHAO Dantong, DING Junqi, et al. Recommendation method of crop disease prescription based on CDSSM [J]. Transactions of Chinese Society for Agriculture Machinery, 2023, 54(3): 308－317.

[13] BARTELS R H M A, DONK R D, VERHAGEN W I M, et al. Reporting the results of Meta-analyses: A plea for incorporating clinical relevance referring to an example [J]. The Spine Journal, 2017, 17(11): 1625－1632.

[14] BOURSIANIS A D, PAPADOPOULOU M S, DIAMANTOULAKIS P, et al. Internet of Things (IoT) and agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review [J]. Internet of Things, 2020: 100187.

[15] KIM S, LEE M, SHIN C. IoT-based strawberry disease prediction system for smart farming [J]. Sensors, 2018, 18(11): 4051.

[16] 高云君, 葛丛丛, 郭宇翔, 等. 面向关系型数据与知识图谱的数据集成技术综述[J]. 软件学报, 2023, 34(5): 1－27.
GAO Yunjun, GE Congcong, GUO Yuxiang, et al. Survey on data integration technologies for relational data and knowledge graph [J]. Journal of Software, 2023, 34(5): 1－27.

[17] CONG K, LI T, LI B B, et al. KGDetector: Detecting Chinese sensitive information via knowledge graph-enhanced BERT [J]. Security and Communication Networks, 2022: 4656837.

[18] LIN J J, ZHAO Y Z, HUANG W Y, et al. Domain knowledge graph-based research progress of knowledge representation [J]. Neural Computing and Applications, 2021, 33(2): 681－690.

[19] 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.

[20] ACHEAMPONG F A, NUNOO-MENSAH H, CHEN W. Transformer models for text-based emotion detection: A review of BERT-based approaches [J]. Artificial Intelligence Review, 2021: 1－41.

[21] ZHANG S C. Challenges in kNN classification [J]. IEEE Transations on Knowledge and Data Engineering, 2022, 34(10): 4663－4675.

[22] 孙敏, 罗卫红, 相林, 等. 设施果菜病害诊断的知识表达与推理模型[J]. 农业工程学报, 2012, 28(17): 149－156.
SUN Min, LUO Weihong, XIANG Lin, et al. Knowledge expression and disease diagnostic reasoning model for greenhouse fruit and vegetable crops [J]. Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(17): 149－156.

[23] 龚建兴, 朱雷, 王华兵, 等. 基于功能图的作战体系关键节点分析[J]. 系统工程与电子技术, 2022, 44(8): 2515－2521.
GONG Jianxing, ZHU Lei, WANG Huabing, et al. Analysis of key nodes in combat system based on function graph [J]. Systems Engineering & Electronics, 2022, 44(8): 2515－2521.

Association mining of crop diseases and intelligent diagnosis based on electronic medical records

XU Chang[1], ZHANG Lingxian[1*], QIAO Yan[2]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*;
2. *Beijing Plant Protection Station*, *Beijing* 100029, *China*)

Abstract: Research on crop diseases has become a hot topic of the application of artificial intelligence technology in smart agriculture. Existing studies are restricted by the factors such as difficulties in collecting data, high cost of technology implementation, and complex crop disease trends. Plant electronic medical records (PEMRs) formed by Beijing Plant Clinic provides a new idea for the diagnosis and prevention of crop diseases and pests. PEMRs are stored in the form of multi-modal data, containing a wealth of plant information, disease information, and environmental information. Therefore, how to mine PEMRs information and utilize it to assist follow-up research is an urgent problem to be solved. In view of the information representation ability of knowledge map, the mining ability of machine learning and the feature extraction ability of deep learning, structured data is used to construct the knowledge graph of crop diseases and pests, unstructured data and domain knowledge are used for knowledge enhancement according to the characteristics of PEMRs. Further, Neo4j graph database and graph data science (GDS) combined with machine learning algorithm was employed to conduct association mining from three dimensions of popular point discovery, link discovery and similar disease discovery. At the same time, text feature extraction and disease diagnosis were realized by using unstructured text data based on the bidirectional encoder representation from transformers (BERT) with convolutional neural network (CNN) model, and intelligent service was realized by simulating plant doctor. The comprehensive accuracy of 20 common diseases was up to 93.13%. This study can provide theoretical support for timely diagnosis, symptomatic control, scientific medication guide and assistant decision-making of crop diseases and pests, and innovate a new model and new business of social service of agricultural science and technology.

Keywords: crop diseases and pests; knowledge graph; association mining; plant electronic medical record; pest diagnosis; machine learning