李学霖, 容芷君, 但斌斌,等. 基于 mRMR 与改进 MOPSO 的糖尿病特征选择方法[J]. 智能计算机与应用,2026,16(2):141-146. DOI:10.20169/j.issn.2095-2163.24040601

基于 mRMR 与改进 MOPSO 的糖尿病特征选择方法

李学霖, 容芷君, 但斌斌, 付 婷, 庞奥康, 杨 鑫
(武汉科技大学 机械自动化学院, 武汉 430081)

摘 要: 糖尿病特征具有多模态、高维度、冗余复杂等特点,影响糖尿病的预测精度。针对此问题,本文提出基于特征交互与改进多目标粒子群优化算法的糖尿病特征选择方法。首先,对多模态数据进行融合,利用文本与数值数据共同构建糖尿病特征集;然后,通过 mRMR 与改进多目标粒子群优化算法删除冗余特征,筛选出与糖尿病相关性高的重要特征;最后,基于相关性分析构建组合指标特征集。采用 SVM、随机森林、逻辑回归和决策树等 4 种预测模型对其进行分类评估,结果显示组合指标特征集的预测准确率为 90%。对糖尿病预测有较高的准确率。

关键词: 糖尿病预测; 特征提取; 组合指标; mRMR; MOPSO
中图分类号: R319;TP181 **文献标志码**: A **文章编号**: 2095-2163(2026)02-0141-06

Diabetic feature selection method based on mRMR and improved MOPSO

LI Xuelin, RONG Zhijun, DAN Binbin, FU Ting, PANG Aokang, YANG Xin
(College of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China)

Abstract: The paper proposes a diabetes feature selection method based on feature interaction and improved multi-objective particle swarm optimization algorithm to address the characteristics of multimodality, high dimensionality, redundancy, and complexity in diabetes prediction. Firstly, multimodal data is fused to construct a diabetes feature set using both textual and numerical data. Then, redundant features are eliminated through mRMR and the improved multi-objective particle swarm optimization algorithm to select important features highly correlated with diabetes. Finally, a composite indicator feature set is constructed based on correlation analysis. Four prediction models including SVM, Random Forest, Logistic Regression and Decision Tree are employed for classification evaluation. The results show that the prediction accuracy of the composite indicator feature set reaches 90%, indicating a high accuracy in diabetes prediction.

Key words: diabetes prediction; feature extraction; composite index; mRMR; MOPSO

0 引 言

根据国际糖尿病联盟最新发布的数据,中国糖尿病患者人数现位居全球首位[1],其高患病率及各类并发症给病患造成了巨大的疾病负担。由于 2 型糖尿病最初的无症状性和复杂的危险因素组成,因此对糖尿病准确的早期预测,能够对糖尿病的防控和治疗起到很大作用。

机器学习(ML)已被积极用于建立复杂慢性疾病的计算机辅助诊断。机器学习研究的很大一部分致力于构建更好的疾病分类系统[2-3],这些系统的一个重要方面是特征选择。因为糖尿病具有多因素风险组成,风险因素高维复杂,且各风险因素高度相关。近年来,不同的机器学习方法被应用于选择重要特征以改善模型训练和分类结果。主要分为 3 种类型,分别是:过滤器、包装器和嵌入式方法[4]。其中,过滤器根据预定义的统计测试/度量单独评估每个特征与目标的关联[5]。李国豪等学者[6]结合 Relief 系列算法对糖尿病影像指标进行特征选择。李占山等学者[7]提出基于互信息的距离测度 RReliefF,以获得性能更优的滤波标准。程雨轩等学者[8]提出了一种基于邻域互信息的特征选择算法。包装器方法在特征空间中迭代搜索,以提取最大分类器精度的相关特征子集[9]。耿焕同等学者[10]利用遗传算法进行特征选择。嵌入式方法将特征空间搜索的排他步骤结合到分类器训练中,作

作者简介: 李学霖(2000—),女,硕士研究生,主要研究方向:大数据分析。
通信作者: 容芷君(1974—),女,博士,教授,主要研究方向:医疗大数据。Email:1832077275@qq.com。

收稿日期: 2024-04-06

为一种优化方法来减少搜索和分类开销[11]。戴贵洋等学者[12]提出了随机森林和模糊系统的二次筛选的特征选择模型。柯东等学者[13]基于随机森林的特征重要性排序构建疾病预测模型。

近年来,基于互信息的滤波特征选择技术得到了广泛的应用。基于互信息的传统方法只关注于确定特征是相关的、还是冗余的,两者都代表双向交互。然而,在生物系统中,多种危险因素可能相互作用,产生不同的生理或病理行为,这就需要分析超越2个特征的相互作用。尽管包装器和嵌入式方法可以通过子集选择间接捕获特征交互,但由于分类器基础的影响,彼此之间并不完全构成实际的相互依存关系。

过滤法虽然时间复杂度低,但需要预先设定选择特征的阈值,因此其特征选择的主观性可能会对模型性能产生影响。包装法从原始特征集中选择出使最终任务模型性能最佳的特征子集,包装法直接以模型性能为评价指标,所选特征集预测性能良好,但由于原始特征集的数量通常较多,其时间复杂度要远高于过滤法。嵌入法包括基于树模型的特征选择和基于模型中惩罚项的特征选择[14]。为充分利用其优势,一些学者提出了混合特征选择方法。文武等学者[15]结合信息增益与萤火虫算法确定最优特征子集。郑睿程等学者[16]提出了一种基于正交化最大信息系数与特征协同法与随机森林递归消除结合的混合特征选择方法。熊玲珠等学者[17]利用最大信息系数保留相关特征,再利用XGBoost迭代删除冗余特征。

基于此,本文提出了一种基于特征相关性、冗余度和交互准则的混合特征选择方法。该方法是对现有最小冗余最大相关滤波方法的扩展,同时基于包装方法反复训练模型并评估特征子集的性能来选择最佳特征子集特征。

1 糖尿病组合指标特征选择模型

现代的疾病预测主要是靠大数据、机器学习等技术通过病人电子病历相关的数据特征,构建一种用于定量描述数据特性和疾病发病率的关联的疾病预测模型,以达到对特定疾病的预测。考虑到医生提供的描述性和评论性的文本能够更好反映病人健康状况,首先利用体检数据中的文本数据和数值数据共同构建糖尿病特征集。然后,利用基于特征交互的mRMR与MOPSO混合特征选择方法删除冗余特征,筛选出与糖尿病相关性高的重要特征。

研究表明血脂参数与糖尿病发病机制有较强的相关性,同样能对2型糖尿病的预测提供参考。因此基于Logistics回归的相关性分析构建血质参数比,得到含重要特征的组合指标特征集。糖尿病组合指标特征选择模型如图1所示。

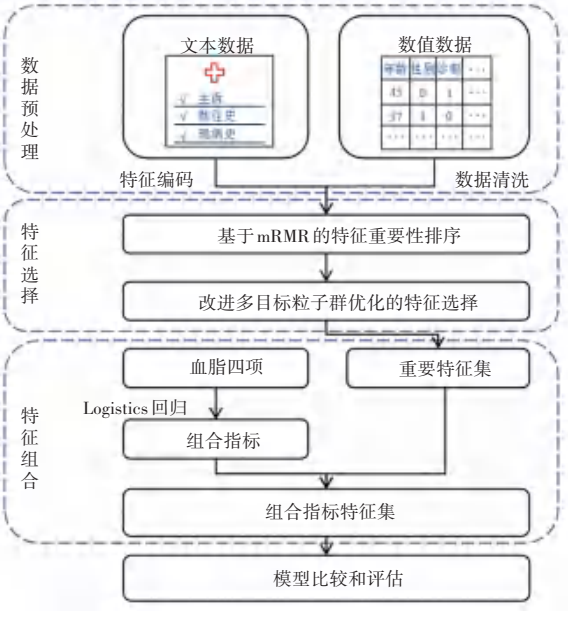

图1 糖尿病特征选择模型

Fig. 1 Diabetes feature selection model

2 文本与数值特征组合

2.1 文本特征提取

本研究采用襄阳市一家基层医疗机构的体检资料作为研究对象,采集到的糖尿病数据集结构复杂,同时包含文本数据和数值数据。糖尿病文本数据多为医生问诊描述与一些对化验检查单的描述,其中包含些有价值的、潜在的重要信息,因此需要对文本数据进行特征提取与编码,将其与数值数据一起进行分析。

通过自然语言处理方法,可以实现对文本数据到模型可识别特征向量的转化。本文采用Jieba对糖尿病患者的文本病历进行分词、去停用词等处理,然后使用TF-IDF对处理后的文本进行计算。TF-IDF是一种基于词袋模型的算法,用于评估一个词语在文档集合中重要性[18]。通过将一个词语的权重分为2个部分:词频(TF)和逆文档频率(IDF)。对于第i位患者的文档d_i,词w_j的指标参数值计算公式分别如下:

$$\mathrm{TF}_{ij} = \frac{n_{ij}}{|d_i|} \tag{1}$$

$$\mathrm{IDF}_j = \log \frac{|\mathrm{DOC}|}{|i:w_j \in d_i| + 1} \tag{2}$$

$$\mathrm{TF-IDF}_{ij} = \mathrm{TF}_{ij} \times \mathrm{IDF}_j \tag{3}$$

得到了TF-IDF矩阵、并从TF-IDF矩阵中提取

文本的关键词权重，接下来对计算得到的关键词权重进行倒序排列，删除权重接近 0 的特征，最终得到排名前 25 个词汇作为糖尿病文本特征关键词，如图 2 所示。

为便于文本特征与数值特征的进一步选择研究，对所选取的 25 个文本特征进行 one-hot 编码向量化表示为数值形式，例如将有头晕特征的样本在头晕特征一栏记为 1，无头晕文本样本记为 0。

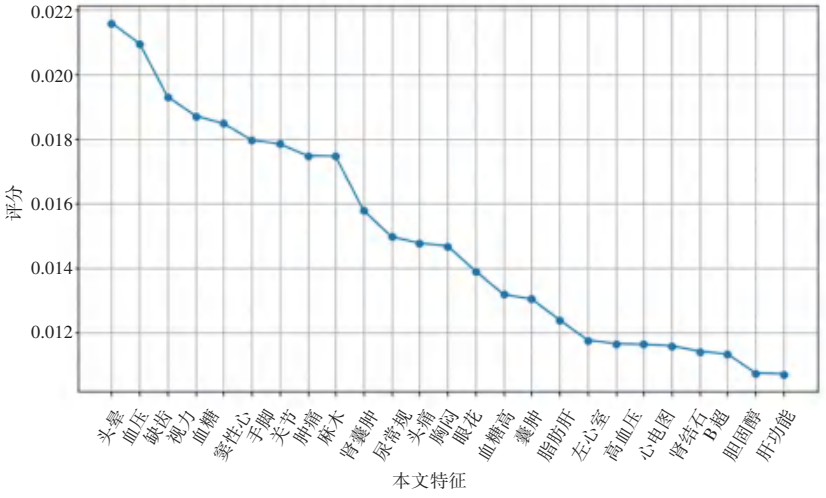

图 2　基于 TF-IDF 的文本特征评分

Fig. 2　Text feature score based on TF-IDF

2.2　数值数据清洗

数值数据集中包含了一般个人的基本情况、个人常规检验项目以及个人生活方式等影响因素。最后形成的数据集共计 3 886 个体检数据[其中有 1 943 例阴性(50%)和 1 943 例阳性(50%)]，37 个特征变量。获取到原始数据集后，部分数据缺失严重。首先对数据进行清洗和转化。

在缺失值的处理过程中，对于糖尿病初步采集的数据，难以判断该特征对糖尿病是否完全没有价值，故在分类和数字类型变量中，分别用众数和均值插值两种方法，对缺失信息比较严重的部分则全部剔除。在异常值处理中，异常值即在医疗数据集中存在不合理的值。对离散型数值数据进行简单识别，删除异常值。对连续型数值，根据 3 Sigma 原则，在 3 Sigma 范围之外出现的数据样本被认为是异常值，直接剔除。

对于由字符型、二分类组成的属性类型，例如性别、吸烟情况、诊断等进行赋值转换。选用整型的数据来表示原始的属性值含义，变量的赋值见表 1。

表 1　赋值说明

Table 1　Assignment description

特征	赋值说明
性别	女为 0，男为 1
诊断	不患有糖尿病为 0，患有糖尿病为 1
吸烟状况	从不吸烟为 0，过去吸烟为 1，吸烟为 3
年龄	<18 为 1，[18,35)为 2，[35,45)为 3，[45,60)为 4，≥60 为 5
锻炼频率	不锻炼为 1，偶尔锻炼为 2，经常锻炼为 3

对数值数据进行清洗后加入经过特征编码的文本特征，得到预处理后的 62 个数值特征构成糖尿病初始特征集。

3　基于混合特征选择的组合指标特征集构建

3.1　基于特征交互的 mRMR 重要性排序与特征选择

mRMR 基于特征相关性和特征冗余两个标准，利用互信息度量来选择最优子集[19]。考虑一个输入特征集 $F=\{f_1, f_2, f_3, \cdots, f_n\}$ 的特征选择问题，其中 f_1、f_2、f_3、$\cdots$、f_n 为 n 个输入特征，h 为目标特征。设 S = NULL 为要选择的特征子集。在每一步中，mRMR 迭代地在输入特征集 F 中搜索一个最优特征，该特征集 F 优化以下 2 个目标：

(1)最大相关性。最初将具有最高信息增益的特征添加到子集 S 中，然后在每次迭代中考虑输入特征集 F 中使子集 S 已选到特征的平均相关性最大化的特征 i。由此推得的公式为：

$$\max V_I, V_I = \frac{1}{|S|}\sum_{i\in S} I(h,i) \tag{4}$$

(2)最小冗余。同时在每次迭代中，特征 i 与子集 S 中的其他候选特征进行最小冗余检查。研究推得的公式为：

$$\min W_I, W_I = \frac{1}{|S|^2}\sum_{i,j\in S} I(i,j) \tag{5}$$

基于 mRMR 的特征重要性排序步骤如下所示：

(1)对输入特征集中的每个特征与目标变量之

间的信息增益进行计算,并按照信息增益的降序对特征进行排序。

(2)在每一次迭代中,选择具有最大信息增益的特征,并将其添加到特征子集中。然后对于剩余的每个特征,计算其与已选择特征之间的特征冗余和交互增益。

(3)如果 2 个特征之间的交互增益为正,则说明这 2 个特征相互作用会提供额外的信息。然后将特征的个体增益添加到这 2 个特征的交互增益中,形成累积相关性。

(4)通过将累积相关性除以特征子集上的平均冗余来计算特征的重要性得分/排名,得分越高,特征越重要。

(5)选择具有最高得分的特征,并将其添加到特征子集中。随即将重复上述步骤,直到所有特征都被排序。

3.2 基于改进的多目标粒子群优化的特征数量确定

基于特征交互的 mRMR 只能输出特征的重要性排序,不能选择出最佳的特征子集,因此引入改进的多目标粒子群优化算法确定所选择特征的数量与准确率的最优解,可以使多目标粒子群优化算法加速收敛,缩短特征选择时间[20]。改进的多目标粒子群优化算法步骤如下:

(1) 初始化惯性权重系数的最大值 $w_{\max}$ 和最小值 $w_{\min}$, 飞行时间的初始值 H_o, 学习系数的最大值和最小值 $C_{1\max}$、$C_{1\min}$、$C_{2\max}$、$C_{2\min}$。

(2) 定义改进的算法,以实现参数动态调整。

(3) 定义优化问题与算法,执行优化过程,设置终止条件是:最大迭代次数为 100。

(4) 在每一代迭代中,算法将执行如下操作:

① 计算每个粒子在目标空间中的适应度值。

② 根据粒子的个体最优解和群体最优解来更新每个粒子的速度和位置。

③ 根据每个粒子的适应度值更新非支配解集,确保其中的解决方案是互不支配的。

④ 重复以上步骤,直到达到最大迭代次数为止。

在上述迭代过程中,通过自动调整惯性权重系数、飞行时间和学习系数,结合粒子群算法的基本原理,不断优化搜索空间以找到最优解的近似解集。

其中,惯性权重系数的更新公式为:

$$w = w_{\max} - \left(\frac{w_{\max} - w_{\min}}{t_{\max}} t\right) \tag{6}$$

飞行时间的更新公式为:

$$h = H_0\left(1 - \frac{t}{t_{\max}}\right) \tag{7}$$

学习系数的更新公式为:

$$c_1 = c_{1_{\max}} - \left(\frac{c_{1_{\max}} - c_{1_{\min}}}{t_{\max}} t\right) \tag{8}$$

$$c_2 = c_{2_{\max}} - \left(\frac{c_{2_{\max}} - c_{2_{\min}}}{t_{\max}} t\right) \tag{9}$$

对改进 MOPSO 算法进行 100 次迭代,输出结果如图 3 所示,准确率最高时对应的特征数量为 11,因此选择 mRMR 重要性排序的前 11 位特征,如图 4 所示。

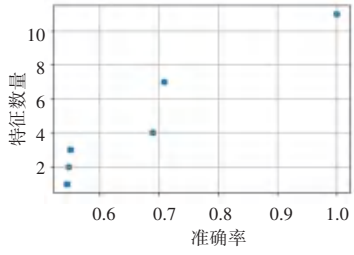

图 3 基于 MOPSO 的特征数量确定

Fig. 3 Determining the number of features based on MOPSO

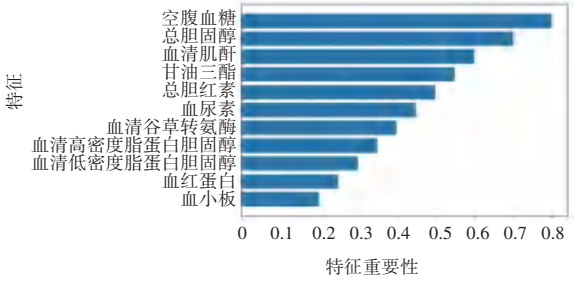

图 4 前 11 位特征重要性排序

Fig. 4 Top 11 feature importance ranking

因此选择特征排名前 11 名组成重要特征集:[空腹血糖、总胆固醇、血清肌酐、甘油三酯、总胆红素、血尿素、血清谷草转氨酶、血清高密度脂蛋白胆固醇、血清低密度脂蛋白胆固醇、血红蛋白、血小板]。

3.3 组合指标构建

有研究表明,胰岛素抵抗、胰岛 β 细胞是 2 型糖尿病发病的重要因素,与血脂参数比值这一指标有较强的相关性[21],同样能对 2 型糖尿病的预测提供参考。因此利用 Logistic 回归分析不同血脂参数对于预测糖尿病的贡献,同时通过比值解释自变量组合的相对影响程度,基于多因素 Logistic 回归构建血脂参数比值。对血脂四项指标进行多因素 Logistic 回归分析,得到回归模型系数、p 值等见表 2。

表 2 中结果显示,总胆固醇(TC)、甘油三酯(TG)、血清低密度脂蛋白胆固醇(LDL-C)系数为正,与目标变量之间成正相关,血清高密度脂蛋白胆

固醇(HDL-C)系数为负,与目标变量之间成负相关,差异具有统计学意义($p < 0.05$)。说明总胆固醇、甘油三酯、血清低密度脂蛋白胆固醇是糖尿病发生的危险因素,血清高密度脂蛋白胆固醇是糖尿病的保护因素,因此构建新的组合指标 TC/HDL-C、TG/HDL-C、LDL-C/HDL-C,将组合指标添加到重要特征集中,构建组合指标特征集。

表 2　血脂四项多因素分析表

Table 2　Analysis of four factors of blood lipid

指标	系数	标准误差	Z 值	$P>\|Z\|$
TC	0.264 1	0.039	6.794	0.000
TG	0.043 1	0.036	1.207	0.027
HDL-C	-0.168 9	0.055	-3.082	0.002
LDL-C	0.281 3	0.086	3.284	0.001

4　实验验证

为了验证组合指标特征集的分类性能,文中将原始特征集与经过特征选择与特征组合的组合指标特征集分别输入 SVM、随机森林、决策树和逻辑回归预测模型中,按照 8∶2 随机划分为训练集和测试集带入模型进行训练,得到的分类精确率、召回率和 $F1$ 值分别见表 3~表 5。

表 3　分类精确率

Table 3　Classification accuracy rate

特征选择	SVM	随机森林	决策树	逻辑回归
原始特征集	0.672	0.688	0.660	0.679
重要特征集	0.745	0.784	0.727	0.743

表 4　分类召回率

Table 4　Classification recall rates

特征选择	SVM	随机森林	决策树	逻辑回归
原始特征集	0.692	0.639	0.673	0.676
重要特征集	0.615	0.719	0.750	0.653

表 5　分类 $F1$ 值

Table 5　Classification $F1$ values

特征选择	SVM	随机森林	决策树	逻辑回归
原始特征集	0.671	0.665	0.658	0.671
重要特征集	0.709	0.763	0.727	0.711

4 种预测模型中,随机森林模型的分类 $F1$ 值最高,因此利用随机森林模型对构建的糖尿病组合指标特征集进行分类预测,ROC 曲线如图 5 所示。由图 5 可见组合指标特征集对糖尿病预测有较高的准确率,即基于组合指标的特征选择方法后的糖尿病预测模型具有较优的预测性能。

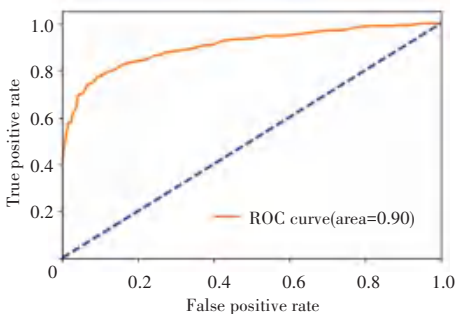

图 5　基于随进森林的组合指标特征集 ROC 曲线

Fig. 5　ROC curve of combination index feature set based on Random Forest

5　结束语

本研究在糖尿病多模态数据的基础上,针对糖尿病预测数据的冗余性和数据复杂性,采用 mRMR 与改进 MOPSO 两阶段混合特征选择方法删除冗余特征,寻找出最有效的预测特征子集;并在常规糖尿病危险因素的基础上进一步扩大了样本量,纳入了更多非常规的脂质参数,基于 Logistic 回归分析血脂参数与糖尿病患病结局的相关性构建比值指标,探讨了组合指标对糖尿病预测的影响。实验结果表明,该特征选择方法能够有效降低数据复杂与冗余,且组合指标特征集在评估和预测未来糖尿病风险方面有较高的预测准确率。

参考文献

[1] 陈楠, 张建伟, 李博一, 等. 浅谈中国医保与糖尿病经济负担[J]. 实用糖尿病杂志,2020,16(6):138-140.

[2] MOON S, JANG J Y, KIM Y, et al. Development and validation of a new diabetes index for the risk classification of present and new-onset diabetes: Multicohort study[J]. Scientific Reports, 2021, 11(1): 15748.

[3] PEI Dongmei, GONG Yang, KANG Hong, et al. Accurate and rapid screening model for potential diabetes mellitus[J]. BMC Medical Informatics and Decision Making, 2019, 19:41.

[4] SNEHA N, GANGIL T. Analysis of diabetes mellitus for early prediction using optimal features selection[J]. Journal of Big Data, 2019, 6(1):13.

[5] ZHOU Peng, WANG Ni, ZHAO Shu. Online group streaming feature selection considering feature interaction[J]. Knowledge-Based Systems, 2021, 226: 107157.

[6] 李国豪, 杨豪, 刘彦, 等. 基于 Relief 系列算法的脑网络特征选择与分类[J]. 计算机仿真, 2022, 39 (10): 354-358.

[7] 李占山, 杨云凯, 张家晨. 基于熵权法的过滤式特征选择算法[J]. 东北大学学报(自然科学版), 2022, 43(7): 921-929.

[8] 程雨轩, 毛煜, 张小清, 等. 基于次相关特征和邻域互信息的在线多标记特征选择算法[J]. 山东大学学报(理学版), 2024, 59(5): 70-81.

[9] BERNARDINI M, MORETTINI M, ROMEO L, et al. Early

temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: a multiple instance boosting approach[J]. Artificial Intelligence in Medicine, 2020, 105: 101847.
[10]耿焕同, 戴中斌, 沈阳. 基于遗传算法的特征选择方法在短时强降水预报中的应用[J]. 气象科学,2023,43(1):126-134.
[11] REMESEIRO B, BOLON-CANEDO V. A review of feature selection methods in medical applications[J]. Computers in Biology and Medicine, 2019, 112: 103375.
[12]戴贵洋, 綦秀利, 余晓晗. 融合人类知识的随机森林特征选择方法研究[J]. 计算机技术与发展,2022,32(7):155-160.
[13]柯东, 晏峻峰. 基于 GA-XGBoost 算法的肺癌预测研究[J]. 计算机时代,2023(11):131-135.
[14]焦龄霄, 周凯, 张子熙, 等. 基于 mRMR-IPSO 的短期负荷预测双阶段特征选择[J]. 重庆大学学报,2024,47(5):98-109.
[15]文武, 赵成, 赵学华, 等. 基于信息增益和萤火虫算法的文本特征选择[J]. 计算机工程与设计, 2019, 40(12): 3457-3462.
[16]郑睿程, 顾洁, 金之俭, 等. 数据驱动与预测误差驱动融合的短期负荷预测输入变量选择方法研究[J]. 中国电机工程学报, 2020,40(2):487-500.
[17]熊玲珠, 邱伟涵, 罗计根, 等. 基于最大信息系数和迭代式 XGBoost 的混合特征选择方法[J]. 计算机应用与软件,2023, 40(1):280-286.
[18]黄春梅, 王松磊. 基于词袋模型和 TF-IDF 的短文本分类研究[J]. 软件工程,2020,23(3):1-3.
[19] TUPPAD A, PATIL S D. An efficient classification framework for type 2 diabetes incorporating feature interactions[J]. Expert Systems with Applications, 2024, 239: 122138.
[20]杨宝杰, 石凯元, 陈佳凯, 等. 基于改进粒子群算法的电力工程数据多目标优化方法[J]. 电子设计工程,2024,32(5):95-99.
[21] SHENG Guotai, KUANG Maobin, YANG Ruijuan, et al. Evaluation of the value of conventional and unconventional lipid parameters for predicting the risk of diabetes in a non-diabetic population[J]. Journal of Translational Medicine, 2022, 20(1): 266.