

基于机器学习的早期糖尿病风险预测

摘要: 糖尿病疾病是一个日益严重的医学问题,它是一种代谢疾病,身体内的葡萄糖长期处于一个高水平的状态,会产生尿频、口渴、饥饿程度加剧等症状,从而导致肾衰竭、中风、视力受损等并发症的产生。糖尿病的识别往往是病人询问医生或者是到诊断中心询问,会使诊断过程过于繁琐。但是逐步上升的机器学习方法解决了这一问题。本文的目的是采用机器学习方法,预测患者患糖尿病的可能性。因此采用四个机器学习分类算法,即支持向量机、决策树、随机森林及 XGBoost,来检测早期糖尿病。实验采用的是从孟加拉国锡尔赫特的锡尔赫特医院患者那里收集的直接问卷。这四个算法的性能评估采用准确率、精确率、召回率、F1 分数、AUC 五个指标来进行评估。实验显示 XGBoost 的预测效果优于其他算法。

关键词: 糖尿病;风险预测;支持向量机;决策树;随机森林;XGBoost

1 引言

据国际糖尿病联合会 2019 年报告显示,全球糖尿病患者数约为 4.63 亿,中国患病人数高达 1.16 亿,居世界第一。近年来,随着生活水平的提高,我国糖尿病患者数量不断增加,其防治已成为我国重要的公共卫生问题^[1]。糖尿病是继心脑血管疾病、恶性肿瘤之后第三大威胁人类健康的慢性病,糖尿病患者可能会出现严重并发症,如脑血管意外、视网膜脱落、肾脏损伤等^[2],不仅给患者的生活带来严重影响,还给社会带来了沉重的经济负担。对于大多数糖尿病患者,如果及早发现并开始治疗,并发症将很容易控制,甚至可以避免发生。因此,进行早期糖尿病风险预测,对于降低糖尿病及其并发症的发病率、节约国家医疗资源具有重要意义。

机器学习的分类算法广泛应用与医学领域的数据分类。糖尿病受身高、体重、遗传和胰岛素功能等功能的影响,我们考虑的主要因素就是血糖浓度。早期识别是唯一远离并发症的补救方法^[3]。许多研究者进行疾病诊断实验时,会使用各种分类的机器学习算法,例如:支持向量机(SVM)、朴素贝叶斯、决策树、逻辑斯蒂回归、神经网络等等。数据挖掘和机器学习方法对于来自不同数据源的数据的疾病诊断处理具有强大的能力^[4]。在研究糖尿病, Nai-Arun 等人^[5]提出了一种分类集成学习来研究糖尿病,利用增益比特征选择技术对数据进行分析。Orabi 等人^[6]介绍了一种通过提高预防措施警报来帮助人们的方法。它是糖尿病疾病的预测系统,它将预测是否成为候选人以及在什么年龄。该系统基于机器学习概念,使用决策树技术,通过添加带有随机化代码的回归技术来预测年龄。Bamnote 等人^[7]提出了一种使用遗传编程(GP)检测糖尿病的分类器,使用分类表达式创建分类器。使用仅算术运算符的简化函数池,允许在交叉和突变期间进行较少的验证和宽大处理。Nai-Arun 等人^[8]首先研究了四个众所周知的分类模型,即决策树、人工神经网络、逻辑回归和朴素贝叶斯。然后,研究了袋装和增压技术以提高此类模型的鲁棒性。诸如对糖尿病的研究还有很多很多,如 Vijiya Kumar 等人^[9]提出的使用随机森林对糖尿病进行预测;Sisodia 等人^[10]研究的是使用分类算法预测糖尿病等等。

本次的研究工作是关注早期糖尿病这一疾病。在这项工作中采取了支持向量机、决策树、随机森林和 XGBoost,这四种机器学习方法来对早期糖尿病进行预测。在四种机器学习方法下,都取得了良好的精度。

2 相关理论及方法

2.1 支持向量机 (SVM)

SVM 是一种有监督学习算法,可用于解决数据挖掘或模式识别领域中的数据分类问题。其基本思想是建立一个最优决策超平面,使得该平面两侧距平面最近的两类样本之间的距离最大化,从而使得该模型用于分类问题时能具有良好的泛化能力。SVM 适用于样本较小、非线性及高维空间问题,可与其他机器学习算法联合使用。SVM 通过引入拉格朗日常数解决凸二次优化问题,表示为:

$$L(\omega, \lambda, \alpha) = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^N \lambda_i \{y_i(\omega x_i + b) - 1\} \quad (1)$$

式中, $\|\omega\|$ 为正常超平面的范数, b 为常数, λ_i 为拉格朗日乘数, $x_i (i = 1, 2, \dots, n)$ 为线性可分的向量, y_i 为输出类。

2.2 决策树

决策树(Decision Tree)是一个监督机器学习算法, 主要用于研究分类问题。一个决策树学习算法需要包含特征选择、决策树生成和决策树剪枝过程。本文使用的是 CART 算法, 它的核心是“最小化基尼指数”——基尼指数越小, 样本集的纯净度越高, 特征的区分能力越强。基尼指数也就是: $Gini_index = \sum_{v=1}^V \frac{|D_v|}{|D|} Gini(D_v)$, 其中, D_v 是特征 A 取值为 v 时对应的子样本集, $|D|$ 是总样本集样本数, $|D_v|$ 是子样本集样本数。

决策树一般适合处理离散型数据, 计算复杂度不高, 对中间值的缺失不敏感, 可以处理不相关特征数据。但是决策树方法可能产生过度匹配问题, 对连续性的字段比较难以预测。它通常适用于金融分析、医疗诊断、营销推荐、交通安全等。

2.3 随机森林

随机森林(Random Forest)是建立多个决策树并将他们融合起来得到一个更加准确和稳定的模型, 是 bagging 思想和随机选择特征的结合。随机森林构造了多个决策树, 当需要对某个样本进行预测时, 统计森林中的每棵树对该样本的预测结果, 然后通过投票法从这些预测结果中选出最后的结果。

随机森林适用于高维数据, 不容易产生过拟合。对于大部分数据遗失, 仍然可以维持高准确度。对于数据集的适应能力强, 既能处理离散型数据, 也能处理连续性数据, 数据集也无需规范化。

2.4 XGBoost

XGBoost 算法是一个优化的分布式梯度增强库, 其以 CART 决策树作为基分类器, 采用新增树形成的新函数拟合之前预测的残差, 然后累加所有树的预测结果, 得到最终预测结果。XGBoost 的目标函数为:

$$\min L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3)$$

式中, n 为训练样本数量, k 为决策树数量, f_k 为基学习器。损失函数 l 用于衡量真实分数与预测分数的差距。正则化项 Ω 包含两个部分, 其中 T 表示叶子节点数量, w 表示叶子节点分数; γ 和 λ 表示惩罚力度, 可控制叶子节点数量并限制节点分数, 防止模型过分贴合训练数据而损失预测效果导致过拟合。

3 实验设计

3.1 数据集说明

本次使用的数据是早期糖尿病风险预测数据集, 从孟加拉国锡尔赫特的锡尔赫特糖尿病医院的患者那里收集的直接问卷。问卷中包含了年龄、性别、多尿等等, 如表 1 所示, 共收集了 520 名患者数据。

3.2 数据预处理

将 520 个样本按照 7:3 的比例分为训练集和测试集两部分。将患病赋值为 1, 未患病赋值为 0, 组间比较采用卡方检验。选取对糖尿病患病有显著影响的变量作为自变量输入建立预测模型, 同时采样数据标准化操作, 采用有结果标签的训练集对模型进行训练, 最后使用测试数据对预测模型进行分类准确性的评价比较。采用 Python 对其进行数据分析。

对 16 个变量进行分组描述, 并进行差异性检验。如表 2 所示, Age (X_1), Itching (X_{10}), Delayed healing (X_{12}), Obesity (X_{16}) 4 个变量对是否患糖尿病无显著性影响, 因此在构建分类模型时, 采用余下 12 个变量

作为自变量输入。

表 1 变量赋值

变量	赋值说明
Class(Y, 类别)	不患病=0, 患病=1
Age(X_1 , 年龄)	$<50=0$, $\geq 50=1$
Gender(X_2 , 性别)	Female=0, Male=1
Polyuria(X_3 , 多尿症)	No=0, Yes=1
Polydipsia(X_4 , 烦渴)	No=0, Yes=1
Sudden weight loss(X_5 , 体重减轻)	No=0, Yes=1
Weakness(X_6 , 无力)	No=0, Yes=1
Polyphagia(X_7 , 多食症)	No=0, Yes=1
Genital thrush(X_8 , 生殖器鹅口疮)	No=0, Yes=1
Visual blurring(X_9 , 视物模糊)	No=0, Yes=1
Itching(X_{10} , 皮肤瘙痒)	No=0, Yes=1
Irritability(X_{11} , 暴躁易怒)	No=0, Yes=1
Delayed healing(X_{12} , 伤口愈合慢)	No=0, Yes=1
Partial paresis(X_{13} , 部分麻痹性偏瘫)	No=0, Yes=1
Muscle stiffness(X_{14} , 肌肉紧张)	No=0, Yes=1
Alopecia(X_{15} , 脱发)	No=0, Yes=1
Obesity(X_{16} , 肥胖)	No=0, Yes=1

表 2 糖尿病患病情况单因素分析

特征列	χ^2	p值	自由度	是否与糖尿病显著关联($\alpha = 0.05$)
Age	2.1388	0.143614	1	否
Gender	103.0369	0	1	是
Polyuria	227.8658	0	1	是
Polydipsia	216.1716	0	1	是
Sudden weight loss	97.2963	0	1	是
Weakness	29.7679	0	1	是
Polyphagia	59.5953	0	1	是
Genital thrush	5.7921	0.016098	1	是
Visual blurring	31.8085	0	1	是
Itching	0.0462	0.829748	1	否
Irritability	45.2083	0	1	是
Delayed healing	0.9621	0.32666	1	否
Partial paresis	95.3876	0	1	是
Muscle stiffness	7.2887	0.006939	1	是
Alopecia	36.0641	0	1	是
Obesity	2.3275	0.127108	1	否

3.3 模型评价标准

在二分类预测模型评估指标中，混淆矩阵可用于判断分类器分类效能优劣，具体如表 3 所示。采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数 4 个指标评价各模型性能。

（1）准确率（Accuracy, ACC）。该指标为预测结果和真实结果同为正例和同为反例占所有样本的比例，反映分类器对整个样本的判定能力，表示为：

$$ACC = \frac{TP + TN}{TP + TN + FN + FP}$$

(4)

表 3 混淆矩阵

	预测为正例	预测为负例
真实为正例	True Positive(TP)	False Negative(FN)
真实为负例	False Positive(FP)	True Negative(TN)

（2）精确率（Precision）。该指标为预测正确的正例数占预测为正例总量的比例，表示为：

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

(3) 召回率 (Recall)。该指标为预测正确的正例数占真正正例数的比例，表示为：

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

(4) F1 分数。该指标表示为精确率与召回率的调和平均值，表示为：

$$F1 = \frac{2TP}{FP + FN + 2TP} \quad (7)$$

(5) AUC。该指标表示 ROC 曲线下的面积，取值在 0.5~1 之间；ROC 曲线横轴表示负例分错的概率，纵轴表示正例分对的概率。AUC 可以直观地评价分类器性能优劣，其值越大越好。

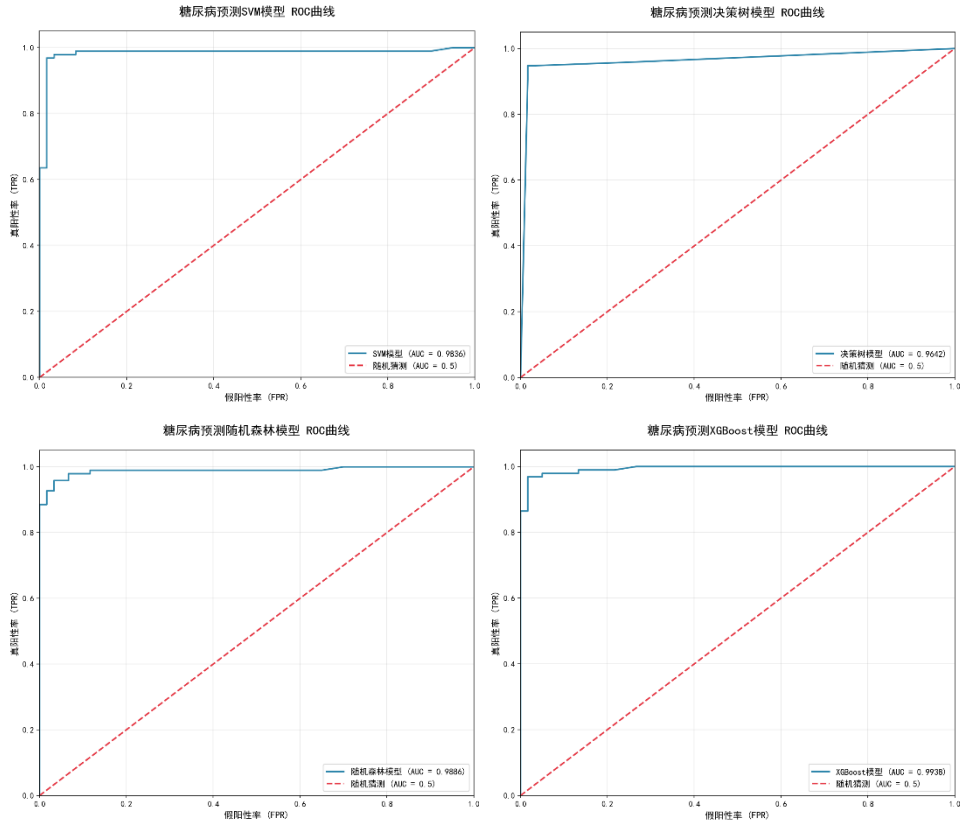


图 1 ROC 曲线

3.4 模型预测性能评价

基于 SVM、决策树、随机森林、XGBoost 算法建立预测模型，4 种分类器的性能比较结果见表 4，ROC 曲线如图 1 所示，决策树可视化如图 2 所示。可以看出，XGBoost 的区分度最好，AUC 达到 99.38%。从准确率、精确率、召回率和 F1 结果可以看出，XGBoost 的预测准确率最高，达到 97.44%，其精确率和召回率在 4 个模型中最佳。集成学习模型的预测性能比单个分类器有所提升，能够更为精确地进行糖尿病风险预测。

表 4 4 个分类器测试集预测结果

算法	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
SVM	97.44	97.45	98.32	97.21
决策树	95.51	95.76	96.82	95.38
随机森林	95.51	95.72	95.44	95.93
XGBoost	97.44	97.66	98.73	97.93

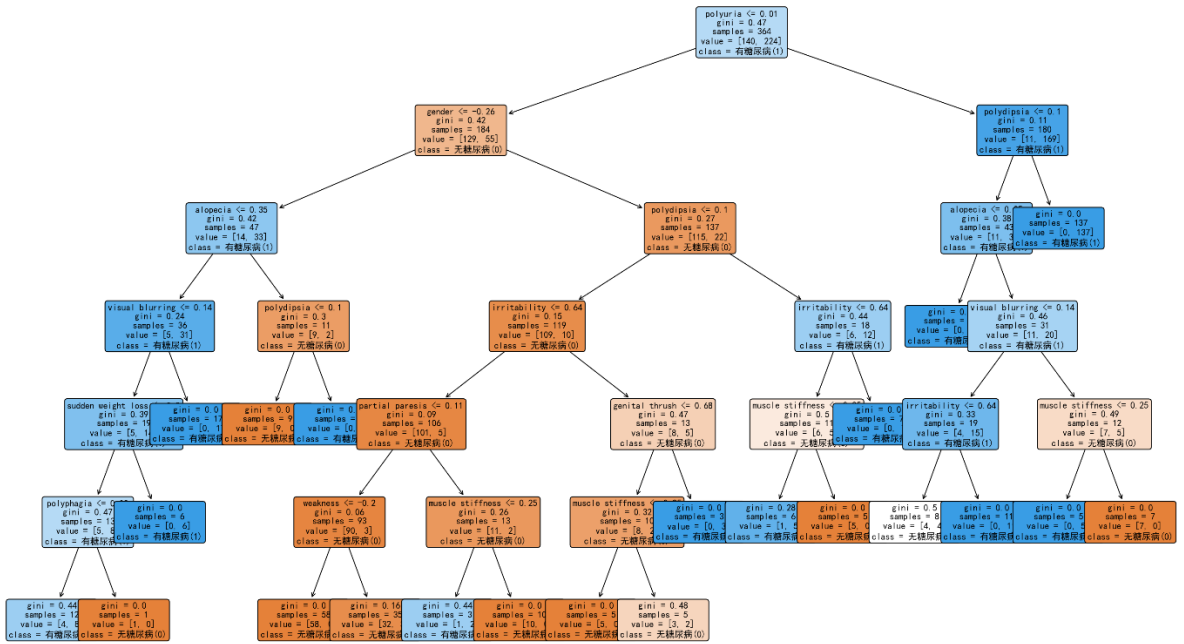


图2 决策树可视化

预测性能最好的 XGBoost 算法给出的变量重要性如图 3 所示。影响权重排名前十的因素依次为烦渴、多尿症、脱发、感到无力疲惫、体重突然减轻等。

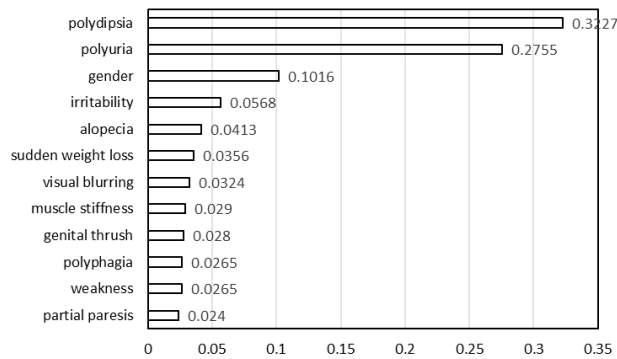


图3 变量权重值

4 总结

本文采用早期糖尿病风险预测数据集，通过使用 SVM、决策树、随机森林及 XGBoost 模型，探讨了早期糖尿病风险因素的预测。风险因素预测在识别摆脱早期糖尿病疾病的风险方面起着重要的作用。使用预测效果最好的模型分析糖尿病风险因素，发现多尿(polyuria)和烦渴(polydipsia)是导致糖尿病的主要因素。

其次从模型预测结果上来分析，四种模型的精度都达到 90%以上，说明这四种模型都能很好的预测早期糖尿病疾病。在所使用模型中 XGBoost 模型的精度最高(97.44%)，所以预测早期糖尿病疾病来说采用 XGBoost 效果更佳。

References:

- [1] GUO L X. Diabetes epidemic situation and coping strategies [J]. Chinese Journal of Clinical Healthcare, 2020, 23(4): 433-436.

- [2] 王琦琪,戴家佳,崔熊卫.基于集成学习模型的糖尿病患病风险预测研究[J].软件导刊,2022,21(04):62-66.
- [3] 练春兰. 基于机器学习方法的早期糖尿病风险预测 [J].Statistics and Application, 2023, 12(04):974-984.DOI:10.12677/sa.2023.124101.
- [4] Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9, 1-16. <https://doi.org/10.4236/jilsa.2017.91001>
- [5] Nai-Arun, N. and Sittidech, P. (2014) Ensemble Learning Model for Diabetes Classification. Advanced Materials Research, 931-932, 1427-1431. <https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427>
- [6] Ying, W.X., Hong, H.S., Quan, L.Z., et al. (2009) The Predictive Role of Diabetes Mellitus with Erectile Dysfunction on Cardiovascular Risk. Chinese Circulation Journal.
- [7] Pradhan, M. and Bamnote, G.R. (2015) Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems & Computing, 327, 763-770. https://doi.org/10.1007/978-3-319-11933-5_86
- [8] Nai-Arun, N. and Moungrmai, R. (2015) Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science, 69, 132-142. <https://doi.org/10.1016/j.procs.2015.10.014>
- [9] Vijiya Kumar, K., Lavanya, B., Nirmala, I. and Sofia Caroline, S. (2019) Random Forest Algorithm for the Prediction of Diabetes. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, 29-30 March 2019, 1-5. <https://doi.org/10.1109/ICSCAN.2019.8878802>
- [10] Sisodia, D. and Sisodia, D.S. (2018) Prediction of Diabetes Using Classification Algorithms. Procedia Computer Science, 132, 1578-1585. <https://doi.org/10.1016/j.procs.2018.05.122>