Statistics and Application 统计学与应用, 2021, 10(5), 787-793
Published Online October 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.105081

Hans汉斯

高斯判别模型对人类心脏疾病的预测分析

沈金辉

云南财经大学，云南 昆明

收稿日期：2021年9月2日；录用日期：2021年9月17日；发布日期：2021年10月8日

摘　要

随着现代化生活方式普及，大部分劳动者利用个人电脑进行办公，长时间坐着办公减少了必要的身体锻炼容易引发心脏疾病。本文从个人身体信息出发，利用分类预测模型建立对个人心脏疾病的预警机制。对于个体的年龄、性别、胸痛类型、静息血压等分类数值以及胆固醇含量、静息血压、最大心率等连续型数值进行描述性统计分析。区别逻辑回归等判别式学习算法，另辟蹊径。从贝叶斯先验角度出发引入了生成模型中推导严谨的高斯判别模型。

关键词

二分类预测，生成算法，高斯判别分析，心脏疾病，混淆矩阵

Predictive Analysis of Gauss Discriminant Model for Human Heart Disease

Jinhui Shen

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Sep. 2nd, 2021; accepted: Sep. 17th, 2021; published: Oct. 8th, 2021

Abstract

With the spread of modern lifestyles and the use of personal computers for office work, sitting for long periods of time reduces the need for physical exercise and can lead to heart disease. In this paper, based on personal body information, classification prediction model is used to establish an early warning mechanism for individual heart disease. Descriptive statistical analysis was performed on age, sex, type of chest pain, resting blood pressure and continuous values such as cholesterol content, resting blood pressure and maximum heart rate. Discriminant learning algorithms, such as logistic regression, provide a new way. From the perspective of Bayesian prior, a rigorous Gaussian discriminant model is introduced in the generation model.

文章引用: 沈金辉. 高斯判别模型对人类心脏疾病的预测分析[J]. 统计学与应用, 2021, 10(5): 787-793.
DOI: 10.12677/sa.2021.105081

Keywords

The Second Class Prediction, Generation Algorithm, Gauss Discriminant Analysis, Heart Disease, Confusion Matrix

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

个人的身体健康一直是社会上最关注的问题。在过去几十年中进行常规体检需要去医院逐项排查，但是面对一些特殊疾病的预防性体检常常给人们带来烦恼。例如，医院的内窥镜检查可以帮助中年人排查胃癌但是伴随巨大的痛苦，对于拥有心血管疾病遗传史的人群进行专项病症排查需要付出巨额的检查费用。因此如何减轻人们在疾病预防时的代价是一个有意义的问题。目前科研人员发现人体内的一些常规指标对于预测心脏疾病有着较强的关联性，因此从人体的常规体检数据去预测分析心脏疾病发生的可能性大小，利用数据分析模型得出一个基本的得病概率认识，可以减少许多不必要的专业检查，从而减轻了检查费用以及医疗系统的压力。

2. 对心脏疾病的描述性统计分析。

2.1. 对样本数据中主要变量的介绍

通过表 1 可以知道，样本数据中主要有两类数据。第一类数据：性别，运动是否心绞痛，胸痛的类型，空腹血糖含量是否正常，静息心电图等离散型数据。第二类数据：年龄，静息血压，胆固醇含量，最大心率等连续型数据。接下去分别对两类数据进行描述性统计分析。

Table 1. Independent variable name explanation
表 1. 自变量含义解释

自变量的名称	自变量含义解释	数值类型
Age	年龄	连续型数值
sex	性别	二分类数值
exang	运动是否引发心绞痛	二分类数值
cp	胸痛的类型	四分类数值
trtbps	静息血压	连续型数值
Chol	胆固醇含量	连续型数值
Fbs	空腹血糖含量是否正常	二分类数值
Rest_ecg	静息心电图	三分类数值
Thalach	最大心率	连续型数值
Target	是否患病	二分类数值

2.2. 离散型数据的描述性统计分析

通过表 2 对样本数据的年龄进行整体分析，发现总样本量为 303。平均年龄为 54.37，标准差较大说

明年龄分布比较分散。由分位数可以看出约有75%的个体年龄在47.5以上，说明数据中的样本以中年人为主，存在少量的青年人和老年人。

Table 2. Patient age table
表 2. 患者年龄统计表

统计指标	age
count	303
mean	54.37
std	9.08
min	29
0.25	47.5
0.5	55
0.75	61
max	77

通过图 1 发现，女性人群中未患病人数高于患病人数，但是两者差距并不大，女性在是否患有心脏疾病上没有明显的特征。

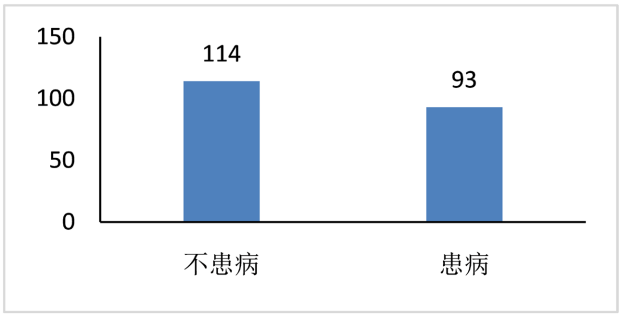

Figure 1. Bar chart of whether women are sick or not
图 1. 女性是否患病人数条形图

通过图 2 发现，男性人群中患病人数远远高于未患病人数，从性别角度来看男性在患心脏疾病的比例高于女性，说明性别在帮助分析患病情况时是一个重要的影响因素。

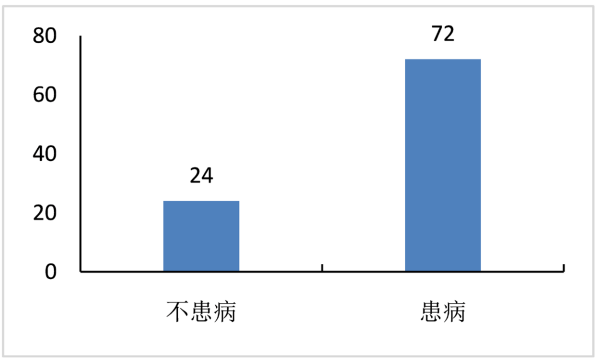

Figure 2. Bar chart of whether men are sick or not
图 2. 男性是否患病人数条形图

给定的数据集中存在一个运动后是否引发心绞痛的变量，经过常识认知一般运动后发生心绞痛可能是人体心脏不好的一种体现，因此为了验证这一猜想画出了条形图。通过图 3 发现，当筛选出在运动后有心绞痛的人群，其中未患病的人数远远高于患病人数，说明运动后引发心绞痛的人群恰恰不会患有心脏疾病，可以认为这一特征对患有心脏疾病具有负效应。

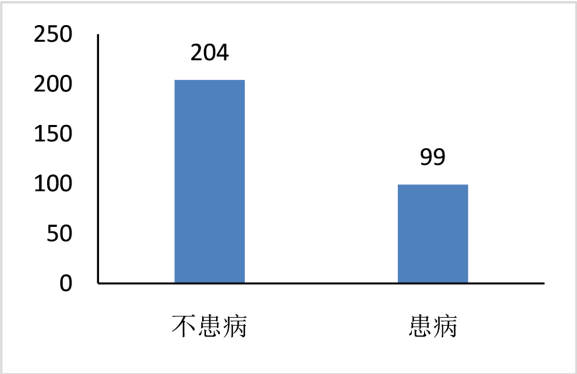

Figure 3. Bar chart of patients with angina pectoris
图 3. 心绞痛下患病人数条形图

2.3. 连续型数据的描述性统计分析

除了大量的离散型特征以外，还有关于血液，心率等大量的连续型指标。离散型特征只能从定性上来分析事物的特征，但是连续型变量可以定量的来描述事物现在的状态。通过图 4 发现，未患病人群的最大心率比患病人群的最大心率小，且患病人群的最大心率集中在 160。可以说明最大心率对不同人群是有区别的，健康人群的最大心率一般比患病人群最大心率小一些。从峰度的角度来说，患病人群的峰度值比较大，而未患病人群的峰度值相对较小。可以认为未患病群体的最大心率分布比较平坦而患病人群的最大心率分布相对集中。

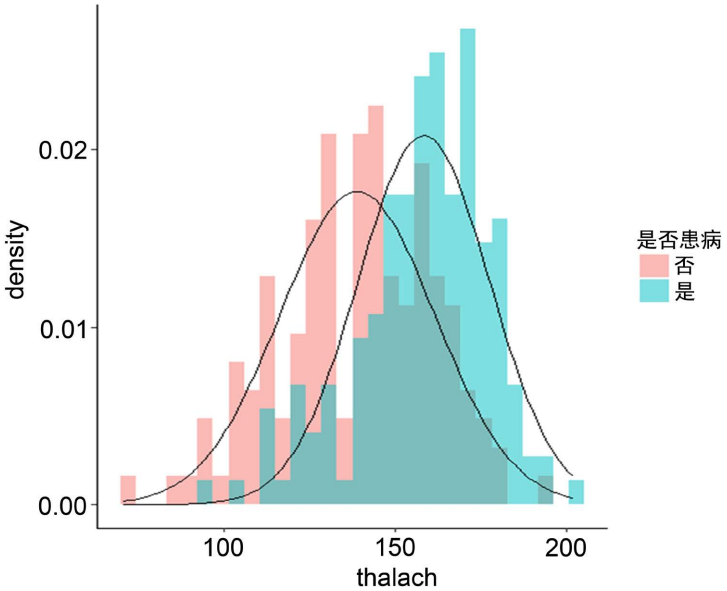

Figure 4. Maximum heart rate and disease histogram
图 4. 最大心率与疾病的直方图

因为静息血压，胆固醇含量，最大心率都是重要的指标，所以在建立模型的时候分析连续型变量之间的相关性是重要的，可以避免连续型变量之间的多重共线性对模型的影响。通过图 5 可以发现，静息血压，胆固醇含量以及最大心率之间的相关系数。静息血压和胆固醇相关系数为 0.12，静息血压和最大心率相关系数为−0.047，可以说明静息血压与胆固醇和最大心率相关程度弱。胆固醇和最大心率相关系数为−0.01，基本可以认为二者之间不存在相关关系。

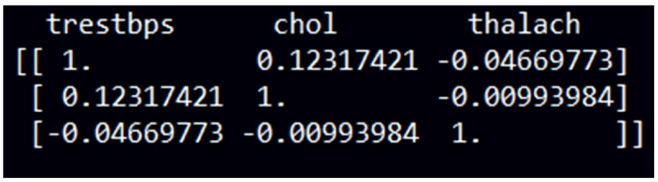

Figure 5. The correlation between continuous variables
图 5. 连续型变量之间的相关性

3. 高斯判别分析

3.1. 模型评价指标介绍

目前面临的疾病预测是一个典型的二分类问题。因此在此本文引入了生成算法中最为经典的高斯判别分析[1] [2]。在介绍高斯判别分析之前，对生成算法的阐述如下:生成算法的本质思想是由贝叶斯分析中引出，基本思想是

$$P(y|x)=\frac{P(x|y)P(y)}{P(x)} \tag{1}$$

$$P(x)=P(x|y_1)P(y_1)+P(x|y_2)(y_2) \tag{2}$$

为了计算在给定自变量的情况下对 $y=1$ 或者 $y=0$ 时的概率，首先由公式(1)可知，只需要对分子部分分别进行建模，由公式(2)可知 $P(x)$ 可以从第一步的建模中计算得出。因此生成算法就是由已知数据估计出 $P(x|y)$ 以及 $P(y)$，从这两个先验条件出发去估计出 $P(y|x)$ 发生的后验概率。这样的建模思路正好区别于判别分析模型的出发思路。

对于一个分类问题而言，虽然模型的建立是解决问题的基本保障，但是关于从模型中得出的结果分析是解决问题最有力的支撑。因此对模型结果建立一套科学的评价指标[3]是关键一步。从问题的起源出发，几十年的数据分析发展，已经对分类问题建立起了一套科学的评价指标。接下来介绍这套指标。目前而言标准化的分类指标是以混淆矩阵(表 3)为核心引出的。

Table 3. Confusion matrix
表 3. 混淆矩阵

	实际情况	
预测情况	TP	FP
	FN	TN

具体有精确率(precision)和召回率(recall)以及 F1 值对分类结果进行衡量。其中，TN (True Negative)表示：预测为假，实际也为假的情况；TP (True Positive)表示：预测为真，实际也为真的情况；FN (False Negative)表示：预测为假，实际上为真的情况。FP (False Positive)表示：预测为假，实际上也为假的情况。

从以上概念出发可以引出一下几个定义。精确率(PPV)如公式(3)所示。表示模型预测正确个数与模型预测总数之间的比例；召回率(TPR)如公式(4)表示模型预测正确个数与实际个数之间的比例；F1 分数表示模型精确率和召回率的一种相互协调的结果。

在分类问题中，TP 与 TN 是正确的预测，FP 与 FN 是错误的预测，因此总体说越接近 0 表示模型效果越差，越接近 1 表示模型效果越好。

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{5}$$

3.2. 模型结果分析

从图 6 中首先可以看到，在利用高斯判别分析训练模型后利用训练集去测试，得到的最终准确率为 81.82%。更详细的情况为，分别考虑正负样本的三个核心指标，首先对于未患有心脏病的样本，在此次训练中的样本总数为 115 个。预测精确率为 86%说明，此模型参数确定后，通过模型预测发现，预测真实未患病个体数量占模型未患病总的预测量的 0.86。高于正负样本混合后的准确率，说明在高斯判别模型对于未患病样本的先验概率的高斯分布模拟较好。对于图 6 中召回率的指标为 74%，由召回率的定义出发可知，预测真实未患病个体数量占实际未患病个体数量的 74%，可以发现召回率，在此模型中存在瑕疵。最后为了综合考虑以上的两个因素，得出了 f1-score 的值为 0.79 非常接近 0.8。由 f1-score 的定义出发可知越接近 0 模型效果越差越接近 1 模型效果越好，因此对于以上两类患病情况的 f1-score 可知，在训练集上的效果较好。同理对于患有心脏疾病的样本来说，有以下结果。预测精确率为 79%说明，此模型参数确定后，通过模型预测发现，预测真实患病个体数量占模型患病总的预测量的 0.79。预测真实患病个体数量占实际患病个体数量的 89%。训练集的总样本量为 242。

Train Result:

Accuracy Score: 81.82%

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.86	0.79	0.82	0.82	0.82
recall	0.74	0.89	0.82	0.81	0.82
f1-score	0.79	0.84	0.82	0.82	0.82
support	115.00	127.00	0.82	242.00	242.00

Figure 6. Experimental results of the training set
图 6. 训练集的实验结果

图 7 中的实验结果是模型在测试集上的表现，真正意义上的体现了高斯判别分析在预测心脏疾病中的预测效果。得到的最终准确率为 86.86%。更详细的情况是，分别考虑正负样本的三个核心指标。首先对于未患有心脏病的样本，在此次训练中的样本总数为 23 个。预测精确率为 86%说明，此模型参数确定后，通过模型预测发现，预测真实未患病个体数量占模型未患病总的预测量的 0.86。对于图 7 中召回率的指标为 78%可知，预测真实未患病个体数量占实际未患病个体数量的 78%。最后为了综合考虑以上的两个因素，得出了 f1-score 的值为 0.82。由 f1-score 的定义出发可知越接近 0 模型效果越差越接近 1 模型

效果越好，因此对于以上两类患病情况的 f1-score 可知，在测试集上的效果较好。同理对于患有心脏疾病的样本来说，有以下结果。预测精确率为 88%说明，此模型参数确定后，通过模型预测发现，预测真实患病个体数量占模型患病总的预测量的 0.88。预测真实患病个体数量占实际患病个体数量的 92%。对于本文核心问题预测二分类的疾病取得了极好的结果。

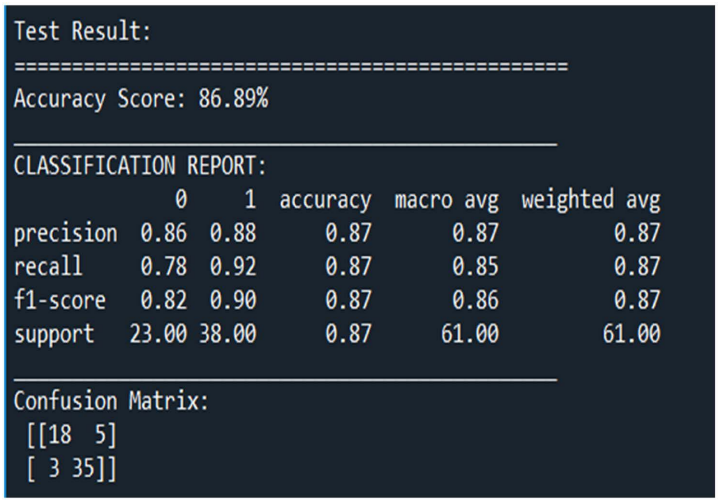

Figure 7. Experimental results of the testing set
图 7. 测试集上的实验结果

4. 结论

本文通过描述性统计首先发现了心脏疾病数据大部分都来自于中年人，中年人由于生活习惯等因素影响已经开始受到心脏疾病困扰。心脏疾病对于不同的性别有着完全不同的影响，女性患病是平均的，但是男性患病的比例高。此类心脏疾病更容易发生在中年男性身上。对于人们一贯认为的运动后引发心绞痛可能是心脏疾病的明显预兆，在数据分析时被完全否定，结果表明运动后容易引发心绞痛的人越不可能患上心脏疾病。静息血压、胆固醇含量和最大心率之间不存在高度的相关性，因此建立模型时不需要考虑多重共线性的影响。最后通过对是否患病的先验性认知建立高斯判别分析模型，发现预测结果在测试集上效果良好，达到了 86.89%，说明这样的模型完全可以从成年人的一般化的体检报告中去预测出是否患有心脏疾病的问题，帮助广大的中年人减少高额专项的心脏疾病检测费用，同时帮助医院减轻了大量的医疗负担。

参考文献

[1] 王学民. 应用多元分析[M]. 上海: 上海财经大学出版社, 2004.
[2] 佚名. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
[3] 吴喜之. 复杂数据统计方法[M]. 北京: 中国人民大学出版社, 2013.