Computer Science and Application 计算机科学与应用, 2021, 11(8), 2108-2116
Published Online August 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.118216

Hans汉斯

应用于心脏病诊断的线性回归决策树模型

闵杰青[1*]，李昕洁[2#]，谭　强[3]，赵　娜[3]，李向娟[1]，王　剑[4]，曾敬勋[5]，刘学承[2]

[1]昆明市儿童医院，云南　昆明
[2]新竹交通大学科技管理研究所，台湾　新竹
[3]云南大学软件学院工程重点实验室，云南　昆明
[4]昆明理工大学信息工程与自动化学院，云南　昆明
[5]英国曼彻斯特大学计算机科学所，英国　曼彻斯特
Email: 24147863@qq.com, #camhero@gmail.com

收稿日期：2021年7月16日；录用日期：2021年8月13日；发布日期：2021年8月20日

摘　要

心脏病是一种十分常见的高发性疾病，已经成为导致人类死亡的主要因素之一。提高心脏病的医疗诊断的准确性，并对其实行更早的干预与治疗是需要关注的问题。在本文中，我们在数据预处理和模型建立前期阶段采用的是python代码实现，最终发现患病比例与性别和年龄也有着一定的联系。然后采用了SPSS对其进行分析，发现R值为0.719，属于0.5~1之间的大效应的情况，因此，模型拟合效果良好。此外，方差分析的显著性值为0，处于0~0.05的范围之内，可以说明各个参数建立的线性关系回归模型具有极显著的统计学意义，即线性关系显著。模型建立的后期阶段采用以决策树为代表的多种预测模型，最终预测准确率如下：基于信息熵的决策树模型为85.6%，基于基尼指数的决策树模型为84.2%，基于基尼指数的决策树(预剪枝)模型为86.6%。我们发现：模型的准确率均在85%左右，其中基于基尼指数的决策树(预剪枝)模型准确率最高。

关键词

变异数分析，线性回归，决策树，智慧医疗

Decision Tree Model Based on Linear Regression for Heart Disease Diagnosis

Jieqing Min[1*], Shin-Jye Lee[2#], Qiang Tan[3], Na Zhao[3], Xiangjuan Li[1], Jian Wang[4], Ching-Hsun Tseng[5], Hsueh-Cheng Liu[2]

[1]Children's Hospital of Kunming, Kunming Yunnan
[2]Institute of Science and Technology Management, National Yang Ming Chiao Tung University, Hsinchu Taiwan
[3]School of Software, Key Laboratory in Software Engineering of Yunnan Province, Yunnan University, Kunming

[*]第一作者。
[#]通讯作者。

文章引用: 闵杰青, 李昕洁, 谭强, 赵娜, 李向娟, 王剑, 曾敬勋, 刘学承. 应用于心脏病诊断的线性回归决策树模型[J]. 计算机科学与应用, 2021, 11(8): 2108-2116. DOI: 10.12677/csa.2021.118216

Yunnan
[4]College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan
[5]Institute of Computer Science, The University of Manchester, Manchester UK
Email: 24147863@qq.com, #camhero@gmail.com

Received: Jul. 16th, 2021; accepted: Aug. 13th, 2021; published: Aug. 20th, 2021

Abstract

Heart disease is a very common high-incidence disease, which has become one of the main factors leading to human death. Improving the accuracy of medical diagnosis of heart disease and implementing earlier intervention and treatment are issues that need attention. In this article, we adopted python code in the early stage of data preprocessing and model establishment, and finally found that the disease ratio is also related to gender and age. Then SPSS was used to analyze it, and it was found that the R value was 0.719, which is a large effect between 0.5~1. Therefore, the model fitting effect is good. In addition, the significance value of the analysis of variance is 0, which is within the range of 0~0.05, which can indicate that the linear regression model established by each parameter has extremely significant statistical significance, that is, the linear relationship is significant. In the later stage of model establishment, a variety of prediction models represented by decision tree are used. The final prediction accuracy is as follows: the accuracy of the decision tree model based on information entropy is 85.6%, the accuracy of the decision tree model based on the Gini index is 84.2%, and the accuracy of the decision tree (pre-pruning) based on the Gini index is 86.6%. We found that the accuracy of the models is around 85%, and the decision tree (pre-pruning) model based on the Gini index has the highest accuracy.

Keywords

Variance Analysis, Linear Regression, Decision Tree, Smart Healthcare

1. 引言

在传统的医疗诊断中，医生往往是根据自己积累的经验以及患者呈现和描述的症状来判断病人病情以及发病原因。然而，这很有可能会导致主观上的判断失误。机器学习主要是基于过去案例的经验进行学习，尤其是基于大数据的机器学习同传统医疗相比有着很大的优势。如果我们将机器学习的分类算法应用于疾病诊断中，那么可以很大程度上提高诊断的准确率，从而帮助人们做出更科学的诊断。

我们使用机器学习可以解决医疗过程中的很多问题，这对患者和医生都有好处。我们经常使用机器学习中的面向图像特征处理的深度神经网络和分类算法应用在医疗诊断领域中。提取医学影像特征、标准化临床数据和转化文本数据等问题都被人工智能有效解决了。

在传统计算机看来，医生的问诊记录、患者的日常护理记录、病理科的检验报告、放射科的 CT 报告等是没有任何意义，而对于人工智能却是有很大意义的。文献[1]使用机器学习的方法诊断糖尿病视网膜病变；文献[2]研究基于集成学习的乳腺癌分类；文献[3]使用感知机算法诊断脊柱病。

本文使用基于线性回归分析的决策树模型，将其应用于心脏诊断，来帮助医生进行治疗。前期利用spss 线性回归来分析数据中变量的关系，最后利用决策树算法来建立模型。但是因为决策树算法有不同的实现方式，因此本文在对比两种不同决策树(一个利用信息熵，一个利用基尼指数)结果后选择表现最优的方式作为最后建立模型的算法。

2. 项目设计

2.1. 总体设计

总体结构设计如图 1 所示：

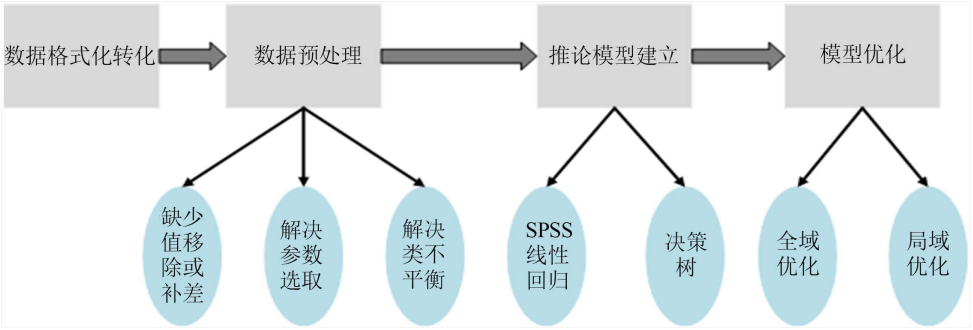

Figure 1. Overall design
图 1. 总体设计流程

2.2. SPSS 数据线性回归分析和结果

本文使用 SPSS 做线性回归分析[4]。

参数说明如表 1 所示：

Table 1. Parameters
表 1. 参数

序号	英文符号	参数
0	age	年龄
1	sex	性别 1 = male，0 = female
2	cp	胸痛类型(4 种)：值 1：典型心绞痛，值 2：非典型心绞痛，值 3：非心绞痛，值 4：无症状
3	trestbps	静息血压
4	chol	血清胆固醇
5	fbs	空腹血糖 > 128 mg/dl，1 = true，0 = false
6	restecg	静息心电图(值 1，2，3)
7	thalach	达到的最大心率
8	exang	运动诱发的心绞痛(1 = yes, 0 = no)
9	oldpeak	相对于休息的运动引起的 ST 值(ST 值与心电图上的位置有关)
10	slope	运动高峰 ST 段的坡度(值 1：uploping 向上倾斜，值 2：float 持平，值 3：downsloping 向下倾斜)
11	ca	主要的血管数量(0~3)
12	thal	一种叫做地中海贫血的血液疾病(3 = 正常，6 = 固定缺陷，7 = 可逆转缺陷)
13	target	生病与否(0 = no, 1 = true)

判定系数[5]一般需要大于60%才行，是判定线性方程拟合优度的重要指标。我们可以用判定系数来解释回归模型因变量变异的能力。我们将判定系数用R表示，其值越接近1越好。表2中第2列即为判定系数。我们得到的结果显示R = 0.719时模型的效果最好。

模型残差独立性[6]检验：如果DW值被包含在无自相关性的值域之中(查询Durbin Watson table)，就认为残差是独立的。本例DW = 1.032，残差是独立的。

Table 2. Linear regression analysis result[b]
表2. 线性回归分析结果[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin Waton
1	0.719[a]	0.518	0.496	0.35419	1.032

a. Predictors: (Constant), thal, restecg, fbs, thalach, chol, trestbps, ca, sex, cp, stope, exang, age, oldpeak. b. Dependent Variable: target.

我们得到的方差[7]分析的显著性值 < 0.01 < 0.05。结果说明线性关系显著，即由自变量与因变量各个参数建立的线性关系回归模型具有显著的统计学意义。分析结果如表3~5所示：

Table 3. ANOVA result[b]
表3. 方差分析结果[b]

Model		Sum of Square	df	Mean Square	F	Sig.
1	Regression	38.893	13	2.992	23.848	0.000[a]
	Residual	36.255	289	0.125		
	Total	75.149	102			

a. Predictors: (Constant), thal, restecg, fbs, thalach, chol, trestbps, ca, sex, cp, stope, exang, age, oldpeak. b. Dependent Variable: target.

Table 4. Significance analysis results of each feature[a]
表4. 各个特征显著性分析结果[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	0.899	0.293		2.830	0.005
	age	−0.001	0.003	−0.015	−0.304	0.761
	sex	−0.196	0.047	−0.183	−4.157	0.000
	cp	0.113	0.022	0.233	5.036	0.000
	trestbps	−0.002	0.001	−0.070	−1.583	0.114
	chol	0.000	0.000	−0.037	−0.838	0.403
	fbs	0.017	0.060	0.012	0.291	0.771
	restecg	0.050	0.040	0.053	1.249	0.213
	thalach	0.003	0.001	0.139	2.671	0.008
	exang	−0.144	0.051	−0.136	−2.804	0.005
	oldpeak	−0.059	0.023	−0.137	−2.564	0.011
	slope	0.079	0.042	0.098	1.863	0.063
	ca	−0.101	0.022	−0.206	−4.603	0.000
	thal	−0.119	0.036	−0.146	−3.339	0.001

a. Dependent Variable: target.

Table 5. Residual statistical results[a]
表 5. 残差统计结果 [a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	−0.3457	1.2750	0.5446	0.35887	303
Residual	−0.94748	0.93509	0.00000	0.34648	303
Std. Predicted Value	−2.481	2.035	0.000	1.000	303
Std. Residual	−2.675	2.640	0.000	0.978	303

a. Dependent Variable: target.

我们需要检验数据是否可以做回归分析。回归分析对数据的要求是十分严格的，所以有必要分析残差。

从图 2 来看，不完全对称的左右两侧；从图 3 来看，散点并没有全部靠近斜线。

综合而言，没有得到最好的残差正态性结果，但是在现实分析当中，理想状态的正态并不多见，接近或者近似就可以接受。

3. 决策树建模和结果

我们最后选择建立模型的算法是：基于信息熵[8]和基尼指数[8]的决策树[9]算法。

决策树算法是一种不断逼近离散函数值的方法，是一种典型的分类算法。第一步先进行数据处理，利用归纳算法生成决策树和可读的规则，然后应用决策分析新数据。从本质上来说，通过一系列规则对数据进行分类的过程就叫做决策树算法。

上世纪 60 年代到 70 年代末，决策树方法产生了。J Ross Quinlan 为了减少树的深度提出了 ID3 算法[10]，但是没有对叶子数目进行研究。C4.5 算法[11]是在 ID3 算法的基础上进行了改进的，大大改进了剪枝技术、预测变量的缺值处理、派生规则等方面，适合于分类问题和回归问题。

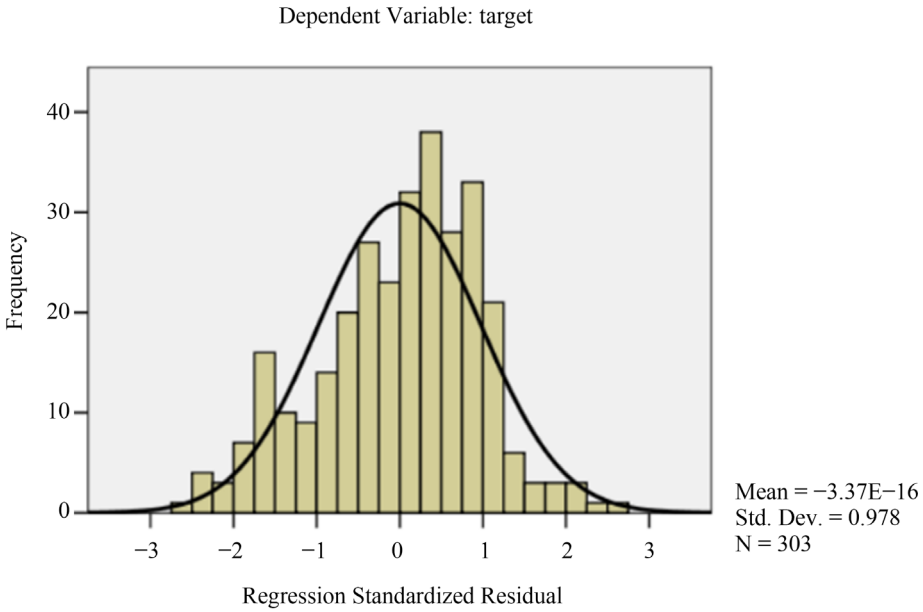

Figure 2. Standardized residual histogram
图 2. 标准化残差直方图

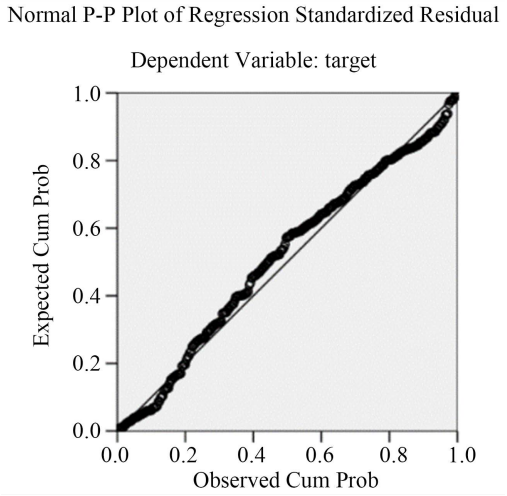

Figure 3. P-P plot of standardized residuals
图 3. 标准化残差的 P-P 图

决策树算法发现数据中蕴涵的分类规则是通过构造决策树实现的，其核心内容是构造精度高且规模小的决策树。构造决策树可以分为两步：第一步，生成决策树：决策树由训练样本集生成。一般情况下，有历史的、达到某个综合水平的和用来进行数据分析处理的数据集作为训练样本的数据集。第二步，对决策树进行剪枝：这个过程实际上是对获得的决策树进行检验、校正和修饰。我们用测试数据集(与原来不同的样本数据集)中的数据测试决策树生成过程中产生的初步规则，剪除那些影响预衡准确性的无用分枝。

模型选择

结果中的信息有一下几点：

1) precision——查准率/准确率

2) recall——查全率/召回率

3) F1-score——基于查准率和查全率的调和平均

4) Support——样本标签(本文中的标签是 0/1)出现的次数

5) Accuracy——模型准确率

6) macro average——宏平均值，即所有标签结果的平均值

7) weighted average——加权平均值，即所有标签结果的加权平均值

通过基于信息熵的决策树模型得到的结果如图 4 所示：

0.8561872989698997

	precision	recall	f1-score	support
0	0.84	0.77	0.80	113
1	0.87	0.91	0.89	186
accuracy			0.86	299
macro avg	0.85	0.84	0.84	299
weighted avg	0.86	0.86	0.85	299

Figure 4. Result of information entropy based on decision tree model
图 4. 信息熵的决策树模型的结果

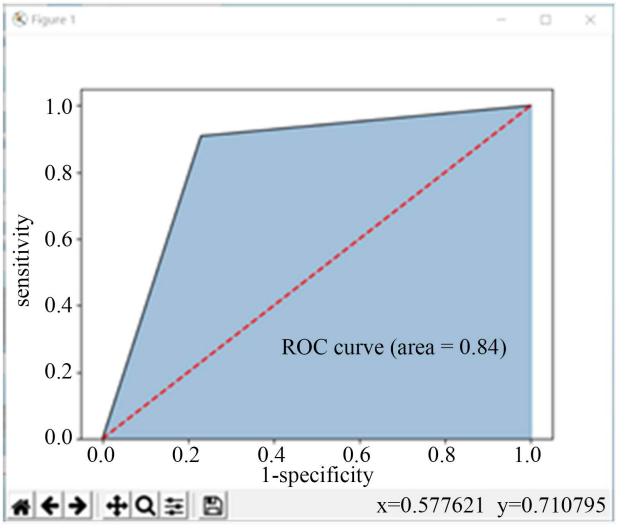

Figure 5. ROC curve of information entropy based on decision tree model
图 5. 信息熵的决策树模型的 ROC 曲线

基于基尼指数的决策树的结果：

0.842809364548495

	precision	recall	f1-score	support
0	0.81	0.77	0.79	113
1	0.86	0.89	0.88	186
accuracy			0.84	299
macro avg	0.83	0.83	0.83	299
weighted avg	0.84	0.84	0.84	299

Figure 6. Result of Gini index based on decision tree model
图 6. 基尼指数的决策树模型的结果

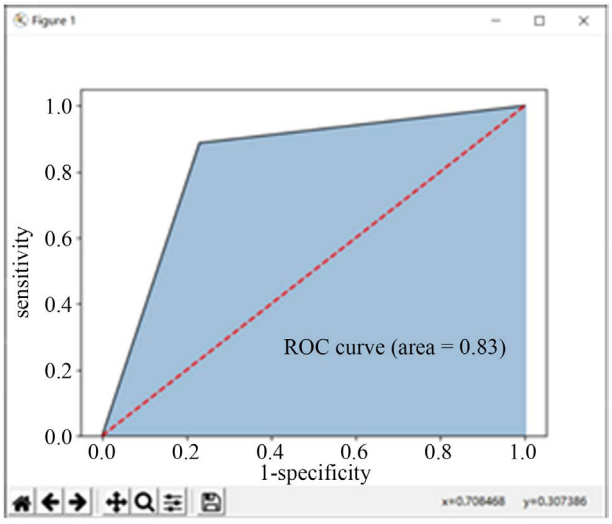

Figure 7. ROC curve of Gini index based on decision tree model
图 7. 基尼指数的决策树模型的 ROC 曲线

基于基尼指数的决策树(预剪枝)的结果：

0.8662207357859532

	precision	recall	f1-score	support
0	0.84	0.80	0.82	113
1	0.88	0.91	0.89	186
accuracy			0.87	299
macro avg	0.86	0.85	0.86	299
weighted avg	0.87	0.87	0.87	299

Figure 8. Result of Gini index based on pre-pruning decision tree model
图 8. 基尼指数的预剪枝决策树模型的结果

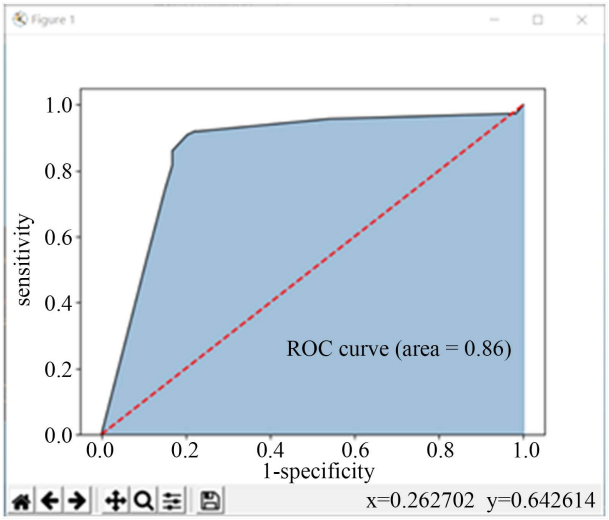

Figure 9. ROC curve of Gini index based on pre-pruning decision tree model
图 9. 基尼指数的预剪枝决策树模型的 ROC 曲线

通过对比准确率、召回率、ROC 曲线图(如图 5~9)可以看到，对于本文的数据，两种决策树的结果比较接近，但最终还是基于基尼指数的决策树的效果要略微好于基于信息熵的决策树。因此，本文选择建立模型的算法是基于基尼指数的决策树。

5. 结语

本文使用基于线性回归分析的决策树模型应用在心脏疾病诊断。我们最后实现的方式是通过决策树算法，这是为了避免过拟合，也因为资料是关联式资料库的关系，经过特征筛选后，我们选择了决策树模型。此模型经过预剪枝后，稳定度提升，准确度与其他机器学习模型相比要稍微好一点。虽然最后的准确度仍然没有达到 90%或者更高，但这些结果也可以说明机器学习在疾病诊断方面的可行性。

此外，本文所用的模型只有一个，要想进一步提高模型对策准确度，我们可以将它与另外的模型相结合(比如 SPSS、朴素贝叶斯、BP 神经网络等等)，应该会取得不错的效果。

基金项目

昆明市卫生健康委员会卫生科研课题项目(2020-09-04-112)；云南省重点研发计划“基于智慧医疗平

台的儿童疾病智能诊疗体系构建及应用示范”；中国博士后科学基金(2020M673312)；云南省博士后基金；云南大学“东陆中青年骨干教师”基金(C176220200)；云南省软件工程重点实验室开放基金资助项目(2020SE311)；云南省自然科学基金项目(202101AT070167)。

参考文献

[1] Cao, K. (2019) Artificial Intelligence on Diabetic Retinopathy Diagnosis: An Automatic Classification Method Based on Grey Level Co-Occurrence Matrix and Naive Bayesian Model. *International Journal of Ophthalmology*, **12**, 1158-1162.

[2] 邓卓, 苏秉华, 张凯. 基于集成学习的乳腺癌分类研究[J]. 中国医疗设备, 2020, 35(12): 59-62.

[3] Yu, Y.X. (2019) The Application of Intelligent Medicine of Perceptron Algorithm in the Diagnosis of Spinal Disease. *China New Telecommunications*, **21**, 229-231.

[4] 使用 SPSS 进行线性回归分析[EB/OL]. https://jingyan.baidu.com/article/b2c186c8055f49c46ef6ff0b.html, 2021-01-30.

[5] R 做线性回归[EB/OL]. https://www.sohu.com/a/230584172_274950, 2021-01-30.

[6] D-W 检验[EB/OL]. https://baike.baidu.com/item/D-W%E6%A3%80%E9%AA%8C/8030379?fr=aladdin, 2021-01-30.

[7] 方差[EB/OL]. https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE, 2021-01-30.

[8] 杜小芳, 陈毅红. Spark MLlib 中决策树算法不同特征选择标准比较[J]. 太原师范学院学报(自然科学版), 2020, 19(4): 37-39+51.

[9] 张振, 田雪飞, 郜文辉, 何凤姣, 邓天好, 宋晓燕, 郑飘, 黄振. 基于决策树及贝叶斯网络建立原发性肝癌肝郁脾虚证诊断模型研究[J]. 中国中医药信息杂志, 2020, 27(9): 115-120.

[10] Ren Y.X., Wang, S.Y., Luo, Y.T. and Chen, S.Y. (2020) ID3 Algorithm-Based Research on College Students' Mobile Game Preferences and Analysis of Circumvention Paths. *Academic Journal of Engineering and Technology Science*, **3**.

[11] Afrianto, E., Suseno, J.E. and Warsito, B. (2020) Decision Tree Method with C4.5 Algorithm for Students Classification Who Is Entitled to Receive Indonesian Smart Card (KIP). *IOP Conference Series*: *Materials Science and Engineering*, **879**, Article ID: 012072.