doi:10.11835/j.issn.1000-582X.2023.02.010

基于 XGBoost 模型的营养成分分析高血压预测方案

蒋淮奘[1],谭　浪[2],李时杰[1],刘　昱[1],王峻峰[3]
(1.天津大学 微电子学院,天津 300072; 2.北京智芯微电子科技有限公司,北京 102200;
3.云南省第一人民医院,昆明 650031)

摘要:高血压是一种常见的慢性病,若能早发现、早采取措施可降低其引发并发症的风险。尽管高血压的产生与发展与诸多因素有关,但饮食被公认为影响高血压的主要因素之一。机器学习模型可以对疾病进行有效预测,并提供辅助治疗。笔者提出一种基于 XGBoost 的通过分析营养成分预测高血压的方案,该方案由数据转换、特征选择、数据清理与标准化、模型搭建、分类与评估 5 部分组成。实验结果表明,XGBoost 在高血压预测中获得了 0.859 的 F1 分数且准确率超过 85%,高于随机森林、支持向量机与人工神经网络。此外通过分析不同营养成分对高血压预测的影响因素,获得了影响高血压的前 10 个营养特征,大部分与医学结论相同,验证了模型的有效性。

关键词:机器学习;高血压预测;营养成分分析;极端梯度提升
中图分类号:TP181; P315.69　　**文献标志码**:A　　**文章编号**:1000-582X(2023)02-119-011

Hypertension predicting scheme by analyzing nutritional ingredients based on xgboost model

JIANG Huaiyan[1], *TAN Lang*[2], *LI Shijie*[1], *LIU Yu*[1], *WANG Junfeng*[3]
(1. School of Microelectronics, Tianjin University, Tianjin 300072, P. R. China; 2. Beijing Smartchip Microelectronics Technology Company Limited, Beijing 100000, P. R. China; 3. The First People's Hospital of Yunnan Province, Kunming 650031, P. R. China)

Abstract: Hypertension is a common chronic disease. Early detection and early measures can reduce the risk of complications. Although the occurrence and development of hypertension are related to many factors, diet is recognized as one of the main factors affecting hypertension. Machine learning models can effectively predict the disease and provide adjuvant therapy. Accordingly, this paper proposes a scheme based on XGBoost to predict hypertension by analyzing nutritional ingredients. The scheme consists of five parts: data conversion, feature selection, data cleaning and standardization, model building, classification and evaluation. The experimental results show that XGBoost obtains an F1_score of 0.859 in the prediction of and the accuracy rate exceeds 85%, which are higher than random forests, support vector machine and

收稿日期:2020-07-09　　**网络出版日期**:2021-12-22
基金项目:国家自然科学基金资助项目(61771338);云南省重点研究资助项目(2018IB007);天津市科技计划项目重大专项资助项目(18ZXRHSY00190)。
Supported by National Natural Science Foundation of China(61771338), Key research plan of Yunnan Province (2018IB007) and Major Special Project of Tianjin Science and Technology Plan(18ZXRHSY00190).
作者简介:蒋淮奘(1994—),女,硕士研究生,主要从事机器学习方向研究,(E-mail)jianghyan@tju.edu.cn。
通信作者:王峻峰,男,云南省第一人民医院,主任医师,主要从事临床医学方向研究,(E-mail) 18987173605@qq.com。

artificial neural network. By analyzing the influence factors of different nutritional ingredients on the prediction of hypertension, we obtain the top 10 nutritional characteristics that affect hypertension, most of which are the same as medical conclusions, verifying the effectiveness of the model.

Keywords: machine learning; hypertension prediction; nutritional ingredients analysis; Extreme Gradient Boosting

高血压是一种严重威胁人类健康的慢性病,根据美国高血压控制委员会制定的标准[1],反复测量的收缩压超过 14 mmHg 或舒张压超过 90 mmHg 可认定为高血压。英国权威杂志《柳叶刀》(The Lancet)2017 年的研究显示[2],1975 年全球高血压患病人数为 6×10^9,到 2015 为止增加至 11×10^9,患病人数几乎翻了一倍;在世界范围内每年有 750×10^4 人死于高血压或由其引发的并发症。中国"十二五"高血压抽样调查结果显示,2017 年中国有 2.45×10^9 的成年人为高血压患者,占成年人比例的 23.2%;有 1.25×10^9 人不知道自己是否患有高血压,此人数超过患者人数的一半;此外 1.5×10^9 的患者未使用药物进行治疗,只有约 $3\ 700\times10^4$ 的高血压患者得到了控制;而处在高血压的边缘人数也达到了 4.35×10^9[3]。目前高血压在中国呈现低知晓率、低治疗率、低控制率的形势。

影响血压状况的因素有很多,如性别、年龄、吸烟、肥胖以及不健康的饮食等,有诸多研究在这方面进行探索[4-6],结果显示不良的饮食是高血压形成与发展的重要影响因素。日常生活中饮食与人密切相关,合理的饮食可以促进身心健康和预防疾病,而饮食可以理解为营养成分摄入,因此不同种类和数量的营养成分摄入会影响疾病的发生以及人们的健康状况。研究也证实了饮食营养与血压值存在一定关系,如高血压患者的血压与膳食中钠摄入成正相关[7];高血压患者血浆中的总包和脂肪酸含量较正常人更高,脂肪与血压成正相关[8];服用维生素 A、C、E 能降低高血压患者的血压尤其是收缩压[9]。

高血压早期因无明显症状而不易被发现,很难引起重视,若能及早发现问题,通过合理饮食及相关医疗措施,可有效控制并避免后期引起并发症。饮食在高血压的发生及发展中都起到很大作用,所以建立一种通过分析营养成分来预测高血压的模型十分必要。近年来,有学者在高血压的风险因素分析以及预测方面进行了研究,但通过饮食营养来预测高血压的研究较少,方法体系还不成熟。如 Dong 等[10]通过改进的反向传播神经网络算法研究了高血压的影响因素,包括遗传因素,生活方式因素,肥胖和合理饮食。Sinkuo chai 等[11]基于数据挖掘技术建立了高血压并发症的预测模型。张伟等[12]提出了一种改进的 C4.5 决策树算法,通过使用住院患者的医疗相关数据来预测高血压,最终获得了 81.58%的准确率。Nimmala 等[13]通过基于 AAA 的 J48 分类器使用年龄、愤怒和焦虑程度来预测高血压,获得了 84.30%的准确率。以上研究成果对高血压的发生机制进行了深入探讨,但研究数据主要为影响高血压的一般特征,且使用的分析预测模型较为单一,对比性不强。

因此,以营养成分为主要特征,以年龄、身形体态等一般特征为辅助特征,结合机器学习、统计学习等相关技术提出了一个高血压预测的五阶段方案,并搭建了基于 XGBoost 的分析营养成分预测高血压模型,结果显示所提出的预测模型具有较高准确率、精确率、召回率与 F1 分数。此外还针对高血压预测中不同营养特征的影响因子完成风险分析,分析结果可以帮助医生以及患者提早发现问题,采取措施或进行治疗,降低医疗成本并提高患者生存率。

1 基本原理及方法

通过对问题进行分析和解构,笔者要实现高血压预测需要经过以下步骤:1)需要将人的饮食数据转换为所需要的营养成分数据,并筛选出有利于模型预测的一般特征;2)处理得到的营养成分和一般特征数据会伴随着缺失等问题,需要对数据进行清理;3)分类模型可分为二元分类和多元分类模型,通过分析人的每天营养成分摄入以及相关特征来预测高血压为二元分类任务,需要搭建相应的二元分类模型来实现预测。据研究提出了一个 5 阶段方案,具体流程如图 1 所示。

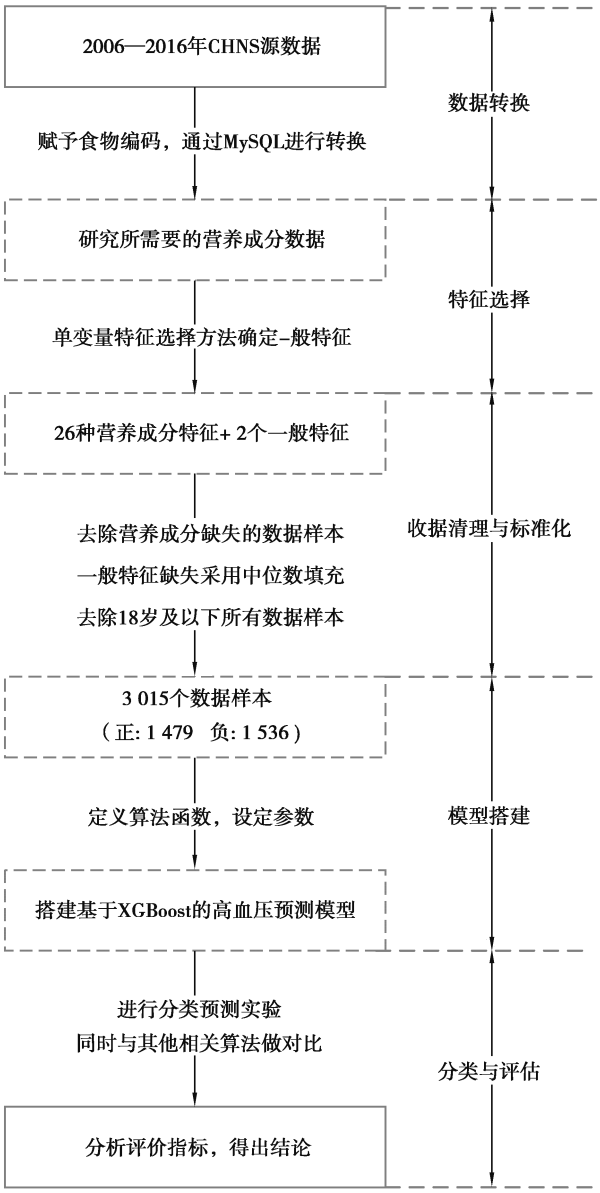

图 1　预测高血压五阶段方案流程图

Fig. 1　Flow chart of five stage scheme for predicting hypertension

1.1　实验数据来源

本次实验的数据来自于中国营养与健康调查(CHNS，china health and nutrition survey)，该项目由中美合作，从 20 世纪 80 年代起对中国多个地区居民的饮食结构和营养状况等变化进行追踪研究，至今共进行了 10 次调查，其所有研究调查数据面向公众开放，详细信息请参见[14]。CHNS 调查时间跨度较大，因此实验只选择从 2006 年开始的最近 3 次调查的数据进行分析。CHNS 数据并未直接提供研究所需的营养成分数据，而是记录了调查期间每个人食用的食物名称及重量，以及相应的身高、体重、臀围、头围等基本信息和每个人的血压值。因此对调查数据做出以下处理：

CHNS 所提供的食物数据包含食物消耗量和相应的食物代码，中国疾病预防控制中心营养与食品安全研究所发布的中国食物成分表包含了每种食物的食物代码，以及每 100g 食物所含有的 26 种营养成分的数量[15]，研究使用 MySQL 搭建数据库建立 2 个字段之间的关系，将 CHNS 食物数据转换为对应的 26 种营养成分数据。

CHNS 同时提供了被调查者的血压值检测数据，据统计约有 60％的人进行了连续 3 次测量，30％进行了

2次测量,10%左右的人只有一次测量结果或者没有记录数据。高血压诊断需要进行多次反复测量,因此只选择了包含3次的测量结果数据。之后对高压与低压分别取平均值,若高压≥140 mmHg或者低压≥90 mmHg,则视为高血压,标记为01作为正样本;否则为非高血压,标记为10作为负样本。

1.2 特征选择

通过前面处理办法,将饮食数据转换为所需的营养成分数据,并首先选择了这26种营养数据作为预测模型的主要特征。考虑到不同年龄段人们的饮食习惯和结构不同,如年轻人由于工作和其他原因更倾向于食用高碳水化合物和高能量的食物,总体摄入量相对较高。老年人则倾向于食用低碳水化合物、高纤维的食物。此外,不同身形的人的饮食摄入也有差异[16]。因此,希望将年龄和身形体态作为预测模型的辅助特征,从而使预测结果更加合理准确。CHNS数据库中给出了每个调查者的身高和体重,通过身高和体重可以计算出每个人的身体质量指数(BMI, body mass index)来表示个人的身形。

为了更好地验证上述想法,需要使用特征选择方法从全部特征中剔除不相关或者冗余的特征来减少特征个数,从而提高模型精确度及减少运行时间,使构建出来的模型更好。本实验为二分类问题,故采用单变量特征选择方法验证年龄与BMI值是否可作为本分类实验的特征。单变量特征选择方法有4种,选择了适合二元分类任务的SelectKBest方法,SelectKBest中的score_func参数选择f_classif,它会计算单变量与训练目标之间的方差分析F值(Anova F-value),F值越大,说明特征影响分类结果越大。选择了CHNS数据库中提供的基本信息,包括参与者的性别、出生年份、上臂围、三头肌皮褶、臀围、腰围以及要验证的年龄和BMI作为变量特征进行验证。最终将输出结果由高到低排序,如表2所示。

表2 SelectKBest特征选择结果

Table 2 SelectKBest feature selection results

特征名称	F值
年龄	374.47
BMI	145.97
腰围	28.35
臀围	26.24
三头肌皮褶	12.23
出生年份	6.52
上臂围	2.23
性别	0.35

从表中可知,年龄与BMI的F值分别为374.47、145.97,明显高于其他特征的F值,说明年龄与BMI可以作为预测高血压分类模型的特征,而性别、臀围等基本信息的F值过低,则直接剔除。最终,预测高血压分类模型选择26种营养成分数据以及年龄与BMI共计28维,作为输入特征。

1.3 数据清理与标准化

在机器学习领域中获得的原始数据通常伴有缺失值,即数据集中某些特征属性的值不完全。为了保证数据完整性,利于模型准确预测,需要判断缺失值的类型并完成填充。机器学习中常用的处理缺失值的方法有人工填写、特殊值填写、均值填充、中位数填充、多重插补等。由于营养数据特征是通过饮食记录转换而来,因此若饮食记录有缺失,数据本身的性质无法使用上述方法进行填充,所以这一部分缺失数据直接删除。年龄和BMI 2个特征本身缺失值比例小于5%,这一部分缺失值对整体模型预测影响不大,故使用中位数填充进行替换。

研究显示,18岁及以下未成年人的血压会随着年龄、身高的增长以及体重的增加在标准范围内升高[17],若非家族遗传,很少患有高血压,所以这一部分数据不具有代表性,为了更好地评估模型准确性,删除了18岁

及以下未成年人的数据。通过整个数据清理过程，最终得到了包含 28 个特征的 1 582 个数据样本，包括 826 个患高血压的正样本与 756 个未患病的负样本，比例接近 1∶1。

由于输入特征主要是每日营养成分摄入量，种类繁多且单位不同，同时某些营养特征的总体方差过大，可能会导致一些机器学习算法的主目标函数阻止参数估计其学习其他特征，造成很难收敛或不能收敛的状况。数据标准化是将数据按比例缩放，使之落入一个小的特定区间，可以将其转化为无量纲的纯数值来去除数据的单位限制，便于不同单位或量级的指标能够进行比较和加权。因此对数据集进行了标准化处理，使每个特征值的平均值为 0，方差为 1，相当于转化成为标准正态分布即高斯分布。标准化的公式如下

$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

其中：x_i 指的是数据集中特定维度的所有数据；$\min(x)$是数据集中同一维度的最小数据；$\max(x)$是数据集中同一维度的最大数据；x'表示标准化数据的值。

1.4　模型搭建

XGBoost(extreme gradient boosting)又称极端梯度提升，由 Chen 等于 2014 年开发和推出[18]，并且在近年来的 Kaggle 比赛中取得非常突出的表现。XGBoost 是基于梯度提升决策树(GBDT)的改进算法，通过 boosting 思想将个体学习器组合在一起，产生依赖关系，同时可以有效构建提升树且并行运行。XGBoost 算法因其运算快速、高效准确、泛化能力强等优点广泛应用于分类与回归领域。其核心概念是通过添加树，拟合最后预测的残差来学习新功能，然后获得样本得分，通过将每棵树的分数相加，可以得出样本的最终预测分数。对于具有 m 个特征的 n 个标记样本，使用 K 个加法函数预测分数的公式如下

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{2}$$

$$F = \{f(x) = w_{q(x)}(q: R^m \to T, w \in \boldsymbol{R}^{\mathrm{T}})\}, \tag{3}$$

其中：F 是回归树的空间；$f(x)$是其中一个回归树；$w_{q(x)}$表示每个 T 叶树的独立结构分数。XGBoost 的目标函数被定义为

$$L = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{4}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2, \tag{5}$$

其中：l 代表了模型的损失函数；Ω 是正则化项；T 表示叶节点的数量；w 是叶节点的分数；γ 与 λ 代表了防止过度拟合的控制系数。当生成第 n 棵树时，预测分数公式可以写成

$$\hat{y}_i(t) = \hat{y}_i(t-1) + f_t(x_i), \tag{6}$$

其中 $\hat{y}_i(t-1)$是前 $t-1$ 轮模型的预测分数，对应的目标函数可以改写为

$$L(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i(t-1) + f_t(x_i)) + \Omega(f_t), \tag{7}$$

为了加速优化，使用泰勒二阶展开式

$$L(t) = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i(t-1)) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_j^2, \tag{8}$$

通过添加样本的损失函数，重新组合样本，最后利用顶点公式求出最优的 w 以及目标函数公式 L 如下

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \tag{9}$$

$$L = -\frac{1}{2} \sum_{i=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T, \tag{10}$$

$$G_i = \sum\nolimits_{i \in I_j} g_i, \tag{11}$$

$$H_i = \sum\nolimits_{i \in I_j} h_i。 \tag{12}$$

XGBoost 在寻找最佳分割点时结合了传统的贪心算法以及近似算法，根据百分位法列举几个可能成为

分割点的候选项，然后根据式(9)、(10)计算出最佳分割点。XGBoost使用多种方法来避免过度拟合，例如引入正则化、行采样以及特征采样，同时还增加了对稀疏数据的处理。此外XGBoost还具有其他的优势，例如能进行并行处理，使速度有了很大提升；具有高度的灵活性，可自定义优化目标与评价标准；内置交叉验证，允许在每一轮boosting迭代中使用交叉验证。综合以上XGBoost在分类算法中的优势，选择了XGBoost作为通过分析营养成分预测高血压的模型。

通过分类算法搭建模型并最终实现高血压预测，需要通过定义算法函数、调用函数搭建网络模型、训练与验证模型、期间调整参数及最后测试与评估模型等步骤。

基于XGBoost的高血压预测模型的设计流程图如图2所示，具体流程如下：首先定义算法函数，调用XGBoost函数搭建网络模型；随后设定初始参数并输入训练集，进行模型训练，每训练一次调整一次权值，直到训练误差最小或达到要求的最高训练次数1 000次；训练后存储当前网络文件，输入验证集，对比评估指标来确定需要人工手动调整的参数是否最优，如此循环直到所有参数全部最优；随后进入测试阶段，对模型评估获得相应指标，完成通过分析营养成分预测高血压的分类实验。

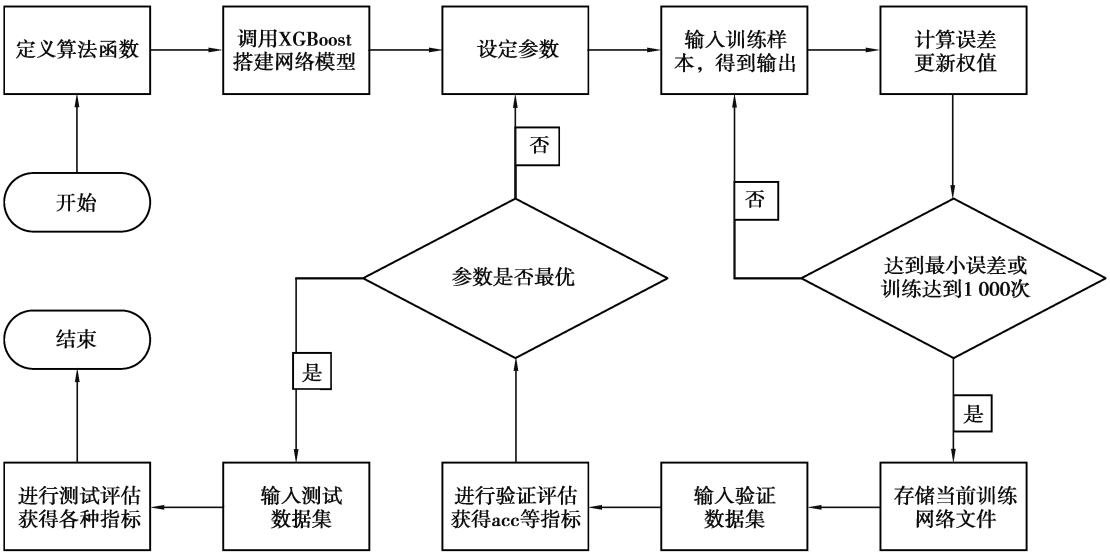

图2 XGBoost建模流程图

Fig. 2 XGBoost modeling flow chart

1.5 分类与评估

目前的研究中很少有通过分析营养成分来预测患高血压风险的模型，提出了通过分析营养成分来预测高血压的5阶段方案，搭建了基于XGBoost的高血压预测模型，同时与多种常见机器学习分类算法进行对比来验证模型的有效性。算法包括随机森林(RFs，random forest)、支持向量机(SVM，support vector machine)以及人工神经网络(ANN，artificial neural network)。RFs主要利用的是集成学习中的bagging算法，将多棵树集成到一同分类，树与树之间关系为并行，互不影响[19]。SVM作为一种机器学习的有监督分类方法是建立在统计学习理论的VC维理论和结构风险最小原理的基础上，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳平衡[20]，它在解决小样本、非线性以及高维模式识别中表现出诸多优势。ANN是一个局部最优解的分类和预测算法，由大量神经元相互连接而成，每个神经元节点都是一些动态的权重参数，ANN的学习过程是对大量样本进行归纳学习，然后内部进行自适应，过程中各个神经元节点调整相应权重，使神经网络处于稳定的范围且权重收敛[21]。这些对比算法的建模过程从形式上与XGBoost有相同部分，不同点在于SVM中有4种核函数，本实验中的特征个数远小于样本个数，故选择了径向基核函数；RFs衡量分裂质量的性能函数选择为entropy，即为信息增益的熵；ANN需要确定隐藏层以及各个层神经元个数，由于输入特征为28个，经过实验最终选择28-56-56-2的神经元结构，其中包括1个输入层、1个输出层和2个隐藏层。

在完成分类模型构建之后，需要对模型的效果进行评估，在二分类问题中评价模型最简单也最常用的是准确率(Accuracy)，但若数据集正负样本不均衡，准确率并不能很好地评估模型的有效性，因此引入精准率(Precision)、召回率(Recall)以及 F1 分数(F1_score)[22]，具体公式如下

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}, \tag{13}$$

$$\text{Precision}=\frac{TP}{TP+FP}, \tag{14}$$

$$\text{Recall}=\frac{TP}{TP+FN}, \tag{15}$$

$$\text{F1\_score}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}。 \tag{16}$$

其中：TP 为真正类，即实际是正类且预测也为正类的个数；FP 为假正类，即实际是负类但预测为正类的个数；TN 为真负类，即实际是负类且预测也为负类的个数；FN 为假负类，即实际是正类但预测为负类的个数。

除了以上指标，还使用 ROC(receiver operating characteristic)曲线，是反应特异性与灵敏度这 2 个连续变量的综合指标，它以平面曲线图的形式来全面且客观地对模型以及系统进行分析和评估[23]。ROC 曲线以真正类率(TPR，true positive rate)为 y 轴，以假正类率(FPR，false positive rate)为 x 轴，x、y 两轴取值范围均为 0～1。当 FPR 值保持不变时，TPR 值越大，曲线越接近(0,1)点，表明模型越好越稳定。ROC 曲线下面积即 AUC(area under roc curve)值，用来直观地反应 ROC 曲线图的情况，AUC 值越接近 1 则代表 ROC 曲线越接近(0,1)点，说明模型的分类性能越好。

以上为提出的高血压预测的五阶段方案，通过数据转换得到了所需要的营养成分数据；通过特征选择确定了以 26 种营养成分为主要特征，年龄与 BMI 为辅助特征的 28 维特征；通过数据清理与标准化去除了数据冗余，提高了模型运算速度和准确度；在模型搭建中选择了运算速度快、准确率高、泛化能力强的 XGBoost 模型；最终通过与其他分类模型一同比较评估来验证提出的方法与模型的有效性。

2　实验结果及分析

2.1　XGBoost 与其他分类算法结果对比

以通过分析人的日常营养成分摄入以及一般信息来预测其患高血压的风险为目标，提出了预测高血压的五阶段方案，搭建了基于 XGBoost 的预测模型，并在方案的最后一步与随机森林(RFs)、支持向量机(SVM)与人工神经网络(ANN)这 3 种分类算法作对比，对比结果使用的是测试集的数据。经过数据清理，最终得到 1 582 个数据样本，将其平均分成 5 份，随机取其中一份作为测试集数据，其他为训练集数据。测试数据共有 316 个数据样本，包括 142 个患高血压的正样本，174 个未患病的负样本。评价指标包括准确率(Accuracy)，召回率(Recall)，精准率(Precision)以及 F1 分数(F1_score)。具体结果如表 3 所示。

表 3　4 种分类算法结果对比

Table 3　Comparison of the results of the four classification algorithms

分类算法	评估指标			
	Recall	Precision	F1_score	Accuracy
RFs	0.873	0.756	0.815	0.816
SVM	0.887	0.663	0.759	0.721
ANN	0.838	0.704	0.765	0.769
XGBoost	0.926	0.802	0.859	0.857

从表中可以看出，包括XGBoost在内的4种分类算法的测试准确率与F1分数均超过0.72，其中准确率最低为0.721，最高为0.857；F1分数最低为0.759，最高为0.859。由SVM和ANN这2个分类算法得到的准确率与F1分数均低于0.80，说明这2种算法解决本分类问题的能力较差；RFs与XGBoost的准确率与F1分数均超过0.80，性能较好。其中与RFs相比，XGBoost获得了0.857的最高的准确率与0.859的最高的F1分数，同时召回率为0.926，精确率为0.804也为最高。因此，综合以上指标可以得出，XGBoost为通过分析营养成分预测高血压的最佳模型。

由于实验测试所用试数据集正负样本并不是完全的1∶1，为了更好地比较这4个分类算法的性能，画出了测试过程中4个分类算法的ROC曲线图，并将其组合到了一张图中，具体如图3所示。其中不同的颜色代表不同分类算法的ROC曲线，蓝色点划线为XGBoost，绿色点划线为RFs，紫色点划线为SVM，红色点划线为ANN。从图中可以看出XGBoost的ROC曲线更接近(0,1)点，分类效果最好，而ANN的ROC曲线离(0,1)点最远，分类效果最差。此外还可以通过对比ROC曲线下面积即AUC值，来更加直观地看出算法分类的效果，具体如表4所示。

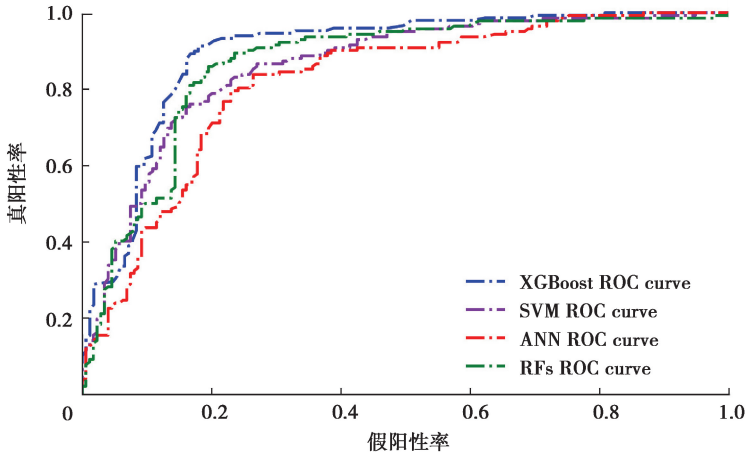

图3 4种分类算法的ROC曲线图

Fig. 3 ROC curve of four classification algorithms

表4 4种分类算法AUC值

Table 4 AUC values of four classification algorithms

分类算法	AUC
RFs	0.864
SVM	0.859
ANN	0.817
XGBoost	0.903

从表中得知XGBoost的AUC值最高且超过了0.9，而其他各个算法的AUC值均未超过0.9。通过观察这4种分类算法的ROC曲线与AUC值得出的结论与表3中评估指标得出的结论相同。搭建的基于XGBoost的通过分析营养成分预测高血压的模型，拥有较高的准确率、精确率、召回率与F1分数，分类效果好，稳定性强。

由于XGBoost中参数较多，借助网络搜索(GridSearchCV)方法来优化XGBoost中需要手动调整的参数。最终XGBoost模型达到最佳效果时的最佳参数如表5所示。

表 5　XGBoost 的最佳参数

Table 5　Best parameters of XGBoost

编号	参数名称	默认值	最佳参数
1	learning_rate	0.3	0.01
2	n_estimators	10	1 000
3	max_depth	6	3
4	min_child_weight	1	5
5	gamma	0	0.3
6	subsample	1	0.75
7	colsample_bytree	1	0.85

表中共展示了 XGBoost 的 7 种可变参数，每个参数都有不同的含义，其中 learning_rate 为算法的学习率，控制每次迭代更新权重的步长，默认值为 0.3，选取的最佳参数为 0.01；n_estimators 为总的迭代次数即基础学习器的个数，通常以树的形式存在，默认值为 10，选择的最佳参数为 1000；max_depth 代表树的最大深度，默认值为 6，典型值为 3～10，值越大越容易过拟合，选取的最佳参数为 3；min_child_weight 是最小叶子权重，默认值为 1，典型值为 2～10，值越小越容易过拟合，选取的最佳参数为 5；gamma 为惩罚项系数，是指定节点分裂所需的最小损失函数下降值，默认值为 0，选取的最佳参数为 0.3；subsample 表示用于训练模型的子样本占整个样本集合的比例，默认值为 1，取值范围为 0～1，选择适当比例可防止过拟合，选择的最佳参数为 0.75；colsample_bytree 代表用于训练模型的特征占全部特征的比例，默认值为 1，取值范围为 0～1，本文选择的最佳参数为 0.85。

2.2　特征分析

通过算法进行分类最重要的部分是用于进行预测的特征，某些特征在预测中 起到非常重要的作用，因此为了进一步探索营养成分与高血压之间的关系，以及验证模型的有效性与稳定性，引用了特征重要性分析。在前面的实验中 XGBoost 模型验证为营养成分预测高血压的最佳模型，而 XGBoost 可以根据结构分数的增益作为某个特征的分割点，特征的重要性得分可以用特征在所有树中被调用出现次数的总和表示。在调参最优的 XGBoost 中，根据特征重要性排序，最终获得了影响高血压分类的前 12 个特征，具体如表 6 所示。

表 6　影响高血压分类的前 12 个特征

Table 6　The top 12 features that affect the classification of hypertension

特征	特征重要性得分
Age	328
BMI	198
Fat	95
VC	85
Fe	72
Na	58
Mg	51
CHO	49

续表6

特征	特征重要性得分
VE	42
Ca	38
VB2	32
CHOL	26

表中特征重要性得分从高到低排序，可以看出影响高血压分类的前2个因素是年龄和BMI值，它们的特征重要性得分分别为301和225，不同年龄和不同身形的人的饮食结构不同，也就是说年龄和BMI值影响着其他营养成分特征，且随着年龄增长和体重增加，高血压的患病率会逐渐上升，所以年龄与BMI值处于前2个位置是合理的。排名在3～12的营养成分特征分别是：脂肪(Fat)、维生素C(VC)、铁(Fe)、钠(Na)、镁(Mg)、碳水化合物(CHO)、维生素E(VE)、钙(Ca)、维生素B2(VB2)、胆固醇(CHOL)。

文献[7-8]中指出膳食中钠的摄入与脂肪的摄入与人的血压成正相关，文献[9]中指出服用维生素C对降低高血压患者的血压值具有一定的作用，而钠、脂肪与维生素C在高血压预测模型中的营养成分特征重要性得分中也排在前5位。文献[24]指出对抑制高血压有积极影响的营养成分有镁、钙、钾和膳食纤维，对抑制高血压有消极影响的营养成分有钠和碳水化合物，其中镁、钙、钠以及碳水化合物这4种，也位于XGBoost模型获得的影响高血压分类的营养成分的前10位。

3 结 论

对高血压的预测问题展开了深入研究，在现有的几种以医疗相关指标与一般信息为主要特征的高血压预测方法的基础上，针对膳食营养与血压值之间的联系，提出了一种通过分析营养成分预测高血压的五阶段方案，搭建了基于XGBoost的高血压预测模型，通过分析个人日常摄入的营养成分信息以及年龄与BMI来预测其是否患高血压。从实验结果来看提出的方法获得了85.7%的准确率以及0.859的F1分数，相比其他分类算法均为最高，验证了提出的高血压预测五阶段方案的可行性；通过特征重要性分析，获得了影响高血压的前10个营养成分，对比各类文献可知，钠、脂肪、维生素C、镁、钙以及碳水化合物对高血压的影响与现有医学结论相同，从而验证了模型的有效性。

参考文献：

[1] Muntner P, Krousel-Wood M, Hyre A D,et al. Antihypertensive prescriptions for newly treated patients before and after the main antihypertensive and lipid-lowering treatment to prevent heart attack trial results and seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure guidelines[J]. Hypertension, 2009, 53(4): 617-623.

[2] Zhou B, Bentham J, Cesare M D, et al. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants[J]. Lancet, 2017, 389(10064): 37-55.

[3] Xing L, Liu S, Tian Y, et al. Trends in status of hypertension in rural northeast China: results from two representative cross-sectional surveys, 2013-2018[J]. Journal of Hypertension, 2019, 37(8):1.

[4] Channanath A M, Farran B, Behbehani K, et al. Impact of hypertension on the association of BMI with risk and age at onset of type 2 diabetes mellitus: age- and gender-mediated modifications[J]. Plos One, 2014, 9(4):e95308.

[5] Maimaris W, Paty J, Perel P, et al. The influence of health systems on hypertension awareness, treatment, and control: a systematic literature review[J]. Plos Medicine, 2013, 10(7):e1001490.

[6] Millett C, Agrawal S, Sullivan R, et al. Associations between active travel to work and overweight, hypertension, and diabetes in India: a cross-sectional study[J]. PLoS Medicine, 2013, 10(6), 1001459.

[7] Zanchetti A. The 2003 guidelines for the management of hypertension of the european society of hypertension and european society of cardiology[C]// Comprehensive Hypertension.New York:Elsevier, 2007:1177-1184

[8] Houston M C. Nutraceuticals, vitamins, antioxidants, and minerals in the prevention and treatment of hypertension[J]. Progress in Cardiovascular Diseases, 2005, 47(6):396-449.

[9] 袁林，李培. 高血压病和脑卒中患者内源性 VitC，E 的测定及其意义[J]. 心血管康复医学杂志，2001，10(1)：5-6.
Yuan L, Li P. Measurement and significance of endogenous VitC and E in patients with hypertension and stroke[J]. Journal of Cardiovascular Rehabilitation Medicine, 2001,10(1): 5-6.(in Chinese)

[10] Dong X. Study on the causes of hypertension with improved BP neural network[C]// 2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT). Shenzhen, China: IEEE, 2010, 1: 21-24.

[11] Chai S, WU LYU Y I, Chang S T, et al. Establish a predictive model of hypertension complications[C] // 2018 International Conference on Machine Learning and Cybernetics (ICMLC). Chengdu, China: IEEE, 2018, 2: 515-520.

[12] Wei Z, Xuan Z, Junjie C. Study on classification rules of hypertension based on decision tree[C] // 2013 IEEE 4th International Conference on Software Engineering and Service Science. Beijing, China: IEEE, 2013: 93-96.

[13] Nimmala S, Ramadevi Y, Sahith R, et al. High blood pressure prediction based on AAA++ using machine-learning algorithms[J]. Cogent Engineering, 2018, 5(1):1-12.

[14] China Health and Nutrition Survey. Survey data online available[EB/OL]. https://www.cpc.unc.edu/projects/china/

[15] 杨月欣，王光亚. 中国食物成分表[M]. 北京：北京大学医学出版社，2002.
Yang Y X, Wang G Y. Chinese food ingredient list[M]. Beijing: Peking University Medical Press, 2002.(in Chinese)

[16] Oleniuc F C, Buliga D M. The impact of eating behaviour and food preferences on nutritional status[C]// 2013 E-Health and Bioengineering Conference (EHB).Romania, Iasi: IEEE,2013: 1-4.

[17] 李凯，耿贯一. 儿童血压变化及其影响因素[J]. 中国公共卫生，1997(1):17-18.
Li K, Geng G Y. Changes of children's blood pressure and its influencing factors[J]. China Public Health, 1997(1): 17-18.(in Chinese)

[18] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015, 1(4): 1-4.

[19] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.

[20] Ertekin S, Bottou L G, C. Nonconvex online support vector machines[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(2):368-381.

[21] Khalil Alsmadi M, Omar K B, Noah S A, et al. Performance comparison of multi-layer perceptron (back propagation, delta rule and perceptron) algorithms in neural networks[C] // 2009 IEEE International Advance Computing Conference. Patiala, India:IEEE, 2009: 296-299.

[22] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation[J]. International Journal of Radiation Biology & Related Studies in Physics Chemistry & Medicine, 2005, 51(5):952-952.

[23] Yang F, Lv J H, Lei S F, et al. Receiver-operating characteristic analyses of body mass index, waist circumference and waist-to-hip ratio for obesity: Screening in young adults in central south of China[J]. Clinical Nutrition, 2006, 25(6): 1030-1039.

[24] Lei Z, Yang S, Liu H, et al. Mining of nutritional ingredients in food for disease analysis[J]. IEEE Access, 2018, 6: 52766-52778.

（编辑　侯　湘）