数据挖掘技术在医疗诊断中的应用——感知机模型诊断心脏病

张一宁

（绍兴市第一中学，浙江绍兴，312000）

摘要：时代在进步，信息、科学技术在快速发展，自从步入大数据时代以来，各个领域的信息量以指数形式增多，人们已不满足于使用传统的统计方法来分析复杂繁多的信息，高效地从大数据中提取出真正有价值的信息是各行各业亟需解决的问题，医疗领域也是如此，机器学习方法应运而生，它是一种极其重要的数据挖掘技术。本研究利用感知机机器学习算法来分析心脏相关数据，最终得出受测人是否有心脏病的判断，该系统既可以使受测人本人能及时判断自己是否具有该疾病，也便于医院医生进行辅助诊断。

关键词：医疗数据挖掘；心脏疾病诊断；机器学习；感知机算法

DOI:10.16589/j.cnki.cn11-3571/tn.2019.04.003

0 引言

现如今各种疾病的出现，严重影响了人们的身体健康，精确并迅速地判断出个体的患病情况是十分重要的，医生的误诊或拖延诊断将使患者付出生命的代价。以一言蔽之，我们需要一个高效、高准确率的医疗诊断系统来为我们出力，而机器学习可以很好地弥补医疗领域发展中的漏洞。

人们对于机器学习的态度从最初的怀疑，到后来的震惊，再到现在的叹服，但大多数人对于机器学习确实如以管窥天，一知半解。机器学习是一种十分重要的数据挖掘技术，将其应用在医疗领域可显著提高疾病诊断的效率和精确率，更好地服务百姓、造福百姓。因此，本研究具有极其重要的现实意义。

随着科技水平的提高和医疗领域的快速发展，传统的医疗诊断分析越来越不适应当前的生活节奏，不能满足大众更好的医疗需求，“人工智能＋医疗”的命题不断浮出水面，逐渐被大众所接受和使用。

到目前为止人工智能在医疗领域已取得了可观的成就。比如，来自瑞士和荷兰的研究人员通过先进的人工智能软件和机械线束，设计了一款以智能辅助技术帮助中风和脊髓损伤患者重新走路的外部装备；IBM 同阿尔伯塔大学的研究人员利用机器学习算法分析功能性磁共振成像，创建了识别大脑与精神分裂症相关的模型，对精神分裂症进行预测达到了 74% 的准确率。由此可以看出，机器学习正一步步走向我们的世界，并改善我们的生活。

本研究将机器学习技术和医疗领域进行了有机的结合，旨在创造一种与受测人进行人机交互的方式，收集受测人的输入数据即受测人的各项身体指标，并通过感知机算法分析该数据，构建心脏病辅助诊断模型，最后系统利用该模型鉴别出受测人是否有潜在的心脏疾病。

机器学习技术顺应“人工智能＋医疗”的发展趋势，它能主动研究和分析数据信息并对样本标签做出高准确率的智能判断，为医疗诊断提供了有效的支持与技术基础，有效提升了医疗效率，改善了医疗质量。

1 训练数据集介绍

■ 1.1 数据来源及训练数据集介绍

本文选用 UC Irvine Machine Learning Repository 网站的心脏病数据作为本文的训练数据集，该数据集包含 270 个受测人样本，每个样本中包含 13 个与心脏病有关的特征，该特征包括了年龄、性别、血压、心率等 13 种个体指标，其中有 7 个连续型特征，6 个离散型特征。可以较全面地反应每个个体的心脏状况，从而训练出更加准确的机器学习模型。该数据集目标为预测受测人是否有心脏病，标签 1 表示不患有心脏病，标签 2 表示患有心脏病。

数据集特征字段及字段解释如表 1 所示。

表1　数据集特征字段解释

特征字段	字段解释
age	受测人的年龄
sex	性别(1表示男性，0表示女性)
cp	胸痛苦类型(1~4分别表示程度由低到高)
trestbps	静息血压(单位为：mm/Hg)
chol	血清胆固醇(单位：mg/dl)
fbs	空腹血糖(苦空腹血糖＞120mg/dl则用1表示，若空腹血糖＜120mg/dl则用0表示)
restecg	静息心电图结果(用0、2表示)
thalach	最大心率(单位为：次/秒)
exang	运动是否诱发心绞痛(若会诱发则用1表示，若不会诱发则用0表示)
oldpeak	运动引起的ST段压低
slope	ST段斜率最大值
ca	荧光检查标注的主要血管数量(单位为：个)
thal	3表示正常；6表示固定缺陷，7表示可逆缺陷

■ 1.2 数据预处理之标准化

数据集由于存在量纲不同的影响，不可直接被机器学习所应用，应将数据集标准化，从而保证数据本身的特点不会影响机器学习的预测。标准化的公式：

$$X' = \frac{X - \bar{X}}{\sigma}$$

其中，X' 表示标准化后的特征值，X 表示原始特征值，

$\bar{X}$ 表示数据所在列的平均值，σ 表示数据所在列的标准差。

2 感知机模型

■ 2.1 感知机算法介绍

20 世纪 50 年代，美国神经学家 Frank Rosenblatt 提出了可以模拟人类感知能力的机器，并称之为“感知机”，感知机算法随之诞生。感知机算法是一个基本的二分类机器学习算法，该算法原理为：找出一个 n 维分离超平面可以基本将训练集特征空间中所有样本点分为正负两类，该平面即对应一个感知机模型，从而利用该模型对新样本进行分类。

■ 2.2 感知机模型的构建

感知机模型即决策函数：

$$y=sigh(w \cdot x+b)$$

其中，w 为权值向量，b 为截距，x 为样本的特征向量，y 为样本的标签，取 +1 或 -1；sigh(m) 为符合函数，当 m>0 时，sigh(m)=+1；当 m<0 时，sigh(m)=-1。

感知机模型示意图如图 1 所示。

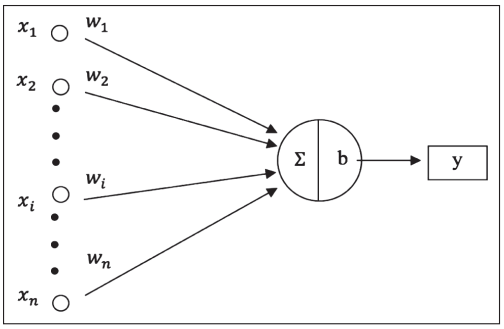

图 1　模型预测示意图

感知机算法的任务即利用心脏病训练数据集来选出使模型构建更准确的参数 (w,b)。

2.2.1 构建损失函数

本研究使用的损失函数为所有被误分类的点到超平面的距离之和，衡量超平面分类失败程度，损失函数的值越大，超平面的分类效果越差。其中误分类点表示训练集中被错误分类的点，表示为 $y_i(w \cdot x_i+b)<0$，从而得出损失函数的表达式为：

$$L(w,b)=-\sum y_i(w \cdot x_i+b)$$

损失函数值越小，误分类点就越少，说明得到的超平面分类效果越好。因此为了找到最好的超平面，需将损失函数取最小值。

2.2.2 梯度下降法

本研究使用梯度下降法来最小化损失函数，从而达到模型参数估计的目标。

梯度下降法是一种迭代方法，它可以通过不断改变参数 (w,b) 值，使得损失函数的值越来越小。梯度方向是使损失函数 L(w,b) 增加最快的方向，因此本文选用梯度方向的负方向作为 (w,b) 的更新方向，因此得到的 (w,b) 就可以使损失函数 L(w,b) 减少最快。根据损失函数表达式，其梯度方向的公式为：

$$\frac{L(w,b)}{w}=-y_i x_i$$

$$\frac{L(w,b)}{b}=-y_i$$

之后在梯度的负方向即 ($y_i x_i, y_i$) 对 w 和 b 进行迭代更新，迭代公式为：

$$w \leftarrow w+\eta \ y_i x_i$$

$$b \leftarrow b+\eta \ y_i,$$

其中，η 表示参数学习率，由人为确定。

不断重复该过程，直到损失函数的值小于规定的阈值，此时样本中基本所有样本点都被正确分类。

■ 2.3 利用感知机模型进行数据分析

根据上文，求得感知机算法模型：y=sigh(w • x+b)，接下来利用该模型对新受测人的输入信息进行分析。收集受测人的各项身体特征，将新数据的特征值代入得到的感知机模型中，根据决策函数计算出该样本对应的标签即可。

■ 2.4 模型评估及训练结果

本研究使用的模型评估方法为交叉验证法，该方法将原始心脏病训练数据集分成训练集和测试集两部分，分别为 80% 和 20%，样本个数分别为 216 和 54。其中训练集用来训练模型，测试集用来衡量模型的质量。模型评估指标选用准确率和召回率，其中准确率表达式为：

$$准确率=\frac{m'}{m},$$

其中，m′ 表示测试集中分类正确的样本数，m 表示测试集总样本数。

当测试集样本正负类样本个数不均匀时，准确率往往不能科学地得到最终结果，因此我们选用召回率作为第二个模型评估指标，召回率表达式为：

$$召回率=\frac{n'}{n},$$

其中 n' 表示测试集分类正确的正类个数，n 表示测试集正类总数。

本研究模型训练结果如图 2 所示。

```
权值向量: w = [[-1.39583333  3.          2.33333333  4.88679245  1.10045662  1.
   0.5        -3.84732824  2.         -0.33870968  1.5         3.33333333
   3.75       ]]
偏置: b = [-6.]
该感知机模型正确率为: 0.78
召回率为: 0.8095238095238095
```

图 2　模型训练结果

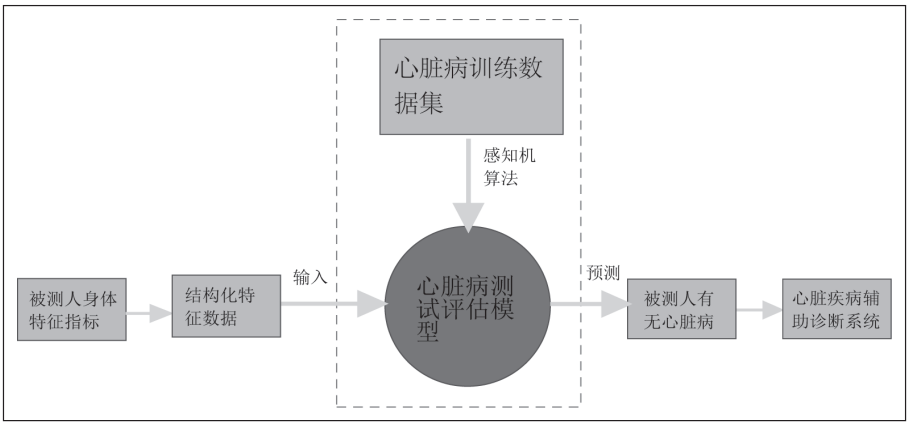

图 3　总流程图

训练得出模型的权值和偏置如图 2 所示，根据该结果可知，模型的正确率稳定在 0.75 ～ 0.80 之间，召回率也相对稳定，根据本文的数据状况，此结果较合理且令人满意。

3 可实现的功能

本研究总流程图如图 3 所示。

本研究总流程为：首先从已知心脏病训练数据集出发，通过感知机机器学习算法构建本研究的核心模型，即心脏病测试评估模型。之后收集被测人的心脏相关身体特征指标数据，模型系统自动将该数据封装到 Excel 表格之中，有利于被模型所识别。最后将其输入心脏疾病测试评估模型中，经过模型分析运算，即可预测出被测人是否患有心脏疾病，系统自动出具分析报告并将其发给心脏疾病辅助诊断系统。

4 结语

机器学习是极其重要的数据挖掘技术，在医疗领域，其能帮助医生从大量已知数据中发现并总结出一定的规律，从而提高医疗诊断的准确率，有效减少误判现象的发生。

本文从人工智能医疗的重要性出发，说明了机器学习在医疗领域的重大意义，之后对现阶段国内外的人工智能医疗现状做了简要介绍，并指出了人工智能在医疗领域已取得的部分成就，随后阐述了机器学习应用于医疗的创新点。该课题对训练数据集进行了详细描述，讲解了机器学习中的感知机算法，从几个方面系统讲述了感知机算法的原理。最后本文通过图文形式对研究总流程进行了总结。

我国科技发展还处于上升阶段，因此各领域数据集不够完善、现存的数据集格式不标准、既懂医疗和计算机技术的复合型人才缺少等都是需要弥补的漏洞，共同导致机器学习模型构建存在误差。但是，笔者相信，随着机器学习的应用范围的不断扩大，在这个科技快速发展的时代，机器学习在医疗方面的应用一定能取得突破，在当代大放光彩，对机器学习的讨论和研究，也一定能助推当代人工智能与科技的发展，它能改变人们的传统观念，带领人们早日进入高度智能化的时代。

参考文献

＊ [1] 周志华 . 机器学习 [M]. 北京：清华大学出版社 .2016.

＊ [2] 李航 . 统计学习方法 [M]. 北京：清华大学出版社 .2012.

（上接第 38 页）

ical editor for dance notation[J].In Robot and Human Interactive Communication,Proceedings,IEEE International Workshop,2002:59-64.

＊ [3]Maranan,D.,Coyle,M.,Calvert,T.:A Tool for Translating Dance Notation to Animation.Proc.Western Computer Graphics Symposium,2002.

＊ [4]Wilke,L.,Calvert,T.,Ryman,R.,Fox,I.:From dance notation to human animation:The LabanDancer Project. Computer Animation and Virtual Worlds,vol.16,2005:201-211

＊ [5]Hachimura K,Nakamura M.Method of generating coded description of human body mo-tion from motion-captured data [C].//Robot and Human Interactive Communication,2001.Proceedings.10th IEEE International Workshop on.IEEE,2001:122-127.

＊ [6]Worawat Choensawat,Minako Nakamura,Kozaburo Hachimura.GenLaban:A tool for generating Labanotation from motion capture data[J].Multimedia Tools and Applications.2014.

＊ [7]Chen H,Qian G,James J.An Autonomous Dance Scoring System Using Marker-based Motion Capture[J].IEEE Workshop on Multimedia Signal Processing,2005:1-4.

＊ [8]Jiaji Wang,Zhenjiang Miao,Hao Guo,Ziming Zhou,Hao Wu,“Using automatic generation of Labanotation to protect folk dance,”J.Electron.Imaging 26(1),011028 (2017),doi: 10.1117/1.JEI.26.1.011028.

＊ [9] 周子鸣 . 基于动态规划的拉班舞谱自动生成研究 [D]. 北京交通大学 ,2017.

＊ [10] 郭浩 . 基于人体运动捕捉数据自动生成拉班舞谱的研究 [D]. 北京交通大学 ,2015.