

基于深度置信网络的 2 型糖尿病微血管并发症预测 *

李瑞瑶¹ 许婧怡² 戴浩宇² 孙慧文³ 鲍 瀛¹ 华履春¹ 吴天星^{2,4}

(¹ 南京鼓楼医院信息管理处 南京 210008 ² 东南大学计算机科学与工程学院 南京 211189

³ 南京医科大学鼓楼临床医学院老年科 南京 210008

⁴ 新一代人工智能技术与交叉应用教育部重点实验室 南京 211189)

〔摘要〕 目的/意义 基于真实世界数据构建预测模型,实现对 2 型糖尿病微血管病变的预测和早期筛查。方法/过程 以南京鼓楼医院 10 年间真实世界数据为资料,将检验检查结果、病历文书等纳入考量,构建基于粒子群算法优化深度置信网络的 2 型糖尿病微血管并发症预测模型。结果/结论 该模型能够预测 2 型糖尿病微血管并发症,性能优于随机森林、支持向量机基准模型,为真实世界数据疾病预测模型研究提供借鉴。

〔关键词〕 2 型糖尿病;微血管并发症;疾病预测模型;临床数据处理;真实世界数据

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.07.012

Prediction of Microvascular Complications in Type 2 Diabetes Mellitus Based on Deep Belief Network

LI Ruiyao¹, XU Jingyi², DAI Haoyu², SUN Huiwen³, BAO Ying¹, HUA Lvchun¹, WU Tianxing^{2,4}

¹ Department of Information Management, Nanjing Drum Tower Hospital, Nanjing 210008, China; ² School of Computer Science and Engineering, Southeast University, Nanjing 211189, China; ³ Department of Geriatrics, Nanjing Drum Tower Hospital Clinical College of Nanjing Medical University, Nanjing 210008, China; ⁴ Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing 211189, China

〔Abstract〕 **Purpose/Significance** A prediction model is constructed based on real-world data to achieve prediction and early screening of type 2 diabetic microvascular complications. **Method/Process** Based on the real world data of Nanjing Drum Tower Hospital in the past 10 years, a particle swarm optimization based deep belief network (PSO-DBN) prediction model for microvascular complications in type 2 diabetes mellitus is constructed by taking test results and medical record documents into consideration. **Result/Conclusion** The PSO-DBN model can predict diabetic microvascular complications, and the performance is better than that of random forest and support vector machine (SVM) benchmark models, it provides references for the research of disease prediction model of real-world data.

〔Keywords〕 type 2 diabetes; microvascular complications; disease prediction model; clinical data processing; real-world data

〔修回日期〕 2024-03-09

〔作者简介〕 李瑞瑶,工程师,发表论文 16 篇;通信作者:吴天星。

〔基金项目〕 国家自然科学基金面上项目(项目编号:62376058);江苏省医院协会医院管理创新研究课题(项目编号:JSYGY-3-2023-616);南京大学中国医院改革发展研究院课题项目(项目编号:NDYGN2023021)。

1 引言

糖尿病是一种以高血糖为特征的慢性代谢紊乱性疾病。糖尿病微血管病变是糖尿病特有的慢性血管并发症，是一种慢性病理过程。其中糖尿病肾脏疾病（diabetic kidney disease, DKD）、糖尿病视网膜病变（diabetic retinopathy, DR）和糖尿病神经病变（diabetic neuropathy, DN）是最常见的微血管并发症，且分别是导致终末期肾损伤、成人失明、非创伤性截肢的主要原因^[1]，若能在发病初期及时发现并针对性治疗，可有效控制或延缓疾病进程^[2]，因此，糖尿病微血管病变早期筛查和预测具有重要意义。

随着人工智能技术在各行业的不断发展，越来越多的机器学习算法得到应用^[3-4]。在医学领域，机器学习相关技术在提高疾病早期筛查效率和准确率方面发挥了极大作用^[5-6]。国内外学者结合糖尿病并发症的相关因素和各类人工智能算法构建了效果良好的预测模型，逐步从经典回归模型向机器学习模型发展^[7]，如利用支持向量机（support vector machine, SVM）^[8]、神经网络^[9]等建立预测诊断模型，通过实验证明人工智能算法模型的优越性和可行性。但现有工作的特征变量选择仅采用单一统计

学方法，预测模型则集中于传统机器学习模型，或者对其中某一方面进行技术更新，尚未有整合性的改进工作。

2 资料与方法

2.1 资料来源

研究资料来源于南京鼓楼医院电子病历系统，数据时间为2012—2022年。在保护患者隐私的前提下，提取患有DKD、DR和DN的234 918例患者数据。原始数据包括历次就诊记录，数据字段包括患者ID、就诊次数、就诊日期、诊断、病历文书、检验检查项目名称、检验检查明细名称、检验检查结果、报告日期等。

2.2 研究方法

2.2.1 技术流程 对234 918例糖尿病微血管并发症患者的数据进行预处理，初步确定模型所需关键指标和正负样本标签。分别对检验检查数据和电子病历文书进行缺失值插补和文本向量嵌入，通过特征变量选择获得模型训练评估所需数据集，最后构建基于粒子群算法优化的深度置信网络预测模型，并与基准模型对比，完成模型评估。技术流程，见图1。

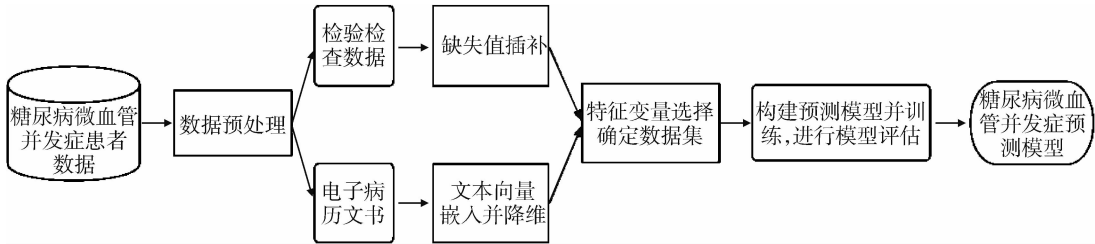


图1 技术流程

2.2.2 数据预处理 主要分为检验检查数据和病历文书数据预处理。对检验检查数据进行研究内容无关数据排除，通过“糖尿病微血管病变”“DKD”“DR”“DN”等关键词进行文献搜索，与临床专家沟通讨论，确保所选指标具有医学和实践依据，同时排除初期就诊即确诊并发症的患者。获得正例样本DKD 6 441例，DR 7 725例，DN 7 163例。在每

种并发症的剩余样本中进行负样本分层采样（正负样本比例范围在 $[1:1, 1:2]$ ），设定负样本为10 000条数据集（比例约为1:1.4）。根据关键指标数据集特征值缺失数量对剩余样本分层，缺失数量分为 $[0, 3]$ 、 $(3, 6]$ 、 $(6, \infty)$ 3组，按照6:3:1的比例分层采样，获得含正负样本标签的数据预处理结果。

2.2.3 缺失值插补 对检验检查数据主要利用缺失值随机森林算法插补缺失值。该算法弥补了“惰性”插补方法（如中位数或平均值插补方法）的单一性，适用于连续值和离散值的混合数据类型^[10]。具体而言，对缺失数据使用中位数或众数填充，将这些值标记为预测值，然后按照变量缺失率从小到大的顺序填补，迭代此过程直至最新填补结果与上一次结果变化很小时停止，获得数据插补后的完整检验检查数据集。

2.2.4 文本向量嵌入并降维 与检验检查数据不同，文书数据常以长文本形式记录，可通过自然语言处理技术进行生物医学文本挖掘。利用 BioBERT^[11] 实现文本向量嵌入，即将诊断文本中的词句映射为向量空间中的向量表示，在保留词语语义信息的同时，保留词语之间的相关性。通过 BioBERT 模型获得的文本向量嵌入存在维度过高的问题，而高维向量可能会弱化原有的检验检查特征对分类结果的影响力，因此进一步使用主成分分析方法对嵌入向量降维处理。

2.2.5 特征变量选择 特征变量是与疾病发生紧密相关的因素，有助于提高模型性能，降低过拟合的可能性，并增强模型的可解释性^[8]。特征选择的目标是从原始特征集中选择最相关的特征，以建立一个简化但高效的并发症数据集。在利用 logistic 回归分析得到统计学特征变量的基础上，基于支持向量机递归特征消除算法进一步对特征排序和筛选，以达到降低特征维度、简化模型的效果^[12]。该方法通过受试者操作特征曲线下面积评价指标进行特征子集评分，直至获得不同并发症对应的最优特征组合。

2.2.6 预测模型构建 构建基于粒子群算法优化的深度置信网络预测模型（particle swarm optimization based deep belief network, PSO-DBN），利用粒子群算法对深度置信网络的隐藏层神经元个数和学

习率进行寻优，见图 2。

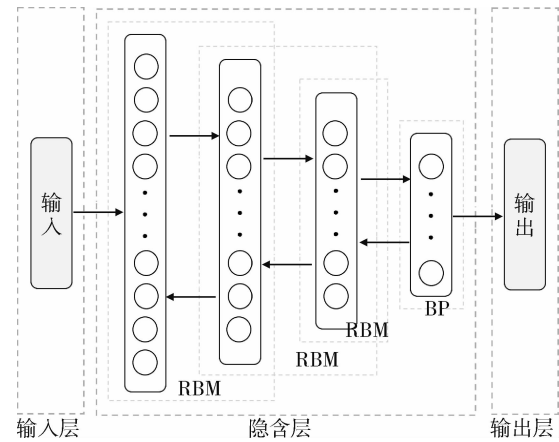


图 2 深度置信网络模型结构

深度置信网络隐藏层由反向传播神经网络（back propagation neural, BP）层和 3 层受限玻尔兹曼机（restricted boltzmann machine, RBM）层构成，一个 RBM 层是由 1 个可视层和 1 个隐含层进行全连接并加以限制组成的。为了寻找模型的最优值，训练过程中在每一层之间都进行最优选择。在模型参数寻优阶段，采用 PSO 算法的优化策略，以避免经验取值带来的过拟合问题或模型过于简单的困扰。粒子每次迭代时都对局部解进行优化，直至产生全局最优解，得到最优神经元个数和学习率等参数。

3 结果

3.1 特征变量选择结果

对关键指标进行 logistic 回归分析，以探究关键指标与 3 种 2 型糖尿病微血管并发症的关联性。设置 $P < 0.05$ 的显著性水平和优势比（odds ratio, OR）不为 1 的统计指标，确定在回归模型中差异具有统计学意义的 17 个特征变量，见表 1。

表 1 2 型糖尿病微血管病变特征变量 logistic 分析结果

特征变量	糖尿病肾脏疾病		糖尿病视网膜病变		糖尿病神经病变	
	OR	P	OR	P	OR	P
病程		0.001		0.001		0.001
年龄	1.006	0.000	1.025	0.000	1.009	0.000
性别	0.934	0.002	0.956	0.005	0.944	0.003

续表 1

特征变量	糖尿病肾脏疾病		糖尿病视网膜病变		糖尿病神经病变	
	OR	P	OR	P	OR	P
糖化血红蛋白	0.966	0.001	1.073	0.000	1.057	0.000
甘油三酯	1.080	0.000	0.968	0.007	0.931	0.000
总胆固醇	1.078	0.000	1.006	0.742	0.918	0.000
HDL-C	0.755	0.000	1.025	0.665	1.331	0.000
LDL-C	1.075	0.001	1.009	0.670	0.893	0.000
空腹血糖	0.990	0.080	1.082	0.000	0.992	0.049
eGFR (MDRD)	0.984	0.000	1.004	0.000	1.005	0.211
尿微量白蛋白/肌酐	1.005	0.000	1.000	0.000	1.000	0.000
空腹 C 肽	1.000	0.000	0.992	0.000	0.990	0.000
尿蛋白	1.001	0.008	1.043	0.029	0.998	0.023
红细胞比容	0.962	0.150	0.975	0.324	1.033	0.001
肌酐	1.002	0.081	1.004	0.041	0.999	0.598
空腹胰岛素	1.006	0.069	0.998	0.037	0.994	0.098
餐后 2 小时血糖	0.987	0.026	1.018	0.002	1.002	0.661

注：以 $P < 0.05$ 为差异具有统计学意义；OR 用来衡量二分类问题中两种取值可能性的比率，即当某个特征的取值增加一个单位时，相应的目标变量取值为 1 的比率相对于目标变量取值为 0 的比率的变化情况，如果 $OR = 1$ ，表示两种可能性的比率没有变化。以 $OR \neq 1$ 为某特征对目标变量存在影响。

在得到统计学特征变量的基础上，利用 SVM - RFE 算法进一步筛选特征变量。结果表明，针对 3 种并发症可以分别得到一个最优特征组合，见表 2。

表 2 2 型糖尿病微血管并发症对应的特征组合

并发症类型	特征组合
糖尿病肾脏疾病	病程、年龄、性别、糖化血红蛋白、甘油三酯、总胆固醇、HDL-C、LDL-C、eGFR (MDRD)、尿微量白蛋白/肌酐、尿蛋白、餐后 2 小时血糖
糖尿病视网膜病变	病程、年龄、性别、糖化血红蛋白、甘油三酯、空腹血糖、eGFR (MDRD)、空腹 C 肽、尿蛋白、肌酐、空腹胰岛素、餐后 2 小时血糖
糖尿病神经病变	病程、年龄、性别、糖化血红蛋白、甘油三酯、总胆固醇、HDL-C、LDL-C、空腹血糖、空腹 C 肽、尿蛋白、红细胞比容

3.2 数据集维度选择结果

对于患者门诊和住院病历文书的处理，使用开源的 BioBERT_v1.1_pubmed 作为文本向量生成的预训练模型，并利用中文糖尿病数据集微调模型^[13]使其适应本实验的中文文本，进而实现每个患者诊断文本的特征向量嵌入。其中，每条描述文本被表示为一个 768 维的向量。为避免过度降维造成向量信息丢失，实验时使用主成分分析方法将向量从 768 维分别降至 100、50、30、10 维。对不同维度的向量进行降维损失和特征保留分析，见图 3。降

维损失随着维度的减少而增加。在后续模型构建阶段，利用基准模型随机森林 (random forest, RF) 算法对不同维度的样本数据集进行测试。对比评价指标，见表 3，权衡不同维度下的性能表现，最终选择使用降维至 50 维的向量数据集。

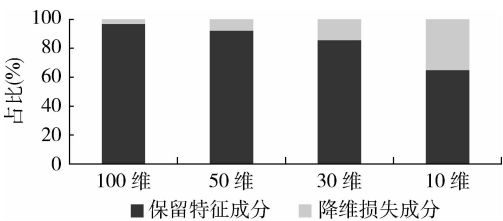


图 3 不同维度向量的降维损失和特征保留分析

表 3 基准模型在不同文本向量维度数据集上的预测结果

模型	糖尿病肾脏疾病		糖尿病视网膜病变		糖尿病神经病变	
	正确率	F1	正确率	F1	正确率	F1
RF - 0d	0. 69	0. 51	0. 61	0. 47	0. 58	0. 38
RF - 30d	0. 70	0. 54	0. 62	0. 50	0. 60	0. 40
RF - 50d	0. 72	0. 56	0. 62	0. 51	0. 60	0. 40
RF - 100d	0. 71	0. 54	0. 61	0. 48	0. 60	0. 37

3.3 预测模型训练结果

获得特征变量选择和文本向量嵌入后的最终数据集，按 4:1 的比例划分为训练集和测试集，完成基于粒子群算法优化的深度置信网络预测模型训练。设定初始种群粒子的个数为 20，进行最多 100 次迭代完成解位置和学习率的更新。在训练过程中，采用 5 折交叉验证的方法找到最优解。当迭代到第 25 次时，DBN 模型在适应度函数下达到最高的分类准确率，分类性能最佳，此时各隐含层的神经元个数分别为（74，83，145，136）。

4 预测效果分析

利用预处理后的测试数据集预测目标患者糖尿病中、后期是否会患有 3 种相关糖尿病微血管并发症，以 RF、SVM 模型作为基准模型，根据前文表 3 选择 50 d 数据集作为最终数据集，将 PSO - DBN 与基准模型对比，针对 DKD、DR、DN 3 种 2 型糖尿病微血管并发症分别展开分类预测。使用正确率和 F1 值作为评估指标^[14]。正确率是常用的分类模型评估指标，表示正确预测的样本数占总样本数的比例。F1 值综合考虑模型的精确度和召回率，是两者的调和平均值。具体实验结果，见表 4。

表 4 不同模型在 2 型糖尿病微血管并发症上的预测结果对比

模型	糖尿病肾脏疾病		糖尿病视网膜病变		糖尿病神经病变	
	正确率	F1	正确率	F1	正确率	F1
RF	0. 72	0. 56	0. 62	0. 51	0. 60	0. 40
SVM	0. 70	0. 55	0. 61	0. 47	0. 58	0. 39
PSO - DBN	0. 73	0. 62	0. 70	0. 62	0. 67	0. 59

在基准 RF 模型下，使用病历文书的预测效果更好，尤其在 50 维文本向量时取得最佳性能；PSO - DBN 模型的正确率和 F1 分数均超过基准 RF 模型、SVM 模型，显示其在真实世界数据下的优越性。相较于传统机器学习模型，本文构建的基于粒子群算法优化的深度置信网络分类模型具有更高的预测准确性。

5 结语

2 型糖尿病微血管病变是一个慢性的病理过程，

在发病初期及时发现并针对性治疗，可有效控制或延缓疾病进程。本文基于真实世界的生物医学领域数据，引入深度学习方法，成功构建了效果良好的预测模型，实现了对 2 型糖尿病微血管病变的预测和早期筛查。

虽然实验结果表现良好，但仍存在一些局限性。首先，模型的性能可能受到实验选用数据的限制，对于 2 型糖尿病并发症患者的特征可进一步拓展数据来源，增加区域范围内的就诊病历数据，整合影像学数据、基因组学信息等多模态数据，细化研究病程的不同阶段，以全面洞察并发症的多层次

特征。其次，由于数据量和模型复杂度的制约，本模型仍需要在更大规模的数据集或更复杂的模型上验证，提高其泛化能力。

作者贡献：李瑞瑶负责论文撰写、数据收集；许婧怡负责论文撰写、数据分析；戴浩宇负责数据分析；孙慧文、华履春负责数据审校；鲍瀛负责论文审校；吴天星负责数据分析、论文审校。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 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.

2 朱大龙, 陆菊明. 《中国 2 型糖尿病防治指南 (2020 年版)》编写说明 [J]. 中华糖尿病杂志, 2021, 13 (4): 312 - 314.

3 唐海国, 张帝, 刘成英, 等. 计及拓扑变换的主动配电网故障风险智能预警 [J]. 电力工程技术, 2022, 41 (5): 193 - 201, 226.

4 施泉生, 黄晓辉, 胡伟, 等. 基于区块链的改进智能合约电力交易模型 [J]. 电力工程技术, 2022, 41 (1): 11 - 18.

5 陈思含, 张云秋. 基于机器学习的 2 型糖尿病并发症预测模型研究 [J]. 中华医学图书情报杂志, 2020, 29 (11): 31 - 38.

6 袁筱祺, 赵英英. 基于多层感知器神经网络的住院病案内涵质量预测模型研究 [J]. 医学信息学杂志, 2023, 44 (11): 35 - 40.

7 KEE O T, HARUN H, MUSTAFA N, et al. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review [J]. Cardiovasc diabetol, 2023, 22 (1): 13.

8 DAGLIATI A, MARINI S, SACCHI L, et al. Machine learning methods to predict diabetes complications [J]. Journal of diabetes science and technology, 2018, 12 (2): 295 - 302.

9 CHU H, CHEN L, YANG X, et al. Roles of anxiety and depression in predicting cardiovascular disease among patients with type 2 diabetes mellitus: a machine learning approach [J]. Front psychol, 2021, 12 (4): 1189.

10 张鑫, 韦哲, 曹彤, 等. 基于 PSO - SVM 的糖尿病并发症预测研究 [J]. 中国医学装备, 2022, 19 (2): 10 - 13.

11 LEE J, YOON W, KIM S, et al. BioBERT: a pre - trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36 (4): 1234 - 1240.

12 HECTOR S, CLARISSA V, ESTEBAN V, et al. SVM - RFE; selection and visualization of the most relevant features through non - linear kernels [J]. BMC bioinformatics, 2018, 19 (1): 432.

13 REN K T, BENJAMIN J J S, HENG Y K, et al. Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review [J]. Journal of diabetes science and technology, 2023, 17 (2): 474 - 489.

14 TAHA A A, HANBURY A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool [J]. BMC medical imaging, 2015, 15 (1): 29.

《医学信息学杂志》版权声明

(1) 作者所投稿件无“抄袭”、“剽窃”、“一稿两投或多投”等学术不端行为，对于署名无异议，不涉及保密与知识产权的侵权等问题，文责自负。对于因上述问题引起的一切法律纠纷，完全由全体署名作者负责，无需编辑部承担连带责任。(2) 来稿刊用后，该稿包括印刷出版和电子出版在内的版权、复制权、发行权、汇编权、翻译权及信息网络传播权已经转让给《医学信息学杂志》编辑部。除以纸载体形式出版外，本刊有权以光盘、网络期刊等其他方式刊登文稿，本刊已加入万方数据“数字化期刊群”、重庆维普“中文科技期刊数据库”、清华同方“中国期刊全文数据库”、中邮阅读网。(3) 作者著作权使用费与本刊稿酬一次性给付，不再另行发放。作者如不同意文章入编，投稿时敬请说明。

《医学信息学杂志》编辑部