人工智能技术应用于医学数据分析 DOI：10.13406/j.cnki.cyxb.003387

基于人工智能的糖尿病预测研究

周乐明[1]，尚明生[2]，王永红[3]，宋景麟[4]，李小松[3]，黄　刚[5]，王　科[6]
（1. 重庆邮电大学计算机科学与技术学院，重庆　400065；2. 中国科学院重庆绿色智能技术研究院大数据中心，重庆　400714；3. 重庆市黔江中心医院检验科，重庆　409099；4. 重庆银行博士后研究中心，重庆　400024；5. 成都市第三人民医院心血管内科，重庆　610031；6. 重庆市永川区人民医院检验科，重庆　402160）

【摘　要】目的：以临床类指标建立基于极限梯度增强（extreme gradient boosting，XGBoost）、基于梯度提升树的分类器（light gradient boosting machine，LightGBM）、自适应增强（adaptive boosting，AdaBoost）、多层感知器（multilayer perceptron，MLP）等4种分类器的糖尿病预测模型，并评价其筛查效果。**方法：**根据病例对照研究设计采集研究组、对照组的99项临床类数据，使用python3.8进行了分析，接着采用线性插补、固有非负隐特征（inherent non negative implicit features，INLF）模型等方法对特征缺失值进行了预测，然后使用4种分类器构建分类模型来检测糖尿病。**结果：**3 241例高血压合并糖尿病患者作为研究组，4 181例高血压患者作为对照组被纳入模型进行分析，包含99个特征，通过基于XGBoost、LightGBM、AdaBoost和MLP等4种分类器的糖尿病鉴别分类准确率分别为0.894 9、0.887 5、0.862 0、0.856 6。**结论：**本研究提出基于INLF预测的分类器模型框架的筛查效果较好，初步解决了通过机器学习来进行糖尿病早期筛查的问题，对临床诊断具有一定的实际意义，可作为一种简单、有效的糖尿病及其并发症筛查的方法。

【关键词】不完备数据；糖尿病并发症；固有非负隐特征；分类器
【中图分类号】R587.1　　**【文献标志码】**A　　**【收稿日期】**2023-11-17

Diabetes prediction based on artificial intelligence

Zhou Leming[1]，*Shang Mingsheng*[2]，*Wang Yonghong*[3]，*Song Jinglin*[4]，*Li Xiaosong*[3]，*Huang Gang*[5]，*Wang Ke*[6]
（*1. College of Computer Science and Technology，Chongqing University of Posts and Telecommunications；2. Big Data Center，Chongqing Institute of Green and Intelligent Technology，Chinese Academy of Sciences；3. Clinical Laboratory，Qianjiang Central Hospital of Chongqing；4.Bank of Chongqing Postdoctoral Research Center；5. Department of Cardiovascular Medicine，The Third People's Hospital of Chengdu；6.Clinical Laboratory，Yongchuan People's Hospital of Chongqing*）

【Abstract】*Objective*：To establish a diabetes prediction model based on four classifiers of extreme gradient boosting（XGBoost），light gradient boosting machine（LightGBM），adaptive boosting（AdaBoost），and multilayer perceptron（MLP） according to clinical indicators，and to evaluate the screening effect. ***Methods*：**According to the case-control study design，99 attributes of clinical data from the study group and the control group were collected，and analyzed by python 3.8. Then the linear interpolation method and an inherent non-negative latent feature（INLF） model were used to predict the feature missing value，and the classification model was constructed using four classifiers to detect diabetes. ***Results*：**Through analyses of 3 241 patients with hypertension combined with diabetes（study group） and 4 181 patients with hypertension（control group） in the model，99 features were included. The accuracy rates of the diabetes classification model based on XGBoost，LightGBM，AdaBoost，and MLP classifiers were 0.894 9，0.887 5，0.862 0，and 0.856 6，respectively. ***Conclusion*：**Our proposed classifier model framework based on INLF prediction has a good screening effect，and preliminarily solves the problem of early diabetes screening through machine learning，which has certain practical significance for clinical diagnosis and can be used as a simple and effective screening method for diabetes and its complications.

【Key words】incomplete data；diabetes complication；inherent non-negative latent feature；classifier

国际糖尿病联合会发布第10版的《全球糖尿病地图》显示，中国是糖尿病人数最多的国家，超过1.4亿人。中国在糖尿病上的医疗支出位居世界第二，医疗和经济负担非常沉重，Chen C等[1]的研究发现，糖尿病患者的直接经济负担约为8 000元，间接经济负担约为2 000元。糖尿病可能会引起心血管和微血管病变，并对多个系统造成危害。Zhong VW等[2]的调查研究发现糖尿病患者的血压、血脂达标率不足三成，Zhou K等[3]发现三酰基甘油酯升高和

作者介绍：周乐明，Email：ywkzlm@126.com，研究方向：大数据智能计算。
通信作者：尚明生，Email：msshang@cigit.ac.cn。
基金项目：重庆市科技局、重庆市卫生健康委联合科研资助项目（编号：2019ZDXM006）。
优先出版：https://link.cnki.net/urlid/50.1046.R.20231227.1633.016（2023-12-29）

高密度脂蛋白降低与糖尿病进展率的增加独立相关。糖尿病经常与高血压共存，《柳叶刀》的1项研究揭示了降低血压是预防新发2型糖尿病的有效策略[4]。因此，早期发现糖尿病发展轨迹的影响因素，有利于早期筛查和规范化管理，对糖尿病合并症进行规范化管理和治疗是非常有必要的。但是原发性高血压合并糖尿病的早期症状不明显，要诊断清楚需要进行多项其他检查，往往延误诊断和治疗。同时由于在临床实践中不可能完成每项检查，因而临床检验的项目通常是不完备的，这对通过检查指标进行早期筛查构成了重大挑战。

因此提出了基于检验数据建立科学的分类模型进行糖尿病早期筛查。Berikol GB等[5]提出了1种基于心电图（electrocardiogram，ECG）、超声心动图和实验室测试的支持向量机诊断急性冠脉综合征的方法，但检验数据局限于肌酸激酶（creatine kinase，CK）、肌酸激酶-MB（creatine kinase isoenzymes，CK-MB）、肌红蛋白（myoglobin，Mb）、肌钙蛋白T（cardiac troponin T，cTnT）和肌钙蛋白I（cardiac troponin I，cTnI）等指标。美国拉格兰德州东俄勒冈大学的Dinh A等[6]提出了利用机器学习预测糖尿病和心血管疾病的数据驱动方法，开发了心血管、糖尿病前期和糖尿病预测模型，但超过50%的缺失值都从数据集中删除，导致可用变量数量进一步减少。Daniels J等[7]研究了迁移学习方法在不同预测范围内的血糖预测的影响，但局限于葡萄糖浓度等少量指标，样本量也仅有12例。张春富等[8]提出基于GA-Xgboost模型的糖尿病风险预测，但对缺失值采取的是删除乙肝类特征，其余缺失值采取均值填充的方法，会造成信息的损失。龚军等[9]提出了基于机器学习算法的原发性高血压并发冠心病的患病风险研究，未针对糖尿病进行研究，且只用了一种基于非参数的随机森林方法进行缺失值填补。Hossain ME等[10]提出了针对2型糖尿病患者心血管疾病预测风险模型，但未基于不完备的检验特征，未运用预测方法进行填补，且对象为患有2型糖尿病和心血管疾病的患者和仅患有2型糖尿病的患者。

虽然以上研究为糖尿病并发症的分类提供了一些可行的方法，同时包括缺失值处理和分类2个步骤，但在预处理的第一步也存在一些不足，不完备数据的处理方法一般是通过预填充统计，这不可避免地造成准确性的损失。随机森林回归作为一种重要的集成算法，其缺点是决策树的相似性较大，掩盖了真实结果；均值填充的缺点是对于数据分布不平稳或有异常值时，会导致填充后的数据失真；简单线性回归的缺点是需要判断变量之间的线性关系，不能很好地拟合非线性数据集。考虑到本研究中的医疗数据基本是数值型，而且有近100个维度，因此采用矩阵分解的方法进行预测后补全，但基于普通矩阵分解（matrix factorization，MF）的方法的一个局限性是采用中间全矩阵来近似高维矩阵，然后对该中间矩阵进行分解，但由于数据的不完备，不完备高维数据的填充需要占用更多的时间和空间，具有较高的计算和存储成本。基于单元素的非负隐特征（non-negative latent factor，NLF）模型通常用于分析这类不完备数据，模型能够高效、准确地表示大规模的不完备矩阵，而不是基于整个补全后的特征矩阵，将原始的高维矩阵分解为2个低维矩阵，具有较高的精度和较低的计算复杂度，但目前的NLF模型局限是依赖于专门的学习策略[11]。

本文提出了一种结合固有非负隐特征分析的糖尿病检测模型框架。主要工作如下：采用非线性函数的INLF模型，从不完备的检查数据中提取非负隐特征矩阵，以准确预测未知的缺失特征；基于不同的检验特征通过不同分类器来检测慢性疾病。本文的主要贡献总结如下：创建了高血压和高血压合并糖尿病2个慢性数据集，数据集中有99个常用指标；提出了糖尿病早期筛查模型；根据特征重要性排序，分析了区分糖尿病的关键特征；根据特征相关性进行了相关性分析。

1 对象与方法

1.1 研究对象

1.1.1 纳入和排除标准 该研究以重庆市黔江中心医院2017年至2019年匿名电子病历（electronic medical record，EMR）首页为基础进行的回顾性研究，所有的诊断都由ICD-10（国际疾病分类，第10版）指定，慢性病检验数据来自于重庆市黔江中心医院A医院实验室信息管理系统（laboratory information system，LIS）中2017年至2019年的历史数据，涉及的检验项目较多，包括血液、体液、生物化学、免疫等项目。为了证明提出的策略的有效性，根据纳入排除标准制作了2个慢性病数据集（高血压病、高血压病合并糖尿病）。研究组定义为诊断为含有原发性高血压（I10）、糖尿病（E10~E14）及其合并症的住院患者共3 241例，对照组为4 181例仅患有原发性高血压（I10）及其合并症（不包括糖尿病）的患者。

1.1.2 数据预处理 数据集成。采用数据关联方法对医院信息系统（hospital information system，HIS）入院诊断数据和LIS临床检验数据进行整合和清洗，基于纳入和排除标准进行了选择。

数据归一化。对2种慢性病的检验数据进行预处理，使用最小-最大法对数据进行归一化，映射到0~1范围之内。

1.2 研究方法

1.2.1 基于INLF的特征提取和预测方法原理 NLF模型可以有效地处理不完备数据，但不能表示非线性特征[12]。为了更有效地分析不完备的医疗数据，本文提出一种基于INLF的糖尿病筛查模型。INLF模型的主要思想是将一个不完全矩阵V分解为多个LF矩阵。然后通过单元素相关的非负映射函数将决策参数与输出LF连接起来进行预测[13]。假设W和Z表示两个实体集，$R^{|W|\times|Z|}$表示目标矩阵，每个元素$r_{w,z}$描述

实体 $w\in W$ 和 $z\in Z$ 之间的一些关系，并通过非负数据进行量化。R 通常是一个高维不完备矩阵，X 是输出的隐特征。引入（$|W|+|Z|$）×d 维向量 L 作为决策参数，以及单元素相关的映射函数 δ，它将 L 中的每个元素映射到 X 中的对应元素，以放松非负性约束。X 和 L 之间的关系可以表述为：

$$\forall w\in W,\forall z\in Z,g\in\{1,2,3,\cdots,e\}:$$

$$x_{(w)g}=\delta(l_{(w)g}),x_{(z)g}=\delta(l_{(z)g}). \quad (1)$$

通过使 δ 满足条件 $\forall l\in R:\delta(l)\geqslant 0$，重构后得到如下目标函数：

$$\arg\min_{L}\varepsilon(X)=\frac{1}{2}\sum_{\forall r_{w,z}\in\Gamma}\left(v_{w,z}-\sum_{g=1}^{e}\delta\left(l_{(w)g}\right)\delta\left(l_{(z)g}\right)\right)^{2} \quad (2)$$

采用的参数选择是：F=60，λ=0.001，η=0.01，F 表示维度，λ 表示步长，η 表示学习率。

1.2.2 不完备检验数据的预测模型 医学检测数据往往高维且不完备，为筛查高血压和高血压合并糖尿病的数据集，需要填补缺失值，并采用一些方法对慢性病进行分类。将 INLF 方法与代表性的填充方法进行了比较，具体介绍如下：

F0：基于固有非负隐特征分析的慢性病检测模型。首先，使用 INLF 模型提取不完备矩阵的隐特征；然后，基于隐特征补全矩阵对缺失值进行预测。*F1*：模型用多次插值填充缺失值，主要通过探索变量之间的关系，建立回归模型来预测缺失值。*F2*：用平均填充缺失值，往往适用于简单的数据填充，填充效果也很小。*F3*：用 k 近邻（KNN）填充缺失值，是一种基于域的学习回归算法。

1.2.3 分类预测模型 选择的代表性的分类器如下：*C1*：XGBoost 是一种梯度增强决策树（gradient boosting decision tree，GBDT）分类器，该方法采用弱分类器决策树进行迭代训练得到最优模型，具有训练效果好，不易过拟合的优点。*C2*：LightGBM，它在传统 GBDT 算法的基础上进行了优化，包括基于直方图的决策树算法、基于梯度的单边采样（gradient-based one-side sampling，GOSS）算法、互斥特征绑定算法（exclusive feature bundling，EFB）算法。*C3*：AdaBoost 方法的自适应含义是使用之前被错误识别的分类器的样本来训练下一个分类器。以上分类器的参数采取了默认的参数设置。*C4*：MLP 分类器是一种通过在数据集上的训练来学习由输入空间到目标空间的映射函数的分类模型，参数设置为（30，15，6），最大迭代次数 500 次。

2 结 果

2.1 基本资料分析

本研究共纳入 4 181 例原发性高血压的患者为对照组，原发性高血压合并糖尿病 3 241 例为研究组，详见表 1。D1A：原发性高血压数据集，来自实验室信息系统，包含来自 4 181 例患者的 99 个检查项目，数据集 D1A 的密度为 71.33%。D1B：原发性高血压伴糖尿病数据集，由实验室信息系统采集，包括 3 241 例患者的 99 个检查项目，数据集 D1B 的密度为 70.03%。D1：合并 D1A 和 D1B 数据集，包括 7 422 例患者，具有 99 个属性维度。按检验类别分，采集的检验特征包括了血糖相关项目（葡萄糖、糖化血红蛋白等），肾功能项目（胱抑素、尿素氮等），血脂项目（总胆固醇、甘油三酯等）和肝功能项目（如谷草转氨酶等）等。

表 1 慢性病数据集

疾病类型	样本数量	样本属性	数据集	类别
原发性高血压	4 181	99	D1A	0
原发性高血压合并糖尿病	3 241	99	D1B	1
糖尿病合并组	7 422	99	D1	2

2.2 预测及分类效果比较

在 D1 数据集上，将实验结果与其他常用算法进行比较。为了说明该方法的有效性，本文从以下 2 部分进行对比实验。首先比较各种填充缺失值的方法，然后比较几种经典的分类算法，包括 XGBoost、LightGBM、AdaBoost 和 MLP。

对 D1 数据集划分为 80% 的训练集与 20% 的测试集，进行 5 次实验，从实验结果可以看出，采用准确率（Accuracy，ACC）进行模型评估，分类器在 INLF 预测前后的效果与其他填充方法相比，使用 INLF 预测的填充效果明显优于使用其他填充方法。

从实验结果来看，INLF 是基于元素的非负矩阵分解，只工作于已知的高维不完备矩阵项，由于 INLF 不断满足非负性要求，它与 SGD（随机梯度下降）等优化器兼容，将输出的隐特征与决策参数分离，因此能够对高维不完备矩阵中缺失值进行较好地预测。见表 2。

表 2 D1 数据集不同分类器的效果比较

填充方法分类模型	F0	F1	F2	F3
C1	0.894 9	0.854 5	0.455 2	0.849 8
C2	0.887 5	0.864 0	0.443 8	0.858 6
C3	0.862 0	0.835 7	0.738 0	0.820 9
C4	0.856 6	0.832 3	0.847 1	0.816 2

特征的重要性分析，通过重要性排序对关键特征进行筛选（图 1），有利于通过更加少量和简便易得的指标提前发现糖尿病患者，在结合 2 种疾病的 D1 数据集上进行实验，得到的结果也符合医学规律，除了列出的血糖（glucose，GLU）、糖化血红蛋白（glycated hemoglobin，HbA1c）指标外，还发现了一些次要特征与糖尿病并发症有关，如 β2-微球蛋白（β2-microglobulin，β2M/BMG）、平均红细胞血红蛋白浓度（mean corpuscular hemoglobin concentration，MCHC）、血清补体（serum complement C1q，C1q）等，提示可以结合次要特征作为早期糖尿病筛查的辅助指标。

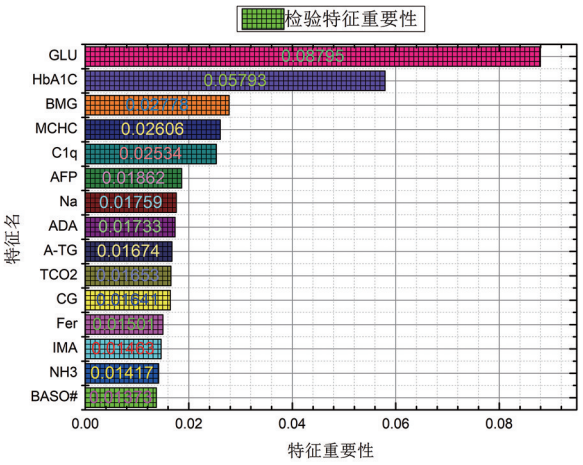

图 1 基于 INLF 预测的 XGBoost 分类器前 15 个重要特征排序

特征相关性分析。对分类器中排名前20位的检验指标进行Pearson相关性分析，发现β2-微球蛋白（BMG）与α1-微球蛋白（α-Microglobulin，α1-MG）的相关系数达到了较强的正相关（相关系数0.79）。HbA1C与GLU呈一定程度的正相关（0.46），钠（Na）与GLU呈一定程度的负相关（-0.33），提示需要对这些指标进行监测和预警，及早发现和预防糖尿病及并发症。

3 讨 论

本文基于固有非负隐特征预测缺失值的方法，并结合4种分类器探索了糖尿病分类模型框架，初步解决了通过机器学习方法进行糖尿病的早期筛查问题，同时探寻了糖尿病各项实验室指标之间的关系，对临床实践具有一定的实际意义。

实验结果发现，不同缺失值预测方法会影响不同的分类结果。表2显示，将2组的99个检验特征填充后通过分类器进行分类，其中均值填充效果最差，使用XGBoost、LightGBM分类器时精度仅为0.455 2、0.443 8，可能是受到数据分布的影响，而基于INLF预测的分类准确率优于KNN法、插值法和均值等方法，基于INLF的XGBoost分类器的准确率达到89.49%，原因是它应用了非线性函数，具有较强的非线性映射能力，不依赖于数据类型或如何分布，也能处理复杂的检验数据。同时，不同分类器及其参数的选择也有一定影响作用，XGBoost由于需要保存数据的特征值，以及保存了特征排序的结果，导致空间、时间消耗大[14]；AdaBoost的缺点是当数据不平衡导致分类精度下降；LightGBM有一些需要进行合适的参数调优，同时对噪声数据敏感[15]。

针对分类器存在参数较多、收敛较慢的问题，基于群体智能算法具有全局优化的优点，将采用先进的群体智能算法[如自适应多目标粒子群优化算法（multiple objective particle swarm optimization，MOPSO）[16]算法、多目标差分进化算法（Multi-Objective Differential Evolution，MODE）]及其改进算法对分类器的参数进行优化，再与分类器组合后进行特征选择，以选择出最有识别性的特征，从而提高分类精度，减少冗余特征干扰。

今后将对关键指标进行逐步缺失观察检测效果，同时也进行按比例缺失患者标签的半监督学习，以减少人为标记的工作量，提高工作效率。为了提高疾病的早期检测能力，还将把研究范围扩展到更广泛的领域，通过联合分析慢性病共病网络的拓扑结构特征，以提早发现糖尿病共病组的共病模式和发展轨迹，并对这些伴随疾病进行及时预防和干预，从而减轻糖尿病及并发症的危害。

参 考 文 献

[1] Chen C, Song JL, Xu XL, et al. Analysis of influencing factors of economic burden and medical service utilization of diabetic patients in China[J]. PLoS One, 2020, 15(10): e0239844.

[2] Zhong VW, Yu DM, Zhao LY, et al. Achievement of guideline-recommended targets in diabetes care in China: a nationwide cross-sectional study[J]. Ann Intern Med, 2023, 176(8): 1037-1046.

[3] Zhou KX, Donnelly LA, Morris AD, et al. Clinical and genetic determinants of progression of type 2 diabetes: a DIRECT study[J]. Diabetes Care, 2014, 37(3): 718-724.

[4] Nazarzadeh M, Bidel Z, Canoy D, et al. Blood pressure lowering and risk of new-onset type 2 diabetes: an individual participant data meta-analysis[J]. Lancet, 2021, 398(10313): 1803-1810.

[5] Berikol GB, Yildiz O, Özcan IT. Diagnosis of acute coronary syndrome with a support vector machine[J]. J Med Syst, 2016, 40(4): 84.

[6] Dinh A, Miertschin S, Young A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning[J]. BMC Med Inform Decis Mak, 2019, 19(1): 211.

[7] Daniels J, Herrero P, Georgiou P. A multitask learning approach to personalized blood glucose prediction[J]. IEEE J Biomed Health Inform, 2022, 26(1): 436-445.

[8] 张春富，王 松，吴亚东，等. 基于GAXgboost模型的糖尿病风险预测[J]. 计算机工程，2020，46(3)：315-320.
Zhang CF, Wang S, Wu YD, et al. Diabetes risk prediction based on GAXgboost model[J]. Comput Eng, 2020, 46(3): 315-320.

[9] 龚 军，杜 超，钟小钢，等. 基于机器学习算法的原发性高血压并发冠心病的患病风险研究[J]. 解放军医学杂志，2020，45(7)：735-741.
Gong J, Du C, Zhong XG, et al. Researches on the illness risk of essential hypertension complicated with coronary heart disease based on machine learning algorithm[J]. Med J Chin People's Liberation Army, 2020, 45(7): 735-741.

[10] Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes[J]. Expert Syst Appl, 2021, 164: 113918.

[11] Luo X, Zhou MC, Xia YN, et al. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems[J]. IEEE Trans Ind Inform, 2014, 10(2): 1273-1284.

[12] Luo X, Zhou MC, Li S, et al. An inherently nonnegative latent factor model for high-dimensional and sparse matrices from industrial applications[J]. IEEE Trans Ind Inform, 2018, 14(5): 2011-2022.

[13] Shang MS, Luo X, Liu ZG, et al. Randomized latent factor model for high-dimensional and sparse matrices from industrial applications [J]. IEEE/CAA J Autom Sin, 2019, 6(1): 131-141.

[14] Ogunleye A, Wang QG. XGBoost model for chronic kidney disease diagnosis[J]. IEEE/ACM Trans Comput Biol Bioinform, 2020, 17(6): 2131-2140.

[15] Punmiya R, Choe S. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing[J]. IEEE Trans Smart Grid, 2019, 10(2): 2326-2329.

[16] Xiang Y, Zhou YR, Yang XW, et al. A many-objective evolutionary algorithm with pareto-adaptive reference points[J]. IEEE Trans Evol Comput, 2020, 24(1): 99-113.

（责任编辑：周一青）