计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):228−232 [doi: 10.15888/j.cnki.csa.006729]

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

# 基于聚类和 XGboost 算法的心脏病预测①

刘　宇[1], 乔　木[2]

[1](南京烽火天地通信科技有限公司, 南京 210019)
[2](武汉邮电科学研究院, 武汉 430074)
通讯作者: 刘　宇, E-mail: 13805167455@qq.com

**摘　要**: 过去十几年来, 心脏病发病率在全球一直呈上升趋势且居高不下. 所以, 如果可以通过计算机手段提取人体相关的体检指标, 且通过机器学习的方式来分析不同特征及其权值对于心脏病的影响, 对于预测和预防心脏病将起到很关键的作用. 因此本文提出一个基于聚类和 XGboost 算法的预测方法. 通过对数据的预处理, 区分特征, 再通过聚类算法如 K-means 对数据集聚类分块. 最后用 XGboost 算法进行预测分析. 实验结果表明, 所提出的基于聚类和 XGboost 算法的预测方法的可行性和有效性, 为就医推荐等应用提供了精准有效的帮助.

**关键词**: 心脏病预测; 聚类; 机器学习; K-means; XGboost

引用格式: 刘宇,乔木.基于聚类和 XGboost 算法的心脏病预测.计算机系统应用,2019,28(1):228–232. http://www.c-s-a.org.cn/1003-3254/6729.html

## Heart Disease Prediction Based on Clustering and XGboost

LIU Yu[1], QIAO Mu[2]

[1](Nanjing FiberHome World Communication Technology Co. Ltd., Nanjing 210019, China)
[2](Wuhan Research Institute of Posts and Telecommunications, Wuhan 430074, China)

**Abstract**: In the past decade, the incidence of heart disease has been on the rise and remains high in the world. If the physical examination indicators related to the human body can be extracted by computer measures, and the influence of different characteristics and their weights on heart disease can be analyzed through machine learning, it will play a key role in predicting and preventing heart disease. Therefore, a prediction method based on clustering and XGboost algorithm is proposed in this study. By preprocessing the data and distinguishing the features, the data sets are clustered by clustering algorithm, such as K-means. Finally, the XGboost algorithm is used to predict and analyze. The experimental results show that the proposed method based on clustering and XGboost algorithm is feasible and effective, which provides accurate and effective help for the application of medical recommendation.

**Key words**: heart disease prediction; clustering; machine learning; K-means; extreme gradient boosting

早在上个世纪初, 心脏病就已经成为人们对于健康隐患关注的焦点. 全世界每年约有 1750 万人死于心脏病及其并发症, 占全部死亡人数的 1/3; 而我国社会开始进入老年化阶段, 老龄化趋势明显, 患有慢性病的人群 (如患有心脏病和血压不稳定人群) 也逐渐增长. 对于慢性病患者而言, 他们的生命体征数据应该进行实时监测, 否则会严重影响医生的准确诊断.

目前, 大部分医疗机构对于心脏病的检测诊断还是以医生个人经验和体检结果为标准. 不仅在人工上花费很高, 也会延误患者最佳就诊时间, 但是, 如果可以以机器学习的方法作为辅助诊断, 为临床诊断提供有效精准的指导帮助, 将会很大程度的提高诊断的科

① 收稿时间: 2018-07-07; 修改时间: 2018-08-09; 采用时间: 2018-08-16; csa 在线出版时间: 2018-12-26

学性. 而且现在大数据时代下的技术融合屡见不鲜, 利用机器学习来辅助心脏病诊断和预测会是一个值得研究的问题.

近年来, 陈天奇对 GBDT (Gradient Boosting Decision Tree) 算法进行改进[1], 提出了一种设计高效、灵活并且可移植性强的最优分布式决策梯度提升库 XGBoost (Extreme Gradient Boosting, XGBoost), 其原理是通过弱分类器的迭代计算实现准确的分类效果, 本文将 XGBoost 引入到心脏病预测中, 分析用户在医疗平台的检查数据信息建立分类预测模型, 从而个作为医生的辅助判断. 实验结果表明, 所提出的基于聚类和 XGboost 算法的预测方法的可行性和有效性, 为就医推荐等应用提供了精准有效的帮助.

## 1 方法

### 1.1 数据预处理

因为数据是从真实场景下获取的, 所以大体上并不完整, 不能够直接应用于实验训练, 因此正需要数据预处理来解决这一问题。所谓预处理就是对不标准的数据进行类型变换, 除去空值等操作[2], 本次实验的数据包括 14 个特征, UCI 开源数据集某医院的 300 个心脏病患者体侧数据和一万个经过数据清洗的某健康 App 的用户调查数据. 而本次算法模型的应用场景是根据这些特征对样本进行预测, 并判断是否患病. 样本是否患病是一个分类问题, 正好符合本文模型的应用标准, 因为本次算法模型选用的是线性模型逻辑回归, 全部特征的值对应的都是 double 型, 且不能为空.

(1) 二值类的数据

这类数据通常只有两个值, 如性别字段有两种表现形式 female 和 male, 我们可以将 female 表示成 0, 把 male 表示成 1.

(2) 多值类的数据

比如表示胸部的疼痛感的 cp 字段, 我们可以通过疼痛的由轻到重映射成 1~4 的数值以及缺陷种类可以划分 0~10 个不同的等级, 并以此处理为训练可用的数据.

(3) 去空的数据

在实际情况下的患者的检查数据中, 会有部分检查项未填或者未进行, 该特征会为空, 在数据中会以?表示, 我们就需要去掉有空值的那行数据, 去噪, 保持训练模型的精准性.

第一步和第二步的处理通过 sql 脚本来实现, 最后一步去空则是在 Python 里遍历每一行进行操作. 处理过后的数据特征包括年龄, 性别, 胸部疼痛, 血压,胆固醇,心电图,最大心跳等 14 个变量. 变量列表如表 1 所示.

表 1 变量名列表

| 变量类型 | 变量名 | 变量范围 |
|---|---|---|
| 医疗特征 | 年龄 | 30~80 |
| | 性别 | 0/1 |
| | 胸部疼痛 | 1~4 |
| | 血压 | 100~200 |
| | 胆固醇 | 100~450 |
| | 空腹血糖 | 0/1 |
| | 心电图 | 0~2 |
| | 最大心跳 | 100~200 |
| | 心绞痛 | 0/1 |
| | 运动比较心压 | 0~5 |
| | 心电图倾斜度 | 1~3 |
| | 血管数 | 0~3 |
| | 缺陷种类 | 0~10 |
| | 是否患病 | 0/1 |

### 1.2 聚类算法

本文引用的聚类算法是 K-means 算法, K-means 算法中的 $K$ 代表类簇个数, means 代表类簇内数据对象的均值 (这种均值表示的是类簇中心)[3]. K-means 算法是一种经典的聚类算法, 此算法以数据对象之间的距离作为聚类标准, 即数据对象之间距离越小则表示这类数据拥有较高的相似度, 就会朝着一个中心点聚集, 距离越大则作用相反. 而数据对象的间距通常用欧氏距离来计算. 算法流程就是一直重复这一计算过程到标准测度函数收敛为止. 给出聚类 $K$ 和数据子集 $T=\{t_1,t_2,\cdots,t_n\}, t_i=(x_i,y_i)$.

K-means 算法的基本思想[4]如下:

步骤 1. 随机选择 $K$ 作为初始质心点;

步骤 2. 对于所剩下的对象, 则根据它们与这些聚类中心距离, 分别将它们分配给与其最相似的聚类;

步骤 3. 计算每个所获新聚类的聚类中心;

步骤 4. 如果满足了标准, 就停止步骤, 否则转回步骤 2, 直到条件满足.

停止条件如下: 没有需要分配的任务到不同的簇, 质心不再发生变化, 或者均方误差值下降幅度很小, 其计算式:

$$E=\sum_{k=1}^{k}\sum_{x\in c_k} d(d(x,m_k))^2 \tag{1}$$

其中, $c_k$ 是第 $k$ 个簇, $m_k$ 是簇 $c_k$ 的质心, $d(x,m_k)$是 $x$ 和质心 $m_k$ 之间的距离[5], 在本文提出的方法中, 欧氏距离是每个目标点到簇中心的距离:

$$d(x,m_k)=\|x,m_k\|=\sqrt{(x_1-m_k)^2+(x_2-m_k)^2+\cdots+(x_n-m_{kn})^2} \tag{2}$$

本文使用 K-means 算法是为了分割数据集, 因为海量数据会对模型精准度产生影响, 根据某个特征做聚类分成若干的小数据集, 可以使模型更加准确.

### 1.3 XGboost 算法

XGBoost 是一种改进的 GBDT 算法, GBDT 是 2001 年 Friedman 等人提出的一种 Boosting 算法. 它是一种迭代的决策树算法, 该算法由多棵决策树组成, 所有树的结论加起来作为最终答案[6]. 而 XGBoost 算法与 GBDT 有很大的区别. GBDT 在优化时只用到一阶导数, XGBoost 则同时用到了一阶导数和二阶导数, 同时算法在目标函数里将树模型复杂度作为正则项,用以避免过拟合[7].

XGBoost 算法目标函数:

$$J(f_t)=\sum_{i=1}^{n}L\left(y_i,\hat{y}_i^{t-1}+f_t(x_i)\right)+\Omega(f_t)+C \tag{3}$$

根据泰勒公式[8]展开:

$$f(x+\Delta x)\approx f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2 \tag{4}$$

同时令:

$$g_i=\frac{\partial L\left(y_i,\hat{y}_i^{t-1}\right)}{\partial \hat{y}_i^{t-1}},h_i=\frac{\partial^2 L\left(y_i,\hat{y}_i^{t-1}\right)}{\partial \hat{y}_i^{t-1}} \tag{5}$$

决策树复杂度[9]计算公式:

$$\Omega(f_t)=\gamma\cdot T_t+\lambda\frac{1}{2}\sum_{j=1}^{T}w_j^2 \tag{6}$$

将式 (4)、(5)、(6) 代入式 (3), 求得目标函数:

$$J(f_t)\approx\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma\cdot T+C \tag{7}$$

对式 (7) 求解, 求得最佳的$\omega_j^*$以及其对应的目标函数最优值[10]. 两个结果对应如下:

$$\omega_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \tag{8}$$

$$J(f_t)=-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma\cdot T \tag{9}$$

不过通过解得的目标函数来寻找出一个最优结构的树, 加入到模型中, 通常情况下枚举出所有可能的树结构是不可能的, 因此最后会使用贪心算法来寻找最优的分割方案, 也就是一种分割寻找算法, 这个算法称为精确贪心算法[11].

## 2 模型流程

### 2.1 流程实施

本文提出的基于聚类和 XGboost 算法的心脏病预测模型, 数据集是取自 UCI 开源数据集某医院的 300 个心脏病患者体侧数据和一万个经过数据清洗的某健康 App 的用户调查数据. 模型流程是将输入值通过 K-means 算法进行聚类, 找到对应某特征不同分段的类, 并将数据离散化分开成单独的簇, 每个簇经过 XGboost 算法训练和测试得到预测的输出值. 算法包括数据预处理、任务构造、数据离散、模型训练和结果分析 5 个部分, 如图 1 所示.

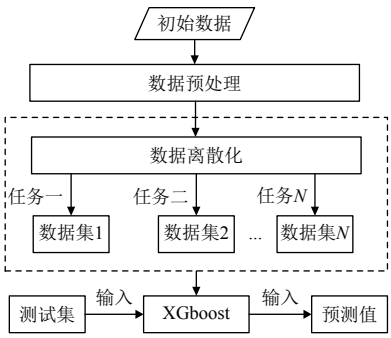


图 1 基于聚类和 XGboost 算法的心脏病预测模型

步骤一. 拿到初始数据后, 对初始数据做预处理,即对数据进行去噪、填充缺失值、类型变换等操作. 获得一个相对完整可用于训练模型的数据集.

步骤二. 选定一个与心脏病判断相关的特征, 如本文选取了胆固醇 (chol) 来作为 K-means 算法的聚类特征, 因为胆固醇特征与本文的心脏病预测模型并无直接联系, 且数据跨越度大, 聚类区分效果明显, 很适合作为实验特征对象. 根据目标估计值构造 $k$ 个学习任务$t+1,t+2,\cdots,t+n,n\leqslant h$,针对每个任务对应的输出值, 分别从数据集通过 K-means 算法聚类. 本文选取 $k$ 值

为 3, 并找到 3 个聚类的中心点. 根据数据离散化, 即统计学习中最大熵的原理, 将数据分成指标程度为高中低的 3 个数据集. 因为训练模型时, 在所有可能的概率模型中, 熵最大的模型是最好的模型, 通常用约束条件来确定概率模型的集合, 所以最大熵原理也可以表述为在满足约束条件的模型集合中选取熵最大的模型[12]. 如本文所确定胆固醇指标的聚类中心在 212 和 291 两处, 所以我们的数据集应该是小于 212, 212 至 291, 大于 291, 分别对应低中高三类数据集.

步骤三. 将 $N$ 个任务的训练集同时训练 XGboost 模型, 得到最终的预测模型.

步骤四. 将离散化过后的数据集以测试集和训练集比例 1:4 分割, 将 $n$ 个任务相应的测试样本输入训练好的 XGboost 模型, 可同时得到 $n$ 个估计值, 最后从中选择需要的估计值$y^*_{t+h}$. 选取需要的估计值之后, 即可以与实测数据进行对比, 从而得出本文算法预测的精确情况.

## 3 实验结果与分析

### 3.1 评价标准

评价分类器性能的优劣通常是用分类准确率 (Accuracy), 精确率 (Precision) 与召回率 (Recall), 准确率 (Accuracy) 表示的是对于给定的测试数据集,分类器正确分类的样本数与总样本数之比[13], 精确率 (Precision) 通常表示的是预测是正例的结果中, 实际为正例的比例. 召回率 (Recall) 表示的是实际为正例样本中, 预测也为正例的比例. 通常以关注的类为正类, 其他类为负类, 分类器在测试数据集上的预测是否准确, 如表 2 所示.

表 2 分类结果混淆矩阵

| 患病状态 | 预测患病 | 预测未患病 |
|---|---|---|
| 实际患病 | *TP* | *FN* |
| 实际未患病 | *FP* | *TN* |

样本预测的 *Accuracy*、*Recall* 和 *F*1 值作为评价指标其公式[14,15]如下:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$Precision = \frac{TP}{TP+FN} \tag{11}$$

$$Recall = \frac{TP}{TP+FP} \tag{12}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{13}$$

其中, $P$ 为阳性样本总数, $TP$ 为正确预测的阳性样本数量, $NP$ 为错误预测的阳性样本数量. 同时, 对于大量样本的数据处理, 算法模型的运算速度也是重要的评价指标. 本实验在个人计算机 (CPU: Intel i7 7700HQ 2.8 GHz; RAM: 16 GB) 上运行, 用 Python 包的 *time.time* () 函数[16]记录时间模型运行前后时间差, 即为运行时间, 方便运行效率比较.

### 3.2 结果比较分析

通过调参后得到的三种算法的最佳模型, 并且拿到测试样本对测试集进行预测, 以分类器评价标准对三种算法的各个方面作比对, 比对结果如表 2 所列. 实验结果表明, 在三种分类算法中, 以 XGBoost 算法训练的模型只是在准确率上稍逊随机森林算法, 而在运行速度方面显著优于随机森林算法. 而以 SVM 算法为基础的模型在原理上相对简单, 尽管运行速度与 XGBoost 算法训练的模型相近, 但是准确率却大大不如前者优秀. 因此经过最后的交叉比较以 XGBoost 算法训练的模型具有准确性较高, 运行速度快等优势. 经过这样的交叉比较, 突出了 XGBoost 算法在分类预测上的高效性.

表 3 运行结果比较

| 算法 | 准确率 | *F*1 | 召回率 | 运行时间 (s) |
|---|---|---|---|---|
| XGboost | 0.83 | 0.81 | 0.75 | 1.15 |
| 随机森林 | 0.99 | 0.87 | 0.71 | 6.35 |
| SVM | 0.81 | 0.76 | 0.73 | 1.18 |

但是单纯的使用 XGboost 算法, 在实际应用场景中可能不能达到最好的预测效果, 因为庞大的数据集也需要模型的准确性, 因此对数据集做了分割, 以高中低三种指标分别训练模型提高精确性, 根据表 4 的直观分析, 分别训练的三种模型在准确率上确实基本比初始模型要高, 在召回率上 L 模型和 H 模型表现的较为优秀. 而 *F*1 值也都差别不大. 因为都是使用的同一算法, 所以运行时间上不会有太大的差别. 综合三个数据集所训练的模型数据, 普遍优于初始数据集的模型, 并且有所提升 1%~2%. 尤其表现在中等指标的数据集训练模型, 其预测效果是最佳的.

表 4 模型准确率对比

| 算法 | 准确率 | *F*1 | 召回率 | 运行时间 (s) |
|---|---|---|---|---|
| 初始模型 | 0.83 | 0.81 | 0.75 | 1.15 |
| L 模型 | 0.83 | 0.80 | 0.77 | 1.15 |
| M 模型 | 0.85 | 0.82 | 0.74 | 1.15 |
| H 模型 | 0.84 | 0.81 | 0.78 | 1.15 |

### 3.3 重要特征分析

通过 XGBoost 的建模可以判断每个特征变量对模型的贡献程度, 从而判断哪些特征变量对于患心脏病的影响更为显著. 并可以以此对医生的判断其参考辅助作用, 分析结果如图 1 所示.

如图 2 所示为 XGBoost 模型的变量重要性结果, 其中, f0、f4、f7 和 f9 这 4 个变量在模型中的重要性排序中在前四位, 这四个变量分别代表着年龄、胆固醇、最大心跳和运动后比较心压. 而重要性最低的变量是心电图倾斜度. 从常识上讲, 年龄越大患心脏病概率越大, 不仅是心脏病, 各种心脑血管疾病患病概率都很大, 毕竟在数据集中高龄患者也是占很多的, 因此不作为主要判断参数. 而胆固醇, 最大心跳, 心压这三个参数是需要经过检测的, 在医生判断的时候根据检测结果, 这三个参数也可以作为主要参考. 就算患者本人体检自测三个参数, 也可以根据异常值推荐就医. 因此不仅仅是预测心脏病, 重要特征的显示也能给医生和患者带来重要的参考意见.

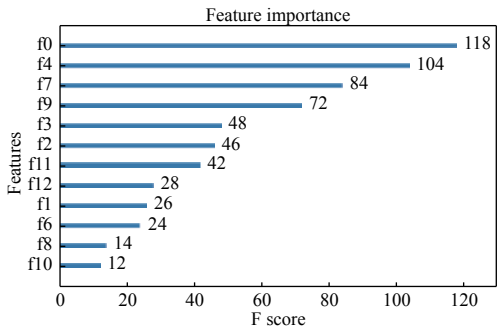


图 2 特征变量重要性

## 4 结论

本文针对医疗行业的心脏病预测问题, 提出了一种聚类和机器学习相结合的心脏病预测方法, 首先从患者数据中提取特征, 将医疗特征作为心脏病预测的输入变量, 然后使用 XGBoost 算法来对心脏病进行预测, 最后将该方法与其他机器学习算法进行比较.

实验结果证明, XGBoost 算法在各个评价标准中普遍优于传统算法, 说明了此模型精准性较高, 论证了使用 XGBoost 算法来对心脏病进行预测的可行性和可靠性. 且通过对变量重要性进行分析, 我们识别了对模型贡献较高的变量. 可以此为依据针对就医, 给医生和患者带来重要的参考意见. 对当前的医疗系统中疾病预防的完善有着重要的现实意义. 虽然模型准确率得到了提高, 但是在实验过程中发现, 改进的算法运行时间要更长, 如何在大数据量的情况下, 降低算法运行时间是接下来的工作.

## 参考文献


1 张潇, 韦增欣, 杨天山. GBDT 组合模型在股票预测中的应用. 海南师范大学学报 (自然科学版), 2018, 31(1): 73–80.
2 蔡旺华. 运用机器学习方法预测空气中臭氧浓度. 中国环境管理, 2018, 10(2): 78–84.
3 华辉有, 陈启买, 刘海, 等. 一种融合 Kmeans 和 KNN 的网络入侵检测算法. 计算机科学, 2016, 43(3): 158–162.
4 周丽娟, 王慧, 王文伯, 等. 面向海量数据的并行 KMeans 算法. 华中科技大学学报 (自然科学版), 2012, 40(S1): 150–152.
5 王娟, 王翰虎, 陈梅. 基于模糊聚类循环迭代模型的心脏病预测方法. 郑州大学学报 (理学版), 2007, 39(4): 137–140. [doi: 10.3969/j.issn.1671-6841.2007.04.034]
6 游德创, 莫赞. 基于模糊 xgboost 算法的银行信用评价研究. 信息通信, 2018, (2): 37–38. [doi: 10.3969/j.issn.1673-1131.2018.02.015]
7 郑凯文, 杨超. 基于迭代决策树 (GBDT) 短期负荷预测研究. 贵州电力技术, 2017, 20(2): 82–84.
8 Xu CX, Yan JF, Yang L, *et al*. Context co-occurrence based relationship prediction in spatiotemporal data. Proceedings of 2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018). Beijing, China. 2018. 63.
9 Xu ZG. Complex production process prediction model based on EMD-XGBOOST-RLSE. Proceedings of 2017 9th International Conference on Modelling, Identification and Control (ICMIC 2017). Kunming, China. 2017. 8.
10 谢冬青, 周成骥. 基于 Bagging 策略的 XGBoost 算法在商品购买预测中的应用. 现代信息科技, 2017, (6): 80–82. [doi: 10.3969/j.issn.2096-4706.2017.06.032]
11 张钰, 陈珺, 王晓峰, 等. Xgboost 在滚动轴承故障诊断中的应用. 噪声与振动控制, 2017, 37(4): 166–170, 179. [doi: 10.3969/j.issn.1006-1355.2017.04.032]
12 杨修德, 王金梅, 张丽娜. XGBoost 在超短期负荷预测中的应用. 电气传动自动化, 2017, 39(4): 21–25. [doi: 10.3969/j.issn.1005-7277.2017.04.005]
13 张昊, 纪宏超, 张红宇. XGBoost 算法在电子商务商品推荐中的应用. 物联网技术, 2017, 7(2): 102–104.
14 柴利达, 薛沁文, 毛娜, 等. 基于大数据的物资价格预测方法探索. 电力大数据, 2017, 20(12): 13–20.
15 Wang WZ, Shi YL, Lyu GF, *et al*. Electricity consumption prediction using XGBoost based on discrete wavelet transform. Proceedings of the 2nd International Conference on Artificial Intelligence and Engineering Applications (AIEA 2017). Wuhan, China. 2017. 14.
16 叶倩怡, 饶泓, 姬名书. 基于 Xgboost 的商业销售预测. 南昌大学学报 (理科版), 2017, 41(3): 275–281. [doi: 10.3969/j.issn.1006-0464.2017.03.015]