Software Engineering and Applications 软件工程与应用, 2021, 10(5), 622-633
Published Online October 2021 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2021.105067

Hans汉斯

基于PCA的Bayes分类器应用于心电图临床诊断

闵杰青[1]，李昕洁[2*]，蒋嘉欣[3]，蒲应明[3]，李向娟[1]，刘凯华[3]，曾敬勋[4]，刘学承[2]

[1]昆明市儿童医院，云南 昆明
[2]杨明交通大学 科技管理研究所，台湾 新竹
[3]云南大学软件学院 云南大学软件学院软件工程重点实验室，云南 昆明
[4]英国曼彻斯特大学计算机科学所，曼彻斯特

收稿日期：2021年8月18日；录用日期：2021年10月11日；发布日期：2021年10月19日

摘　要

心血管疾病(CVD)是一种常见的慢性疾病，初期没有明显症状，发展较慢难以发现，且发病危险性高。因此检查环节十分重要，其中动态心电图采集大量心电数据，大量的心电数据在支持各种心脏疾病诊断的同时，却也提升了人力物力的分析成本，需耗费大量医师人力，大大降低诊断效率，加上人工检查可能因疲劳或分心产生失误，降低可靠性。为了更有效率的运用医师人力、减少人工误差、提高医疗水平质量、解放医师人力与更有效的运用医疗资源以惠及广大患者，设计出了如何利用患者的ECG数据：对资料进行特定的预处理，接着将数据汇入Matlab和SPSS进行主成分分析，之后使用贝叶斯分类器对机器进行训练，并给出运行诊断的结果准确率为75.8%。

关键词

主成分分析，朴素贝叶斯算法，智慧医疗

Naive Bayes Classifier Based on Principal Component Analysis Applied to Clinical Diagnosis Decision of ECG Data

Jieqing Min[1], Shin-Jye Lee[2*], Jiaxin Jiang[3], Yingming Pu[3], Xiangjuan Li[1], Kaihua Liu[4], Ching-Hsun Tseng[4], 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, Xinzhu Taiwan
[3]Key Laboratory in Software Engineering, School of Software, Yunnan University, Kunming Yunnan
[*]通讯作者。

文章引用: 闵杰青, 李昕洁, 蒋嘉欣, 蒲应明, 李向娟, 刘凯华, 曾敬勋, 刘学承. 基于 PCA 的 Bayes 分类器应用于心电图临床诊断[J]. 软件工程与应用, 2021, 10(5): 622-633. DOI: 10.12677/sea.2021.105067

[4]Institute of Computer Science, The University of Manchester, Manchester

Received: Aug. 18th, 2021; accepted: Oct. 11th, 2021; published: Oct. 19th, 2021

Abstract

Cardiovascular disease (CVD) is a common chronic disease with no obvious symptoms at the initial stage, slow development and difficult to detect, and high risk of morbidity. The check process is very important, dynamic electrocardiogram (ecg) collects ecg data, large amounts of ecg data are used in support of various heart disease diagnosis, which promotes the analysis of the manpower cost, costs a lot of physician manpower, greatly reduces the efficiency of diagnosis, and artificial check may cause failure or lower the reliabilitydue to fatigue or distraction. In order to use the physician manpower more efficiently, to reduce the manual error, to improve the quality of medical treatment, to liberate the physician manpower and to use medical resources more effectively to benefit the patients, how to use the ECG data of patients is designed: after a specific preprocessing of the data, the data was imported into Matlab and SPSS for principal component analysis. After that, the machine was trained by Bayesian classifier and the accuracy rate of operation diagnosis was 75.8%.

Keywords

Principal Component Analysis, Naive Bayes Algorithm, Smart Medical

1. 引言

心血管疾病(CVD)是一种常见的慢性疾病，初期没有明显症状，发展较慢难以发现，且发病危险性高，是人类的主要死因之一。

据世界卫生组织统计，心脏疾患是人类的前三大死因，目前全球死亡病例中，有三分之一来自 CVD，而且该比例逐年增加中。1996 年全世界有 1700 万人死于 CVD (冠心病占 4%)，死亡人数是艾滋病的 6 倍，在亚洲国家，心脏疾病死亡比例 CVD 患者占 44%。随着我国人民的饮食生活水平提高，心脏疾病发病的机率也提高了，鉴于心脏疾病产生的死亡率与发病率日亦常见，如何提升心脏病的诊断效率，及早处理提升 CVD 防治已成为当今医学界重要的议题之一[1]。

目前在临床上，心电图(ECG)时常被用于心血管疾病的诊断。传统的心电图诊断需要将心电信号纸上的背景栅格印出，经过医师计算得到各种心电参数，包含心电 R 波的幅值、RR 波间期，P 波、T 波幅值后，然后利用这些参数结合医师自身经验，再辅以病人的其它数据，综合参考后得出诊断结果[2]。从心电图诊断的诊断过程可以发现，现行的心脏病诊断，虽然可以透过现代医疗检查设备提升诊断的正确率，但最终结论，主要还是来自于医生主观的判断。由于心电图包含众多种类、波形变异极大，若未经训练

则不易通过肉眼观察判别，不同的 ECG 心电信号图之间差异很大，同样的疾病产生的心电图可能十分不同，就算是来自同一患者的 ECG 图，也可能存在差异，再加上不同医师在主观判断上可能不同，使得对同一病人，不同的医生可能得到截然不同的结论，进而产生误诊、漏诊，因此欲做出准确诊断，只能依靠医师自身的医学训练、医学知识和大量临床经验。此外，动态心电图采集大量心电数据，大量的心电数据在支持各种心脏疾病诊断的同时，却也耗费大量医师人力，大大降低诊断效率。并且，长时间的分析往往导致视力疲劳和(或)注意力分散造成漏检，因此造成可靠性降低。为了有效改善现有诊断方法，解放医师人力、使患者得到更多更加公正统一的诊断结果、提高医疗资院的运用效率、造福更多患者，我们尝试研究如何在心电诊断图的程序上用计算器与自动化取代部分医师人力[3]。

心电图自动化的主要目标是从大量的动态心电图采集结果中，对一些简单的典型的心电异常情况直接给出诊断信息，作为医生诊断的参考，协助医生将注意力集中到变异波形的分析上，而且心电图自动化诊断可以成为家庭日常应用，给使用者合理的就医建议，避免延误就医[4]。科技日新月异，每天每个领域都有新的进步与研究发现，其中计算机技术发展至今，其不断增益的计算分析能力为各个领域的进一步研究提供了更多，尤其是计算机智能技术这类需要高计算能力的领域，随着计算器演算能力的进步，而有深入发展的可能，因此现在，使用计算机辅助诊断不仅仅是为了处理医疗需求，也拥有足以支持此需求的技术，使用计算机辅助诊断在减少医师工作量的同时，又能加快诊快速度，切合市场需求[5]。心电图自动诊断是计算机智能 AI 技术在医学上的应用较为成功的例子之一，它结合了传感器技术、AI 视觉辨识、信号转换处理技术以及人工智能的逻辑判断技术等最新的研究成果。唯一美中不足的是，目前心电图自动检测及分析系统有许多缺陷尚待克服，例如特征萃取、有效提取信号并去噪、诊断分析方法等，因此本论文在心电图自动诊断上尝试新的方法以提升波形识别的准确率，并满足诊断所需[6]。

2. 相关研究

利用主成分分析降维算法提取特征，再用 K-近邻算法，随机森林，Logistic 回归，支持向量机算法进行分类研究，表明综合预测准确率，召回率，精准率，以及 ROC 曲线模型评价指标来看，支持向量机在模型预测中的表现优于其他 3 种算法[7]。采用 MIT-BIH 数据集进行模型训练并评估算法性能。结果样本扩增和使用带有权重系数的损失函数能够提升模型的召回率和特异性指标，同时保持模型对室性异位搏动(VEB)和室上性异位搏动(SVEB)分类的精确率的指标可以辅助医护人员诊断心脏疾病[8]。

使用基于 DenseNet 的分类模型对多个标签分别训练二分类器，完成多标签分类任务，对数据的正异常识别准确率可以达到 80.13%，灵敏度，特异度和 F1 分别为 80.38%，79.91%和 79.35% [9]。心电图(Electrocardiogram, ECG)被广泛应用于窦性心动过速，室性早搏和心房颤动等心律失常诊断中，进而在心脏疾病诊断分析方面展现巨大的临床应用价值，一种基于 LightGBM 的心电信号分类算法。该算法从心电图中提取单心拍特征，心律波动性特征以及全波形特征建立混合特征集，并采用 LightGBM 实现正常心拍，心房颤动，其他心律不齐，噪声四个类别的分类，该算法的性能指标在 PhysioNet/CinC Challenge 数据集上达到 0.824，优于 CART 和 CatBoost 算法[10]。基于卷积神经网络(Convolutional Neural Network, CNN)的 ECG 分类方法在准确率上已达到很好的标准，LSTM 单元兼顾信号全局特征及局部特征，对于特征提取更具有稳定性和可靠性[11]。

吴恩达团队在 AI 医疗方面取得了革命性突破，完成了心律失常诊断。让 AI 输入心率数据，就可以判断出是否心律失常、具体是哪一种情况。准确度高达 83.7%，超过了人类心脏病医生的 78.0% [12]。人工智能系统进行自我训练，使用 78%的患者数据来寻找发病模式并构建自己的诊断指导系统。机器学习方法预测心脏病发作的准确率优于传统医生诊断标准。被人工智能系统认定为心脏病发作高危因素的严

重神经疾病、口服皮质类固醇等因素都没有在美国心脏病协会的指导方针中[13]。采用基于 CNN 的 SEEG/EEG 脑电数据处理分析，作者根据 7:3 划分训练集和测试集，实验准确率达到 97.7%。给出任意一个的睡眠脑电数据，根据识别方法有比较高的识别率。和心电数据处理方法有类似之处[14]。

采用人工蜂群算法对 SVM 中惩罚因子以及核函数高斯函数的宽度参数进行调优，通过对处理过的 900 × 6 的训练集样本运用 SVM 模型进行训练，通过训练后的模型对测试数据进行分类检测，其检测准确率达到了 98.67%。该方法能够有效的实现对心电信号的常见异常类型检测。实验总结认为该方法对正常心拍的检测准确率较高，但对于左束支传导阻滞和右束支传导阻滞的相互检验以及其他类型病情的检测水平还有待提高[15]。采用人工智能中的深度学习算法对心电信号进行识别分类，以提高心电信号的识别率与实时监测。由 MIT-BIH 数据库中处理获得的 106,428 组心拍被随机分为 70,000 组训练样本和 36,428 组测试样本。测试各神经网络对心电信号识别效率，结果显示，四种深度神经网络对心电信号的总体识别率都达到了 95%以上，表明四种深度神经网络对心电信号的识别分类具有良好的性能，其中尤以 CRNN 对心电信号的识别效果：最好总体识别率达到 98.81%，其泛化能力和收敛性均较好[16]。

利用短时傅里叶变换计算心电波形对应的心电谱图，从而获取信号的时频特性；然后，设计了两种不同的卷积神经网络结构，基于心电谱图对 ECG 信号分类问题进行建模。实验结果表明，相比于经典的 PCA 和 SVM 方法，在 20 类 ECG 分类任务上平均分类准确率可以达到 98%以上，处理单个样本时间 1.4 ms 左右，显示出了较高的精度和效率优势[17]。

结合 BP 神经网络算法和 PanTompkins 算法展开移动式心电监护预警系统的研究，使用 MIT-BIH 国际标准 ECG 数据作为数据源，并首先使用 PanTompkins 算法检测 ECG 信号。该算法针对房早和室早的自动识别分类准确率达到了 93.33%，具有参考意义，可应用到实践中[18]。利用前向多层神经网络的径向基函数算法(Radial-Basis Function)，即 RBF 算法并利用 MIT-BIH (美国麻省理工学院提供的研究心律失常的数据库)心电图数据库训练神经网络，使 RBF 神经网络对未训练过的心电图有较好的分类能力。这种方法用于心电图的分类取得较好的效果，平均分类识别准确率达到 92.6% [19]。

考虑到数据集较小且多变的特点故基于 gcforest 算法对心律异常类型进行分类。通过 MIT-BIH 数据库中的数据，对多任务学习和 gcforest 算法进行了性能评估。多任务学习算法能更好地识别心律异常类型，其中 gcforest 算法的准确率达到了 99.5%，多任务学习的准确率则达到了 99.56% [20]。一种改进遗传 SVM 识别分类器的心拍辅助识别系统，实现了 6 类常规心电信号心拍的辅助识别，进而作为医师的辅助手段去准确判断心律异常的类型，结果表明能较好的对心电信号心拍进行比较准确的识别，但改进遗传 SVM 分类识别的算法准确率更高，对于一组测试数据也能够达到实时识别。心电分类识别达到的平均准确率是 92.7% [21]。

经过不同算法研究对比，贝叶斯网络在 2010 年前为解决精准医疗的机器学习主流模型。鉴于深度学习之高计算成本，以及所谓的梯度不稳定问题处理。本文提出传统的贝叶斯网络及特征选取方法，对 ECG 数据做精准医疗的分类。基于奥克姆剃刀理论基础上，在精确度不降低太多的情况下，尽可能减少计算成本。

3. 背景知识

3.1. 心电图基础知识

3.1.1. 心电图

正常人体，窦房结会引发心脏兴奋，每个心动周期会将兴奋引发的电流导向心脏周围，电变化传播的途径、方向、时间都有一定的规律，心电图的原理即是量测心脏周围电位变化，不同的导电组织和体液间有不同的电位变化，身体表面的电极贴片会捕捉流经身体表面的电流并纪录下来。要注意的

是这些生物电位变化，与心脏的收缩物理活动没有直接关系。简单来说，按照心脏兴奋电流的传导顺序和途径，会被身体表面的电极贴片记录下来，绘制成沿着时间轴的连续波形图即为心电图。心电图可以协助诊断心脏肥大、心肌缺血、心律不整、心肌梗死、判断心肌梗死的部位，判断药物或电解质情况。

3.1.2. 心电图各项指标

心电图纸会沿着时间轴记录心脏兴奋电流的连续波形变化，一个心动周期内呈现一组完整的波形。一组典型的完整心电波形包括 P 波、QRS 波群、T 波、以及在 50%~75% ECG 图中会出现的 U 波、二段波(P-R, S-T)、P-R 间期以及 Q-T 间期。

一、P 波：

P 波是一个心动周期内，出现的第一个偏离波，是心房除极的过程，P 波的波形较矮且钝，时有轻度切迹。由于 P 波电轴综合向量向左下，多在 45˚~50˚之间，故与 aVR 导联方向相反，I、II、aVF、V5~V6 导联直立，V1~V2 导联可直立、倒置或双向，与 III 和 aVL 导联方向有时也会相反。

二、QRS 波群：

P 波后出现的是 QRS 波群，典型的 QRS 波群组是由三个偏离波组成，并非所有的 QRS 波群都具有三个偏离波，但不论存在几个波形，都称为 QRS 波群，其变化复杂且波幅较大，反映心室除极相关的电流。QRS 波群组中，第一个偏离波若为负向波，则称为 Q 波，是第一个负向波，接着第一个正向的偏离波称为 R 波，最后是 S 波，R 波后的负向波称为 S 波，是第二个负向波，如果 S 波之后再出现一个正向波，则称为 R’波；但是在 R’波之后出现的负向波，统一称为 S 波。波形偶有顿挫、切迹或挫折。

三、T 波：

第三个出现的波型是 T 波，代表心室的再极化造成心室舒张的过程。若 T 波越高大，则 T 波时间越长，T 波时间在 0.05~0.25 s 之间。波形方向与 QRS 波群主波方向相同，波峰较钝，波形不对称，上升时较长且缓，下降时较短且陡。同一个心动周期内，T 波不应低于 R 波高度的 1/10，也不应高于 R 波。T 波高耸、T 波倒置(>0.5 V)、T 波双峰……等现象可能是因为血钾过高、早期复极、心肌疾病、低钾血症、心源性猝死、心律失常等警讯。

四、U 波：

一个心动周期内出现的第四个波幅是 U 波，是心室内 Purkinje fiber 再极化的过程，U 波常与 T 波方向一致，但 T 波与 U 波之间约有 0.16~0.25 s 左右的延迟，一个心动周期内振幅最低的通常是 U 波，振幅<0.2 mV，若 U 波高于 T 波称为 TU 倒置，是某些病情的病征之一。U 波有时候不会出现在心电图记录或导联之中，实属常见，学界目前对 U 波发生的机制尚无足够深刻的认识。

五、Ta 波(又称 TP 或 PT 波)：

Ta 波代表心房复极过程所产生的电位变化，Ta 起始于 P 波之后，但方向与 P 波相反，Ta 波较小且 Ta 时刻已经开始心室除极，P 波开始到 Ta 波结束的时间为 0.15~0.45 s，由于 Ta 波在 P-R 段和 QRS 波群的中间，有时会延伸到 ST 段，且 Ta 波波幅很低，常被邻近波群覆盖而不易辨别。Ta 波增大引起 P-R 段移位、传导阻滞之下的 Ta 波出现在 P 波高大的导联上等等都可能有病理含意。

六、P-R 段：

P 波终点至 QRS 波群起点之间的直线时间称为 P-R 段，P-R 段代表心房除极结束至心室除极尚未开始之间的时间段，可以用来评估心房至心室间的传导速度。兴奋电流在 P-R 段的传导非常缓慢，电位差极小，有时小到纪录不出来，由于 P-R 段与 P-R 间期意义近似，因此实验趋向不做常规分析。

七、ST 段：

ST 段，指 QRS 波群结束至 T 波起点之间的与基线等齐的线段，代表心室除极结束到心室复极开始前，心室各部分间没有电位差存在的一段时间。是心脏早期的再极化，ST 段正常会贴近基准线，时而上下移动，上移<0.1 mV、下移<0.05 mV 的范围内尚属正常，但 ST 段波位大幅的上升或下降通常代表有病征，相较起来波段长短则尚无理论发现波段长短上的病理意义。ST 段会受到很多因素影响，例如神经张力、管壁张力、心肌代谢、电解质和药物。

八、P-R 间期：

P-R 间期，指 P 波起点至 QRS 波群起点之间的时间段，代表心房除极开始至心室除极开始前的一段时间，可看作兴奋从窦房结传至新房、心室交界再到达心室，并引起心室兴奋的所需时间，所以也称做房室传导时间，正常值约 0.12~0.20 秒之间，少数人的范围值在 0.11~0.21 秒之间，若 P-R 间期较长，代表房室传导阻滞，P-R 间期随心跳速率、年纪及迷走神经张力的影响而有所不同。

九、Q-T 间期：

Q-T 间期，指 QRS 波群结束到 T 波终点之间的时间，代表心室开始兴奋除极到完全复极再到静息状态所需要的时间，代表整个心缩期的电位变化，Q-T 间期正常时间范围为 0.32~0.44 s，会随心跳速率、年纪及迷走神经张力而产生变化，容易受到药物影响，是药物和离子对心肌影响的指标。心率愈慢 Q-T 间期愈长，反之则愈短。临床应用中常透过 Barett 公式求出：

$$Q-T=0.39\times\left(R-R\right)\pm 0.04s \tag{1}$$

还常用校正的 Q-T 间期(Q-Tc)来纠正心率对 Q-T 的影响：

$$Q-Tc=Q-T/\sqrt{R-R} \tag{2}$$

传统 Q-Tc 的正常上限值为 0.44 s，正常人不同导联间的 Q-T 间期差异最大可达 0.05 s，长度超过则称为 Q-T 时间延长，表示患者有猝死的可能。

3.2. 分类器

分类器应用于如手写数字识别、图片分类、网页分类、垃圾邮件过滤、社交网络用户分组等等，引起了广泛的研究兴趣。分类器是通过在数据空间中寻找分类边界将整个数据空间划分为若干区域，从而将数据依类别特征划出界线，区分成不同分类。通过对训练数据的学习，分类器算法学会如何寻找分类边界，进而能够对未知的测试数据判断所属类别。而训练数据集往往是针对数据样本空间极小规模的采样，很难代表真实的数据在整个样本空间的分布情况。测试数据则往往会分布于整个样本空间。训练样本集中的数据报含的有用信息在很大程度上决定了分类器的能力，但是不同种类的分类器对这些有用信息挖掘能力往往不同。好的分类器需要根据少数的训练样本来对大量的未知的测试资料做出正确预测，这种能力被称为分类器的泛化能力，是衡量分类器表现的一项重要指标。泛化能力强的分类器有更大的实用价值。

在机器学习领域中，许多分类器技术的成熟度都很高：如判定树、人工神经网络、贝叶斯分类器、K 最近邻分类器(KNN)和支援向量机(SVM)等等。这些分类器各有所长，因此不同资料适合不同的分类器，通常需要根据特征空间的维数、特征之间的统计关系、训练集的大小、已知的数据分布的先验知识、时间复杂度的限制等实际应用场景来进行选择。

朴素贝叶斯算法：又称简单贝叶斯法，其假设所有特征在统计上互相独立，亦即所有特征之间彼此都不相关，满足这个假设的前提之下，可以应用统计学的贝叶斯定理来计算条件机率分布，进而完成分类。朴素贝叶斯分类器的优势在于计算量少，模型速度快，由于采用机率的概念，因此在小型数据集也

可表现良好，要注意两点，第一点是贝叶斯机率计算需要满足数据样本之间互相独立，因此若数据样本间相关性高，则分类的准确度会下降。第二点则是朴素贝叶斯算法并无分类规则输出。

M 估计：估计的稳健性(Robustness)概念是指模型调整后的预测结果相应产生的模型误差的波动起伏，因此在宽阔的数据范围内产生的优良估计称为稳健估计。例如在线性模型，若满足独立同分布正态误差，则最小二乘估计(LSE)是有效无偏估计；然而若误差非正态分布时，LSE 不一定是最有效的。但误差分布事先通常不知道，故必须考虑稳健回归的问题。稳健回归(Robust Regression)估计，在面对正态误差时，表现略逊 LSE，但在处理非正态误差时，表现大大优于 LSE。若在在误差项分估计布应用这种稳健特性，能够有效排除异常值干扰。稳健回归中常用的最大似然型的 M 估计来自 DPS。

4. 设计框架

本系统是一个基于机器学习的心脏病诊断系统，根据病人 ECG 数据自动诊断病人是否患有心脏疾病的单机应用。系统主要分为三个模块：数据预处理、特征提取和诊断分类。总体设计流程如图 1 所示。

1) 数据处理模块采用 python 实现，目的是将原始的病人 ECG 信息处理为可供具有特征标注的 xlsx 表格。

2) 特征提取模块使用 matlab 完成，从预处理后的表格中提取用于机器学习训练的特征。

3) 诊断分类模块使用 python 实现，建立分类器模型，使用提取出特征的训练集训练，就可以根据用户给出特征进行机器诊断分类。

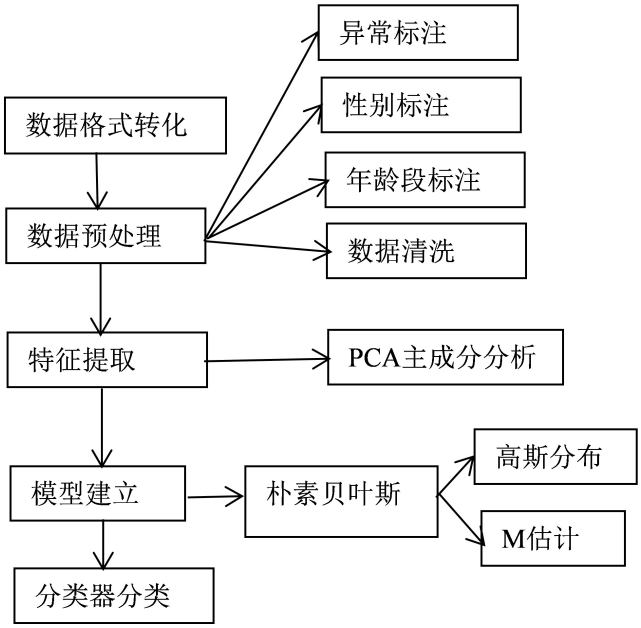

Figure 1. Overall design process
图 1. 总体设计流程

4.1. 数据预处理

1) 先将 xml 档转 csv 档，再把 csv 档转 dat 档。

csv 档是一种含有字符分隔值的文件格式，分隔值可以是逗号或是其他符号，档案存储为内含数据的表格，数据可以是数字或文本，表格储存的是字符序列，而非二元形式的数据。而 dat 文件是二进制文

件，是 ECG 信息记录，网上有关研究的 ECG 数据大量使用的 dat 文件文件保存。

2) 正异常值标注

3) 数据清洗

首先去除与诊断无关的不能作为特征的整列，如患者 ID 等。然后进行缺失值的移除或差补。由于数据庞大且数据的相关性不大，所以缺失值采用了了移除的方法，针对没有分类出正异常的行和有空值的行进行删除。

4) 将数据按年龄进行分段处理。

由于幼儿的心脏功能还不健全，且不同的年龄段有较大的差异，所以按照年龄分类标准进行了分类并标识。

5) 性别的分类。

方便进行数据的读取和模型化。

4.2. 特征提取

向量空间的心电特征提取方法：

1) 主成分分析法 PCA

2) 线性判别分析 LDA

3) 独立成分分析 ICA

采用主成分析法进行特征提取，使用 matlab 进行主成分提取，在此处不考虑性别和年龄两个特征，而对剩余的十三个特征进行提取分析。

将十三个特征进行特征提取，重新生成十三个新的因子，如表 1 所示，第一个因子可以包括原空间的 57.29%，前八个特征可以表示原空间的 99.10%。

Table 1. PCA feature importance
表 1. PCA 特征重要度

编号	还原度
1	0.5729
2	0.7748
3	0.8662
4	0.9071
5	0.9365
6	0.9625
7	0.9776
8	0.9910
9	1.0000
10	1.0000
11	1.0000
12	1.0000
13	1.0000

4.3. SPSS 对原数据进行分析

1) 经过数据预处理得到以下 206796 × 16 的表格，如表 2 所示。

Table 2. Pre-processed data
表 2. 数据预处理后的部分资料

标注	Age	Sex	Heart_rate	PR	QT	QRSDZ	QRS_WIDTH	P_WIDTH	T_WIDTH
正常	8	1	115	108	318	90	82	82	19
正常	6	0	113	116	294	93	78	72	17
异常	6	0	166	112	244	83	62	86	15
正常	9	1	90	124	334	79	78	100	15
异常	6	0	173	100	238	88	82	70	12
异常	7	1	125	108	272	85	70	80	17
正常	9	0	100	120	304	75	88	82	14
异常	10	0	84	108	356	79	94	76	20
正常	10	0	93	118	320	83	88	88	15
异常	5	0	154	110	240	97	58	62	13

从第 D~P 列为心电识别的主要特征，所以分析第 D~P 的因子的降维，在 SPSS 中导入数据并进行分析。

2) 数据的标准化处理，使用 Z-score 标准化。

资料标准化也称为归一化处理，由于不同特征之间可能存在巨大的数值差异，训练模型时可能造成模型误判，为了数据指针之间的可比性，需要进行资料标准化处理。如表 3 所示，得到标准化后的数据。表 3 展示部分数据。

Table 3. Standardized data
表 3. 标准化后的部分数据

ZHEART_RATE	ZPR	ZQT	ZQRSDZ	ZQRS_WIDTH	ZP_WIDTH	ZT_WIDTH	ZQTC	ZP	ZRV1	ZSV5	ZRV5	ZSV1
−0.01460	−0.08735	0.11536	0.21948	0.12768	−0.01893	0.62844	0.26279	1.15289	−0.71094	−1.32986	0.97418	−0.80186

3) 降维因子分析

如表 4，在 KMO 的概率为 0.000，小于显着性水平 0.05，视为拒绝原定假设，与单位矩阵有显着差异，当 KMO 为 0.605，表示较为适合做因子分析。

Table 4. KMO & Bartlett docimasy
表 4. KMO & Bartlett 检定

KMO 和 Bartlett 检定		
Kaiser-Mayer-Olkin 测量取样		0.605
Bartlett 球形检定	大约卡方	966296.471
	df	78
	显著性	0.000

通过陡坡图 2，我们可以看到前八个因子处于比较陡的斜率上，我们选择前八个因子是比较合理的。

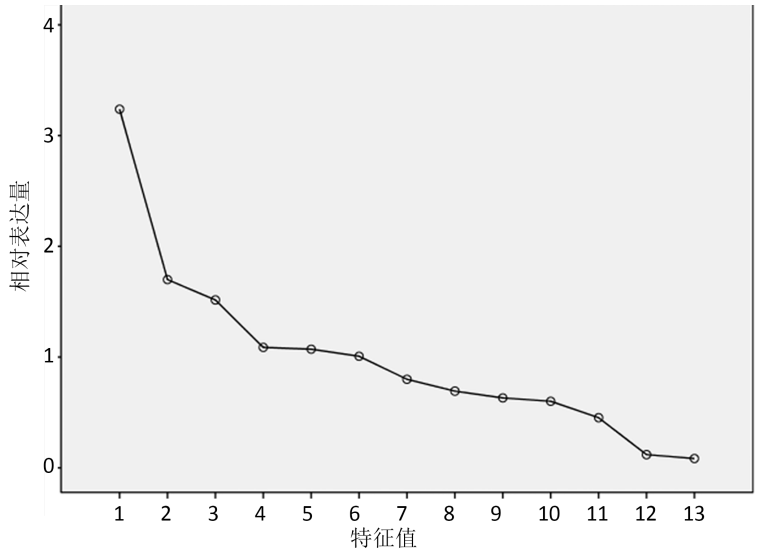

Figure 2. Steslope diagram
图 2. 陡坡图

5. 建立&训练模型

本文采用的模型参照贝叶斯分类器的思想，贝叶斯分类模型如下：

$$P(Y/X)=\frac{P(Y/X)\times P(Y)}{P(X)} \tag{3}$$

其中，X 表示特征集，Y 表示类变量，P(X)为条件原因，P(Y)为先验机率，P(X|Y)为类条件机率，P(Y|X)为后验概率。贝叶斯分类模型就是用前面四项移项计算后得到 P(Y|X)为后验概率，通过比较后验概率进行分类。条件原因 P(X)同时是两个 Y 的后验概率的分母中，因此可以消掉不计。

计算训练集中各个类别所属的训练记录的占比可以轻松得到先验机率 P(Y)。另外使用不同的贝叶斯分类方法，会得到对类条件概率 P(X|Y)不同的估计。

1) 模型训练

按照上面给出的分类器模型，对参数进行训练。当特征是离散型数据时，通过训练集中不同类样本的出现次数可以估计类别先验概率，例如：正常类先验概率 = 正常类样本数量/样本总数，这在程序中对训练集正异常标注列遍历即可统计。类条件概率可以根据类中特征值等于的比例来估计。程序中使用两级字典来统计，第一级字典 Key 分为正常和异常，第二级 Key 是离散的特征值，自动创建。

这样，用户给出的数据后，按特征值检索字典既可以获得样本数并估计该特征的类条件概率。然而，若任一特征的类条件概率为 0，相乘后会使后验概率等于零，大大的影响模型分类结果，为了解决此问题，可以使用 m 估计方法来估计条件概率：

$$P(x_i/y_i)=\frac{n_c\times mp}{n+m} \tag{4}$$

公式(4)中，n_c 是各类中的实例总数，y_i 是各个类别的训练样本中出现的样本数，m 则为等价样本大小的参数，及同为离散特征的个数，p 是自己指定的参数，本程序。m 估计法可以辅助使得概率估计更加完整。训练完成后，用户输入数据代入到朴素贝叶斯模型里，通过比较后验概率的大小就可以进行分

类。样本数共 206,795 条，其中将数据中，70%的样本划入训练集；另外 30%的样本划入测试集。

2) 训练结果

模型训练结果与预期结果进行比对，准确率为 75.8%。

3) 模型比较

Table 5. Models Comparison
表 5. 模型比较

模型	精度	计算成本
C4.5	75.6	66.93
BNT	75.1	8.13
NBTree	75.1	2896.80
KNN	75.6	2398.72
Bagging C4.5	75.9	1696.80
PCA&Bayes	75.8	1547.07

如表 5 所示，经过比对发现，本文数据采用 C4.5 模型的精度为 75.6%，采用 BNT 模型准确率为 75.1%，采用 NBTree 的模型准确率为 75.1%，采用 KNN 模型的准确率为 75.6%，采用 Bagging C4.5 的模型准确率为 75.9%，采用 PCA&Bayes 的准确率为 75.8%。贝叶斯网络在 2010 年前为解决精准医疗的机器学习主流模型。鉴于深度学习之高计算成本，以及所谓的梯度不稳定问题处理。本文提出传统的贝叶斯网络及特征选取方法，对 ECG 数据做精准医疗的分类。基于奥克姆剃刀理论基础上，在精确度不降低太多的情况下，尽可能减少计算成本。

6. 结论

本文通过利用患者的 ECG 数据，对资料进行特定的预处理，接着将数据汇入 Matlab 和 SPSS 进行主成分分析，之后使用贝叶斯分类器对机器进行训练，并给出运行诊断的结果准确率为 75.8%。在未来的工作中将进一步进行多模型的预测，提高预测准确率，并通过用户接口的设计，对结果进行可视化展示，进而实现临床诊断心电科预测系统。

致 谢

感谢以下作者为本文做出的努力：

闵杰青(1978-)，男，云南昆明，医学学士，非 CCF 会员。主要研究方向：儿童心电图诊断以及超声心动图诊断；

李昕洁(通信作者)(1974-)，男，台湾台南人，副教授，博士，主要研究方向：机器学习，人工智能，科技政策(camhero@gmail.com)；

蒋嘉欣(2000-)，女，湖南衡东人，职称：无，学历：本科，主要研究方向：机器学习；

蒲应明(2000-)，男，青海西宁人，职称：无，学历：本科，主要研究方向：计算机视觉；

李向娟(1989-)，女，云南昆明，医学学士，非 CCF 会员。主要研究方向：儿童心电图诊断以及超声心动图诊断；

刘铠华(2000-)，男，山西晋城人，职称：无，学历：本科，主要研究方向：计算机视觉；

曾敬勋(1996-)，男，台湾基隆人，博士研究生，硕士，主要研究方向：深度学习，影像处理；

刘学承(1992-)，男，台湾高雄人，硕士研究生，学士，主要研究方向：机器学习、计算机视觉。

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] 陈砚, 罗妮. 浅析如何提高心血管疾病的护理水平[J]. 基础医学理论研究, 2020, 2(1): 16-17.

[2] 刘洪. 分析冠心病心律失常患者诊断中常规心电图(EcG)和动态心电图(DCG)的诊断效果[J]. 影像研究与医学应用, 2020, 4(5): 48-49.

[3] 吕婷婷, 丁子建, 袁亦方, 等. 深度学习在心电图自动诊断和预测心血管疾病中的应用[J]. 中国心血管杂志, 2021, 26(3): 4.

[4] 刘远致. 一种心电图自动测量方法及设备[P]. 中国, CN110680304A. 2020.

[5] 余燕成, 赵博文, 戴丽雅, 等. 基于容积超声的胎儿心脏计算机辅助诊断技术获取胎儿超声心动图基本诊断切面的应用价值[J]. 中华超声影像学杂志, 2020, 29(4): 314-320.

[6] 王莹莹, 薛超, 殷兆芳, 等. 智能诊断在心电图诊断的发展历程及应用进展[J]. 心血管康复医学杂志, 2019, 28(4): 502-505.

[7] 王官军, 吴婷, 汪龙, 等. 基于机器学习的心电图诊断研究[J]. 实用心电学杂志, 2020, 29(4): 262-268.

[8] 艾文书, 赵兴群. 基于卷积神经网络的心电图心律失常分类方法[J]. 国际生物医学工程杂志, 2021, 44(2): 119-123, 138. https://doi.org/10.3760/cma.j.cn121382-20200520-00206

[9] 赖杰伟, 陈韵岱, 韩宝石, 等. 基于 DenseNet 的心电数据自动诊断算法[J]. 南方医科大学学报, 2019, 39(1): 69-75.

[10] 洪宇光, 王波, 潘湖迪, 等. 基于 LightGBM 的心电信号分类算法[J]. 图像与信号处理, 2020, 9(3): 7.

[11] 胡文博. 基于深度学习心电诊断系统的研究与实现[D]: [硕士学位论文]. 北京: 北京邮电大学, 2019.

[12] 吴恩达新研究: AI 看心电图, 诊断心律失常准确率超过人类医生[EB/OL]. https://zhuanlan.zhihu.com/p/54657157

[13] 人工智能自学预测心脏病发作[EB/OL]. https://baijiahao.baidu.com/s?id=1564933388961558&wfr=spider&for=pc

[14] 唐贤伦, 李伟, 马伟昌, 等. 基于条件经验模式分解和串并行 CNN 的脑电信号识别[J]. 电子与信息学报, 2020, 42(4): 1041-1048.

[15] 张凯. 机器学习在心电数据分析中的研究和应用[D]: [硕士学位论文]. 北京: 北方工业大学, 2019.

[16] 马金伟. 基于深度学习的心电信号识别分类算法研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2018.

[17] 田婧, 张敬, 马雪, 等. 基于卷积神经网络的 ECG 信号识别方法[J]. 杭州电子科技大学学报, 2018, 38(6): 66-70.

[18] 刘翰林, 万相奎, 徐俊, 等. 基于 BP 神经网络的 ECG 信号房早室早分类算法[J]. 湖北工业大学学报, 2018, 33(5): 5-7.

[19] 史航瑞, 梁英. 基于 RBF 神经网络的心电分类识别算法研究[J]. 电脑知识与技术: 学术交流, 2017, 13(19): 137-139.

[20] 沈艺珊. 心电信号异常识别分类研究与实现[D]: [硕士学位论文]. 厦门: 厦门大学, 2018.

[21] 王见, 陈义, 邓帅. 基于改进 SVM 分类器的动作识别方法[J]. 重庆大学学报, 2016, 39(1): 12-17.