基于Stacking融合模型与知识图谱的糖尿病诊断系统研究

吕金萍 张晓梅
宿州学院信息工程学院，中国·安徽 宿州 234000

摘要：糖尿病是全球范围内高发的慢性代谢性疾病，做好早期筛查与并发症预防工作，是减轻医疗负担的关键。传统诊断方式过度依赖医生的主观经验，单一机器学习模型的泛化能力不强，AI 诊断决策缺乏可解释性，且并发症预警机制不够完善。针对这些实际问题，本文设计并实现了一套融合多算法集成学习与医疗知识图谱的糖尿病智能辅助诊断系统。研究过程中，构建了标准化的数据预处理流程，对原始数据进行清洗、异常值处理及缺失值填充；提出基于 Stacking 策略的多模型融合预测方法，整合多种异构树模型，以逻辑回归作为元学习器进行二次决策，有效提高了预测的稳定性与综合性能。同时，依据临床指南与电子病历数据，构建糖尿病领域知识图谱，开发并发症推理引擎，实现了从风险预测到病理路径溯源的可解释性诊断。该系统整合了风险评估、并发症预警、健康管理与辅助决策等功能，可在基层医疗机构辅助应用，具有较好的临床实用价值。

关键词：糖尿病辅助诊断；Stacking 集成学习；知识图谱；可解释性人工智能；系统设计

Research on Diabetes Diagnosis System Based on Stacking Fusion Model and Knowledge Graph

Lv Jinping, Zhang Xiaomei
School of Information Engineering, Suzhou University, China Anhui Suzhou 234000

Abstract: Diabetes is a highly prevalent chronic metabolic disease worldwide. Early screening and complication prevention are crucial for reducing the medical burden. Traditional diagnostic methods rely excessively on doctors' subjective experience, single machine learning models have weak generalization ability, AI diagnostic decisions lack interpretability, and the complication early warning mechanism is imperfect. To address these practical problems, this paper designs and implements an intelligent auxiliary diagnosis system for diabetes integrating multi-algorithm ensemble learning and medical knowledge graph. In the research process, a standardized data preprocessing pipeline is constructed to clean, handle outliers and fill missing values of raw data. A multi-model fusion prediction method based on Stacking strategy is proposed, which integrates multiple heterogeneous tree models and uses logistic regression as the meta-learner for secondary decision-making, effectively improving the stability and comprehensive performance of prediction. Meanwhile, a domain knowledge graph for diabetes is built according to clinical guidelines and electronic medical record data, and a complication reasoning engine is developed to realize interpretable diagnosis from risk prediction to pathological path tracing. The system integrates functions such as risk assessment, complication early warning, health management and auxiliary decision-making, and can be applied in primary medical institutions with good clinical practical value.

Keywords: Diabetes auxiliary diagnosis; Stacking ensemble learning; Knowledge graph; Explainable artificial intelligence; System design

1 绪论

1.1 研究背景

糖尿病作为全球范围内患病率高、病程长，且需终身干预的慢性代谢性疾病，其日益增长的疾病负担已构成严峻的公共卫生挑战。据国际糖尿病联盟最新数据显示，全球成年糖尿病患者人数已突破 5 亿，中国作为糖尿病大国，患病率呈逐年上升且年轻化趋势明显。然而，在实际临床诊疗中，尤其是基层社区卫生服务中心，仍面临诸多挑战：一是早期筛查手段单一，多依赖空腹血糖等单项指标，易受生理波动影响导致漏诊或误诊；二是医生资源匮乏，难以对海量体检数据进行精细化风险评估；三是并发症预警滞后，往往在出现明显临床症状时才被发现，错过

了最佳干预窗口。

1.2 研究现状

近年来，人工智能技术在糖尿病预测与健康管理领域的应用日益广泛。传统风险评分工具操作简单，但预测精度有限；单一机器学习模型在处理复杂临床数据时存在明显局限，面对数据中的非线性关系和多特征交互时，适应性较差；深度学习模型虽能在一定程度上提高拟合效果，但其“黑箱”特性使得医生难以理解决策依据，影响了临床应用中的信任度。在知识图谱应用方面，通用医学知识图谱已发展得较为成熟，但针对糖尿病领域的细粒度知识图谱构建仍不够完善，多数研究仅停留在静态知识展示层面，缺乏与患者实时数据结合的动态推理能力。如何将高精度的预测模型与可解释的推理过程相结合，打造一套“预测准确、解释清晰”的辅助诊断系统，是目前该领域研究的重点和难点。

1.3 研究意义

围绕糖尿病早期筛查与并发症预警的实际需求，本研究主要开展了以下几方面工作：构建标准化的数据预处理流程，提升数据质量；设计基于 Stacking 集成策略的融合预测模型，增强诊断的可靠性；构建糖尿病专科知识图谱，实现并发症的智能推理与可解释输出；研发完整的智能辅助诊断系统，并推动其落地应用。该研究成果能够有效提升基层糖尿病筛查效率，改善并发症早防早治的水平，具有较强的实用价值。

2 数据工程与预处理方法

本研究采用公开糖尿病数据集及合作医疗机构提供的脱敏临床数据，涵盖糖尿病发病相关生理特征与基本信息。原始数据存在缺失值、异常值及量纲不统一等问题，会影响模型训练效果，因此需构建完整的数据预处理流程。

在数据清洗阶段，重点修正异常数值、剔除噪声数据。缺失值处理过程中，对比了多种填充方式的效果，最终选择更适合医疗偏态数据的填充策略，在保留足够样本量的同时，维持了数据原有的分布特征。特征标准化处理则有效消除了不同特征量纲差异带来的影响，使模型训练更加稳定，收敛速度也得到了提升。经过完整的预处理后，数据质量得到显著改善，为后续的模型训练提供了可靠保障。

3 核心模型构建与方法论

3.1 多模型融合预测模型设计

单一机器学习模型很难同时兼顾预测的偏差与方差，在处理复杂临床数据时，容易出现过拟合或泛化能力不足的问题。基于此，本研究结合 Stacking 集成学习思想，构建了多层级的融合模型。模型第一层选用多种具有代表性的树模型作为基学习器，充分利用不同模型的优势形成互补，提升对数据特征的整体捕捉能力。第二层以逻辑回归作为元学习器，对基学习器的输出结果进行整合与二次决策，进一步修正预测偏差，从而提高最终诊断结果的可靠性。

在模型训练过程中，采用交叉验证方式生成基学习器的输出特征，通过参数优化提升各模型的性能，确保融合模型在不同的数据分布下都能保持稳定的表现，有效解决了单一模型适应性差的问题。

3.2 基于知识图谱的并发症推理

为了提升 AI 诊断的可解释性，本研究构建了糖尿病领域知识图谱。参考权威临床指南，明确了疾病、症状、检查指标、并发症、药物、治疗方案等六大类实体类型，抽取各类实体之间的语义关系，形成结构化的知识网络。基于该知识图谱，开发了并发症推理模块，当系统判定患者存在糖尿病风险时，会自动提取患者的关键特征，在知识图谱中进行路径搜索与关联计算，推导出潜在的并发症及其病理路径，实现从“风险结果”到“原因溯源”的可解释性诊断，让医生能够清晰理解预警的依据。

3.3 可解释性分析

为进一步提高模型的透明度，本研究引入可解释人工智能方法，对融合模型进行解析，量化各项特征对预测结果的影响程度。通过全局分析与局部分析相结合的方式，清晰展示影响糖尿病诊断的关键因素，并为每位患者生成个性化的解释依据，让模型的决策过程更加直观可见，显著提高了临床医生对模型的接受度。

4 系统设计与实现

4.1 系统架构

本系统采用 B/S 架构与前后端分离的设计模式，具有良好的扩展性和维护性。前端采用主流开发框架设计界面，支持多设备访问，交互设计简洁易用；后端主要负责实现业务逻辑、权限管理以及数据流转等功能；算法服务以微服务形式独立部署，通过接口为系统提供预测与推理能力。数据层采用混合存储模式，其中关系型数据库用于存储用户信息与诊断记录，图数据库则用于知识图谱的存储与快速查询，保障系统能够高效稳定运行。

4.2 核心功能模块

本系统主要包含四大核心功能模块。智能风险评估模块支持用户输入相关生理指标，快速输出糖尿病风险等级，

并展示关键风险因子以及各项指标的偏离情况。并发症预警与溯源模块在风险评估的基础上，结合知识图谱给出并发症提示、病理路径解释以及针对性的干预建议，为医生快速决策提供支持。健康档案管理模块为用户建立长期电子健康档案，记录各项指标的变化情况与历史评估结果，支持趋势查看与报表生成。个性化健康管理模块根据用户的风险等级与身体状况，提供饮食、运动等方面的个性化指导，帮助用户更好地进行自我健康管理。

同时，系统配备了完善的数据安全机制，采用角色权限控制方式，对敏感信息进行加密存储与传输，符合医疗数据安全相关规范要求。

5 结语

针对糖尿病早期筛查与并发症预防的实际需求，本研究提出了融合 Stacking 集成学习与知识图谱的智能辅助诊断方案，构建了稳定可靠的融合预测模型与可解释并发症推理引擎，并开发出完整的 Web 应用系统。该系统不仅提高了糖尿病风险预测的可靠性，还解决了传统 AI 模型难以解释的问题，实现了可追溯、可理解的辅助诊断输出。整套系统结构轻量化、部署便捷、操作简单，适合在基层医疗机构推广使用。

未来，将进一步扩大多中心临床数据的采集范围，完善知识图谱的覆盖内容，接入连续健康监测设备以实现动态风险评估，同时探索与大语言模型的结合方式，为用户提供更自然、更具个性化的健康管理服务，持续提升系统的实用性与智能化水平。

参考文献：

[1] 许明哲. 基于门控树图注意力网络的糖尿病实体类型预测研究[D]. 北京：北京交通大学，2023.

[2] 刘勘，周晓英. 基于知识图谱的医疗并发症关联关系可视化分析[J]. 计算机应用，2022,42(S1):112-118.

[3] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南（2020 年版）[J]. 中华糖尿病杂志，2021,13(4):315-409.

基金项目：项目课题：宿州学院资助省级大学生创新创业训练计划项目：基于知识图谱与端边云硬件的糖尿病智能辅助诊断系统；项目编号：S202510379147。安徽省教育厅自然科学重点项目：融合注意力与全局上下文信息的图神经网络会话推荐算法研究（2024AH051810）。

作者简介：吕金萍（2004-），女，汉族，安徽亳州人，本科在读，学生，研究方向：计算机应用技术。