李阳, 高海林, 李子杨, 等. 数据挖掘技术在糖尿病风险预测中的应用[J]. 智能计算机与应用,2024,14(12):133-138.
DOI:10.20169/j.issn.2095-2163.24091301

数据挖掘技术在糖尿病风险预测中的应用

李 阳, 高海林, 李子杨, 张蓓蓓, 武 丹
(南京医科大学康达学院 医学信息工程学部, 江苏 连云港 222000)

摘 要: 对糖尿病风险的准确预判能够促使病患及时干预与治疗,减轻痛苦与降低感染并发症的风险,本文探讨数据挖掘技术在糖尿病风险预测与辅助诊断的适用性。基于疾病特征选择剖析特征在疾病诊断中的重要性,进一步利用随机森林等机器学习算法预测糖尿病患病风险。针对 Sylhet 糖尿病数据集的分析发现,利用 SVM-RFE 筛选出 7 个疾病相关特征,特征数量降低 56%;以疾病相关特征为输入运用 5 类机器学习算法训练实现对糖尿病进行风险预测,各类算法预测准确率在 90%以上,其中随机森林算法表现最优,在十折交叉验证中准确率为 95.3%, *F*1 - *Score* 为 96.2%,*AUC* 为 0.980。实验结果表明,机器学习算法能够准确地预测糖尿病患病风险;以疾病特征选择获得诊断特征作为算法输入,在保证准确率的同时提升了诊断效率。

关键词: 糖尿病; 数据挖掘; 风险预测; 辅助诊断; 特征选择
中图分类号: R587.1;TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)12-0133-06

Application of data mining techniques in diabetes risk prediction

LI Yang, GAO Hailin, LI Ziyang, ZHANG Beibei, WU Dan
(Medical Information Engineering Department of Kangda College, Nanjing Medical University, Lianyungang 222000, Jiangsu, China)

Abstract: Predicting the risk of diabetes accurately can intervene and treat diabetes in a timely manner, reduce the risk of patients' pain and infection complications. This paper focuses on exploring the applicability of data mining technology in the diagnosis of diabetes risk prediction and auxiliary diagnosis. Disease feature selection is used to analyze the relevance of features in disease diagnosis, and several machine learning algorithms such as random forests are further utilized to predict the risk of diabetes. The analysis based on the Sylhet diabetes dataset found that 7 disease-related features are screened out by using SVM-RFE, reducing the number of features by 56%. With these disease-related features as input, the training of 5 machine learning algorithms achieves risk prediction for diabetes with prediction accuracy rates of over 90% for each algorithm. Among them, the Random Forest algorithm performs the best, with an accuracy rate of 95.3% in ten-fold cross-validation, an *F*1 - *Score* of 96.2%, and an *AUC* of 0.980. It is demonstrated that machine learning algorithms can accurately predict the risk of diabetes. Using diagnostic features obtained from disease feature selection as algorithm input not only ensures accuracy, but also improves diagnostic efficiency.

Key words: diabetes; data mining; risk prediction; assisted diagnosis; feature selection

0 引 言

糖尿病是一种由胰岛素分泌不足或机体对胰岛素的反应减弱所引发的慢性代谢疾病,导致血糖水平异常上升,严重威胁着公众健康。糖尿病被分为 1 型糖尿病、2 型糖尿病、妊娠期糖尿病以及其他特定类型的糖尿病,常伴有多尿、愈合延迟、视力模糊、脱发等症状,并且随着糖尿病病情周期的深入,患者常常易患并发症,如心血管疾病、肾脏问题、高血压、中风、眼部疾患,甚至可能导致下肢截肢等[1-2]。当前,中国糖尿病病情不容乐观,成为日益突出的公共卫生问题。糖尿病早期的准确诊断有助于患者及时调整生活方式并接受更有效的治疗,减轻病患痛苦与感染并发症的风险,同时可缓解治疗糖尿病的医疗负担。但糖尿病早期检测指标波动较大,检测率低且诊断流程较为复杂,因此对糖尿病的防治研究

基金项目: 南京医科大学康达学院第二期品牌专业建设工程资助项目(JX206000302); 南京医科大学康达学院科研发展基金课题项目(KD2023KYJJ024)。
作者简介: 李 阳(1991—),男,硕士研究生,主要研究方向:智能计算,生物信息。Email: liyangbocai@163.com。
收稿日期: 2024-09-13

力度依然有待增强。

在糖尿病的诊治过程中,需从大量临床样本与数据中,抽取与挖掘糖尿病的诊治信息与知识,协助医生为糖尿病的诊断提供新的方案。近年来,随着数据挖掘技术的快速发展与应用,将其应用于医疗领域成为研究的热点,关于糖尿病诊断与预测的相关研究一直在推进[3],诊断方法主要分为:

(1)以统计学为基础,包含 Cox 回归、Logistic 回归等方法。

(2)以机器学习为主,涵盖了决策树、K 近邻算法(KNN)、支持向量机、XGBoost 以及人工神经网络等多种算法。

Tabaei 等学者[4]建立了基于 Logistic 回归用于筛选糖尿病患者的模型,通过 ROC 曲线进行验证,证明了模型获得了较高的预测准确性。受到疾病风险预测研究的启发,Wilson 等学者[5]使用不同的临床特征组合分别建立模型,使用多种方法进行建模预测,结果发现使用简单临床模型中的特征组合即可达到较好的预测水平,由于特征易得,使得 FOS 糖尿病风险评估模型(FOS DM risk score)得到了广泛的临床应用。李娟等学者[6]建立基于支持向量机的糖尿病预测模型,得出基于线性核函数建立的模型对 2 型糖尿病发生的预测效果较好,并且综合环境和遗传因素共同作用对 2 型糖尿病的预测准确率要高于仅考虑环境因素的预测效果。Xue 等学者[7]采用支持向量机、朴素贝叶斯分类器和 LightGBM 等有监督的机器学习算法对 520 例糖尿病患者和 16~90 岁潜在糖尿病患者的实际数据进行训练,实验证明支持向量机针对该数据集效果最好。Palimkar 等学者[8]利用不同的机器学习算法进行早期糖尿病预测,即逻辑回归、随机森林分类器、支持向量机、决策树、K-最近邻、高斯过程分类器、AdaBoost 分类器和高斯朴素贝叶斯。针对 Pima 糖尿病公开数据集,Kandhasamy 等学者[9]比较了机器学习分类器(J48 决策树、K-近邻、随机森林、支持向量机)对糖尿病患者的分类,对算法的性能进行了测量,并在准确性、灵敏度和特异性方面进行了比较。李佳思[10]同样探索了多种机器学习算法,引入 SHAP 方法进行特征重要性分析。

现有的糖尿病诊断研究将临床上大量的指标特征全部或者通过人工筛选输入到糖尿病诊断模型中[11],其诊断结果易受到症状波动影响,使得糖尿病诊断率较低且已处于不可治愈阶段。利用糖尿病早期会出现尿频、烦渴等临床特征,通过问诊的形式可识别患有糖尿病的风险,但糖尿病临床特征较多,易造成糖尿病问诊过程繁琐且不准确。本文在糖尿病患病初期或存在某些临床症状基础上,挖掘糖尿病最相关临床特征,简化问诊流程,进一步探索机器学习算法在糖尿病潜在风险预测中的适用性。

1 资料与方法

1.1 数据集

Sylhet 糖尿病诊断数据集(https://github.com/OladosuO)[12]是由孟加拉国 Sylhet 糖尿病医院通过对最近患上糖尿病的个体或尚且未患糖尿病但存在某些症状的个体直接进行调查而创建。数据集包含 520 个样本,分别是 320 名糖尿病患者和 200 名非糖尿病患者,涉及可能患有糖尿病的 16 个临床特征,包括多尿症、烦渴、体重减轻、体弱、多食症等,数据集详细描述见表 1。

表 1 糖尿病早期诊断数据集属性描述

Table 1 Description of attributes in diabetes early diagnosis dataset

特征	含义	特征描述/取值
Age	年龄	16 - 90 (years)
Gender	性别	0: Female; 1: Male
Polyuria	多尿症	0: No; 1: Yes
Polydipsia	烦渴	0: No; 1: Yes
Sudden weight loss	体重减轻	0: No; 1: Yes
Weakness	体弱	0: No; 1: Yes
Polyphagia	多食症	0: No; 1: Yes
Genital thrush	生殖器疮口	0: No; 1: Yes
Visual blurring	视觉模糊	0: No; 1: Yes
Itching	瘙痒	0: No; 1: Yes
Irritability	烦躁	0: No; 1: Yes
Delayed healing	延迟康复	0: No; 1: Yes
Partial paresis	部分偏瘫	0: No; 1: Yes
Muscle stiffness	肌肉紧张	0: No; 1: Yes
Alopecia	脱发	0: No; 1: Yes
Obesity	肥胖	0: No; 1: Yes
Class	是否患有糖尿病	0:Negative; 1:Positive

1.2 疾病特征选择

糖尿病在临床上特征较多,选择能够有效判别糖尿病风险特征构成特征子集,其过程为疾病特征选择[13],特征选择从统计学检验与特征选择两个层面进行分析。

(1)统计检验特征在糖尿病样本组间是否具有显著性差异[14]。从医学资料划分的角度,将临床特征划分为定性属性与定量属性,采用卡方检验判定

定性属性与是否患有糖尿病的相关情况，核心思想在于评估理论频数与实际频数之间拟合程度检验，利用显著性概率判定疾病特征是否相关，进一步利用列联系数描述疾病与临床特征之间的关联强度；采用 F 检验判定定量特征与是否患有糖尿病间相关情况，剖析临床症状与疾病间的关系。

(2)基于递归特征消除的疾病相关特征选择，以疾病数据集特征全集为搜索起点，以分类器的预测精度为评价标准，通过循环迭代，消除最不相关的特征，完成相关特征的排序。以 SVM 为分类器的 SVM-RFE[15]能够有效地的分析出对诊断结果影响最相关特征。

1.3　基于数据挖掘的糖尿病风险预测

根据已有文献的研究，利用典型的机器学习算法预测糖尿病风险[9-10,16]，主要包括：线性模型(逻辑回归)[17]、支持向量机[18]、决策树[19]、K 邻近算法[20]和随机森林[21-22]等经典模型。针对糖尿病相关数据集，采用数据挖掘技术对糖尿病风险进行预测分析，诊病流程如图 1 所示。研究对此展开分述如下。

(1)数据预处理。涉及一定类型特征的独热编码与缺失数据的检测替换。

(2)疾病诊断特征选择分析。利用统计检验与递归特征消除法选择疾病判别最相关属性。

(3)模型构建。分为模型训练与模型测试。以样本特征为输入，对机器学习算法进行训练；利用测试集对已构建的模型进行预测检验。

(4)模型评估。主要使用准确率、精确率、召回率、*F*1 - *Score*、*AUC* 等指标评估模型在糖尿病风险预测中的性能。

(5)模型应用。针对新样本的诊断报告，可通过构建的模型，对其是否存在患有糖尿病风险进行预测。

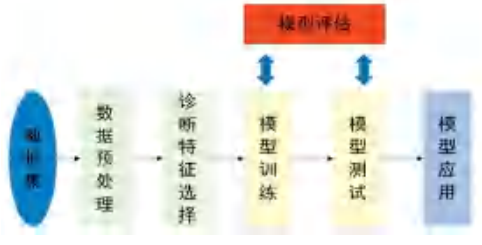


图 1　基于数据挖掘的糖尿病诊断实现流程

Fig. 1　Diabetes diagnosis implementation process using data mining

基于数据挖掘技术对糖尿病风险预测过程可概述为：将数据集按 8∶2 随机划分成训练集与测试集，分别利用样本全部特征和选择特征作为模型输入、是否患有糖尿病作为模型的输出，对支持向量机等 5 类机器学习模型逐一进行训练，并利用测试集对其进行检验。

1.4　模型评价指标

在模型训练与测试过程中，采用准确率、精确率、召回率、*F*1 - *Score*、*AUC* 等多个指标对模型预测效果进行评价[9,14]。

数据集样本按照真实类别与分类模型预测的类别进行统计汇总，形成 True Positive、False Negative、False Positive、True Negative 四个统计值，构建混淆矩阵。其中，True Positive (*TP*) 表示样本的真实类别是正类，并且模型识别的结果也是正类；False Negative (*FN*) 表示样本的真实类别是正类，但是模型将其识别为负类；False Positive (*FP*) 表示样本的真实类别是负类，但是模型将其识别为正类；True Negative (*TN*) 表示样本的真实类别是负类，并且模型将其识别为负类。基于形成的混淆矩阵，评价指标的计算如下。

(1)准确率 (*Accuracy*)。具体公式为：

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

(2)查准率 (*Precision*)。具体公式为：

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

(3)查全率 (*Sensitive*)。具体公式为：

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

(4) *F*1 - *Score*。具体公式为：

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

2　糖尿病应用实验与分析

2.1　糖尿病实验结果

Sylhet 糖尿病数据集，包括 Gender、Polyuria、Polydipsia 等 15 个定性特征与一个定量特征 Age，利用特征选择剖析疾病症状关联程度。针对定性特征，统计疾病与特征之间的频率分布形成的疾病-属性之间的列联表如图 2 所示。图 2 中，包括是否患有糖尿病与性别之间的关系、与体重减轻之间的关系等。并且利用卡方检验与 F 检验判别疾病与特诊之间关系的显著性以及获得其相关程度。设定检验水平为 0.05，判断特征与疾病之间的相关情

况。利用 SVM-RFE 递归特征消除法计算获得特征在疾病判别过程中的重要程度,结果见表 2。

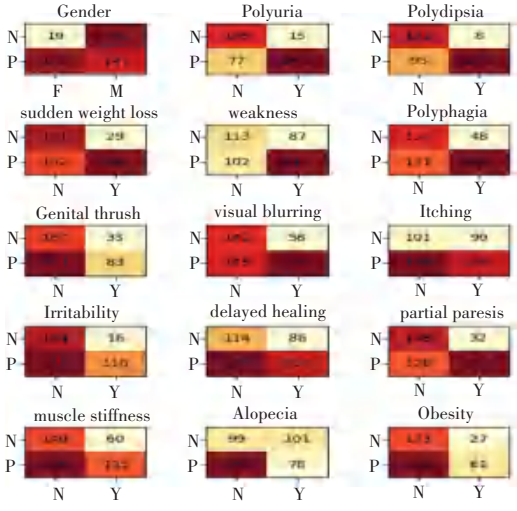


图 2 疾病症状频数统计结果

Fig. 2 Disease symptoms frequency statistics results

考虑到特征与疾病间的相关性和特征重要性排序,选择 7 个特征作为疾病诊断相关特征集,包括年龄、性别两个患者信息与多尿症、烦渴、体重减轻、烦躁和部分偏瘫五个临床特征。将数据集按 8∶2 随机划分成训练集与测试集,训练集包括 416 个样本,测试集包括 104 个样本,分别利用样本 16 个特征属性与构建的疾病诊断特征集 7 个属性特征作为模型的输入,实现了逻辑回归等 5 类模型训练,构建的糖尿病诊断决策树如图 3 所示。

表 2 疾病特征相关分析

Table 2 Disease symptoms correlation analysis

特征	检验方法	显著性概率	是否相关	关联强度	特征重要性排序
Age	F 检验	0.013 0	是	0.106	3
Gender	χ^2 检验	3.290e-24	是	0.406	4
Polyuria	χ^2 检验	1.741e-51	是	0.551	2
Polydipsia	χ^2 检验	6.187e-49	是	0.541	1
Sudden weight loss	χ^2 检验	5.969e-23	是	0.397	5
Weakness	χ^2 检验	4.870e-08	是	0.232	15
Polyphagia	χ^2 检验	1.165e-14	是	0.320	10
Genital thrush	χ^2 检验	0.016 1	是	0.104	14
Visual blurring	χ^2 检验	1.701e-08	是	0.240	11
Itching	χ^2 检验	0.830 0	否	0.009	13
Irritability	χ^2 检验	1.771e-11	是	0.282	7
Delayed healing	χ^2 检验	0.327 0	否	0.042	8
Partial paresis	χ^2 检验	1.565e-22	是	0.393	6
Muscle stiffness	χ^2 检验	0.007 0	是	0.117	12
Alopecia	χ^2 检验	1.909e-09	是	0.254	9
Obesity	χ^2 检验	0.127 0	否	0.066	16

针对多种机器学习算法对糖尿病风险预测的预测结果,采用准确率、精准率、召回率、*F*1 - *Score*、*AUC* 指标对算法进行评价,比较结果见表 3。利用测试集进行检验,得到的 5 类算法预测 ROC 曲线如图 4 所示。

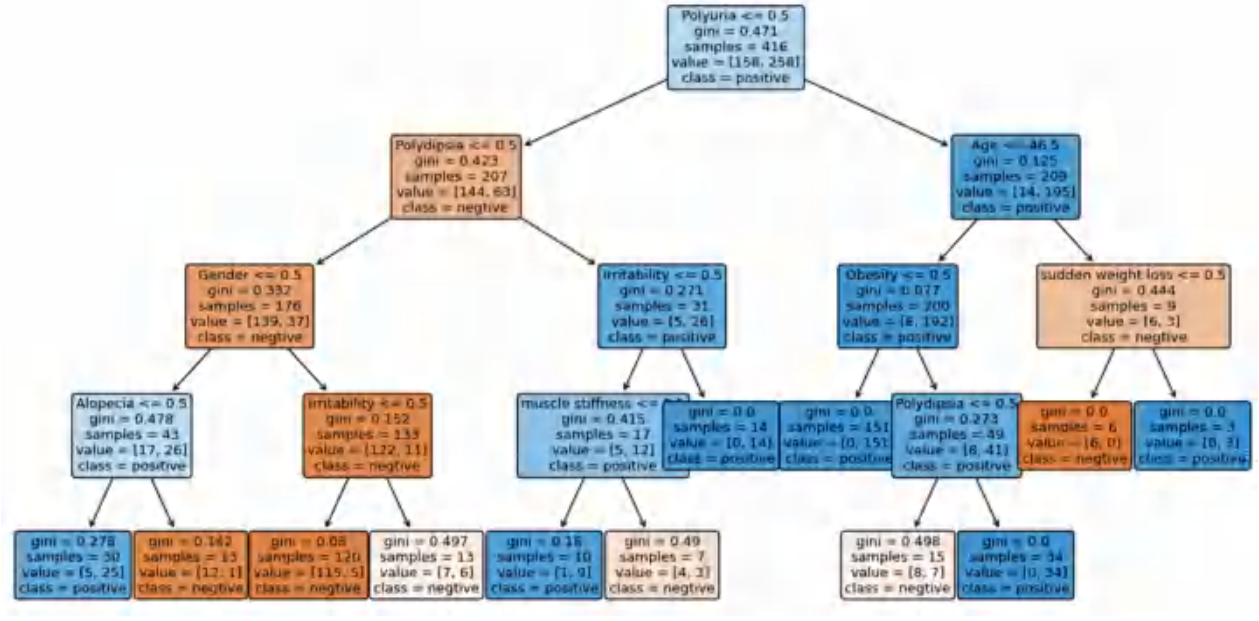


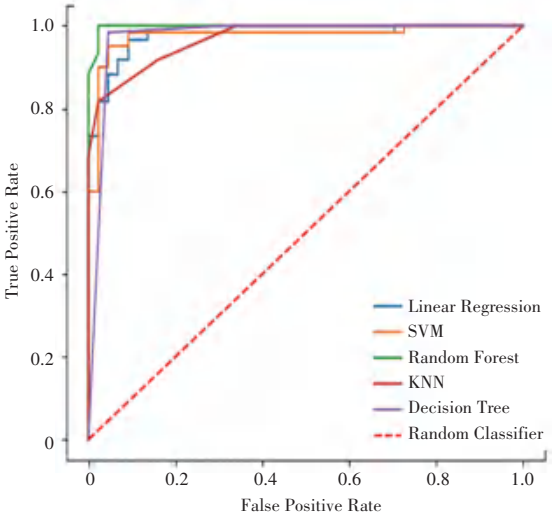
图 3 糖尿病诊断决策树构建

Fig. 3 Construction of decision tree for diabetes diagnosis

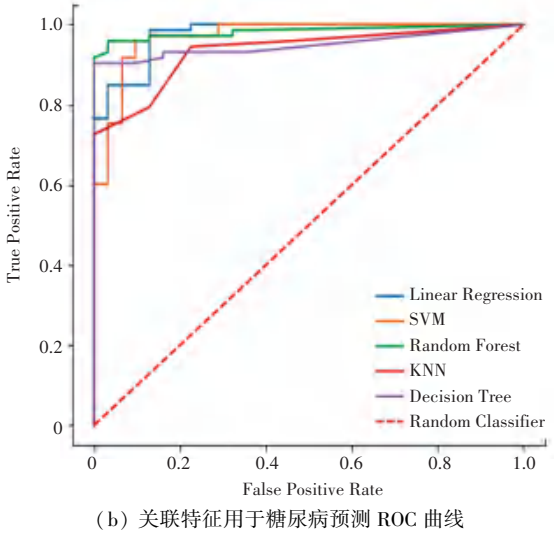
表 3　多种算法对糖尿病风险预测结果评价

Table 3　Evaluation of diabetes risk prediction results by multiple algorithms

模型	指标	所有特征					疾病判别相关特征				
		逻辑回归	SVM	决策树	KNN	随机森林	逻辑回归	SVM	决策树	KNN	随机森林
训练集	*Acc*	0. 937	0. 944	0. 995	0. 923	1. 000	0. 927	0. 925	0. 966	0. 915	0. 983
	Precision	0. 941	0. 949	0. 996	0. 974	1. 000	0. 935	0. 928	0. 987	0. 952	0. 990
	Recall	0. 956	0. 960	0. 996	0. 897	1. 000	0. 943	0. 947	0. 955	0. 902	0. 979
	F1 - Score	0. 949	0. 954	0. 996	0. 934	1. 000	0. 939	0. 937	0. 971	0. 927	0. 985
	AUC	0. 932	0. 940	0. 994	0. 930	1. 000	0. 924	0. 920	0. 968	0. 918	0. 983
测试集	*Acc*	0. 923	0. 942	0. 961	0. 875	0. 980	0. 923	0. 942	0. 923	0. 817	0. 961
	Precision	0. 967	0. 968	1. 000	0. 981	1. 000	0. 945	0. 958	0. 985	0. 935	0. 985
	Recall	0. 909	0. 939	0. 939	0. 818	0. 969	0. 945	0. 958	0. 904	0. 794	0. 958
	F1 - Score	0. 937	0. 953	0. 968	0. 892	0. 984	0. 945	0. 958	0. 942	0. 859	0. 972
	AUC	0. 928	0. 943	0. 969	0. 895	0. 984	0. 908	0. 931	0. 935	0. 832	0. 963



（a）所有特征用于糖尿病预测 ROC 曲线



（b）关联特征用于糖尿病预测 ROC 曲线

图 4　糖尿病预测 ROC 曲线

Fig. 4　ROC curve for diabetes predictions

进一步，研究基于相同的模型设定，对 Sylhet 糖尿病数据集采用 10 折交叉验证得到了 5 类机器学习算法在糖尿病风险预测上的效果，模型综合评估结果见表 4。

表 4　10 折交叉验证多种算法对糖尿病预测效果

Table 4　10-fold cross validation of multiple algorithms for diabetes predictions

模型	*Acc*	*Precision*	*Recall*	*F1 - score*	*AUC*
逻辑回归	0. 913	0. 935	0. 928	0. 930	0. 963
SVM	0. 898	0. 908	0. 934	0. 919	0. 958
决策树	0. 944	0. 966	0. 943	0. 954	0. 952
KNN	0. 898	0. 951	0. 878	0. 912	0. 952
随机森林	0. 953	0. 973	0. 953	0. 962	0. 980

2.2　实验结果分析

Sylhet 糖尿病数据集以定性数据为主，其特征部分为糖尿病早期临床症状。通过表 4 发现，利用卡方检验与 F 检验探究疾病与特征的关联关系，检验发现 13 个特征与是否患有糖尿病是相关的。其中，多尿症（Polyuria）和烦渴（Polydipsia）特征与糖尿病关联强度最大，瘙痒（Itching）、延迟康复（Delayed healing）、肥胖（Obesity）与是否患有糖尿病无统计相关。进一步基于 SVM-RFE 方法特征重要性排序与相关程度总体上一致，Delayed healing 特征对应的重要性程度排在第 8 位，但是检验样本组间无显著性差异，所以仅选择 7 个特征作为疾病判别相关特征，特征数量减少 56%。根据表 3 与表 4 结果讨论可知，利用机器学习算法对糖尿病风险

预测总体表现较好,准确率在 90%以上,表明 SVM、逻辑回归、决策树与随机森林等算法能够利用有效的糖尿病临床信息进行疾病诊断。比较全部特征与关联特征作为输入进行算法学习,在多项指标上均无显著性差异。综合比较各类算法,随机森林糖尿病预测表现最优,利用疾病判别相关特征作为随机森林的输入,在 10 折交叉验证中准确率为 95.3%,精准率为 97.3%,召回率为 95.3%,*F*1 - *Score* 为 96.2%,*AUC* 为 0.98。

3 结束语

本文将机器学习算法应用于糖尿病风险预测,并且探讨了糖尿病与临床特征之间的相关程度,筛选出疾病诊断相关特征以提升问诊效率。基于 Sylhet 糖尿病数据集进行数据挖掘实验,从 16 个样本属性筛选出 7 个疾病诊断相关特征;通过随机构建训练和测试集与 10 折交叉验证两种形式完成多种模型的疾病诊断评估,机器学习算法用于糖尿病诊断在准确率、精确率、召回率、*F*1 - *Score*、*AUC* 等指标上均取得了较好的结果。并且利用样本全部特征与关联特征分别作为模型输入,在糖尿病诊断效果上无显著性差异。针对类似 Sylhet 糖尿病数据集形式的问诊样本,糖尿病与其他疾病的诊断过程类似决策树的构建,决策树能够逐步挑选出最有效的分类指标来对疾病做出诊断,能够提高问诊效率与准确率,为糖尿病初步诊断提供技术支撑。将多个决策树进行组合形成的随机森林在疾病诊断中具有更好的表现。因此,随机森林等集成学习算法在疾病风险预测与诊断中具有较强的适用性。

参考文献

[1] GENUTH S M, PALMER J P, NATHAN D M. Classification and diagnosis of diabetes[M]. Bethesda, MD: National Institute of Diabetes and Digestive and Kidney Diseases, 2021.

[2] 关子安. 现代糖尿病学[M]. 天津: 天津科学技术出版社, 2000.

[3] JAISWAL V, NEGI A, PAL T. A review on current advances in machine learning based diabetes prediction[J]. Primary Care Diabetes, 2021, 15(3): 435-443.

[4] TABAEI B P, HERMAN W H. A multivariate logistic regression equation to screen for diabetes: development and validation[J]. Diabetes Care, 2002, 25(11): 1999-2003.

[5] WILSON P W, MEIGS J B, SULLIVAN L, et al. Prediction of incident diabetes mellitus in middle-aged adults: The framingham offspring study[J]. Archives of Internal Medicine, 2007, 167(10): 1068-1074.

[6] 李娟, 吴疆, 卢莉, 等. 基于支持向量机建立环境和遗传因素对 2 型糖尿病的预测模型[J]. 中华疾病控制杂志, 2012, 16(2):171-175.

[7] XUE Jingyu, MIN Fanchao, MA Fengying. Research on diabetes prediction method based on machine learning[J]. Journal of Physics: Conference Series, 2020, 1684(1): 012062.

[8] PALIMKAR P, SHAW R N, GHOSH A. Machine learning technique to prognosis diabetes disease: Random forest classifier approach[C]// Proceedings of ICACIT on Advanced Computing and Intelligent Technologies. Cham : Springer, 2022: 219-244.

[9] KANDHASAMY J P, BALAMURALI S. Performance analysis of classifier models to predict diabetes mellitus[J]. Procedia Computer Science, 2015, 47: 45-51.

[10] 李佳思. 基于机器学习的糖尿病预测及 SHAP 特征分析[J]. 智能计算机与应用, 2023, 13(1):153-157.

[11] 王成武, 晏峻峰. 早期糖尿病风险预测模型的比较研究[J]. 智能计算机与应用, 2021, 11(1):64-68.

[12] ISLAM M M, FERDOUSI R, RAHMAN S, et al. Likelihood prediction of diabetes at early stage using data mining techniques[C]//Computer Vision and Machine Intelligence in Medical Image Analysis. Piscataway, NJ: IEEE, 2020: 113-125.

[13] LIU Huan, YU Lei. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491-502.

[14] 徐勇勇. 医学统计学[M]. 北京:高等教育出版社, 2014.

[15] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46: 389-422.

[16] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.

[17] NUSINOVICI S, THAM Y C, YAN M Y C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases[J]. Journal of Clinical Epidemiology, 2020, 122: 56-69.

[18] SWEILAM N H, THARWAT AA, MONIEM N K A. Support vector machine for diagnosis cancer disease: A comparative study[J]. Egyptian Informatics Journal, 2010, 11(2): 81-92.

[19] MAJI S, ARORA S. Decision tree algorithms for prediction of heart disease[C]// Proceedings of the Third International Conference on Information and Communication Technology for Competitive Strategies (ICTCS 2017). Cham: Springer, 2019: 447-454.

[20] UDDIN S, HAQUE I, LU H, et al. Comparative performance analysis of K-Nearest Neighbour (KNN) algorithm and its different variants for disease prediction[J]. Scientific Reports, 2022, 12(1): 6256.

[21] BREIMAN L. Random Forests[J]. Machine learning, 2001,45: 5-32.

[22] BIAU G. Analysis of a random forests model[J]. The Journal of Machine Learning Research, 2012, 13(1): 1063-1095.