计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(9):215−218 [doi: 10.15888/j.cnki.csa.007054]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 CatBoost 算法的糖尿病预测方法①

苗丰顺[1,2], 李　岩[2], 高　岑[2], 王美吉[2], 李冬梅[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
通讯作者: 苗丰顺, E-mail: ucasmfs@163.com

摘　要: 近几十年来, 人们生活水平显著提高, 但是健康意识依旧薄弱, 不良的生活习惯和饮食习惯导致糖尿病发病人数急剧增加, 由糖尿病导致的各种并发症严重威胁了人们的健康. 由于糖尿病具有知晓率低的特点, 很多糖尿病患者未能及时发现病症, 导致出现并发症. 本文通过分析糖尿病的特点, 针对医疗数据样本量小、容易缺失的特点, 选择 IV 值分析进行特征选择、使用一种新型的 Boosting 算法 CatBoost 进行糖尿病患者预测, 取得了显著的预测效果.

关键词: 糖尿病; IV 值分析; 特征选择; 集成学习; CatBoost

引用格式: 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.计算机系统应用,2019,28(9):215–218. http://www.c-s-a.org.cn/1003-3254/7054.html

Diabetes Prediction Method Based on CatBoost Algorithm

MIAO Feng-Shun[1,2], LI Yan[2], GAO Cen[2], WANG Mei-Ji[2], Li Dong-Mei[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: In recent decades, people's living standards have improved significantly, but health awareness is still weak. Poor living habits and eating habits have led to a sharp increase in the number of people with diabetes. The complications caused by diabetes are a serious threat to people's health. Because awareness rate of diabetes is low, many patients with diabetes fail to detect the disease in time, leading to complications. In this study, by analyzing the characteristics of diabetes, according to the characteristics of small sample size and easy to be missing, the IV value analysis is used for feature selection, and CatBoost, a new type of Boosting algorithm, is used to predict diabetes patients and achieves significant predictive effects.

Key words: diabetes; IV value analysis; feature selection; ensemble learning; CatBoost

近些年来, 随着我国经济的迅速发展和国民生活方式以及饮食结构的改变, 我国糖尿病患者人数正在以惊人的速度增长. 糖尿病呈现发病率高、知晓率、治疗率和达标率低的现象[1], 严重的威胁了患者的身体健康, 同时给家人和社会带来了沉重的经济负担[2]. 目前大部分医疗机构对糖尿病的诊断还是依靠医生的个人经验和体检数据为准, 这种诊断方式对医生的要求很高, 具有很强的主观性, 容易出现误诊和漏诊的情况, 错失了预防和治疗的最佳时机, 将严重影响到病人的身心健康. 如果将糖尿病和机器学习结合, 采用机器学习算法来辅助医生诊断, 将会很大程度上提高诊断和科学性, 有效的克服医生凭经验诊断的主观性的问题. 因此, 使用机器学习的方法对糖尿病患者进行预测, 具有很大的现实意义.

① 收稿时间: 2019-02-28; 修改时间: 2019-03-22; 采用时间: 2019-03-27; csa 在线出版时间: 2019-09-05

关于疾病预测, 国内外教育专家已经做过很多尝试, 研究出了很多算法, 如: logistic 回归、BP 神经网络模型、COX 比例风险模型、决策树模型[3–5]. 这些方法在疾病预测方面有较好的效果, 但是也有各种各样的缺点, 比如要求数据量大、要求一定时间段的的连续数据、泛化能力太弱、过度拟合、陷入局部最小值、对随机性和波动性数据不敏感、对不平衡数据预测 效果不理想等问题等等. 现有的研究大多采用单个全局优化模型, 单分类器模型性能有限, 存在泛化能力弱和容错性较差等问题.

本文使用集成学习模型进行预测, 它是使用一系列的学习器进行学习, 并使用某种规则把各个学习器的学习结果进行整合, 从而获得比单个学习器更好的学习效果的一种机器学习方法[6]. 集成学习模型现在主要分为 Bagging 和 Boosting. 基于 Bagging 的模型的方差较小, 但是偏差较大, 故对基分类器的准确性要求较高. Boosting 可以降低模型偏差, 它通过迭代地训练一系列的分类器, 每个分类器采用的样本分布都和上一轮的学习结果有关, 对基分类器的准确性要求较低. CatBoost 在 2017 年被 Yandex 首次提出, 是 boosting 的一种实现方式, 它采用对称树的方式, 并且用特殊的方式来处理 categorical features, 从而有效的避免了过拟合的问题, 提高了泛化能力, 提高了模型的鲁棒性, 特别适合样本量小、数据不平衡的情况. 目前该算法在糖尿病预测方面还没有应用[7–9].

通过以上分析, 本文决定采用一种基于特征选择和集成学习算法的模型来进行糖尿病的预测. 通过 IV 值分析进行特征选择, 有效的去除冗余特征, 确定最后的最优特征子集来训练模型; 使用 CatBoost 有效的避免过拟合的问题, 提高模型的泛化能力和鲁棒性, 最终达到良好的预测效果.

1 算法描述

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.

1.1 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[10]. 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$L(y,F^{t-1}(x))$, 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. 式 (1) 表示的是本轮迭代的目标函数h^t.

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

GBDT 使用损失函数的负梯度来拟合每一轮的损失的近似值, 式 (2) 中$g^t(x,y)$表示的是上述梯度.

$$g^t(x,y) = \frac{\partial L(y,s)}{\partial s}|_{s=F^{t-1}(x)} \tag{2}$$

通常用式 (3) 近似拟合h^t.

$$h^t = \arg\min_{h\in H} E(-g^t(x,y)-h(x))^2 \tag{3}$$

最终得到本轮的强学习器, 如式 (4) 所示:

$$F^t(x) = F^{t-1}(x)+h^t \tag{4}$$

1.2 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$\sigma_1$到$\sigma_n$一次遍历随机序列, 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:

$$\frac{\sum_{j=1}^{p}[x_{j,k}=x_{i,k}]\cdot Y_i + a\cdot P}{\sum_{j=1}^{n}[x_{j,k}=x_{i,k}]+a} \tag{5}$$

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 的算法流程如图 1 所示.

由图 1 可知, 为了得到无偏梯度估计, CatBoost 对每一个样本x_i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.

Ordered boosting

Input: $\{(x_k, y_k)\}_{k=1}^{n}$, I:
$\sigma \leftarrow randompermutation\ of\ [1, n]$;
$M_i \leftarrow 0\ for\ i=1..n$;
for $t \leftarrow 1$ **to** I **do**
  for $i \leftarrow 1$ **to** n **do**
    $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_i)$;
  end for
  for $i \leftarrow 1$ **to** n **do**
    $\Delta M \leftarrow$
    $LearnModel((x_j, r_j): \sigma(j) \le i)$
    $M_i \leftarrow M_i + \Delta M$;
  end for
end for
return M_n

图 1 Ordered boosting 流程

2 实验分析

对于整个糖尿病预测的研究包括以下几部分: 数据采集、数据预处理、特征选择、模型预测、结果分析. 如图 2 所示.

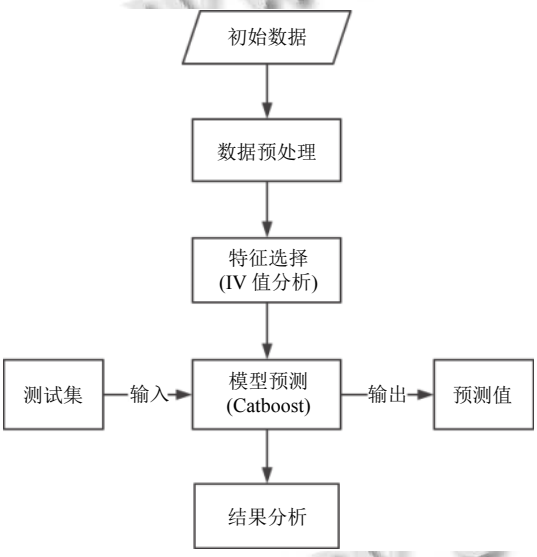

图 2 糖尿病预测模型

2.1 数据采集

本研究的数据来源于实 验室项目: 骕派智慧社区. 数据集是沈阳某医院的 2018 年的一次脱敏体检数据, 包含 2377 条数据, 数据维度高达 63.

2.2 数据预处理

由于原始数据规模太过庞大, 数据不完整、重复、杂乱等问题显著, 数据预处理对于后期的建模预测影响显著. 原始数据存储在不同的数据表中, 并且存在很多缺失项以及数据冗余, 需要对原始数据进行预处理. 本文在数据预处理方面做了以下工作: 丢弃缺失项过多的样本、丢弃缺失值过多的特征、采用均值进行缺失值填充、对标称属性、二元属性进行数据转换、数据集成以及去除冗余.

2.3 特征选择

特征选择对后期的建模预测起着关键性的作用, 尤其是在样本量小、特征多的情况下, 剔除噪音、正确的选择特征会对模型整体的准确性和稳定性有着质的提升.

IV 值分析是常见的处理特征值的方法, 它衡量了某个特征对目标的影响程度. 其基本思想是根据该特征所命中黑白样本的比率与总黑白样本的比率, 来对比和计算其关联程度, 计算公式如下:

$$IV = \sum_{i}^{n} (P_{yi} - P_{ni}) * \ln \frac{P_{yi}}{P_{ni}} \tag{6}$$

其中, n 代表样本在该特征上分成的组数, P_{ni}表示该样本第 i 组数据中白样本占所有白样本的比例, P_{yi}表示该样本第 i 组数据中黑样本占左右黑样本的比例.

本文采用 IV 值分析的方法进行特征选择, 最终选出 23 个对糖尿病有影响的特征变量作为模型的输入变量, 其中包括性别、年龄、体重指数、收缩压、舒张压、胆固醇、甘油三酯、尿素/肌酐等.

2.4 模型预测

经过前面的分析, CatBoost 算法适合用于对糖尿病的研究, 本文就选择 CatBoost 模型来建模. 经过数据预处理和特征选择, 把 23 个特征变量作为输入变量输入模型进行预测, 将 70% 作为训练样本, 30% 作为测试样本.

分类阈值对模型的准确性影响重大, 本文利用样本的区间准确率来确定最后的阈值: 将样本按照预测值进行排序, 然后没 5% 分为一个区间, 找到最后一个区间准确率大于 50% 的区间, 将该区间的端点值作为最后的阈值.

基于准确率、召回率、$F1$ 值等评价指标, 通过与随机森林、XGBoost 等模型进行对比分析, 能够得出 CatBoost 在处理该问题上具有更好的效果.

2.5 评价标准

选取评价指标是整个实验环节的重要一环, 直接影响到后期的结果分析. 分类器性能的优劣通常使用准确率 (*accuracy*)、精确率 (*precision*) 和召回率 (*recall*) 来评价. 准确率 (*accuracy*) 指的是被分类正确的样本数占总样本数的比值, 精确率 (*precision*) 表示的是预

测为正的样本中, 实际为正的所占的比例, 召回率 (*recall*) 表示的是实际为正的样本中, 被预测为正的所占的比例. 分类结果混淆矩阵如表 1 所示.

表 1 分类结果混淆矩阵

	Positive	Negative
True	TP	TN
False	FP	FN

本文选取精确率 (*precision*)、召回率 (*recall*) 和 *F*1 值作为评价指标.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$precision = \frac{TP}{TP+FP} \tag{8}$$

$$recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1 = \frac{2*precision*recall}{precision+recall} \tag{10}$$

其中, *P* 为阳性样本总数, *TP* 为预测正确的阳性样本数, *TN* 是预测错误的阳性样本数, *F*1 值是精准率和召回率的加权调和平均值, 为 1 时最优, 为 0 时最差. 对于复杂的模型和大量的数据, 计算速度也是衡量模型优劣的一个指标, 本实验在服务器上运行, 使用 Python 包的 time. time() 函数记录模型运行的时间.

2.6 试验结果与分析

经过以上步骤的数据处理、特征选择和模型预测, 得出了使用 CatBoost 模型的预测结果, 并且和使用随机森林、lightGBM 模型得出的结果进行了对比, 结果如表 2 所示.

表 2 运行结果分析

模型	精确率	召回率	*F*1 值	运行时间 (s)
随机森林	0. 93	0. 85	0. 79	4. 36
XGBoost	0. 88	0. 86	0. 82	0. 89
CatBoost	0. 89	0. 90	0. 985	0. 91

从表 2 可以看出, 在模型精确率方面, 随机森林略高于 XGBoost 模型和 CatBoost 模型, 在召回率和 *F*1 值方面, CatBoost 都高于 XGBoost 和随机森林, 在运行时间方面, CatBoost 和 XGBoost 明显高于随机森林.

3 总结

本文以糖尿病发病人数多、发现率少以及医疗数据维数高、缺失量大为背景, 依托于实验室骕派智慧社区项目, 实现了对糖尿病的预测模型.

本文使用一种新型的 Boosting 算法 CatBoost 进行糖尿病预测, 并且取得了良好的预测结果. 在文中, 首先对原始数据进行预处理, 然后采用 IV 值分析的方法进行特征选择, 并采用适用于该问题的集成学习模型 CatBoost 进行预测, 最终得出较好的预测结果. 通过实验分析, CatBoost 在各项评价指标上比其他模型都具有明显的优势, 说明 CatBoost 在糖尿病预测方面具有很好的应用价值. 通过本文的研究, 可以对糖尿病预测提供有效的指导, 对保护人们的健康具有非常积极的意义.

参考文献

1 白碧玉, 于琦, 苏闫兵, 等. 中国糖尿病研究论文合作分析. 中国药物与临床, 2017, 17(11): 1619–1621.

2 王海鹏. 我国诊断糖尿病疾病经济负担趋势预测研究[博士学位论文]. 济南: 山东大学, 2013.

3 苏萍, 杨亚超, 杨洋, 等. 健康管理人群 2 型糖尿病发病风险预测模型. 山东大学学报 (医学版), 2017, 55(6): 82–86. [doi: 10.6040/j.issn.1671-7554.0.2017.347]

4 罗森林, 成华, 张铁梅, 等. 多维 2 型糖尿病实测数据的预处理技术. 计算机工程, 2004, 30(17): 178–181. [doi: 10.3969/j.issn.1000-3428.2004.17.071]

5 吴海云, 潘平, 何耀, 等. 我国成年人糖尿病发病风险评估方法. 中华健康管理学杂志, 2007, 1(2): 95–98. [doi: 10.3760/cma.j.issn.1674-0815.2007.02.012]

6 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. 中国实验诊断学, 2018, 22(3): 408–412. [doi: 10.3969/j.issn.1007-4287.2018.03.008]

7 Bottou L. Large-Scale machine learning with stochastic gradient descent. Proceedings of the 19th International Conference on Computational Statistics Paris France. Keynote. 2010. 177–186.

8 Tan PN. Receiver operating characteristic. Liu L, Özsu MT. Encyclopedia of Database Systems. New York, NY, USA: Springer, 2013. 2349–2352.

9 Sau A, Bhakta I. Screening of anxiety and depression among the seafarers using machine learning technology. Informatics in Medicine Unlocked, 2018. [doi: 10.1016/j.imu.2018.12.004]

10 Yang T, Chen WT, Cao GT. Automated classification of neonatal amplitude-integrated EEG based on gradient boosting method. Biomedical Signal Processing and Control, 2016, 28: 50–57. [doi: 10.1016/j.bspc.2016.04.004]