文章编号：2095-2163(2020)09-0109-04 中图分类号：TP391.7 文献标志码：A

基于 Xgboost 和 Keras 框架的多疾病风险预测

黄 旭[1]，贺 松[2]，席欢欢[1]，张 硕[1]，张 慧[1]
（1 贵州大学 大数据与信息工程学院，贵阳 550025；2 贵州大学 医学院，贵阳 550025）

摘 要：慢性病已成为人类健康的主要威胁，找出慢性病发生直接的或间接的因素，做好疾病风险预测有着重要意义。本文采用 3 种集成学习算法 RF、GBDT 和 Xgboost 对 3 种慢性病进行分类，采用分类效果最好的 Xgboost 进行特征选择，使用 Keras 深度学习框架构建神经网络进行多疾病风险预测，采用问题转化中 BR 和 LP 二种方法将多疾病风险预测转化为多标签分类问题。

关键词：慢性病；特征选择；Xgboost；多标签分类；Keras

Multi-disease risk prediction based on Xgboost and Keras frameworks

HUANG Xu[1], HE Song[2], XI Huanhuan[1], ZHANG Shuo[1], ZHANG Hui[1]
(1 College of Big Data and Information Engineering, Guiyang 550025, China;
2 College of Medical, Guizhou University, Guiyang 550025, China)

【Abstract】 Chronic diseases have become a major threat to human health. It is important to find out the direct or indirect factors of chronic diseases to predict the risk of multiple diseases. In this paper, RF, GBDT and Xgboost are used to classify the three chronic diseases, and the Xgboost with the best classification effect is used for feature selection. The Keras deep learning framework is used to construct a neural network to predict the risk of multiple diseases. The multiple disease risk prediction is transformed into multi label classification problem by using Binary Relevance (BR) and Label Powerset (LP) in problem transformation.

【Key words】 chronic disease; feature selection; Xgboost; multi-label classification; Keras

0 引 言

随着计算机技术广泛应用于医疗领域，医院信息化建设得到快速发展。医院信息系统中存储了大量的数据资源，运用数据挖掘技术可以有效的分析、整合和利用这些数据，以达到辅助诊疗的目的。利用数据挖掘技术对疾病风险预测，对疾病的管理、预防、干预等有着重要意义。

慢性病非传染性疾病通常具有发病潜伏期长、病因复杂、反复发作、难以彻底治愈的特点，导致的负担占总疾病负担的 70% 以上，成为制约健康预期，寿命提高的重要因素[1]。因此，在慢性病还没有显现时，就应当做好疾病的预测，提前发现，及时处理，将疾病的影响降到最低。利用数据挖掘技术，找出慢性病发生直接或者间接的危险因素，就能够通过监测这些因素，及时采取预防措施，从而降低发病率。

1 理论方法

1.1 分类算法

集成学习本质上是通过学习并结合多个弱分类器来获得比单一分类器优越的泛化性能[2]。RF、GBDT 和 Xgboost 的弱学习器都是树模型，利用多棵决策树对样本数据训练、分类和预测。在对数据分类时，给出各特征的重要性得分，评价各特征在分类中的作用。

随机森林（Ramdom Forest，RF）是一种基于 Bagging 算法改进的模型，做分类时，各弱分类器之间无强依赖关系，相互独立，每一个弱分类器都有一个分类结果，最后根据森林内决策树投票，按照少数服从多数的原则，对最终结果进行判定。随机森林具有二个特点：数据随机和特征随机。数据随机指的是对训练集有放回的采样，这样不同的树用到的训练集就会有所差异；特征随机指的是每次从所有特征中随机选择特征子集进行划分，可以增强数据的适应能力，优化高维特征的训练速度。

梯度提升决策树（Gradient Boosting Decision TREE，GBDT）是一种将弱学习器限定为分类回归树（CART）模型的前向分布算法，是 boosting 算法中的一种，由多棵决策树组成，每一棵决策树模型的建立

基金项目：贵州省数字健康管理工程技术研究中心项目（黔科合 G 字[2014]4002 号）。

作者简介：黄 旭（1995-），男，硕士研究生，主要研究方向：医学信息、数据挖掘；贺 松（1974-），男，硕士，副教授，硕士生导师，主要研究方向：医疗大数据；席欢欢（1994-），男，硕士研究生，主要研究方向：医学图像处理；张 硕（1993-），男，硕士研究生，主要研究方向：计算机应用与网络安全；张 慧（1994-），女，硕士研究生，主要研究方向：医疗大数据、数据分析。

收稿日期：2020-06-08

是为了不断学习之前树的结果,拟合残差,最后将所有决策树的预测值结合起来得到最终答案。GBDT 可处理离散或连续型数据,进行少量参数调优,可以达到很好的预测效果,具有较快的运算速度和较强的泛化能力。

Xgboost(eXtreme Gradient Boosting)也称为极端梯度提升算法,是在 GBDT 基础上的改进,其特点是模型能自动利用 CPU 进行多线程并行计算,提高运算速度,并且对损失函数进行泰勒公式二阶展开,使得预测精度更高。在损失函数后面增加正则项,可以约束损失函数的下降和模型整体的复杂度[3]。Xgboost 的目标函数 Obj 为式(1):

$$Obj = \sum_{i=1}^{m} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (1)$$

其中,第一项代表传统损失函数,衡量真实标签 y_i 与预测值 $\hat{y}_i$ 之间的差异,第二项代表模型的复杂度,引入正则惩罚项,整体求最优解,实现模型表现和运算速度的平衡。

1.2 超参数优化

网格搜索参数寻优法是一种最基本的参数优化算法。其核心思想是利用穷举搜索,遍历设定参数范围内所有的值,并以验证系统中的评分结果作为指标,得到最优参数。该算法是对参数的每一组情况进行试算,因此,当网格划分的比较密集时,每多一个参数,计算量就会呈几何倍增张,网格搜索寻优法就会非常耗时。

随机搜索区别于网格搜索的暴力搜索方式,采用随机在参数空间中采样的方式,只要随机次数够多,总能找到最优或者近优参数,尽管每次随机结果不一致,但大大提高了高维参数的寻优速度。

1.3 神经网络

神经网络由输入层、隐藏层和输出层构成,如图 1 所示。隐藏层的层数是任意的,夹在输入层和输出层之间,每层由神经元组成,输入层由训练集的实例特征向量传入,数据集有几个特征输入层就有几个神经元,本文对 3 种疾病进行二分类,所以输出层是 3 个神经元。对于每一个神经元模型,当前神经元的输出 y_j 都与上一层神经元 x_i 有式(2)的关系:

$$y_j = f\left(\sum_i W_{ij} x_i + \theta_j\right). \qquad (2)$$

其中,W_{ij} 为神经元 j 与上一层神经元 i 的权重;θ_j 为神经元的偏向;f 为激活函数,引入激活函数增强了神经网络的表达能力。

1.4 Keras 框架

Keras 是一个高层神经网络 API,能够使用 Tensorflow 和 Theano 任一平台作为后端,快速完成深度学习的开发。用户的体验始终是 Keras 考虑的首要内容,具有易使用、可抽象、兼容性和灵活性的特点[4]。Keras 具有许多模块,网络层、损失函数、优化器、参数初始化、激活函数、正则化方法都是独立的模块,可以使用这些模块来构建自己的模型,大大加快了建模速度[5]。

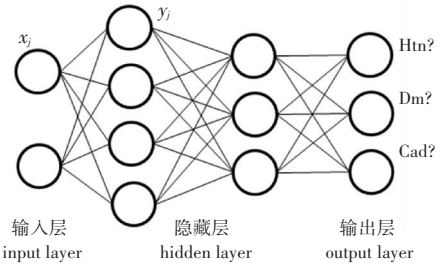

图 1 神经网络

Fig. 1 Neural network

2 特征选择

本文基于训练模型分类性能为标准,以特征重要性排序为主的包裹式选取[6]。

使用前面介绍的 3 种分类算法(参数为默认值)对几种慢性病建模,对比得到一个准确率较高的算法,重新对几种慢性病建模(使用随机搜索进行参数调优),最后将每种慢性病排名前 10 的特征选择出来。

2.1 数据预处理

本研究采用的是 UCI 机器学习存储库:Chronic_Kidney_Disease 数据集,选择其中高血压 Htn、糖尿病 Dm 和冠状动脉疾病 Cad 三条属性做分类研究,将“?”做空值处理,将二分类属性做 0 和 1 处理,每条属性的均值填入空值处,删掉 3 种疾病属性中原含有空值的样本,得到一个 397×25 的数据集。

2.2 模型建立

采用 Python 语言建模,将数据集按照 7:3 的比例划分,70%作为训练数据来训练模型;30%作为测试数据,用来检测模型的性能。使用机器学习库 sklearn 中的 RF、GBDT 和 Xgboost3 种分类算法,不对各算法调参下建模。

使用准确率、精确率(查准率)、召回率(查全率)、F1 值作为评价标准对分类器模型进行性能评价,由表 1~表 3 可知,Xgboost 在 3 种疾病上的预测准确率最高,并降低了误诊率和漏诊率。

2.3 超参数优化

GridSearchCV(网格搜索)和 Random-izedSearchCV(随机搜索)都是 sklearn 包中自动调参的方法,系统

地遍历多种参数组合，通过交叉验证确定最佳效果参数。这里选择寻优速度更好的随机搜索法。

Xgboost 中有 6 项主要参数，不同参数有不同的功能，这些参数设定是否合理，对于模型的好坏有重要影响[7]。选择的 6 项参数以及搜索范围如表 4 所示，以“F1”为准确度评价标准，随机搜索次数 n_iter 设为 1000。由于随机搜索的特性，记录 10 次调参结果，取最优的结果，见表 4、表 5。

表 1　Htn 各分类器性能比较

Tab.1　The performance comparison of different classifiers for Htn

模型类型	准确率/%	Precision	Recall	F1 值
GBDT	83.33	0.79	0.71	0.75
RF	85.83	0.82	0.76	0.79
Xgboost	86.67	0.84	0.76	0.80

表 2　Dm 各分类器性能比较

Tab. 2　The performance comparison of different classifiers for Dm

模型类型	准确率/%	Precision	Recall	F1 值
GBDT	82.50	0.75	0.73	0.74
RF	84.17	0.79	0.73	0.76
Xgboost	85.00	0.77	0.80	0.79

表 3　Cad 各分类器性能比较

Tab. 3　The performance comparison of different classifiers for Cad

模型类型	准确率/%	Precision	Recall	F1 值
GBDT	90.83	0.43	0.30	0.35
RF	91.67	0.50	0.10	0.17
Xgboost	92.50	0.56	0.50	0.53

表 4　Xgboost 最佳参数及调参范围

Tab. 4　The best parameters and range of parameters for Xgboost

参数名	Htn	Dm	Cad	调参范围	步长
learning_rate	0.4	0.5	0.2	[0.1,0.5]	5
n_estimators	90	136	206	[50,500]	450
max_depth	3	14	6	[3,20]	18
min_child_weight	2	3	1	[1,5]	5
gamma	0.9	0.6	0.1	[0,1]	11
colsample_bytree	0.5	1.0	0.8	[0.5,1]	6

表 5　Xgboost 调参性能

Tab. 5　Xgboost parameter tuning for performance

疾病种类	准确率/%	Precision	Recall	F1 值
Htn	88.33	0.85	0.81	0.83
DM	87.50	0.82	0.80	0.81
Cad	92.50	0.56	0.50	0.53

2.4　特征重要性

使用 xgboost 模块中 plot_importance 函数对特征进行重要性排序，使用特征在所有树中被用作分割样本的特征的次数“weight”作为特征重要程度的判断指标，留下排名前 10 的特征。得到 3 种慢性疾病的主要危险因素，如图 2～图 4 所示。

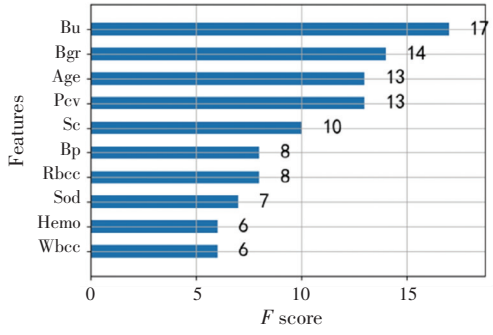

图 2　Htn 特征重要性

Fig. 2　Feature importance of Htn

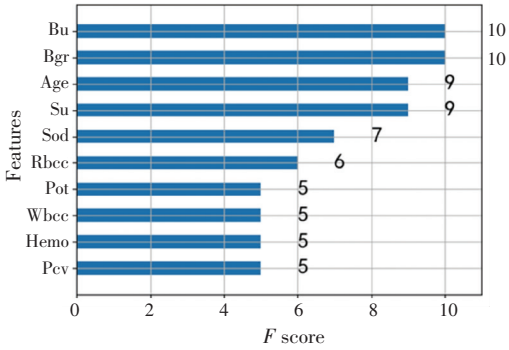

图 3　Dm 特征重要性

Fig. 3　Feature importance of Dm

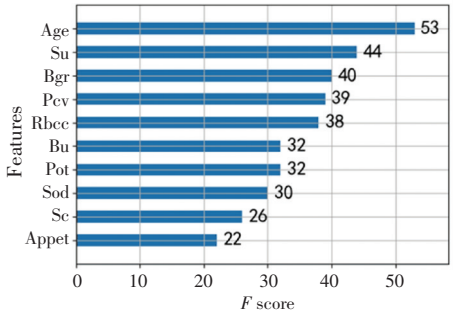

图 4　Dm 特征重要性

Fig. 4　Feature importance of Cad

对 3 种慢性疾病的主要危险因素做并集处理，得到 Bu 血液尿素、Bgr 血糖、Age 年龄、Pcv 红细胞压积、Sc 血清肌酐、Bp 血压、Rbcc 红细胞计数、Sod 钠、Hemo 血红蛋白、Wbcc 白细胞计数、Su 糖、Pot 钾、Appet 胃口，共 13 种主要危险因素，其中 6 项 3 种疾病共同拥有；5 项 2 种疾病共同拥有，只有两项是单独拥有，因此慢性病在预测时不好判断，将医生多年的诊断经验模型化，利用限特征进行疾病预测。

3 多疾病风险预测

一种疾病危险因素可能是多种疾病的判断标准。因此,采用相同危险因素对多种疾病同时进行风险预测,用问题转化的方法将多疾病风险预测问题转化为多标签分类问题。

3.1 多疾病标签转换

问题转化方法中 Binary Relevance(BR)和 Label-Powerset(LP)是两种具有代表性的方法[8]。BR 方法是将多标签转化为多个相互独立的单标签二分类问题;LP 方法是将多个标签转化为多分类单标签问题,3 个标签就有2^3个类。本文选择神经网络作为基础算法,利用深度学习框架 Keras 建立神经网络模型,选择损失函数 binary cross entropy 和 categorical cross entropy 分别对应问题转化中的 BR 和 LP 方法。

3.2 构建神经网络模型

神经网络采用序贯模型。输入层有 13 个神经元;经过多次试验,隐藏层设置为 4,神经元都为 10,后面 3 层引入 Dropout 和 L2 正则化来防止模型过拟合,激活函数选择使用非线性函数 ReLU 函数。ReLU 函数是分段线性函数,在梯度下降算法运算中拥有较好的性能;输出层,BR 方法神经元设置为 3,配合使用 sigmoid 函数作为激活函数,LP 方法神经元设置为 8,配合使用 softmax 函数作为激活函数。对于已有数据集标签,BR 方法可直接使用;对于 LP 方法,需要对 3 个标签组合,再将二进制数转化为 10 进制数,最后进行 onehot 编码。

3.3 训练模型

将数据按 8:2 划分,80%用于网络训练,20%用于精度测试,将数据归一化处理,避免数值量级差异引起权值过大或过小,并提升了运算速度。优化方法选择了一种自适应梯度下降方法 Adam,与其他优化算法相比,其收敛速度较快,学习效果更为有效;二种方法都以 acc 作为度量标准。经过多次实验,模型训练的轮数 epochs 设置为 100,指定权重更新的每个批次所使用实例的个数为 3。

3.4 结果分析

使用 evaluation 函数来评估模型的准确率,由训练返回值 history 作 acc-loss 曲线,如图 5 所示。得到 LP 方法的 acc 为 67.50%,loss 为 0.93;BR 方法的 acc 为 86.25%,loss 为 0.33,可以看出 BR 方法的准确度更高,损失值更趋近于 0,预测效果更好。但是这种方法不考虑疾病间的关联性,模型简单,需要进一步优化增加标签之间关联性才能用于多疾病预测;LP 方法考虑了标签与标签之间的关联,增加了复杂度,使得准确率降低,使用此方法对某一样本预测时,输出端 8 个值的和为 1,其中最大值对应索引即为预测结果,最大值就是预测概率。

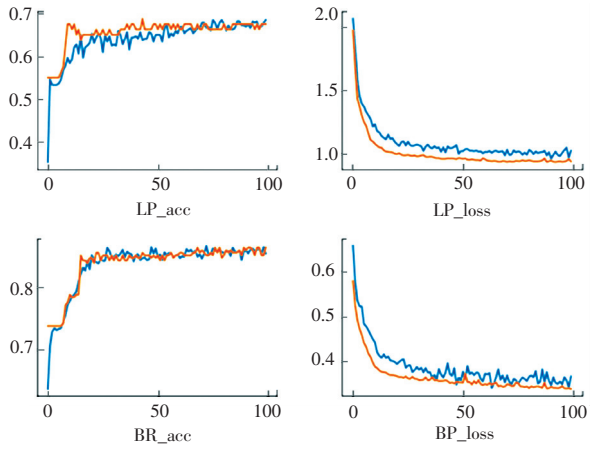

图 5 acc-loss 曲线

Fig. 5 The acc-loss curve

4 结束语

本文阐述了随机森林、GBDT 和 Xgboost3 种集成学习算法的原理及特点,基于 3 种慢性病,使用寻优速度较好的随机搜索法对其中分类效果最好的 Xgboost 进行超参数优化,分别得到 3 种慢性病特征重要性排名前 10 的危险因素。使用 keras 框架构建神经网络模型,对这些危险因素进行多疾病风险预测,对比分析了 BR 和 LP 两种问题转化方式得到的结果。在今后的研究中,要进一步丰富数据,采用更好的数据处理方法,构建更为复杂的神经网络;对损失函数进行研究,在 BR 方法的基础上考虑标签之间的关联性。

参考文献

[1] 中华人民共和国中央人民政府. 健康中国行动(2019—2030 年)[EB/OL].
[2] 钟晓玲. 基于标签相关性和类不平衡性的多标签分类算法[D]. 华南理工大学,2019.
[3] 孙予舒,黄芸,梁婷,等. 基于 XGBoost 算法的复杂碳酸盐岩岩性测井识别[J]. 岩性油气藏,2020,32(4):98-106.
[4] 马湧,王晓鹏,马莎莎. 基于 Keras 深度学习框架下 BP 神经网络的热轧带钢力学性能预测[J]. 冶金自动化,2019,43(2):6-10.
[5] 魏贞原. 深度学习:基于 Keras 的 Python 实践[M]. 北京:电子工业出版社,2018.
[6] 周庆岸. 基于遗传 XGBoost 模型的个人网贷信用评估研究[D]. 江西财经大学,2019.
[7] 张春富,王松,吴亚东,等. 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.
[8] ZHANG Xiaoqing, ZHAO Hongling, ZHANG Shuo, et al. A Novel Deep Neural Network Model for Multi-Label Chronic Disease Prediction.[J]. Frontiers in genetics,2019,10.