第三届中国研究生人工智能创新大赛

[数据驱动下糖尿病并发症智能混合预测模型的研究]

项目文档

[1.0.0]

[2021.08.26]

[希望]

[应用创新]

摘　　要：

随着大数据技术的飞速发展，我国医疗信息化建设取得了很大的进步。与此同时，数据驱动的挖掘方法在生物医学中的应用比以往任何时候都更加迫切，因为这种方法可以将所有可用的信息转化为有价值的知识。

首先，对数据进行筛选和预处理。然后，选取可能会影响到糖尿病并发症的因素，用最大信息系数法（MIC）、随机森林 (RF) 和 CatBoost 组合得出各指标的综合得分，得出最终的主要影响指标。之后分别在 XGBoost 和 RF 算法的基础上，用遗传算法进行超参数调优。最后对 GA-XGBoost 和 GA-RF 进行组合，建立基于集成学习方法的 2 型糖尿病并发症智能混合预测模型，为 2 型糖尿病并发症的预防和早期筛查提供理论参考。

关键词：糖尿病，并发症，XGBoost 算法，智能混合预测

目录

1. 前言

1.1 研究背景

进入 21 世纪以来，随着卫生医疗信息化的发展，我国在医疗信息化建设方面取得了很大的进步。科学研究、卫生服务和管理实践的进步，令信息系统积累了大量的电子病历数据。但是对于大多数医院来说，电子病历系统更多的是一种医疗管理工具，数据资源并没有充分利用。与此同时，人工智能 (Artificial Intelligence, AI) 技术日新月异、蓬勃发展，数据驱动的挖掘方法在生物医学中的应用需求比以往任何时候都更加的迫切，因为该方法可以把所有可用的信息转化为有价值的知识。因此挖掘电子病案系统中的海量数据已成为医学信息化研究的一个趋势。

糖尿病（Diabetic Mellitus, DM）是以高血糖为特征的一种多病因代谢性疾病，伴随因胰岛素分泌不足或缺失而引发的糖、脂、蛋白代谢异常。近年来，糖尿病发病率日趋上升，已然成为 21 世纪全球最严峻的公共卫生问题之一。根据 2015 年 11 月 11 日国际糖尿病联合会（International Diabetes Federation，IDF）发布的第七版数据显示 [1]，糖尿病患者死亡率高于结核病、疟疾和艾滋病死亡率的总和，大约每 6 秒就会出现 1 例糖尿病患者死亡，目前糖尿病成年患者的数量已经从 2013 年的 3.82 亿增加到 4.25 亿，预计到 2040 年全球糖尿病患者将达到 6.42 亿人。中国糖尿病患者数量居于首位，2019 年糖尿病患者达到 1.16 亿人，患病率约为 10.9%，糖尿病患者人数仍有提升趋势。中国仅 2019 年就支出了约 1090 亿美元在糖尿病的相关医疗，糖尿病带来了巨大的社会经济负担。并发症对患者及社会分别造成了很大的身体痛苦及经济负担。根据粗略统计，经济负担对于患有并发症的糖尿病病患相较无并发症患者来说，是数倍甚至几十倍。糖尿病常见的并发症包括糖尿病肾病、视网膜病变、神经病变、糖尿病昏迷和心血管疾病。并发症的早期发现困难，药物难以治愈，因此并发症的预测成为一个研究领域热点。

根据上述所言，AI 技术和数据挖掘与临床医疗诊断的结合已经越来越紧密。这种形势也是大势所趋，在这些新兴技术的推动下，人们已经有能力从这些海量的医疗数据中得到有价值的信息，从而更好的去指导预防治疗等。医疗数据挖掘领域的蓬勃发展，促使更多实验室的研究逐步走向大众的生活，那么对更加智能化的医疗辅助诊断系统的需求也会更加迫切。根据查阅相关文献发现，目前对于糖尿病并发症的预测基本都还停留在数据模型阶段，没有从实验室走向医院得到实际应用，故而开发一款针对糖尿病并发症的预测系统是具有一定实际应用价值的。

大数据时代，数据结构日益多样，数据特征纷繁复杂，数据量级不断增加，传统的、单一的预测模型通常受制于现实数据的复杂特性，如非平稳、异方差性、混沌性以及非线性，使得预测建模已经成为一个极具挑战性的难题。没有哪种预测技术适用于所有场景，单一预测模型只能捕获时间序列的单一特征不能全方面地捕获数据的各个特征，传统预

测模型通常局限于模型的线性假设，机器学习模型，如 ANNs 模型容易陷入局部最优和出现过拟合现象，SVM 对参数选择较敏感，深度学习模型参数众多，微调困难等。基于上述问题，基于集成学习方法的智能预测模型顺势而出。混合预测模型 = 预测模型（如神经网络，ARIMA，支持向量机等）+ 智能优化算法（如粒子群优化算法、布谷鸟优化算法等）+ 数据预处理技术（如异常值检验、经验模态分解，变分模式分解等），充分利用各模型和算法优势来提高模型的预测性能。而智能预测引用集成学习的思想，将多个混合预测模型进行组合优化，目的明确，过程灵活，形成的混合模型有着更好的整体性能，能达到更高的预测效果，是近年来预测方向的重要发展趋势。

1.2 研究意义

此次研究可以应用在糖尿病并发症的预测，通过模型的延申，还可应用到其他疾病的预测及各类预测问题。

研究的意义如下：

(1) 针对医院数据存在的问题，设计一套针对性的数据预处理方案，有效解决由多种原因造成的数据冗余、缺失、异常等问题，对糖尿病并发症的临床数据规范管理和统一标准具有一定意义。

(2) 以智能算法和集成学习方法为基础构建糖尿病并发症智能混合预测模型，提高了糖尿病并发症的预测效果，并为糖尿病并发症早期预防提供了科学的依据，具有一定的理论价值。

(3) 通过设计实现了糖尿病并发症预测系统，更直观的展示了糖尿病数据整合、预处理、查询、统计分析以及对糖尿病并发病的预测结果，为糖尿病并发症的早期干预提供了辅助工具，便于该预测模型的实施和推广，具有一定的应用价值。

1.3 研究现状

Ahmad[2] 比较了神经网络中多层感知（Multi-Layer Perception，MLP）与 ID3 和 J48 算法的预测精度，结果表明，经过修剪的 J48 树的准确度较高为 89.3%。马卡诺塞德尼奥 [3] 提出以多层感知器人工化塑性作为糖尿病预测模型，获得的最好的结果为 89.93%。Mustafa S. Kadhm[4] 提出使用决策树（Decision Tree，DT）在应用 KNN 算法消除不需要的数据之后可将每个数据样本分配给其适当的类，得到更优的分类预测结果。由参考文献 [5] 可知，作者使用 k-means 聚类来识别和消除异常值，遗传算法和相关特征选择用于相关特征提取，最后使用 KNN 算法进行糖尿病患者的分类。Calisir 和 Dogantekin 提出（Linear Discriminant Analysis Morlet Wavelet Support Vector Machine）LDA-MWSVM，一种用于糖尿病诊断的系统 [6]。系统使用线性判别分析（Linear Discriminant Analysis，LDA）方法执行特征提取和缩减，然后使用 Morlet 小波支持向量机（Morlet Wavelet Support Vector Machine，MWSVM）分类器进行分类。Gangji 和 Abadeh [7] 提出一种基于蚁群的分类系

统来提取一组模糊规则，命名为 FCS-ANTMINER，用于糖尿病诊断, FCS-ANTMINER 具有新特征，使其与使用蚁群优化（Ant Colony Optimization，ACO）进行分类任务的现有方法不同, 获得 84% 的准确率。在文献 [8] 中，提出并发布了一个关于数据流挖掘的实时临床决策支持系统（rt-CDSS）的概念设计，引入新系统，可以分析医疗数据流并进行实时预测。

集成学习方法是将基学习器进行组合优化，有两个突出的贡献，一是把众多的简单模型通过一定策略进行组合优化，扬各家之所长，避单一之所短，再者是集成学习方法可以让研究者根据不同问题来组合不同的基学习器，以获得更好的解决方案 [9]。集成学习在实际应用中也越来越广泛，而且不断有科技公司开发出相关的集成学习包进一步推动了集成学习的发展，如 Yandex 公司开发的 CatBoost 和 Microsoft 开发的 LightGBM，使得集成学习在数据竞赛领域发展地如火如荼。伴随着数据挖掘领域的蓬勃发展，愈来愈多的研究者致力于将机器学习中的集成学习方法应用到医疗辅助诊断上面来，从而使得之前的单一模型来预测疾病的方式不断转向到多模型集成来预测疾病。王荣政等人融合线性回归、梯度提升决策树、随机森林等模型对血糖进行预测 [10]。陈佩华等人选择了两种常见的增强算法 Adaboost.M1 和 LogitBoost，以建立糖尿病诊断的机器模型 [11]。Beatriz López 等人使用基于 Bagging 集成思想的随机森林方法来探讨单核苷酸多态性与 2 型糖尿病风险预测之间的关系 [12]。张洪侠根据收集的数据采用基于 Boosting 集成思想的 XGBoost 构建 2 型糖尿病预测模型，发现对模型贡献前三名的变量依次是血糖、甘油三酯和 SLC30A8 基因 rs13266634-C 位点的等位基因 [13]。张玉玺采用了 KNN、支持向量机、逻辑回归、随机森林、集成学习五种方法对糖尿病数据进行预测 [14]。崔书华等人在记录大脑皮层厚度的图像数据基础之上，采用 Adaboost 集成学习方法来判断是否有人存在认知障碍 [15]。Heung-Il Suk 等人采用深度集成学习算法来诊断脑部疾病 [16]。霍东雪等人采用 Adaboost 集成学习方法对大量的医院儿科患者病历进行数据挖掘，从数据样本中提取儿童病患的病症，建立随机森林、支持向量机、逻辑回归三种基模型的融合方法来构建模型，以此来预测未知样本的患病可能性 [17]。

1.4 技术路线图

本文的技术路线图如下：

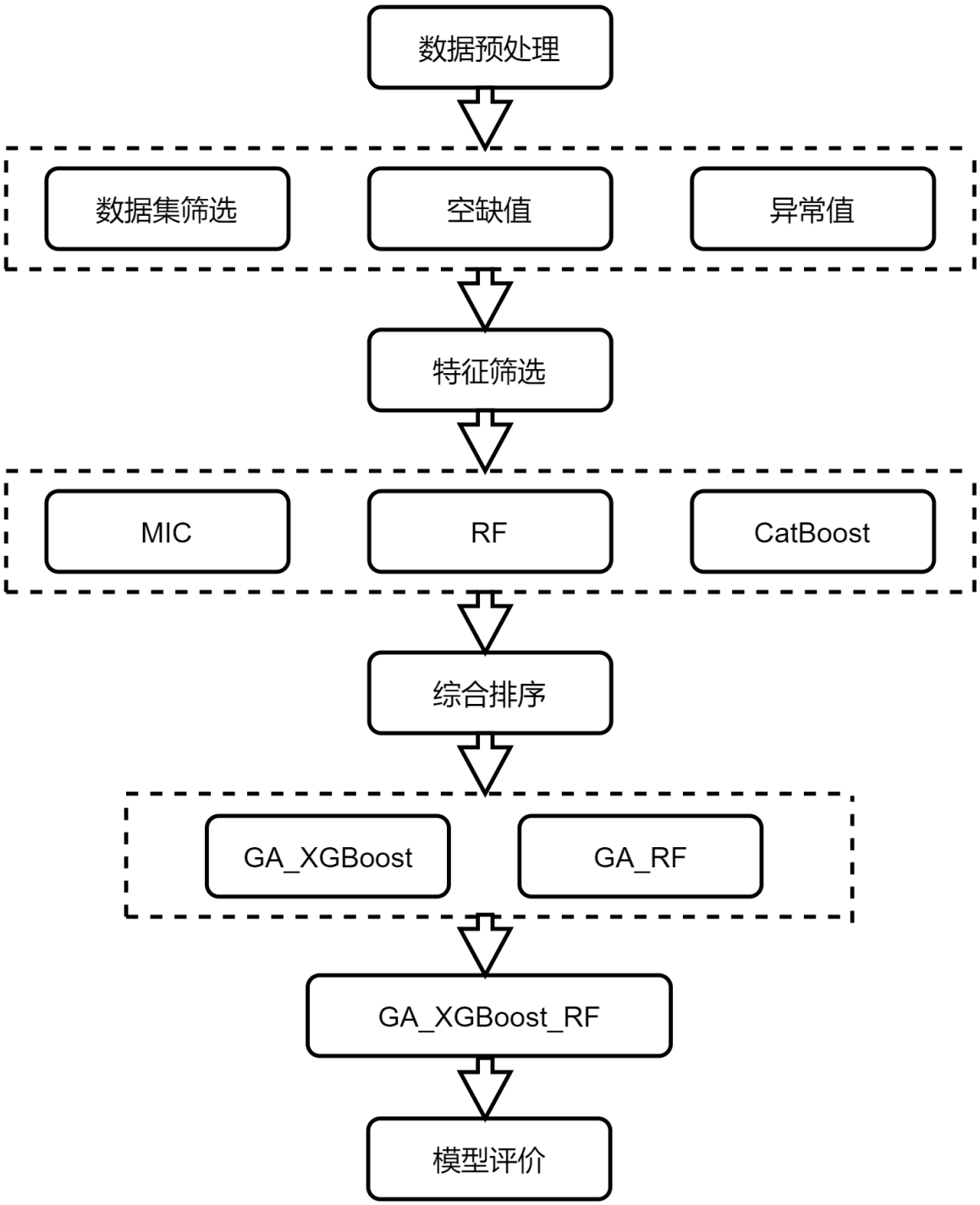

图 1　技术路线图

2. 数据预处理

项目数据来源于国家临床医学科学数据中心的《糖尿病并发症预警数据集》中的 3000 例糖尿病患者的数据，包含患者编号、基本信息、诊断信息、尿常规检查和生化检查

等，具体指标包括各项诊断结果、各项身体基本指标及糖化血红蛋白、甘油三酯等 88 个变量。

对本文的数据进行分析，发现文中数据存在空缺值以及异常值。

空缺值的处理主要是：对于空缺值过多和无实际意义的指标，直接删除；对于空缺值较少的指标，删除缺失的记录。再采用互信息法 [18, 19, 20]（Mutual Information）选取重要指标，进行空缺值处理；对剩余指标的空缺值进行基于链式方差的多次插补方法 [18, 21, 22]（Multivariate Imputation by chained Equations，MICE）进行填补替换。

对于异常值的处理，本文首先通过箱型图找出异常值，将异常数据用就近临界值替代；然后将处理后的数据通过基于库克距离的多变量异常值检验，若不再异常则保留替代方案，否则直接删除。

经过预处理后得到包含 76 个特征指标的 1027 条数据记录。

3. 特征筛选

将数据预处理完成之后，有价值的特征需要被输入机器学习的算法及模型中进行训练。我们一般从以下两方面来考虑是否选择该特征：

1、特征发散程度：若某个特征方差接近于 0，即该样本的这一特征基本无差异，即该特征不发散，那么我们就认为这一特征不能区分该样本。

2、特征与目标的相关性：特征与目标的相关性越高，越应该被选择。

常见的特征筛选方法大致分为：过滤式（filter）、包裹式（wrapper）和嵌入式（embedding）。

过滤式

直接用统计方式对某些特征打分，分数越高的特征，其价值越大。过滤式筛选方法更容易利用传统统计学方法来快速度量特征的重要性，这一筛选方法的结果有统计学理论作为保证，且这一筛选方法，在筛选的过程中没有涉及学习模型，因此筛选速度较快。但是过滤式特征筛选仅能够独立考察每个特征与目标变量间的相关性，从而忽略了不同特征间的关联性及组合效果。常见的过滤式筛选方法：方差选择法，卡方检验，皮尔森相关系数（Pearson Cerrelation Coefficient，PCC），最大信息系数（Maximal Information Coefficient）等。

包裹式

包裹式特征选择：依据最终要使用的机器学习模型、评测性能的指标来选择特征或排除特征。通常包裹式特征筛选要比过滤式特征筛选效果好，但是包裹式特征筛选法通常训练时间较长，系统开销也比较大。最典型的包裹型算法为递归特征删除算法，其原理是使用一个基模型（如：随机森林、逻辑回归等）进行多轮训练，每轮训练结束后，消除若干权值系数较低的特征，再基于新的特征集进行新的一轮训练。其能够直接针对特定学习器

进行优化，考虑到特征之间的关联性，因此通常包裹式特征选择比过滤式特征选择能训练得到一个更好性能的学习器。但是包裹式筛选方法选出特征通用性较弱。在筛选过程中，如果需要更改学习算法，必须针对该学习算法再次进行特征选择。每次对子集进行评价都需要重新对分类器进行训练和测试，所以这一算法很复杂，特别是遇到大规模数据居，算法的运行时间很长。常见的包裹式特征选择方法有：递归特征消除法，特征干扰法。

嵌入式

嵌入式特征筛选：是通过机器学习的算法、模型来分析某一特征是否重要，从而选出最重要的几个特征。嵌入式方法吧特征筛选过程和模型训练过程结合起来，从而更快速的找到最佳特征集合。其可以考虑到不同特征组合产生的组合效果来更好的评价特征重要性，但是计算时间开销很大。常见的嵌入式特征筛选方法有：基于惩罚项的特征选择法，随机森林（Random Forest,RF) 和 CatBoost 集成树模型。

3.1 最大信息系数法

最大信息系数是以信息论中的互信息为基础的，所以我们首先引入信息论中信息熵、条件熵和互信息量的概念及理论。

在如今数据驱动的时代，各行各业每天每时每刻都在产生无穷尽的数据，但是并不是数据越多越好或者如此多的数据可能有很多都不会给我们提供有用的信息。信息论作为衡量不确定性的一种方法，在机器学习算法中的特征选择方面，人们常常将不确定性较小的特征作为最优特征，但是问题在于如何将信息进行量化。基于此问题，Shannon 于 1948 年提出信息熵的概念，将特征的不确定性用数值的形式反映出来，随着信息熵的应用的发展，基于信息熵在特征选择算法上也更具有适用性, 本文把互信息最大信息系数法作为我们特征筛选的方法之一。

(1) 信息熵

信息熵将特征的不确定性用数值的形式反映出来，它是由 Shannon 于 1948 年提出来的。信息熵越大就表明特征的不确定性越大，反之。由于离散信源的明了和易懂性，以下所介绍的信息熵都是基于离散信源 [19]。信息熵的公式为：

$$H(X) = -\sum p(x_i)\log(p(x_i)), i = 1, 2, ..., n$$

其中 X 表示随机变量，信源概率为 $p(x_i)$,$i = 1, 2, ..., r$。

基于上式计算出的信息量为：

$$I(x_i) = -\log(p(x_i))$$

其中信息量 Ix_i 越高，$p(x_i)$ 越小，不确定度越高。

(2) 条件熵

设 X_1, X_2 是两个离散型随机变量，随机变量 X_1 给定的条件下随机变量 X_2 的条件熵 $H(X_2|X_1)$ 表示在已知随机变量 X_1 的条件下随机变量 X_2 的不确定性。公式推导如下：

$$H(X_2|X_1) = \sum_{x_1 \in X_1} p(X_1) H(X_2|X_1 = x_1) = -\sum_{x_1 \in X_1} \sum_{x_2 \in X_2} p(X_2|X_1) \log(p(X_2|X_1))$$

(3) 互信息量

变量 X_1，X_2 的信息熵分别是 $H(X_1), H(X_2)$,X_1，X_2 联合分布的信息熵为 $H(X_1, X_2)$。如果 X_1，X_2 相互独立，则

$$H(X_1, X_2) = H(X_1) + H(X_2)$$

互信息计算公式如下：

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$$

特征选择的最大信息系数法步骤:

设已知有包含 k 个类别的样本数据集，且这 k 个类别中的每类有 m 个样本，每个样本点包含 N 个特征 $\{{x_1}^{(1)}, ..., {x_m}^{(1)}, ..., {x_1}^{(k)}, ..., {x_m}^{(k)}\}$，其中 ${x_i}^{(k)} = (x_{i,1}, ..., x_{i,N})$ 表示类别 k 的第 i 个样本的特征组成的向量。

此方法根据单独最优特征组合搜索法的观点，从 N 个特征中搜索出有效性判据值最大的前 n 个特征作为特征子集用于分类。把最大信息系数作为有效性判据，用于单独最优特征组合搜索法中做特征选择，具体步骤如下：

步骤一：记对照样本为 $y^{(1)}, y^{(2)}, ..., y^{(k)}$, 它们是从每个类中分别随机选取的，其余样本即为训练样本；

步骤二：将对照样本的第 j 个特征与对应的 $m-1$ 个训练样本的第 j 个特征一一对应，k 个类别一共构成与特征 j 有关的 $k(m-1)$ 个变量对；

步骤三：求出步骤二中所得变量对的 MIC ，N 个特征共得到个 N 个 MIC ；

步骤四：最后，对全部特征的 MIC 值由大到小进行排序，选择 MIC 最大的前 n 个特征进入特征子集。

3.2 随机森林

随机森林（RF）是由 Leo Breiman 于 2001 年提出的将决策树中 CART 算法和 Bagging 算法相结合的一种新算法，它利用 bootsrap 重采样方法从原始样本中抽取多个样本，对每个 bootsrap 样本进行决策树建模，再通过多棵决策树的组合，最终以投票的方式得出预测

结果 [23]。大量的研究证明，随机森林算法具有很高的预测能力，且较传统的预测算法，不容易出现过拟合现象。同时，对于小样本非线性、高维模式的识别等问题有着其特有的优势。随机森林主要应用于分类问题和回归问题，对于分类问题，以最终的投票数决定最后的预测结果；而对于回归问题，将所有回归决策树输出值的平均值作为最终的预测值。

随机森林的流程：

（1) 对 N 组样本数据采用 bootstrap 抽样法进行有放回的随机抽样，抽取出 M 个样本，以取出的 M 个样本形成 M 棵能够进行模型训练的决策树，剩余的 $N-M$ 个样本作为袋外数据（out of bag，OOB）用来测试模型的准确性。

（2）假设原始数据样本有 P 个变量，则在每棵决策树的每个节点随机抽取 K 个变量作为备选分枝变量，依据分枝优度准则选择最佳分枝。

（3）每棵决策树开始自顶向下进行递归分枝，叶节点的最小尺寸设定为 5，以此作为决策树生长的终止条件且确保模型建立的准确性。

（4）将生成的 M 棵决策树组成随机森林回归模型，模型的回归效果采用袋外数据（OOB）预测的残差均方进行评价。

3.3 CatBoost

CatBoost 是 Boosting 策略的一种实现方式, 它和 lightGBM 与 XGBoost 类似, 都属于 GBDT 类的算法。CatBoost 在 GBDT 的基础上主要做了两点改进: 处理标称属性和解决预测偏移的问题, 从而减少过拟合的发生 [25]。

GBDT

GBDT 算法是通过一组分类器的串行迭代, 最终得到一个强学习器, 以此来进行更高精度的分类。它使用了前向分布算法, 弱学习器使用分类回归树 (CART)。

假设前一轮迭代得到的强学习器是 $F^{t-1}(x)$，损失函数是 $L\left(y, F^{t-1}(x)\right)$，则本轮迭代的目的是找到一个 CART 回归树模型的弱学习器 h^t，让本轮的损失函数最小。下式表示的是本轮迭代的目标函数 h^t。

$$h^t = \arg\min EL\left(\left(y, F^{t-1}(x) + h(x)\right)\right)$$

GBDT 使用损失函数的负梯度来拟合每一轮的损失的近似值, 下式中 $g^t(x,y)$ 表示的是上述梯度。

$$g^t(x,y) = \frac{\partial L(y,s)}{\partial s}\Big|_{s=F^{t-1}(x)}$$

通常用下式近似拟合 h^t。

$$[h^t = \arg\min_{h \in H} E\left(-g^t(x,y) - h(x)\right)^2$$

最终得到本轮的强学习器, 如下式所示:

$$F^t(x) = F^{t-1}(x) + h^t$$

Catboost

标称属性的一般处理方法是 one hot encoding (独热编码), 但是会出现过拟合的问题, CatBoost 在处理标称属性时使用了更有效的策略, 可以减少过拟合的发生。为训练集生成一个随机序列, 假设原来的顺序是 $\sigma = (\sigma_1, \cdots, \sigma_n)$。从 σ_1 到 σ_n 依次遍历随机序列, 用遍历到的前 p 个记录计算标称特征的数值。$\sigma_{p,k}$ 用如下数值替换:

$$\frac{\sum_{j=1}^{p}[x_{j,k} = x_{i,k}] \cdot Y_I + a \cdot P}{\sum_{j=1}^{n}[x_{j,k} = x_{i,k}] + a}$$

这里添加了一个先验值 P 和参数 $a > 0$。 这是一种常见做法, 它有助于减少从低频类别中获得的噪音。

预测偏移经常是困扰建模的问题, 在 GDBT 的每一步迭代中, 损失函数使用相同的数据集求得当前模型的梯度, 然后训练得到基学习器, 但这会导致梯度估计偏差, 进而导致模型产生过拟合的问题。CatBoost 通过采用排序提升 (ordered boosting) 的方式替换传统算法中梯度估计方法, 进而减轻梯度估计的偏差, 提高模型的泛化能力, Ordered boosting 的算法流程如表1所示。

表 1　Ordered boosting 流程

Ordered boosting	
Input$\{(x_k, y_k)\}_{k=1}^{n}, I$;	for i←1 to n do
σ←random permutation of$[1, n]$;	M←
M_i←0 for $i = 1...n$;	Learn Model$(x, r) : (j)$≤i
for t←1 to I do	M_i←$M_i + M$
for t←1 to n do	end for
r_i←y_i-$M_{\sigma(i)-1}(x_i)$	end for
end for	return M_n

由表1可知, 为了得到无偏梯度估计, CatBoost 对每一个样本都会训练一个单独的模型 M_i, 模型 M_i 由使用不包含样本 x_i 的训练集训练得到。我们使用 M_i 来得到关于样本的梯度估计, 并使用该梯度来训练基学习器并得到最终的模型。

得到这些重要程度并进行排序，最后按三个排名的均值排序，各项评分情况及排序如表2：

表 2　各模型排序

	MIC	MIC 排序	RF	RF 排序	CatBoost	CatBoost 排序	综合排序
GSP	0.227	4	2.193	9	2.438	9	1
UPR_24	0.355	2	1.814	14	2.133	14	2
BMI	0.122	22	3.232	6	2.859	6	3
SCR	0.256	3	1.773	16	1.982	16	4
BU	0.216	5	1.775	15	2.114	15	5
FBG	0.159	10	1.942	13	2.148	13	6
WEIGHT	0.084	35	5.811	3	6.799	3	7
BP_HIGH	0.089	33	4.169	4	5.034	4	8
AGE	0.045	40	13.265	1	12.138	1	9
GLU	0.105	28	3.156	7	2.693	7	10
HBA1C	0.101	30	2.194	8	2.636	8	11
HEIGHT	0.040	42	10.243	2	8.924	2	12
TC	0.114	25	1.973	11	2.319	11	13
LDL_C	0.117	23	1.961	12	2.220	12	14
CP	0.179	8	1.522	20	1.679	20	15
SUA	0.154	13	1.667	18	1.849	18	16
TG	0.105	29	2.020	10	2.363	10	17
BP_LOW	0.041	41	3.297	5	3.293	5	18
ESR	0.202	7	1.504	22	1.599	22	19
HB	0.135	17	1.596	19	1.774	19	20
TBILI	0.156	11	1.502	23	1.571	23	21
ALB	0.206	6	1.426	26	1.485	26	22
UCR	0.112	26	1.756	17	1.947	17	23

4. 智能混合预测模型

4.1 遗传算法

遗传算法（GA）又叫基因算法或进化算法，是一种启发式搜索算法。遗传算法中每个个体都是独立的一个解，通过选择、交叉、变异操作模拟生物的进化过程，从而产生一群更适应环境的个体，重复该过程不断繁衍进化，最后得到一群最适应环境的解 [26]。

遗传算法，又称遗传算法或进化算法，是霍兰德教授根据达尔文进化思想提出的一种启发式搜索算法。遗传算法中的每个个体都是一个独立的解。通过选择、交叉和变异操

作，模拟生物的进化过程，从而产生一组更适应环境的个体。在重复这个过程之后，它们不断地繁殖和进化，最终得到一组最适合环境的解决方案。遗传算法的伪代码如下，主要有五个步骤。

1、编码

遗传算法通过某种编码机制（如二进制编码、实数编码、灰度编码、有序编码等），将研究对象抽象成特定的符号，并按一定的顺序排列。从生物学的角度来看，染色体是由排列成直线的基因组成的。遗传算法中需要操作的对象是基因，因此需要将输入对象编码到相应的基因序列中。以最常用的二进制代码为例，每个个体由特定长度的二进制代码字符串表示，其等位基因值由 0 或 1 组成。

2、适应度计算

适合度计算的目的是评估个体的好坏。在优化问题中，适应值通常是最终目标函数。适应值高的个体有更大的机会存活到下一代，因此本文采用均方误差作为适应值。

通过计算适合度可以评估个体的好坏。通常在优化问题中，适应值是最终目标函数。适应值高的个体则表明该个体较好，因此本文采用均方误差作为适应值。

3、选择

选择的目的是从群体中选择优秀个体进行繁殖，体现了适者生存的原则。适应度函数值越大，则表明解的质量越高，个体的适应度越高，越容易生存和繁殖。适应度函数是遗传算法进行自然选择的唯一标准。轮盘赌算法通常用于随机选择个体，但也有一些情况下，优秀的个体被淘汰。因此，本文结合精英保留策略，将每一代的最优解直接复制到下一代，避免了进化过程中产生的最优解被交叉和变异破坏。

4、交叉

交叉是指以一定的概率交换染色体片段以产生两个新的后代，这是遗传算法中产生新个体的主要途径。交叉是将群体中的个体随机配对，并以一定的概率在它们之间交换一些染色体。整个过程体现了信息交流的理念。

5、变异

在繁殖过程中，群体中各个体的某些基因存在一定概率进行突变，该过程成为变异。变异操作的目的是使基因发生突变，在优化算法中，可以防止算法陷入局部最优。首先在群体中随机选择一定数量的个体，然后以一个较低的概率改变染色体的某个基因片段，为新个体的产生提供了机会。

在生殖过程中，种群中个体的某些基因有一定的突变概率，这一过程称为突变。变异操作能够使基因发生突变。在优化算法中，通常用变异操作来防止算法陷入局部最优。首先，从群体中随机选择一定数量的个体，然后以较低的概率改变染色体的某个基因片段，这一操作为新个体的出现提供了机会。

遗传算法的伪代码如表3:

表 3　遗传算法伪代码

遗传算法伪代码	
输入：	种群规模 P, 迭代次数 T，交叉概率 PC，变异概率 PM 等参数
1.	遗传代数 $t=0$
2.	初始化种群 $P(t)$
3.	计算 $P(t)$ 适应度
4.	while（不满足停止准则）do
5.	$t=t+1$
6.	从 $P(t-1)$ 中选择 $P(t)$
7.	按照 PC 进行交叉
8.	按照 PM 进行变异
9.	产生新的种群 $P(t)$
10.	计算 $P(t)$ 适应度
11.	end while
输出：	最佳个体

4.2 XGBoost

XGBoost 是“极端梯度提升”的缩写，它是在 GBDT 的基础上改进的。XGBoost 是通过 Boosting 将基函数和权重结合起来形成的一种综合算法。XGBoost 算法快速、高效、通用，适合处理大规模表格数据。自 2015 年被陈天齐提出后，它就被广泛应用于统计和机器学习领域 [26]。

对于 n 条 m 维的数据集：

$$\mathrm{D}=\left\{\left(\mathrm{x_i},\quad \mathrm{y_i}\right)\right\}\left(\mathrm{x_i}\in \mathrm{R^m}, \mathrm{y_i}\in \mathrm{R}, \mathrm{i}=1,2,\ldots \mathrm{n}\right)$$

XGBoost 模型表示为：

$$\widehat{y_l}=\sum_{k=1}^{K} f_k\left(x_i\right), f_k\in F(i=1,2,\ldots n)$$

上式中，K 代表树的棵数，x_i 表示第 i 个数据点的特征向量，f_k 表示一棵具体的 CART 决策树，F 是所有 CART 决策树的结构集合。

目标函数包含两部分：

$$Obj=\sum_{i=1}^{n} l\left(y_i, \widehat{y_l}\right)+\sum_{k=1}^{K} \Omega\left(f_k\right)$$

上式中：l 代表训练中存在的误差，是预测值和目标值之间按照特定标准计算的差异程度。Ω 则代表所有 CART 树的复杂度之和。

由于 Xboost 模型不是一个显式函数，无法用传统的方法进行优化，Xboost 使用加性训练进行优化，也就是每次优化一棵树，直到所有树都得到优化。程序如下：

$$\hat{y}_i^{(0)}=0$$

$$\widehat{y}_i^{(1)}=\mathrm{f}_1\left(x_i\right)=\hat{y}_i^{(0)}+\mathrm{f}_1\left(x_i\right)$$

$$\widehat{y}_i^{(2)}=\mathrm{f}_1\left(x_i\right)+\mathrm{f}_2\left(x_i\right)=\hat{y}_i^{(1)}+\mathrm{f}_2\left(x_i\right)$$

$$\cdots\cdots$$

$$\hat{y}_i^{(t)}=\sum_{\mathrm{k}=1}^{\mathrm{t}} \mathrm{f}_{\mathrm{k}}\left(x_i\right)=\hat{y}_{\mathrm{i}}^{(\mathrm{t}-1)}+\mathrm{f}_{\mathrm{t}}\left(x_i\right)$$

其中，$\hat{y}_i^{(t)}$ 为第 t 次模型的预测值，$\mathrm{f}_{\mathrm{t}}\left(x_i\right)$ 为第 t 次加入的新函数。每一轮加入新函数是为了尽可能让目标函数值最大程度的减小。将 $\hat{y}_i^{(t)}$ 带入公式 $Obj=\sum_{i=1}^{n} l\left(y_i, \widehat{y_l}\right)+\sum_{k=1}^{K} \Omega\left(f_k\right)$ 得到第 t 次模型的目标函数：

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right)+\sum_{i=1}^{t} \Omega\left(f_i\right)$$

$$=\sum_{i=1}^{n} l\left(y_{i,} \hat{y}_i^{(t-1)}+f_i\left(x_i\right)\right)+\Omega\left(f_t\right)+constant$$

采用泰勒展开近似定义误差函数 l，得到目标函数为：

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y}_i^{(t-1)}\right)+g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right]$$

$$+\Omega\left(f_t\right)+\ constant$$

其中，$g_i=\partial_{\widehat{y_l}(t-1)} l\left(y_{i,}, \widehat{y_l}^{(t-1)}\right)$，$h_i=\partial_{\widehat{y_l}^{(t-1)}}^2 l\left(y_i, \hat{y}_l^{(t-1)}\right)$，即误差函数 l 的一阶导数和二阶导数。把不影响优化的常数项移除后，得到最终目标函数：

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right]+\Omega\left(f_t\right)$$

XGBoost 作为集成学习中典型的 Boosting 思想算法，同样适用于糖尿病的预测和诊

断。Zhang HX 选择 53 名糖尿病患者和 93 名健康人检测糖尿病相关基因和常见临床指标，并使用 XGBoost 算法构建糖尿病预测模型。评估患糖尿病的风险。XGBoost 模型的预测精度为 86%。实验表明，基于 XGBoost 的糖尿病预测模型具有较好的预测能力。于力正在研究基于因子分析 XGBoost 模型的血糖回归和分类预测，利用因子分析降低特征维数，并通过回归和分类预测天池精准医疗大赛提供的数据。实验表明，基于因子分析的 XGBoost 模型在回归预测和分类预测中表现良好。

4.3 GA_XGBoost_RF 算法

本文利用遗传算法 (GA) 和随机森林 (RF) 算法与极端梯度提升 (XGBoost) 算法进行组合。XGBoost 算法是近年来兴起的一种高效集成学习方法 [27], 已在众多预测领域中取得了应用 [28]。随机森林是一种基于 Bagging 的集成学习方法 [29],XGBoost 算法属于 Boosting 集成学习方法, 将两种不同的集成学习方法进行组合可以结合两种方法的优点。使用遗传算法对 XGBoost 进行调参在运行时间和调参效果上优于网格搜索和随机游走 [30]。并且遗传算法对问题的可行解进行编码, 通过适应度来选择判断基因优劣, 对目标函数没有连续和可导的要求, 因此简化了组合模型参数调优和权值调优的复杂程度。

利用遗传算法良好的全局搜索能力和灵活性 [30], 对 XGBoost 算法和 RF 算法的参数进行优化, 然后利用遗传算法确定 GA_XGBoost 算法和 GA_RF 算法的权值, 最终建立 GA_XGBoost_RF 组合预测模型。下面将分别介绍参数调优和权值调优的具体内容。GA_XGBoost_RF 算法的流程如图2所示。

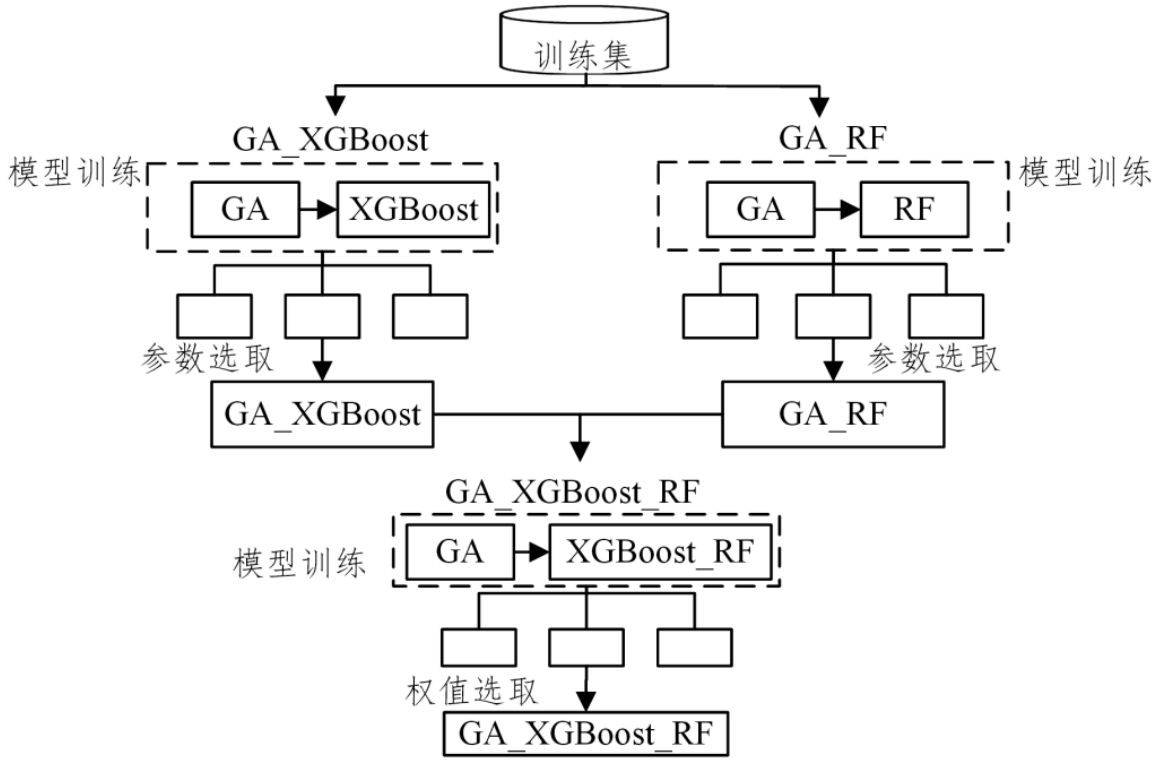

图 2　GA_XGBoost_RF 算法流程图

4.3.1 参数优化

XGBoost 算法和 RF 算法的参数较多, 参数选取直接影响算法的精度, 合理的参数设置可以明显提升模型的预测精度。本文利用遗传算法的全局寻优能力对 XGBoost 和 RF 模型进行参数选择, 使用交叉验证的平均得分作为适应度函数值,XGBoost 采用 5 折交叉验证,RF 因为有一定的随机性所以采用 7 折交叉验证, 建立 GA_XGBoost 和 GA_RF 模型进行算法参数优化。GA_XGBoost(GA_RF) 算法的伪代码如下：

GA_XGBoost (GA_RF) 算法

输入：群体规模 P，迭代次数 T，参数数 N，优秀个体数 u

输出: 参数的最佳组合

1. 初始化算法参数 $(\theta_{i1}, \theta_{i2}, \cdots, \theta_{iN})$
2. 当不满足终止条件时
3. 用 XGBoost(RF) 进行交叉验证
4. 计算适应度
5. 根据适应度选出 M 组最好算法参数组合
6. 进行遗传、变异等运算
7. 生成新的算法参数组合
8. end while

4.3.2 权值优化

对调参后的 GA_XGBoost 模型和 GA_RF 模型建立变权组合预测模型, 其中最重要的是两个模型各自权值的确定, 本文利用遗传算法来确定两个模型的权值。

首先利用参数优化中得到的参数组合在训练集上进行训练, 构建 GA_XGBoost 和 GA_RF 模型。其次使用遗传算法优化权值, 设置权值参数的范围、遗传算法的迭代次数、初始种群的数量, 然后随机生成 P 组初始值, 若不满足停止条件, 则将种群中的每个个体作为组合模型的权值对训练集进行预测, 以交叉熵为适应度函数, 以权值之和等于 1 为约束条件, 从种群中选取 u 个优秀个体。

$$fit = \sqrt{\frac{1}{m}(w_1\hat{y}_{XGB} + w_2\hat{y}_{RF} - y_{True})^2}$$

$$w_1 + w_2 = 1$$

其中,m 为训练集中的样本个数,w_1 和 w_2 分别为 XGBoost 和 RF 模型的权值,$\hat{y}_{XGB}$ 和 $\hat{y}_{RF}$ 分别为 XGBoost 和 RF 模型对训练集的预测值,y_{True} 为训练集的真实值。

对选取的优秀个体进行交叉、变异, 从而产生新的个体, 循环这个过程直到满足条件时停止, 从历代种群中选择最优值作为最终结果, 得到组合模型的权值组合, 建立 $GA\_XGBoost_RF$ 模型。$GA\_XGBoost_RF$ 算法如下：

$GA\_XGBoost_RF$ 算法

输入: 人口规模 P、迭代次数 T、杰出个人数量 U

输出: 最佳权重

1. 训练 $GA\_XGBoost$ 和 GA_RF 模型
2. 初始化权重
3. 当不满足终止条件时
4. 预测训练集
5. 计算适应度
6. 根据适应度选出 u 组最好权重组合
7. 进行遗传、变异等运算
8. 生成新的权重组合
9. end while

4.4 模型评价

T 和 F 代表 True 和 False，是形容词，代表预测是否正确。

P 和 N 代表 Positive 和 Negative，是预测结果。混淆矩阵如表4

表 4　混淆矩阵

	预测结果为阳性 Positive	预测结果为假阳性 Negative
预测结果正确 True	TP	TN
预测结果错误 False	FP	FN

1、准确率（accuracy) 是预测正确的样本占全部样本的比例，公式为：

$$accuracy(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\hat{y}_i = y_i\right)$$

2、精确率（precision）是预测正确的个数占总的正类预测个数的比例，公式为：

$$\frac{TP}{TP+FP}$$

3、召回率（recall）是针对原始样本而言的指标，表示真正样本中有多少预测对了，公式为：

$$\frac{TP}{TP+FN}$$

4、F1 分数 (F1-score) 是 precision 和 recall 的调和平均数（倒数平均数），将它们看成同等重要。通常 precision 和 recall 不能兼得，查准率高了查全率可能会偏低，查全率高了查准率可能会偏低，使用 F1 分数可以综合考虑它们二者。

$$F_1 = \frac{2}{\frac{1}{precision}+\frac{1}{recall}} = \frac{2\times precision\times recall}{precision+recall}$$

5、ROC 曲线, 绘制 ROC 曲线需要计算真正例率（TPR) 和假正例率（FPR), 其定义为：$TPR = \frac{TP}{TP+FN}$，$FPR = \frac{FP}{TN+FP}$。ROC 曲线以图形方式组合 TPR 和 FPR，可以直观反映分类器的真正例率和假正利率的关系。但是在实际应用中 ROC 曲线可能出现交叉情况，无法直观判断孰好孰坏，所以通常依据 ROC 曲线下方的面积 AUC 来判断，AUC 越靠近 1，说明模型性能越好。

基于数据集和上述评价指标, 设计回归预测实验, 将 GA_XGBoost_RF 与默认参数的 XGBoost 和 RF 算法、用遗传算法调参之后的 GA_XGBoost 算法和 GA_RF 算法进行比较, 结果如表5所列。

表 5　模型比较

	XGBoost	RF	GA_XGBoost	GA_RF	GA_XG_RF
AUC	72.81%	73.45%	76.38%	74.94%	74.94%
Accuracy	83.50%	85.92%	84.95%	86.89%	86.89%
Recall	90.85%	94.51%	90.85%	95.12%	95.12%
F1	89.76%	91.45%	90.58%	92.04%	92.04%
Precision	88.69%	88.57%	90.30%	89.14%	89.14%

其更直观地比较见图3：

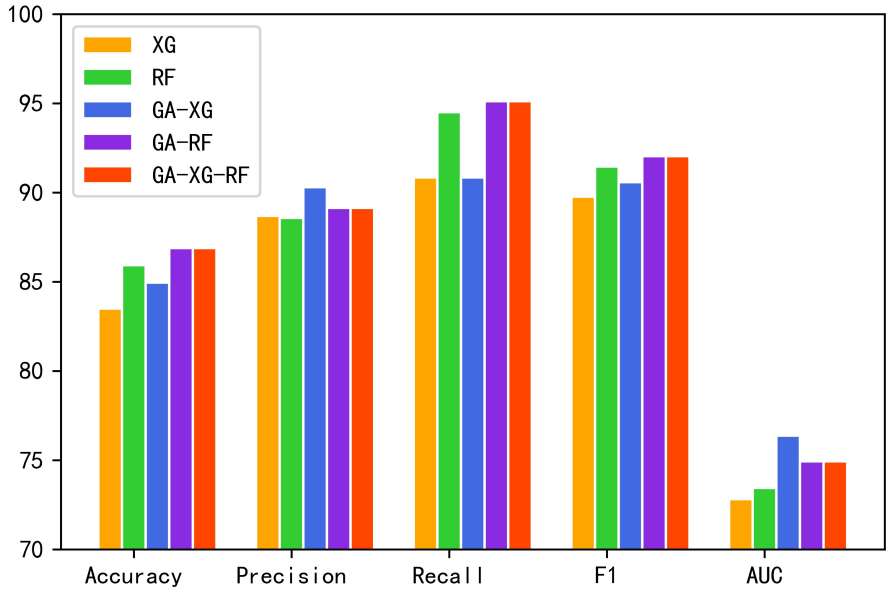

图 3　模型比较

由表5和图3可以看出,用遗传算法可以有效地提高 XGBoost 和 RF 的预测精度，而最后建立的智能组合预测模型 GA_XGBoost_RF 也可以更好地提高数据集的预测精度。

5. 模型评价及展望

本文在特征筛选时使用 MIC、RF 和 CatBoost 三种方法，所筛选出的特征可靠性较高，拟合得到的模型经度高，效果好。

本文将筛选出的特征输入到 GA_XGBoost_RF 模型中，再设置 AUC 为目标函数值用遗传算法对其超参数进行优化得到 AUC、Accuracy 、Recall、F1 和 Precision 四个分数均

较高的模型，有效地避免了人为调整超参数的繁琐性以及不准确性，得到了性能优良的模型。

展望：此次研究可以应用在糖尿病并发症的预测，通过模型的延申，还可应用到其他疾病的预测及各类预测问题。

参考文献

[1] Xu Y, Wang LM, He J, et a1. Prevalence and control of diabetes inChinese adults[J]. JAMA, 2013, 310(9): 948-959.

[2] Ahmad A，Mustapha A，Zahadi E D，et al．Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus[J]．Communications in Computer and Information Science，2011，188：537-545．

[3] Alexis Enrique Marcano Cedeño，Torres J，Fuente D A D L．A Prediction Model to Diabetes using Artificial Metaplasticity[J]．Lecture Notes in Computer Science，2011，6687：418-425．

[4] Mustafa S．Kadhm，Ikhlas Watan Ghindawi，Duaa Enteesha Mhawi．An accurate diabetes prediction system based on K-means clustering and proposed classification approach．Int J Appl Eng Res，2018，13(6)：4038-4041．

[5] Karegowda A G，Jayaram M A，Manjunath A S．Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients[J]．International Journal of Advanced Technology and Engineering Exploration，2012，1(3)：147-151．

[6] D．Çalisir，E．Dogantekin．An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine ClassifierExpert Syst．Appl．2011，38(7)：8311-8315．

[7] Ganji M F，Abadeh M S．A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis[J]．Expert Systems with Applications，2011，38(12)：14650-14659．

[8] Fong S，Zhang Y，Fiaidhi J，et al．Evaluation of Stream Mining Classifiers for Real-Time Clinical Decision Support System：A Case Study of Blood Glucose Prediction in Diabetes Therapy[J]．BioMed Research International，2013(4)：1-16．

[9] 徐继伟, 杨云. 集成学习方法: 研究综述 [J]. 云南大学学报 (自然科学版), 2018, 40(06):1082-1092.

[10] 王荣政, 廖贤艺, 陈湘萍, 等. 基于集成学习融合模型的血糖预测 [J]. 医学信息学杂志, 2019, 40(01):59-62+84.

[11] Chen P, Pan C D. Diabetes Classification Model based on Boosting Algorithms[J]. BMC Bioinformatics, 2018, 19(1):109-121.

[12] Beatriz L, Ferran T F. Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction[J].Artificial Intelligence in Medicine, 2018, 21(85):43-49.

[13] 张洪侠, 郭贺, 王金霞, 等. 基于 XGBoost 算法的 2 型糖尿病精准预测模型研究 [J]. 中国实验诊断学, 2018, 22(03):408-412.

[14] 张玉玺, 贺松, 尤思梦. 集成学习在糖尿病预测中的应用 [J]. 智能计算机与应用, 2019, 9(05):176-179.

[15] 崔书华, 胡斌, 胡涛. 阿尔茨海默病在脑皮层厚度中的集成分类方法研究 [J]. 小型微型计算机系统, 2017, 38(12):2652-2657.

[16] Heung I S, Seong W L, Dinggang S, et al. Deep Ensemble Learning of Sparse Regression Models for Brain Disease Diagnosis[J]. Medical Image Analysis, 2017, 37(11): 39–51.

[17] 霍东雪, 刘辉, 尚振宏, 等. 一种异构集成学习的儿科疾病诊断方法研究 [J]. 计算机应用与软件, 2018, 35(06):54-57+157.

[18] Machine learning modules in Azure Machine Learning Studio (classic)[EB/OL]. [2020-12-30].

[19] 梁潇. 互信息理论及其在战略情报研究中的应用 [J]. 现代情报，2007，27(12)：5-8.

[20] 李占山，吕艾娜. 基于新冗余度的特征选择方法 [J]. 东北大学学报：自然科学版，2020，41(11)：1550-1556.

[21] Slade E, Naylor M G. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations[J]. Statistics in Medicine, 2020, 39(8):1156-1166.

[22] Resche-Rigon M, White I R. Multiple imputation by chained equations for systematically and sporadically missing multilevel data[J]. Statistical Methods in Medical Research, 2018, 27(6):1634-1649.

[23] 韩建军，南少伟，郭呈周，李建平. 基于随机森林的粮仓气密性评价模型 [J]. 现代食品, 2018：187-190.

[24] 崔力娟. 快速支持向量机算法研究 [D].2017

[25] 苗丰顺, 李岩，高岑，王美吉，李冬梅. 基于 CatBoost 算法的糖尿病预测方法 [J]. 计算机系统应用，2019，28（9）：215-218

[26] 张春富. 基于参数优化的机器学习算法在糖尿病预测中的应用 [D].2020

[27] LI H,ZHU Y．Improving XGBoost Based on Gradient Distribu- tion Regulation Strategy[J]．Journal of Computer Applications,2020(1):1-6.

[28] CHEN Z Y,LIU J B,LI C,et al．Ultra Short-term Power Load Forecasting Based on Combined LSTM and XGBoost Model[J]．Power System Technology,2020(2):1-8.

[29] LIU Z X,WANG X．Flight Delay Prediction Based on Random Forest Regression[J]．Modern Computer,2019(15):20-24.

[30] ZHANG C F,WANG S,WU Y D,et al．Diabetes Risk Predic- tion Based on $GA\_XGBoost$ Model[J]．Computer Engineering, 2020(3):315-320.