

基于特征选择和神经网络的糖尿病预测模型研究^{*}

宁莉燕 陈建荣

董建成

苏建彬

（南通市第一人民医院（南通大学第二附属医院） 南通 226001）

（南通大学医学院 南通 226000）

（南通市第一人民医院（南通大学第二附属医院） 南通 226001）

〔摘要〕 采用特征选择算法和人工神经网络建立糖尿病预测模型，阐述模型构建及评价步骤、方法。以灵敏度、特异度、准确率、ROC - AUC 为指标评估模型的预测性能，并与其他算法模型进行对比分析，实验结果表明基于特征选择和人工神经网络的糖尿病预测模型对临床指标未知且复杂的数据集具有更好的抗干扰能力和预测性能。

〔关键词〕 糖尿病预测；特征选择；参数优化；神经网络

〔中图分类号〕 R - 058 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673 - 6036. 2023. 02. 008

Study on Diabetes Prediction Model Based on Feature Selection and Neural Network *NING Liyan, CHEN Jianrong, Nantong First People's Hospital (The Second Affiliated Hospital of Nantong University), Nantong 226001, China; DONG Jiancheng, Medical School, Nantong University, Nantong 226000, China; SU Jianbin, Nantong First People's Hospital (The Second Affiliated Hospital of Nantong University), Nantong 226001, China*

〔Abstract〕 A diabetes prediction model is constructed based on feature selection algorithm and artificial neural network, and the steps and methods of model construction and evaluation are expounded. Sensitivity, specificity, accuracy and ROC - AUC are used as evaluation metrics to evaluate the performance of the proposed model, and the proposed model is compared with other algorithms through experiments. The result indicates that the diabetes prediction model based on feature selection and artificial neural network has better anti - interference capability and predictive performance, which is more suitable for complex datasets with unknown influencing indicators.

〔Keywords〕 diabetes prediction; feature selection; parameter optimization; neural network

1 引言

糖尿病属于常见的慢性病，无论对发达国家还

是发展中国家而言都是严峻的挑战^[1]。我国是世界上糖尿病人口最多的国家之一，其中 20 ~ 79 岁的成人糖尿病患者数约有 1. 164 亿^[2]。《中国防治慢性病中长期规划（2017—2025 年）》制定的目标是

〔修回日期〕 2022 - 11 - 19

〔作者简介〕 宁莉燕，副研究馆员，发表论文 15 篇；通信作者：陈建荣，教授，硕士生导师。

〔基金项目〕 江苏省医院协会医院管理创新研究课题“医联体背景下基于特征选择的糖尿病辅助诊断模型研究”（项目编号：JSYGY - 3 - 2021 - 421）；南通市科技计划项目指令性课题“医联体背景下基于边缘计算的糖尿病辅助诊疗预测模型关键技术研究”（项目编号：JC2020045）；南通市科技计划项目“基于大数据的糖尿病队列数据库建立及预测模型构建研究”（项目编号：JC2022155）。

到 2025 年慢性病危险因素得到有效控制,实现全人群全生命周期健康管理。目前糖尿病发病机理尚未明确,国内缺乏系统的糖尿病风险预测模型^[3-4]。因此,进一步研究糖尿病危险因素,构建并优化风险预测模型非常重要,尤其是针对糖尿病高危人群,准确的预测可以为早发现、早治疗提供辅助^[5]。近年来,研究者应用逻辑回归(logistic regression, LR)、支持向量机(support vector machine, SVM)、随机森林(random forest, RF)、梯度提升树(gradient boosting decision tree, GBDT)、神经网络(artificial neural network, ANN)等多种机器学习算法,研究构建糖尿病预测模型^[5-14]。其中裴修侗等^[9]通过实验表明, SVM 传统算法模型精度较低,利用布谷鸟、灰狼仿生算法优化后精度都有所提升。郭奕瑞等^[10]利用 ANN 和 LR 算法建立 2 型糖尿病预测模型,得出 ANN 模型较 LR 模型具有更好的预测效果。余丽玲等^[11]融合了自回归积分滑动平均(autoregressive integrated moving average model, ARIMA)和 SVM 两种模型,并将融合方法与 ARIMA 模型、SVM 模型、ANN 模型进行对比,实验表明组合模型的预测精度相比单一预测模型有明显提高。

上述对糖尿病特征发现及风险预测的研究取得了一定成果,其中 ANN 模型对临床指标的高维非线性特征表现出较好的拟合性^[10,14],组合算法进一步提升了预测准确度^[5,8-9,13],但对糖尿病临床数据复杂性仍未充分考虑^[15]。鉴于此,本文提出一种特征选择和神经网络组合算法来构建糖尿病预测模型,并基于 Pima 印第安人数据集增加干扰数据,对模型的抗干扰能力和预测性能进行对比研究。

2 数据来源和扩展

本文所使用的 Pima 印第安人数据集(简称 Pima 数据集)来自美国国家糖尿病、消化和肾脏疾病研究所,是机器学习领域重要的公开数据集。该数据集包含 768 个样本,每个样本包含 8 个临床指标和 1 个疾病诊断(标签属性),具体数据可通过

GitHub 查看和获取,见表 1。

表 1 数据集属性

属性名称	属性描述
怀孕次数 (Pregnancies)	怀孕次数
血糖值 (Glucose)	血浆葡萄糖浓度
血压 (Blood Pressure)	舒张压 (mmHg)
皮脂厚度 (Skin Thickness)	三头肌组织褶厚度 (mm)
血清胰岛素 (Insulin)	2 小时血清胰岛素 (mu U/ml)
体重指数 (BMI)	体重指数 (kg/m ²)
糖尿病谱系功能 (Diabetes Pedigree Function)	糖尿病谱系功能
年龄 (Age)	年龄
疾病诊断 (Target)	疾病诊断(标签属性), 0 代表未患糖尿病, 1 代表患有糖尿病

为了验证拟构建预测模型的抗干扰能力和数据适应性,本文将 Pima 数据集作为标准数据集,并在此基础上进行扩展,通过程序生成 3 个检验检查指标,扩展后的数据集包含 11 个临床指标。因此,在实验部分本文所使用的数据集包括:未进行扩展的 Pima 标准数据集和 Pima 扩展数据集。

3 预测模型构建及评价

3.1 糖尿病预测模型 KANN 结构

本文提出一种新的糖尿病预测模型 KANN (KBest-ANN),见图 1。该模型主要包括:数据预处理(包括缺失值处理、数据标准化)、数据集划分、基于 KBest 算法的特征选择、神经网络、网格搜索算法和模型评估等 6 个部分,并在实验部分基于 Python 3.7 和 sklearn 库实现模型的有效性验证。

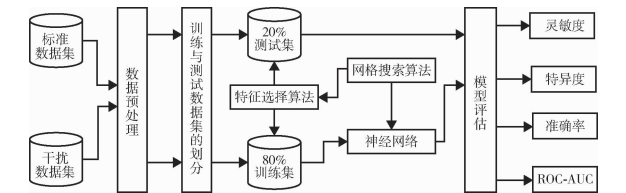


图 1 KANN 糖尿病患病风险预测模型

3.2 数据预处理

3.2.1 缺失值处理 考虑到 Pima 数据集的样本量不大，故采用均值插补法进行缺失值填充。对于异常值也用均值代替。将处理后的数据集合成一个新的数据集并用于后续处理。

3.2.2 数据标准化 为避免临床指标数据波动和分布差异对拟构建的预测模型产生影响，对数据集进行标准化处理，即通过如下线性变换公式，将各指标映射到统一量级和区间^[16]。

3.3 特征选择

由于数据集中临床指标众多，为了提高模型

的预测性能，需要对作为模型输入参数的临床指标进行筛选，使其满足临床指标之间相互独立、临床指标与标签属性之间显著相关。因此，引入如下公式计算皮尔逊相关系数，相关系数 $\rho_{X,Y_{Target}}$ 越接近 1 时，临床指标 X 和目标值 Y_{Target} 之间的相关性越高，反之则越低。计算获得基于皮尔逊相关系数的特征关系热力图，见图 2（a）。由上述结论可知，相关系数越大，对应的热力图颜色越浅，反之越深。

$$\rho_{X,Y_{Target}} = \frac{\text{cov}(X,Y_{Target})}{\sigma_X * \sigma_{Y_{Target}}} = \frac{E(XY_{Target}) - E(X) * E(Y_{Target})}{\sqrt{E(X^2) - E^2(X)} * \sqrt{E(Y_{Target}^2) - E^2(Y_{Target})}} \tag{2}$$

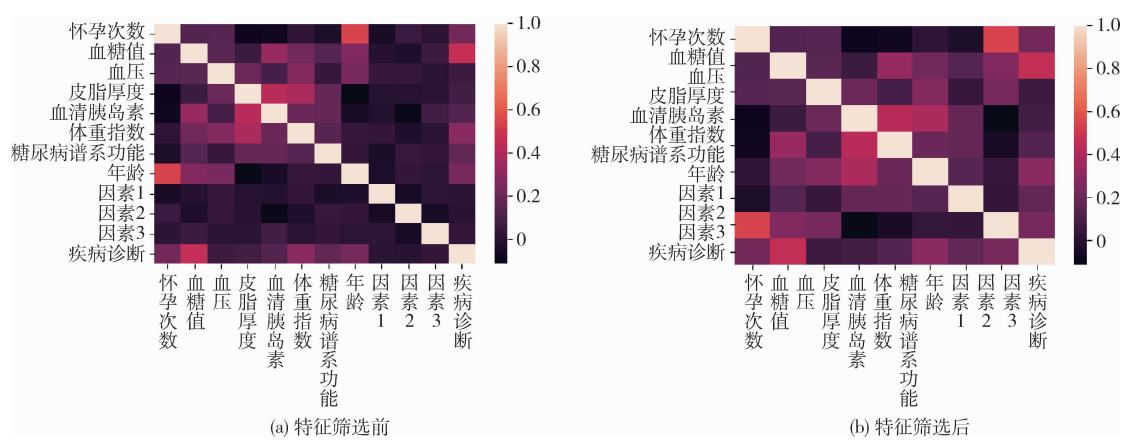


图2 临床指标和标签属性的相关性

由图 2（a）可知，血糖值（glucose）和体重指数（body mass index, BMI）等临床指标与标签属性相关性较高，而影响因素 1（Factor1）、影响因素 2（Factor2）等指标与标签属性相关性较低，同理得出临床指标之间的相关性。

在属性相关性分析的基础上，本文采用基于 Top - K 策略的 KBest 算法来实现具体临床指标的选择，计算方式如下。其中 InParams 表示有效临床指标的集合， K 表示相关性阈值，属于超参。基于 Top - K 策略，得到特征选择的结果，见图 2（b）；保留原有的 8 个临床指标，得出对标签属性的影响程度，见图 3。

$$f(\rho_{X,Y}) = \begin{cases} \rho_{X,Y} \geq K, & \text{临床指标与标签属性} \\ \rho_{X,Y} \leq K, & \text{临床指标之间} \end{cases} \tag{3}$$

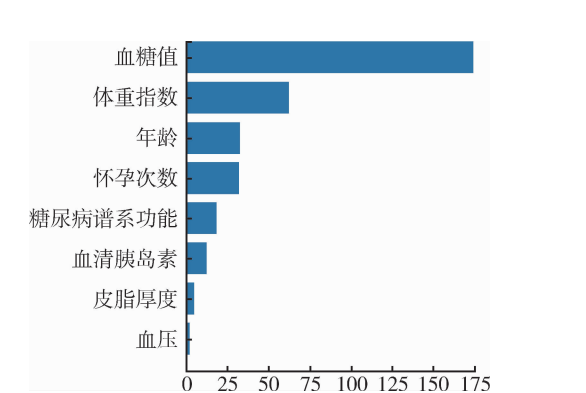


图3 临床指标对标签属性的影响程度

3.4 神经网络

ANN 遵循人工智能的学习原理和人类神经元的自然反应，是一种可纳入多维特征的非线性技术，能够有效学习复杂数据中的高维非线性特征^[14,17]。本模型中，ANN 以特征选择算法 KBest 筛选出的 8 个临床指标作为输入变量，构建输入层，然后以全连接的方式将输入层、隐藏层和输出层进行连接。其中，各节点的计算方式如下。

$$Y_i = F[\sum_{i=1}^p w_i x_i + b_i]$$

(4)

其中 Y_i 为输出变量， x_i ($i=1, 2, \cdots, p$) 为输入变量， w_i ($i=1, 2, \cdots, p$) 为对应输入变量的权重， b_i ($i=1, 2, \cdots, p$) 为偏置项， F 为激活函数。

3.5 模型训练

基于 5 倍交叉验证将两个数据集分为 80% 的训练数据集和 20% 的测试数据进行模型训练与评估。首先利用网格搜索算法确定可变参数的中心值，然后再通过区间搜索对中心值进行微调，避免异常值导致的拟合差异问题，从而确定最优可变参数。考虑 L2 范数正则化 (L2 - norm regularization) 对异常值和离群点的过于敏感，故在神经网络的损失函数中引入 L1 范数正则化 (L1 - norm regularization)，得出优化后 ANN 的最优可变参数取值，见表 2。

表 2 模型最优可变参数

可变参数	值	可变参数	值
激活函数	relu	最大迭代次数	100 000
正则化项参数	1	初始学习率	0.000 1
收敛误差	1e - 20	优化器	adam
隐藏层节点数量	150	-	-

3.6 模型评价

本文以灵敏度、特异度、准确率、受试者工作特征 - 曲线下面积 (receiver operating characteristic - area under the curve, ROC - AUC) 作为模型性能的评价指标，其中 ROC - AUC 通过机器学习库 Sci-kit - learn (sklearn) 的 “roc_auc_score” 函数计算，其他各指标计算方式如下。

灵敏度 = $\frac{TP}{TP + FN}$

(5)

特异度 = $\frac{TN}{TN + FP}$

(6)

准确率 = $\frac{TP + TN}{TP + FP + TN + FN}$

(7)

其中，TP 为正确预测阳性样本的数目；TN 为正确预测阴性样本的数目；FP 为错误预测阳性样本的数目；FN 为错误预测阴性样本的数目。

4 结果与讨论

利用测试集对已训练模型进行评估，实验结果表明，在基于 Pima 标准数据集的实验中，本模型与 GBDT 的评价指标均具有最优表现、性能接近；但在基于 Pima 扩展数据集的实验中，本模型的评价指标均优于其他对比算法，对干扰因素具有更好的泛化能力，更适合临床指标复杂的应用场景。KANN 的灵敏度比 GBDT 高 1 个百分点；特异度比 Logistic 高 4 个百分点；准确率比 SVM 高 3 个百分点，见表 3。在某些部分 GBDT 和 Logistic 所围面积大于本模型；但从整体来看，本模型的 ROC - AUC 仍然是最优的，见图 4。

表 3 基于 Pima 标准和扩展两个数据集的评价指标

模型名称	灵敏度		特异度		准确率		ROC - AUC	
	标准数据集	扩展数据集	标准数据集	扩展数据集	标准数据集	扩展数据集	标准数据集	扩展数据集
支持向量机	0.85	0.86	0.60	0.62	0.77	0.79	0.72	0.74
逻辑回归	0.85	0.85	0.62	0.64	0.78	0.79	0.73	0.74
随机森林	0.87	0.86	0.60	0.55	0.79	0.77	0.73	0.71
梯度提升树	0.88	0.87	0.68	0.60	0.82	0.79	0.78	0.73
无特征选择的本模型	0.88	0.85	0.68	0.60	0.82	0.77	0.78	0.72
本模型	0.88	0.88	0.68	0.68	0.82	0.82	0.78	0.78

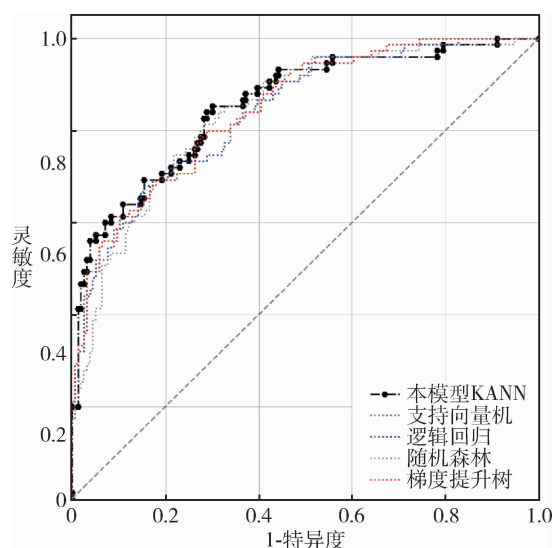


图 4 各分类模型 ROC-AUC 曲线

5 结语

本文针对糖尿病临床指标复杂性提出一种基于特征选择和神经网络的糖尿病风险预测模型 KANN。该模型先对输入数据进行标准化,避免数据波动和分布差异引起的模型拟合误差;通过 KBest 特征选择对临床指标的有效性进行判断和筛选,避免冗余和噪声临床指标对模型训练的干扰;将特征选择算法 KBest 和神经网络 ANN 进行整合,构建糖尿病风险预测模型,并以网格搜索、adam 优化算法、L1 范数正则化对构建的模型进行优化;基于 5 倍交叉验证和评估指标进行一系列实验,将本模型与其他常用模型进行对比,验证了本模型的优越性。由于本实验仅基于公开的 Pima 数据集,尚未在其他数据集中进行验证,下一步拟采用真实临床环境数据集对模型进行验证与优化。

参考文献

- 1 MISRA A, GOPALAN H, JAYAWARDENA R, et al. Diabetes in developing countries [J]. Journal of diabetes, 2019, 11 (7): 522-539.
- 2 CHO N H, SHAW J E, KARURANGA S, et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045 [J]. Diabetes research and clinical practice, 2018, 138 (4): 271-281.

- 3 朱亚楠. 基于数据挖掘技术的 2 型糖尿病患病风险预测研究 [D]. 秦皇岛: 燕山大学, 2017.
- 4 American Diabetes Association. 2 classification and diagnosis of diabetes: standards of medical care in diabetes - 2021 [J]. Diabetes care, 2021, 44 (S1): S15-S33.
- 5 HASAN K, ALAM A, DAS D, et al. Diabetes prediction using ensembling of different machine learning classifiers [J]. IEEE access, 2020 (8): 76516-76531.
- 6 VIJAYAN V V, ANJALI C. Prediction and diagnosis of diabetes mellitus - A machine learning approach [C]. Trivandrum: 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2015.
- 7 王成武, 晏峻峰. 早期糖尿病风险预测模型比较研究 [J]. 智能计算机与应用, 2021, 11 (1): 64-68.
- 8 SANTHANAM T, PADMAVATHI M S. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis [J]. Procedia computer science, 2015, 47: 76-83.
- 9 裴修侗, 周晓东, 陈凯祥, 等. 基于支持向量机的糖尿病诊断优化算法研究 [J]. 技术与市场, 2021, 28 (8): 9-11.
- 10 杨美洁, 唐建军. 基于随机森林算法的糖尿病预测研究 [J]. 医学信息学杂志, 2019, 40 (9): 47-49.
- 11 母东生, 蔡梦杰, 王桂娟, 等. 基于生活习惯的糖尿病预测与可视分析 [J]. 计算机技术与发展, 2021, 31 (10): 154-160.
- 12 郭奕瑞, 李玉倩, 王高帅, 等. 人工神经网络模型在 2 型糖尿病患病风险预测中的应用 [J]. 郑州大学学报 (医学版), 2014, 49 (2): 180-183.
- 13 余丽玲, 陈婷, 金浩宇, 等. 基于支持向量机和自回归积分滑动平均模型组合的血糖值预测 [J]. 中国医学物理学杂志, 2016, 33 (4): 381-384.
- 14 ZHANG Y, YIN C, WU Q, et al. Location-aware deep collaborative filtering for service recommendation [J]. IEEE transactions on systems, man, and cybernetics: systems, 2021, 51 (6): 1-12.
- 15 徐壮. 基于遗传算法的糖尿病预测模型优化与实现 [D]. 南京: 南京邮电大学, 2021.
- 16 李仪, 林建君, 朱习军. 基于改进 DNN 的糖尿病预测模型设计 [J]. 计算机工程与设计, 2021, 42 (5): 1418-1424.
- 17 车前子, 郑启文, 陈思, 等. 基于人工神经网络算法的 2 型糖尿病发病风险预测模型的构建 [J]. 中国慢性病预防与控制, 2020, 28 (4): 274-279.