Software Engineering and Applications 软件工程与应用, 2023, 12(3), 474-484
Published Online June 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.123047

Hans汉斯

基于改进狮群算法优化神经网络的糖尿病风险预测

惠亚楠，冯慧芳

西北师范大学数学与统计学院，甘肃 兰州

收稿日期：2023年3月20日；录用日期：2023年6月8日；发布日期：2023年6月19日

摘　要

糖尿病风险评估预测有助于早期发现糖尿病，降低发病率和并发症。针对糖尿病风险预测问题，提出一种基于改进狮群算法优化神经网络的糖尿病风险预测模型。引入非线性扰动因子改进狮群算法，使得算法既能加强全局优化能力，避免陷入局部最优，又能保证局部优化能力，提高算法的收敛速度。利用改进狮群算法(ILSO)的寻优能力优化神经网络的权重和偏置参数，建立基于ILSO-BP神经网络的预测模型。同时，采用少类样本合成过采样技术和递归特征消除方法对糖尿病数据进行预处理，提高模型预测能力。在真实糖尿病数据集PIMA上的实验结果表明，基于ILSO-BP神经网络的糖尿病风险预测模型，其预测性能优于基线模型，也优于基于遗传算法、鲸鱼优化、粒子群优化等算法优化的神经网络预测模型，对糖尿病风险具有良好预测能力，能够对糖尿病早期筛查起到辅助作用。

关键词

糖尿病预测，改进狮群优化算法，BP神经网络，合成少数类过采样技术，递归特征消除

An Optimized Neural Network Prediction for Diabetes Risk Based on Improved Lion Swarm Algorithm

Yanan Hui, Huifang Feng

College of Mathematics and Statistics, Northwest Normal University, Lanzhou Gansu

Received: Mar. 20th, 2023; accepted: Jun. 8th, 2023; published: Jun. 19th, 2023

Abstract

Diabetes risk prediction helps to detect diabetes early and reduce the incidence and complications.

文章引用：惠亚楠，冯慧芳．基于改进狮群算法优化神经网络的糖尿病风险预测[J]．软件工程与应用，2023，12(3)：474-484. DOI: 10.12677/sea.2023.123047

To address the problem of diabetes risk prediction, a diabetes risk prediction of the neural network based on improved lion swarm optimization (ILSO) algorithm is proposed. A nonlinear perturbation factor is introduced to improve the lion swarm algorithm, so that the algorithm enhances the global search capability to avoid falling into local optimum, and enhances the local search capability to provide convergence speed. The weights and bias parameters of the neural network are optimized by using the optimization ability of the Improved Lion Swarm optimization algorithm (ILSO), and a prediction model based on the ILSO-BP neural network is developed. Meanwhile, a synthetic minority oversampling technique and recursive feature elimination are employed for pre-processing diabetes data to enhance the model prediction capability. The experimental results on the real diabetes dataset PIMA show that the diabetes risk prediction based on ILSO-BP neural network has better prediction performance than the baseline models, and better than the neural network model based on genetic algorithm, whale optimization and particle swarm optimization. The proposed model has good prediction ability for diabetes risk and can play an auxiliary role in early diabetes screening.

Keywords

Diabetes Risk Prediction, Improved Lion Swarm Optimization Algorithm, BP Neural Network, Synthetic Minority Oversampling Technique, Recursive Feature Elimination

1. 引言

糖尿病 DM (Diabetes Mellitus)是由于胰岛素分泌不足或胰岛素抵抗导致血糖水平或血糖偏高而引起的慢性代谢性疾病。如果不尽早介入治疗，糖尿病会严重影响个人的健康问题，如乳酸性酸中毒、肾衰竭、低血糖昏迷、重度感染等。国际糖尿病联盟 IDF (International Diabetes Federation)于 2021 年 12 月 6 日正式发布全球糖尿病地图(IDF Diabetes Atlas)第 10 版[1]，根据最新报告显示，2021 年全球约有 5.37 亿成人(20~79 岁)患有糖尿病，占全球该年龄段人口的 10.5%，其中约 90%是 II 型糖尿病患者。糖尿病的患病率仍在上升，预计 2045 年将增加到 7.83 亿，患病率为 12.2%。中国有 1.41 亿人的成年糖尿病患者居世界首位，患病率也比世界水平略高，约为 13% [2]。

糖尿病已经成为全世界威胁人类健康的重大慢性疾病之一，预防和控制糖尿病发生已经成为全世界致力解决且迫在眉睫的问题。医疗大数据的增加和机器学习算法的发展，为检测及诊断糖尿病提供了新的途径和方法。近年来，许多的研究工作者基于数据驱动的方法对糖尿病分类预测进行了研究。目前，基于数据驱动的糖尿病分类研究主要包括基于传统机器学习和基于深度学习的分类预测研究。Malviya 等人使用随机森林 RF (Random Forest)、AdaBoost、支持向量机 SVM (Support Vector Machine)、极限树(ExtraTree)等传统机器学习模型对糖尿病数据进行分类预测，研究结果表明 ExtraTree 效果优于其他分类模型[3]。Tigga 等人基于收集的数据和 PIMA 数据集，采用 logistic 回归、朴素贝叶斯、随机森林等机器学习算法来预测糖尿病并对比分析实验结果，得出随机森林分类效果最好的结论[4]。随着机器学习的发展和在医疗数据中的广泛使用，研究者通过嵌入优化算法和集成多个单分类器等手段提高模型的预测精度。张春富等提出的基于遗传算法的 Xgboost (GA-Xgboost)模型，使用 GA 的强全局搜索能力优化 Xgboost，提高其收敛速度，使得 GA-Xgboost 模型在糖尿病风险预测任务中性能优于线性回归、DT、SVM 及 NN

等模型[5]。Ali 等使用 KNN、朴素贝叶斯 NB (Nave Bayes)、线性判别分析、决策树作为基本分类器，随机森林作为元分类器，构建堆叠的集成分类器进行了糖尿病风险研究[6]，也有研究者采用其他的模型集成方法基于各种机器学习方法构建糖尿病分类模型，发现集成的分类器性能比单个分类器显著[7] [8]。尽管基于传统机器学习的糖尿病预测已取得了相当好的准确率，支持向量机具有严格的理论基础和数学基础，可避免局部最小问题，且小样本学习具有较强的泛化能力。然而，支持向量机的分类性能对参数选择比较敏感，实际应用中存在依靠经验或试算方法确定参数的局限性。这在一定程度上限制了模型的泛化能力。传统的机器学习模型架构简单、参数有限，适合比较简单的分类预测场景，不能挖掘大数据中深层的、隐含的时空相关性，故对复杂数据的分类预测能力有限。相对来说，深度学习通过监督学习或者无监督学习自动学习复杂数据中的模式并提取数据特征，其强大的分层特征学习能力在各种机器学习任务中取得了巨大的成功[9] [10]。

深度学习是实现人工智能的重要技术之一。近年来，深度学习被广泛应用于计算机视觉、机器翻译与语音识别、信息检索和医学诊断等领域。深度学习模型是通过卷积神经网络结构提取数据深层次特征，实现网络端到端的训练，通过不断优化网络参数使得模型具有更好预测能力。Joseph 等开发了一个基于贝叶斯优化、引入注意力机制的可解释的 TabNet 模型进行糖尿病的预测分类[11]。TabNet 架构可进行特征重要性分析，用于确定对糖尿病分类最重要的特征，该模型结合贝叶斯优化和 TabNet 的优点，对有关诊断提供有贡献的潜在因素的见解，以确保医生和患者理解模型预测的原因。该模型的缺点是训练时间长且仅适用于表格数据。Liu 等人针对糖尿病分类预测，提出一种支持图卷积神经网络的双流学习架构，通过临床数据证明了所提方法的有效性[12]。Bala 等采用极限树和随机森林模型对糖尿病的特征进行筛选，然后构建基于深度神经网络的分类预测模型，以 PIMA 糖尿病数据集为基础验证了模型的有效性，并与多种基线模型进行比较，结果表明所提模型分类效果显著[13]。Kannadasan 等使用堆叠的自动编码器训练网络层，提取数据中的隐藏特征，然后加入 softmax 层对数据集进行分类，最后利用训练数据集的监督反向传播方式对网络进行微调，构建了一个基于深度神经网络的糖尿病风险预测模型[14]，此模型最大程度集成了自动编码器和 softmax 的优点，从而实现良好的分类性能。

虽然深度学习模型对复杂数据具有良好的分类性能，但是，在实践中也存在一些挑战，比如模型优化算法、网络结构、超参数、激活函数等设置对深度学习模型的性能和效果会产生不同的影响。神经网络的训练中所遇到的问题对深度学习模型优化的挑战更大，科研人员也做了很多研究工作，提出了多种优化算法。张鑫等人选择 SVM 作为分类模型，通过粒子群优化算法 PSO (Particle Swarm Optimization)对 SVM 的参数误差惩罚因子和核函数进行优化，输出最佳参数组合，用于构建糖尿病性视网膜病变和神经病变的并发症分类预测模型[15]。Karegowda 等人采用决策树和相关性 Genetic 特征选择方法进行特征筛选，并建立了基于 GA-BP 神经网络的糖尿病分类模型，混合 GA-BP 可以得到最优的网络参数和权值[16]。Aljarah 等使用鲸鱼优化算法 WOA (Whale Optimization Algorithm)优化多层神经网络的权重和偏置，建立了基于多层神经网络的糖尿病分类模型[17]。Si 等以 15 个医学数据集为基础，研究了不同优化算法在人工神经网络分类中的效果[18]。尽管近年来基于深度学习的糖尿病分类研究取得了长足的进步，但神经网络在参数训练过程中容易陷入局部最优，影响网络模型的预测性能，因此，针对神经网络优化方法的研究仍亟待深入展开。群智能优化算法及其应用是当前优化领域的重要分支和研究热点，也是交叉学科的前沿性研究方向之一。本文提出基于改进狮群算法优化神经网络的糖尿病风险预测模型。首先，采用合成少数类过采样 SMOTE (Synthetic Minority Oversampling Technique)技术消除数据集不平衡性问题，采用递归特征消除 RFE (Recursive Feature Elimination)方法进行特征选择，然后，引入非线性扰动因子改进狮群算法 LSO 收敛因子，得到一种改进狮群算法 ILSO，采用 ILSO 对 BP 神经网络的网络参数进行优化。最后，利用优化后 BP 模型对真实糖尿病数据进行分类预测。

2. 改进的狮群优化算法

2.1. 狮群优化算法

狮群优化算法(Lion Swarm Optimization, LSO)是刘生建等人于 2018 年提出的新的群智能优化算法[19]。相比 PSO、人工蜂群算法 ABC (Artificial Bee Colony)、引力搜索算法 GSA (Gravitational Search Algorithm)等，LSO 算法收敛速度最快，最容易跳出局部最优值，可以更好地解决全局最优问题。

2.2. 改进狮群算法

在 LSO 算法中，狮王的最佳位置受到参数 β、α_f、α_c 的影响，其中 α_c 在迭代过程中线性递减，但是狮群算法在实际搜索过程是非线性变化的。为了更好地平衡算法的全局搜索与局部寻优，我们引入了非线性扰动因子对 α_c 进行改进，得到改进 LSO (ILSO)算法。α_c 的改进公式为：

$$\alpha_c = 0.1\left(\overline{h}-\overline{l}\right)\left(1+\sqrt{1-\left(\frac{k}{T}\right)^2}-\left(\frac{e^{\frac{k}{T}}-1}{e-1}\right)^2\right) \tag{1}$$

扰动因子 α_c 在改进前后的变化趋势见图 1。改进后的扰动因子 α_c 在迭代初期变化速率较慢，保证充分的全局搜索能力，在迭代后期，衰减速率较快，这不仅增强了局部搜索能力，同时提高寻优速度，使得算法快速收敛。整体而言，改进后的 α_c 能更好的权衡全局优化与局部优化之间的关系，提高算法优化能力和收敛速度。

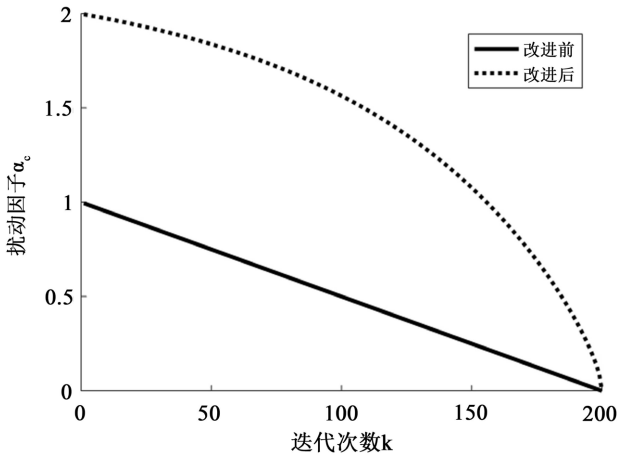

Figure 1. The variation trend of disturbance factor before and after improvement
图 1. 改进前后扰动因子的变化趋势

2.3. ILSO-BP 预测模型

BP 神经网络既有着自学习、自组织等优势，也存在容易陷入局部最优、收敛速度缓慢、训练精度不高等缺点。为了改进这些缺点，本文将改进的狮群优化算法 ILSO 引入 BP 神经网络的参数寻优过程，用 ILSO 优化算法取代传统梯度下降法优化 BP 神经网络的权重和偏置，并构建基于 ILSO-BPNN 的糖尿病分类预测模型。

基于 ILSO-BP 的糖尿病分类预测模型的算法流程见图 2，主要步骤包括：

1) 原始糖尿病数据 PIMA 属于非平衡数据，使用合成少数类过采样技术 SMOTE (Synthetic Minority

Oversampling Technique) [20]消除不平衡性；为避免特征冗余，使用递归特征消除 RFE (Recursive Feature Elimination) [21]方法进行特征选择。

2) 确定预测 BP 神经网络结构。根据预处理数据集的特征，确定 BP 神经网络结构，包括输入层神经元个数 I 、输出层神经元个数 O 、隐藏层神经元个数 H 等。

3) 种群初始化。设置种群数量 N，寻优空间的维度 D ，最大迭代次数 T ，种群初始位置 $x_i=\left(x_{i1},x_{i2},\cdots,x_{iD}\right),i=1,2,\cdots,N$ ，成年狮数量 β 。其中寻优空间的维度 D 由 BP 神经网络结构决定 $D=H\times I+H\times 1+O\times H+O\times 1$ 。

4) 确定适应度函数。使用训练数据的预测值 y 与真实值 y 之间的交叉熵函数 L 作为适应度函数。设 BP 神经网络的第 l 层第 j 个节点的输入为：

$$net_{jk}^{l}=\sum_{i}W_{ji}^{l}O_{ik}^{l-1}+b_{j}^{l} \tag{2}$$

激活函数为 σ ，第 l 层第 j 个节点的输出为：

$$O_{jk}^{l}=\sigma\left(net_{jk}^{l}\right) \tag{3}$$

则适应度函数为：

$$L\left(O_{jk}^{l},y\right)=y\log O_{jk}^{l}+\left(1-y\right)\log\left(1-O_{jk}^{l}\right) \tag{4}$$

其中，W_{ji}^{l} 为第 $l-1$ 层第 i 个神经元与第 l 层第 j 个神经元之间的连接权重，b_{j}^{l} 表示第 l 层第 j 个神经元的偏置。

5) 分配狮王、母狮、幼狮位置。狮群个体位置代表待寻优的 BP 神经网络的初始权重和偏置。

6) 狮群位置更新。依据(5) (6) (7)式更新狮王、母狮、幼狮的位置进行全局搜索，计算适应度值并排序。

$$x_{i}^{k+1}=g^{k}\left(1+\gamma\left\|p_{i}^{k}-g^{k}\right\|\right) \tag{5}$$

$$x_{i}^{k}=\frac{p_{i}^{k}+p_{c}^{k}}{2}\left(1+\alpha_{f}\gamma\right) \tag{6}$$

$$x_{i}^{k}=\begin{cases}\dfrac{p_{i}^{k}+g^{k}}{2}\left(1+\alpha_{c}\gamma\right),q<\dfrac{1}{3}\\ \dfrac{p_{i}^{k}+p_{m}^{k}}{2}\left(1+\alpha_{c}\gamma\right),\dfrac{1}{3}\le q<\dfrac{2}{3}\\ \dfrac{\overline{g}^{k}+p_{i}^{k}}{2}\left(1+\alpha_{c}\gamma\right),\dfrac{2}{3}\le q<1\end{cases} \tag{7}$$

其中 $\alpha_{f}=0.1(\overline{h}-\overline{l})\exp\left(-\dfrac{30k}{T}\right)^{10}$ ，p_{i}^{k} 为第 i 个狮子在第 k 代的历史最优位置，g^{k} 为第 k 代群体最优位置，γ 是服从正态分布 $N\left(0,1\right)$ 的随机数，p_{c}^{k} 为第 k 代母狮在群体中随机挑选的捕食伙伴的最优位置，α_{f} 为扰动因子，$\overline{l}$ 和 $\overline{h}$ 分别表示狮子活动范围空间各维度的最小值均值和最大值均值，α_{c} 为幼狮位置移动范围扰动因子，p_{m}^{k} 为幼狮跟随母狮的第 k 代的最佳位置，$\overline{g}^{k}$ 为第 i 个幼狮在捕猎范围内被驱赶的位置，$\overline{g}^{k}=\overline{h}+\overline{l}-g^{k}$ ，q 为概率因子，q 服从均匀分布 $U\left[0,1\right]$ 。

7) 判断是否达到优化条件。若迭代次数未到达 T ，且相邻两次训练结果的误差未小于阈值 ε，则返回步骤 6)；否则，输出最优的狮群位置向量，并转步骤 8)。

8) 获得 BP 神经网络最优参数。狮群算法寻优产生的最优值解码得到 BP 神经网络的最优权值和偏置。

9) 建立基于 BP 神经网络的糖尿病预测模型并进行分类预测。将相关数据输入到 ILSO-BP 网络模型，输出结果即为分类预测值。

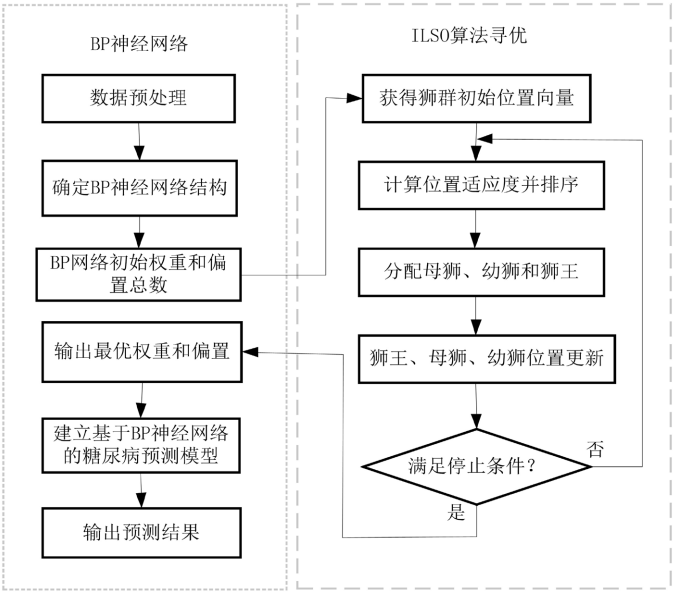

Figure 2. Flowchart based on ILSO-BP diabetes prediction model
图 2. 基于 ILSO-BP 糖尿病预测模型流程图

3. 实验结果与分析

3.1. 数据预处理

本文所选择的数据集是 PIMA 糖尿病数据集，来源于美国国家糖尿病和消化肾脏疾病研究所[22]。表 1 为数据集的基本特征。数据集共 768 行 9 列，8 个特征 1 个类别，其中阴性样本 500 条，阳性样本 268 条，数据类别不平衡。为了提高模型分类精度，对数据做预处理，包括过采样和特征选择。我们选择 70% 训练数据做 SMOTE 过采样处理来增加阳性样本数，减少噪声对阳性样本分类效果的影响。过采样的数据有 686 条，阳性样本和阴性样本各 343 条。使用递归特征消除法 RFE 消除特征之间的冗余，选取最优特征组合，降低特征维数。通过实验发现 PIMA 数据中的 8 个特征不存在冗余。

3.2. 模型性能评价指标

本文选择准确率 A、精确率 P、召回率 R、$F1$ 和 AUC 作为评价指标。其中 A、P、R 和 $F1$ 由公式(8)~(11)计算得到：

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$P=\frac{TP}{TP+FP} \tag{9}$$

$$R=\frac{TP}{TP+FN} \tag{10}$$

$$F1=\frac{2}{\frac{1}{P}+\frac{1}{R}} \tag{11}$$

其中，*TP* 指将正类预测为正类的数目，*FN* 指将正类预测为负类的数目，*FP* 指将负类预测为正类的数目，*TN* 将负类预测为负类的数目。

AUC (Area under the ROC Curve)值是指 ROC (Receiver Operating Characteristic Curve)曲线下的面积，其中 ROC 曲线以假正类率为横坐标，真正类率为纵坐标的绘制的曲线。AUC 越接近 1，预测精度越高，当 AUC 为 0.5 时，预测无意义。

Table 1. Basic characteristics of PIMA dataset
表 1. PIMA 数据集基本特征

序号	特征	缺失值	均值	标准差	最大值/最小值
1	Pregnancies	111	3.8451	3.3696	17/0
2	Glucose	5	120.8945	31.9726	199/0
3	Blood Pressure	35	69.1055	19.3558	122/0
4	Skin Thickness	227	20.5365	15.9522	99/0
5	Insulin	374	79.7995	115.2440	846/0
6	BMI	11	31.9926	7.8842	67.1/0
7	Diabetes Pedigree Function	0	0.4719	0.3313	2.42/0.078
8	Age	0	33.2409	11.7602	81/21

Table 2. Forecast performance comparison
表 2. 预测性能比较

模型	准确率(%)	精确率(%)		召回率(%)		F1 (%)	
		y = 0	y = 1	y = 0	y = 1	y = 0	y = 1
KNN	76.63	81.91	71.11	74.76	79.01	78.17	74.85
SVM	72.83	74.77	70.13	77.67	66.67	76.19	68.35
RF	79.35	86.52	72.63	74.76	85.19	80.21	78.41
NB	77.17	83.50	69.14	77.48	76.71	80.37	72.73
DT	82.61	84.47	80.25	84.47	80.25	84.47	80.27
ILSO-BP	**87.50**	**93.20**	**80.25**	**85.71**	**90.28**	**89.30**	**84.97**

3.3. ILSO-BP 模型参数设置

BP 神经网络输入层节点为 8，2 个隐含层，每层神经元数为 50，输出层节点为 1，最大迭代次数为 5000，学习率 0.1，激活函数 tanh，损失函数为交叉熵损失函数。狮群数量 30，β 为 0.2，最大迭代次数 500，l 和 h 的初值分别为−0.1 和 0.1。实验环境采用英特尔第十代酷睿 i5-10210U 处理器，显卡型号 2500U。算法模型采用 python3.9 编程，绘图工具采用 matplotlib 绘图模块。

3.4. 实验结果分析

3.4.1. ILSO-BP 与基线模型的性能比较

本文选择的基线模型包括基于网格搜索优化超参数的 K 近邻(KNN)、支持向量机(SVM)、随机森林(RF)方法和朴素贝叶斯(NB)、决策树(DT)方法。ILSO-BP 模型和所有基线模型的评价指标见表 2。由表 2

可知，我们所提出的 ILSO-BP 模型分类准确率达到 87.50%，优于基线模型。不论是标签为 0 的样本还是标签为 1 的样本，其精确率、召回率、F1 值都明显大于其他基线模型的各个指标。实验结果证明本文所提出的 ILSO-BP 模型对糖尿病分类具有较高的预测精度。

图 3 是各分类器的 ROC 曲线图。ROC 曲线图下的面积是 AUC，AUC 越接近 1，模型检测效果越好。由图 3 对 AUC 排序：ILSO-BP > DT > RF > KNN > NB > SVM，且 ILSO-BP 的值为 0.8673，明显大于其基线模型的 AUC 值。依据 AUC 的一般判断标准说明我们所提的 ILSO-BP 模型对 PIMA 数据的分类效果很好，可用于糖尿病风险预测。

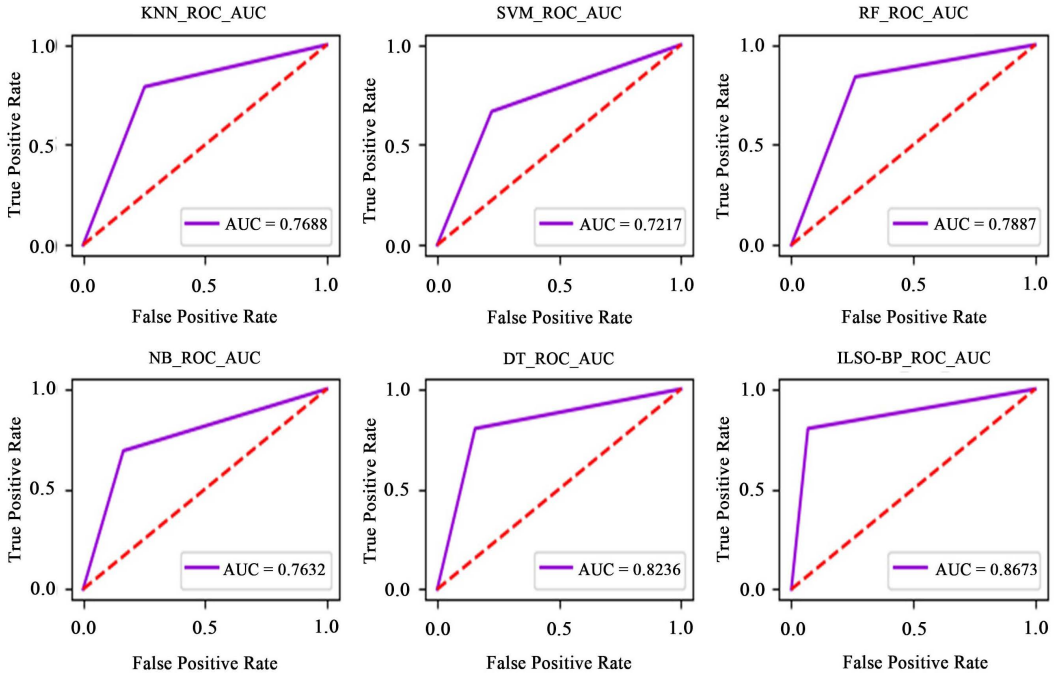

Figure 3. ROC plot
图 3. ROC 曲线图

3.4.2. ILSO-BP 与其他糖尿病预测模型性能比较

表 3 给出本文所提出的预测模型与其他已有预测糖尿病预测模型的比较。所有文献都采用公共数据集 PIMA。文献[23] [24]和[25]分别采用了粒子群优化、多目标优化及粗糙集与蝙蝠优化的模糊理论分类预测模型，他们的准确率分别为 82.32%、81.50%和 85.33%。Aljarah 等提出基于 WOA、PSO 优化 BP 的模型分类准确率分别为 77.86%、79.77% [17]。文献[16]提出的基于 GA 优化 BP 模型的准确率为 84.71%。使用本文提交的预测模型的准确率是 87.50%，该模型在预测糖尿病方面有较高的准确率。

3.4.3. 消融实验

通过消融实验检验基于 ILSO-BP 模型中的每个组件对预测效果的影响，实验结果见表 4。BP 模型的准确率分别为 85.33%，增加了狮群优化算法的 LSO-BP 的准确率提高到 86.41%，基于改进狮群优化算法的 ILSO-BP 的准确率达到了 87.50%，与 BP 和 LSO-BP 相比较，ILSO-BP 的准确率分别提高了 2.17%和 1.09%。精确率、召回率和 F1 等指标的值也说明了改进的预测模型 ILSO-BP 具有良好的预测性能。所提的 ILSO-BP 模型和 LSO-BP、BP 模型的实验耗时分别为 129 s，126 s 和 154 s，这一事实说明引入狮群算法确实提高了神经网络的参数寻优效率。

3.4.4. 数据预处理方法对模型性能的影响

正确合理的数据预处理对模型的预测性能也有较大影响。数据预处理阶段不仅采用 SMOTE 过采样技术消除样本不平衡问题，而且使用递归特征消除法 RFE 消除特征之间的冗余。通过实验发现 PIMA 数据中的 8 个特征不存在冗余，故在模型训练和预测中仍保持 8 个特征。表 5 为不同数据预处理时模型性能，表中 A_{SRIB} 代表了(SMOTE) + (RFE) + (ILSO-BP)模型的准确率，A_{RIB} 代表了(RFE) + (ILSO-BP)模型的准确率。如果选择其中 4 个特征进行建模预测，准确率 A_{RIB} 和 A_{SRIB} 分别为 68.83%和 64.13%，我们发现过采样 SMOTE 降低了预测性能。5 个特征预测也有类似结果。当选取特征数为 6 和 7 时，过采样技术对模型预测性能几乎没有影响。当 SMOTE 与 RFE 相结合后，准确率 A_{RIB} 和 A_{SRIB} 分别为 83.12%和 87.50%，预测性能有显著提高。

Table 3. Performance comparison of ILSO-BP with other diabetes prediction models
表 3. ILSO-BP 与其他糖尿病预测模型性能比较

方法	模型核心思想	准确率(%)	参考文献
PSO-Nefclass	基于粒子群优化(PSO)的神经模糊分类模型(Nefclass)	82.32	Mostafa *et al.* [23]
FRBCSs-MOEOAs	基于多目标进化优化(MOEOAs)的模糊规则的分类系统(FRBCS)	81.50	Marian *et al.* [24]
RST-BatMiner	基于粗糙集理论(RST)和蝙蝠优化(BA)的混合决策支持系统	85.33	Cheruku *et al.* [25]
WOA + BP	基于鲸鱼优化(WOA)的 BP 神经网络	77.86	Aljarah *et al.* [17]
PSO + BP	基于粒子群优化(PSO)的 BP 神经网络	79.77	Aljarah *et al.* [17]
GA + BP	基于遗传算法(GA)的 BP 神经网络	84.71	Karegowda *et al.* [16]
ILSO-BP	基于改进狮群优化(ILSO)的 BP 神经网络	87.50	本文提出的模型

Table 4. Ablation test results
表 4. 消融实验结果

模型	准确率(%)	精确率(%)		召回率(%)		F1(%)	
		y = 0	y = 1	y = 0	y = 1	y = 0	y = 1
BP	85.33	79.61	92.59	93.19	78.13	85.86	84.75
LSO-BP	86.41	81.55	92.59	93.34	79.79	87.05	85.71
ILSO-BP	87.50	93.20	80.25	85.71	90.28	89.30	84.97

Table 5. Impact of data preprocessing on ILSO-BP
表 5. 数据预处理对 ILSO-BP 的影响

特征序号								准确率(%)	
1	2	3	4	5	6	7	8	ARLB	ASRIB
	√				√	√	√	68.83	64.13
√	√				√	√	√	69.48	63.59
√	√	√			√	√	√	74.68	74.46
√	√	√	√		√	√	√	75.97	75.54
√	√	√	√	√	√	√	√	83.12	87.50

4. 结语

研究证明，越早了解糖尿病患病风险并进行早期干预可降低糖尿病致死率。本文提出一种基于改进狮群算法优化神经网络的糖尿病风险预测模型。首先，引入非线性扰动因子改进初始狮群算法收敛因子，在初始阶段，非线性扰动因子递减缓慢，提升算法跳出局部搜索的能力，从而避免算法陷入局部最优，最终达到增强全局优化能力，在后期阶段，扰动因子快速递减，这样可增强算法局部搜索能力，提高算法的收敛速度。利用改进狮群算法(ILSO)的寻优能力优化神经网络参数，提高神经网络的精度与收敛速度。构建基于 ILSO-BP 神经网络的预测模型。同时，采用过采样技术 SMOTE 和递归特征消除方法 RFE 对糖尿病数据进行预处理，提升模型预测能力。以皮马印第安人糖尿病数据集 PIMA 为基础进行了广泛实验，实验结果表明，经过改进狮群算法优化 BP 神经网络预测模型对糖尿病风险具有良好预测能力，为糖尿病的辅助诊断提供支持。下一步将在更多数据集上验证模型的有效性，同时尝试其他优化算法或数据预处理方法，以提高模型的预测精度。

基金项目

国家自然科学基金/National Natural Science Foundation of China (71761031)。

参考文献

[1] International Diabetes Federation (2022) IDF Diabetes Atlas. 10th Edition, International Diabetes Federation. https://diabetesatlas.org/

[2] Sun, H., Saeedi, P., Karuranga, S., *et al*. (2022) IDF Diabetes Atlas: Global, Regional and Country-Level Diabetes Prevalence Estimates for 2021 and Projections for 2045. *Diabetes Research and Clinical Practice*, **183**, 109-119. https://doi.org/10.1016/j.diabres.2021.109119

[3] Malviya, L., Mal, S., Lalwani, P., *et al*. (2021) Diabetes Classification Using Machine Learning and Deep Learning Models. In: Bajpai, M.K., Singh, K.K. and Giakos, G., Eds., *Machine Vision and Augmented Intelligence—Theory and Applications*, *Lecture Notes in Electrical Engineering*, Vol. 796, Springer, Singapore, 487-503. https://doi.org/10.1007/978-981-16-5078-9_40

[4] Tigga, N.P. and Garg, S. (2020) Prediction of Type 2 Diabetes Using Machine Learning Classification Methods. *Procedia Computer Science*, **167**, 706-716. https://doi.org/10.1016/j.procs.2020.03.336

[5] 张春富, 王松, 吴亚东, 王勇, 张红英. 基于 GA-Xgboost 模型的糖尿病风险预测[J]. 计算机工程, 2020, 46(3): 315-320.

[6] Ali, M., Haider, M.N., Lashari, S.A., *et al*. (2022) Stacking Classifier with Random Forest Functioning as a Meta Classifier for Diabetes Diseases Classification. *Procedia Computer Science*, **207**, 3459-3468. https://doi.org/10.1016/j.procs.2022.09.404

[7] Kalagotla, S.K., Gangashetty, S.V. and Giridhar, K. (2021) A Novel Stacking Technique for Prediction of Diabetes. *Computers in Biology and Medicine*, **135**, Article ID: 104554. https://doi.org/10.1016/j.compbiomed.2021.104554

[8] Kumari, S., Kumar, D. and Mittal, M. (2021) An Ensemble Approach for Classification and Prediction of Diabetes Mellitus using Soft Voting Classifier. *International Journal of Cognitive Computing in Engineering*, **2**, 40-46. https://doi.org/10.1016/j.ijcce.2021.01.001

[9] Alex, S.A., Nayahi, J.J.V., Shine, H., *et al*. (2022) Deep Convolutional Neural Network for Diabetes Mellitus Prediction. *Neural Computing and Applications*, **34**, 1319-1327. https://doi.org/10.1007/s00521-021-06431-7

[10] 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.

[11] Joseph, L.P., Joseph, E.A. and Prasad, R. (2022) Explainable Diabetes Classification Using Hybrid Bayesian-Optimized TabNet Architecture. *Computers in Biology and Medicine*, **151**, Article ID: 106178. https://doi.org/10.1016/j.compbiomed.2022.106178

[12] Liu, Y.C., Liu, W., Chen, H.R., *et al*. (2021) Graph Convolutional Network Enabled Two-Stream Learning Architecture for Diabetes Classification Based on Flash Glucose Monitoring Data. *Biomedical Signal Processing and Control*, 69, Article ID: 102896. https://doi.org/10.1016/j.bspc.2021.102896

[13] Bala, M.K.P., Srinivasa, P.R., Nadesh, R.K. and Arivuselvan, K. (2020) Type 2: Diabetes Mellitus Prediction Using

Deep Neural Networks Classifier. *International Journal of Cognitive Computing in Engineering*, **1**, 55-61. https://doi.org/10.1016/j.ijcce.2020.10.002

[14] Kannadasan, K., Edla, D.R. and Venkatanareshbabu, K. (2019) Type 2: Diabetes Data Classification Using Stacked Autoencoders in Deep Neural Networks. *Clinical Epidemiology and Global Health*, **7**, 530-535. https://doi.org/10.1016/j.cegh.2018.12.004

[15] 张鑫, 韦哲, 曹彤, 王能才, 张海英. 基于 PSO-SVM 的糖尿病并发症预测研究[J]. 中国医学装备, 2022, 19(2): 10-13.

[16] Karegowda, A.G., Manjunath, A.S. and Jayaram, M.A. (2011) Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of PIMA Indians diabetes. *International Journal on Soft Computing*, **2**, 15-23. https://doi.org/10.5121/ijsc.2011.2202

[17] Aljarah, I., Faris, H. and Mirjalili, S. (2018) Optimizing Connection Weights in Neural Networks Using the Whale Optimization Algorithm. *Soft Computing*, **22**, 1-15. https://doi.org/10.1007/s00500-016-2442-1

[18] Si, T., Bagchi, J. and Miranda, P.B.C. (2022) Artificial Neural Network Training Using Metaheuristics for Medical Data Classification: An Experimental Study. *Expert Systems with Applications*, **193**, Article ID: 116423. https://doi.org/10.1016/j.eswa.2021.116423

[19] 刘生建, 杨艳, 周永权. 一种群体智能算法——狮群算法[J]. 模式识别与人工智能, 2018, 31(5): 431-441.

[20] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2011) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[21] Mundra, P.A. and Rajapakse, J.C. (2010) SVM-RFE with MRMR Filter for Gene Selection. *IEEE Transactions on NanoBioscience*, **9**, 31-77. https://doi.org/10.1109/TNB.2009.2035284

[22] https://aistudio.baidu.com/aistudio/datasetdetail/107922/0

[23] Daho, M.E.H., Settouti, N., Lazouni, M.E.A. and Chikh, M.A. (2013) Recognition of Diabetes Disease Using a New Hybrid Learning Algorithm for NEFCLASS. *Proceedings of IEEE Conference on* 8*th International Workshop on Systems, Signal Processing and Their Applications* (*WoSSPA* 2013), Algiers, 12-15 May 2013, 239-243.

[24] Gorzałczany, M.B. and Rudziński, F. (2017) Interpretable and Accurate Medical Data Classification—A Multi-Objective Genetic-Fuzzy Optimization Approach. *Expert Systems with Applications*, **71**, 26-39. https://doi.org/10.1016/j.eswa.2016.11.017

[25] Cheruku, R., Edla, D.R., Kuppili, V. and Dharavath, R. (2018) RST-BatMiner: A Fuzzy Rule Miner Integrating Rough Set Feature Selection and Bat Optimization for Detection of Diabetes Disease. *Applied Soft Computing*, **67**, 764-780. https://doi.org/10.1016/j.asoc.2017.06.032