编号：A0516

基于集成学习与深度学习的冠心病预测

目 录

表格和插图清单

摘 要

近年来，中国居民心血管疾病的患病率持续上升，死亡率居于首位，冠心病为其中常见疾病之一。与此同时，2020 年疫情的突发提升了居民对互联网医疗的认知度与认可度，为缓解线下医院压力、提高诊治效率，各地区有关部门积极推进信息技术在医疗行业的应用。

本文首先对患者的特征信息建立混合模型，此模型仅借助患者自身特征、症状和病史进行冠心病预测模型的构建，使用神经网络、随机森林和朴素贝叶斯三种不同的算法分别建立预测模型，然后使用集成学习下的 logistic 回归建立混合模型提升预测的准确率。之后，本文通过使用 ResNet34 网络模型进行心电图图像分类的特征学习，并对其进行验证，从另一角度增加了冠心病预测的准确性与可靠性。建立恰当的模型之后，本文可以有效对患者进行方便快捷、成本低廉的冠心病预测，为冠心病的诊断提供有效的辅助支持，这可以为冠心病患者尽早发现病情提供帮助。

关键词： 冠心病预测 神经网络 集成学习 深度学习 图像分类

一、引言

（一）研究背景

1.冠心病现状及病因

在 21 世纪，心血管疾病已然成为威胁人类健康的主要因素，据世界卫生组织统计，冠心病居于新世界十大死因排序之首。

而在中国，冠心病已成为重大的公共卫生问题,防治心血管疾病刻不容缓。党的十八大以来，我国全面推进健康中国建设，并将其作为关系我国现代化建设全局的战略任务。《中国心血管病报告 2018》估算心血管病患者达到 2.9 亿,其中的冠心病患者人数就有 1100 万。

冠状动脉粥样硬化性心脏病（简称冠心病），是指冠状动脉粥样硬化导致心肌缺血、缺氧而引起的心脏病。随着胆固醇、脂肪等沉淀物逐渐积累在动脉内壁，严重时动脉甚至完全阻塞，向心肌提供的血液和氧气便随之减少。

随着我国经济社会发展和人们生活方式快速变化，人们饮食习惯发生巨大改变，摄取的动物脂肪和高胆固醇的食物已远远超量。同时，现代社会快节奏的生活和日益增加的竞争，使部分人内分泌功能变得紊乱，引起高血压，同时还可造成脂肪代谢紊乱。这些因素都极大增加了冠心病患病的可能性。

2. 冠心病诊断方法

在医学上，诊断冠心病主要有两种方法：实验室检查与辅助检查（主要是各类图像）。

实验室检查一般为抽血化验，因为高血压，高血糖，高血脂是引起冠心病

的主要因素，但也会根据不同的临床类型进行不同的生化学检查： 如出现心肌梗死时血清肌红蛋白、肌钙蛋白会增高，所以进行血清心肌酶检查。

辅助检查中冠状动脉造影是目前诊断冠心病的“ 金标准 ”，但有创心血管检查成本相对昂贵且易对患者带来身心伤害，因此一般选用心电图反映心脏的电活动，这在临床上对冠心病出现的心肌缺血、心肌梗死、心律失常的诊断有较高的敏感性和重要意义。

（二）文献综述

冠心病作为一种普遍且高死亡率的疾病，各国都对此进行了广泛深入的研究，近年来，随着互联网医疗的兴起，远程会诊、治疗的热度持续上升，若能够对疾病建立高准确度的预测模型，不仅可以缓解医疗资源紧缺带来的医疗压力，也可以避免病情的延误，实现快速就诊，并且降低诊疗费用。因此各国的医院和研究机构都在通过建立各类预测模型对该疾病进行尽可能准确的预测。

如今在全球范围内已建立了众多针对冠心病的预测模型，如 Aniruddha Dutta 等(2020)使用美国国家健康与营养调查（NHANES）的数据建立具有卷积层的高效神经网络用于冠心病的预测。然而李方舟(2021)指出，经典的冠心病诊断预测模型虽然有中等的预测效能，但在不同人群间预测效能的差异性较大，由于模型的建立基于欧美人群的数据，所以在中国人群中效能不佳。当然，近期也不乏有国内的专家开始针对冠心病进行预测研究，如李雨洁、郑锐龙等（2020）在使用数据挖掘技术建立冠心病早期预测模型，但对特征信息进行处理时，仅将其按缺陷种类数量进行分类，并未明确具体的缺陷类型。

随着图像处理的兴起，心电图作为冠心病诊断的有力依据也被列为又一重点研究对象，但许多研究是针对不同种类心电图之间的对比，或是心电图联合超声心动图等对冠心病进行诊断，如戴小毛(2017)探讨了冠状动脉成像对冠心

病的诊断价值。国内在此方面也有多方面的研究，如黄健、林进意等(2021)指出，这两种图像的联合可以及时发现患者病情，并且对冠心病诊断的准确率达到了较高水平。本文在前人研究的基础上，对心电图进行图像分类处理并且将其与特征变量的诊断结果进行综合。

当前有众多针对冠心病的研究模型，然而大部分仅针对少数特征变量或按其数量组合进行建模，并未单独将它们作为独立的变量，且使用国内患者数据的模型较少，地域性的偏差会降低模型的预测效能。因此本文根据疾病特点对经典模型进行优化，且使用国内某地区患者数据，使其更好地服务于国内冠心病的预测。同时，本文也对心电图进行图像处理，从而在对患者进行是否患有冠心病的判断时，可以综合特征判断和图像处理的结果，以达到更符合真实情况的诊断结果。

二、数据来源与分析

（一）患者信息数据

1.数据清洗与脱敏

患者信息数据来源为某医院 2012-2021 年电子病历，共 13699 条数据。由于部分数据的缺失，以及对患者信息保密的需求，首先对数据进行了脱敏和清洗，最终得到 13298 条可用性高的数据。为实现建立高准确率冠心病预测模型的目标，在直接使用病例中部分信息的同时，也从西医诊断中提取患者的病史和临床症状，选择出现频率高且在医学角度与冠心病相关性较强的疾病作为自变量，共得到 18 组自变量。

2.变量提取

为了更好的研究导致冠心病发作的因素，避免自变量过多导致的多重共线性，我们对自变量进行进一步的筛选。由于大部分数据的类型为分类数据，所以选择使用 AIC 准则结合逐步回归进行变量筛选， AIC 的数学表达式为：

$$\mathrm{AIC}=2p+n*\log(\frac{SSE}{n}) \tag{1}$$

其中 p 是模型中的自变量个数， n 为样本量， SSE 是残差平方和。在 n 固定的情况下， p 越少， AIC 越小， SSE 越小， AIC 越小， p 越少说明模型越简洁， SSE 越小说明模型拟合度越高，由此可得， AIC 越小，模型就越简洁和精准。

逐步回归通过剔除变量中重要程度低且和其他变量高度相关的变量，降低变量间多重共线性程度。首先，对引入的变量进行 F 检验，检查该变量是否对模型造成显著的变化。若发生显著变化，则对所有变量进行 T 检验；若没有显著变化，则该变量被消除，直到没有显著变量进入方程，也没有显著变量被消除为止。此时得到最优变量集。

本文通过逐步选取对判断是否患有冠心病拟合最好的解释变量，最终保留了 10 个自变量，所有变量的定义见表 1。

表 1 各变量定义及具体赋值

变量名称	符号	变量类型	分类变量赋值
年龄（*age*）	x_1	连续型	——
脑梗死（*w.brain*）	x_2	离散型	患病=1，不患病=0
心功能（*c.heart*）	x_3	离散型	患病 1 级=1，2 级=2，3 级=3，4 级=4，不患病=0
高血压（*c.highblood*）	x_4	离散型	患病 1 级=1，2 级=2，3 级=3，不患病=0
脂肪肝（*w.fattyliver*）	x_5	离散型	患病=1，不患病=0

房颤（*w.af*）	x_6	离散型	患病=1，不患病=0
住院天数（*period*）	x_7	连续型	——
心绞痛（*w.angina*）	x_8	离散型	患病=1，不患病=0
糖尿病（*w.diabetes*）	x_9	离散型	患病=1，不患病=0
心律失常（*w.arrhythmia*）	x_9	离散型	患病=1，不患病=0
冠心病（*w.CHD*）	y	离散型	患病=1，不患病=0

对变量的筛选过程输出结果见表 2。

表 2　逐步回归筛选结果

	估计值	标准误	t 值	显著性
截距	-0.1251376	0.0157100	-7.965	1.78e-15***
x_1	0.0051587	0.0002347	21.978	<2e-16***
x_2	-0.0383567	0.0078853	-4.864	1.16e-06***
x_3	0.0219053	0.0026864	8.154	3.83e-16***
x_4	-0.0061931	0.0022309	-2.776	0.00551**
x_5	-0.0287093	0.0089707	-3.200	0.00138**
x_6	-0.0195213	0.0094817	-2.059	0.03953*
x_7	0.0011838	0.0006354	1.863	0.06247.
x_8	0.7180782	0.0065975	108.841	<2e-16***
x_9	0.0190372	0.0077486	2.457	0.01403*
x_{10}	-0.0211210	0.0096676	-2.185	0.02893*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

得到回归方程如下：

$$\begin{aligned} w.CHD = {} & 0.1251376 + 0.0051587x_1 - 0.383567x_2 + 0.0219053x_3 \\ & -0.061931x_4 - 0.0287093x_5 - 0.0195213x_6 + 0.0011838x_7 \\ & +0.7180782x_8 + 0.0190372x_9 - 0.00211210x_{10} \end{aligned} \tag{2}$$

由(2)式可得患高血压(x_4)与患冠心病的相关性较强，原因是：患者的血压升高，低密度脂蛋白向血管内皮下移动，导致动脉粥样硬化和血管腔狭窄，最终导致冠心病。此外，肥胖将会导致血脂升高，进一步导致患者出现脂肪肝(x_5)，糖尿病(x_9)等多种代谢性疾病，以及出现冠心病，脑梗死(x_2)等心脑血管

疾病。除此之外，心功能状况(x_3)、房颤(x_6)等变量为心脏相关疾病，也属于冠心病前兆或并发症，其中心绞痛(x_8)的系数达到 0.7180782，与临床诊断时医生重点关注的相关疾病相吻合。由此可知，此方法筛选出与疾病有关的变量，在医学上皆与冠心病的发作相关。

最后，在建立模型的过程中，需要先使用大量样本对模型进行训练，再用模型对其余样本进行检验，以达到最高的预测准确率，因此所有数据被分为两个数据集。我们按照 7：3 的比例将样本数据随机分为训练集和测试集，分别为 9309 条和 3989 条数据。

（二）心电图图像来源

图像来自 kaggle，心电图是诊断冠心病的一种简单而常用的方法，冠心病患者的心电图大多表现为 ST 段水平型或下斜型的压低和 T 波的倒置。它也能发现与冠心病相关疾病的症状。如心绞痛发作时 ST 段低，异性心绞痛出现短暂性 ST 段抬高，不稳定型心绞痛有明显的 ST 段压低和 T 波倒置；心肌梗死时表现为异常 Q 波。

三、基于患者信息预测模型的建立

（一）基于神经网络的预测模型

1.神经网络方法介绍

本文采用多层全连接前向网络模型，在网络中，每个神经元接受前一层的输入并输出到下一层，这种网络实现了信号从输入空间到输出空间的转换，每个神经元接受线性组合的输入后，先进行简单的线性加权，之后对每个神经元使用非线性的激活函数后输出。由于它处理信息的方式仅为简单非线性函数的

多重组合，因而比较容易实现。在模型的训练之后，参数会根据训练数据进行调整，使得模型能够拟合数据。

神经网络结构图如下，基本为 $wx+b$ 的形式。其中 x_1, x_2 表示输入节点，w_1, w_2 为它们各自的权重，每个输入都会被赋予一个权重。1 为偏置节点，它是默认存在的，一般情况不会被明显画出，在神经网络的每个层次中，除输出层之外，都会含有这样一个偏置单元，通常设这些参数值为向量 b，称之为偏置。$g(z)$ 为激活函数，a 为输出。图 1 所示的神经网络的矩阵运算为：

$$g(b+x_1*w_1+x_2+w_2)=z \tag{3}$$

$$g(z)=a \tag{4}$$

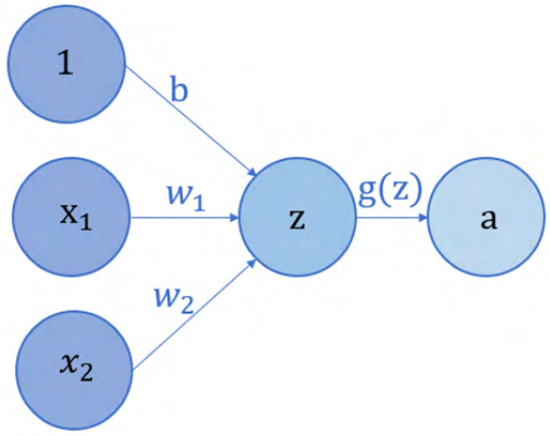

图 1　神经网络矩阵运算示意图

由于标签为二分类型数据，我们选用二分类 logistic 回归模型，并且使用激活函数 sigmoid，函数的功能是将实数压缩至 0 到 1 之间，该函数表达式为

$$g(z)=\frac{1}{1+e^{-z}} \tag{5}$$

函数图像如图 2 所示：

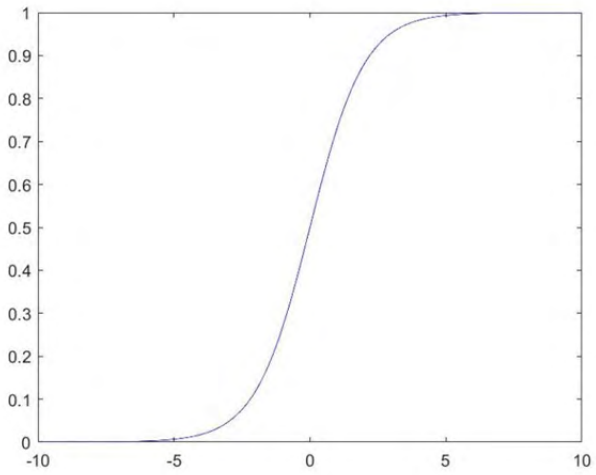

图 2　sigmoid 函数图像

当 z 为大于 5 的正数时，$g(z)$的值趋向 1，当 z 为小于-5 的负数时，$g(z)$的值趋近于 0。将激活函数的函数值视为概率，当激活函数的函数值大于 0.95 时，即有 95%以上的概率为正样本。

在训练期间权重更新的量被称为“学习率”，建立模型时希望得到一个学习率，既可以极大地减少网路损失，也可以使训练时间较少。模型通过每一次逐步提高小批量的学习速率，并记录每一次增量之后的损失，由于损失函数下降最陡处为学习速率最优处，模型主要关注损失函数图的坡度，最终将学习率设置为 0.1。

2.神经网络模型的建立

此过程使用训练集的数据。首先我们采用标准归一化的方法将初始数据集变为均值为 0 且方差为 1 的新数据集，以期消除由于指标量纲不同造成的影响，将它们置于同一数量级后直接进行对比。归一化公式如下：

$$x^{*}=\frac{x-u}{\delta} \tag{6}$$

其中 μ 为所有样本数据的均值，δ为所有样本数据的标准差。

之后对模型进行训练，得到神经网络结构图，如图 3 所示。

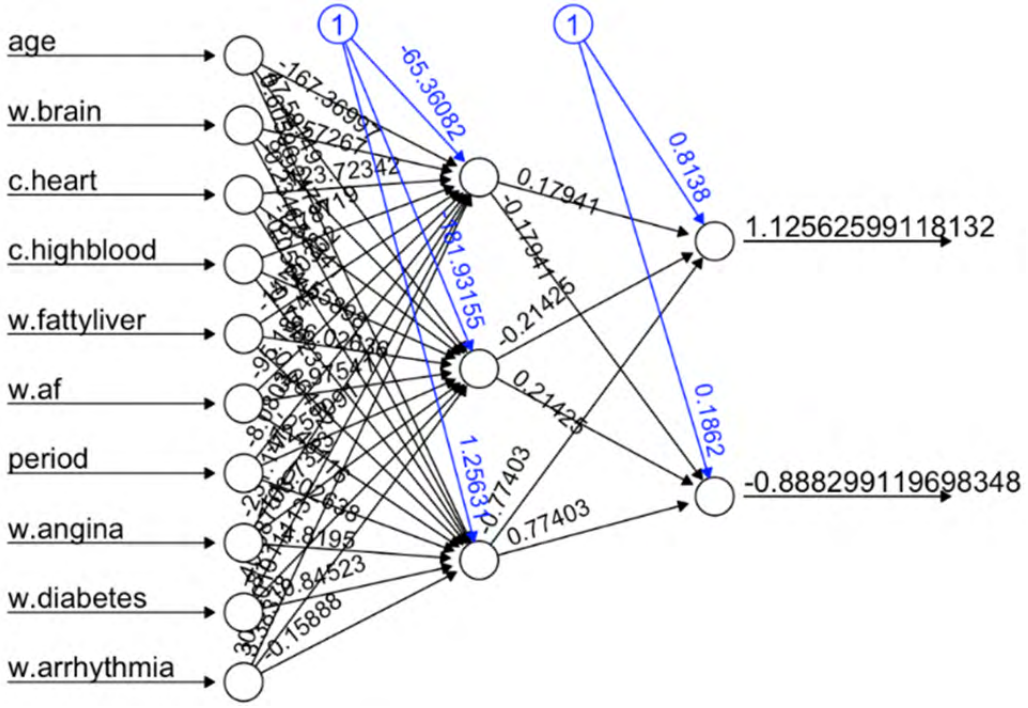

图 3　神经网络结构图

隐藏层的 3 个神经元各自接受来自 10 个输入的权重，这 3 个神经元又在自身各自不同权重的影响下成为输出层的输入，最终由输出层输出最终结果。输出结果中V_1代表未患冠心病的概率，V_2代表患有冠心病的概率。

同时，我们对 10 个自变量的重要性进行排序，结果如图 4 所示：

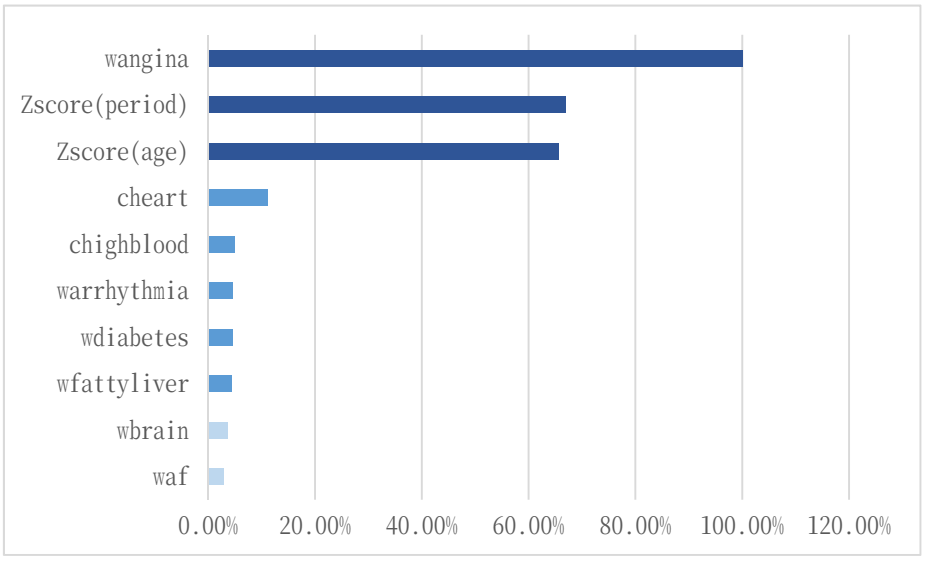

图 4　自变量重要性分析图

由图 4 可得，是否患有心绞痛是冠心病预测的最重要因素，住院时长和年

龄的重要性虽低于心绞痛，但也是预测过程中对结果影响较强的因素。患者是否患有脂肪肝、脑梗死和房颤对预测是否患冠心病并没有很显著的作用。

3.神经网络模型的检验

将模型应用于测试集进行是否患病的预测，整体预测正确率达到 85.1%，对已患冠心病的预测准确率达到 71.5%。

进一步通过受访者工作特征曲线（简称 ROC 曲线）对预测结果进行分析，首先做如下定义：

表 3　ROC 曲线符号定义表

		预测	
		1	0
实际	1	True Positive(TP)	False Negative(FN)
	0	False Positive(FP)	True Negative(TN)

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{TP + TN} \tag{8}$$

图 5 纵轴数值为*TPR*，横轴数值为*FPR*，由公式(1.7)、(1.8)可知 ROC 曲线的横坐标和纵坐标是没有相关性的，因此应当将 ROC 曲线视作无数个点，每个点都代表一个分类器，其横纵坐标表征了这个分类器的性能，左上角的点为完美分类点，它代表所有的分类全部正确，即归为 1 的点全部正确（*TPR*=1），归为 0 的点没有错误（*FPR*=0）。对于 ROC 曲线，一个重要的特征是它的面积，面积为 0.5 是随机分类，识别能力为 0，面积越接近于 1 识别能力越强，面积等于 1 为完全识别。下图曲线下的面积在 0.5 与 1 之间且靠近左上角，说明该模型有较高的预测价值。

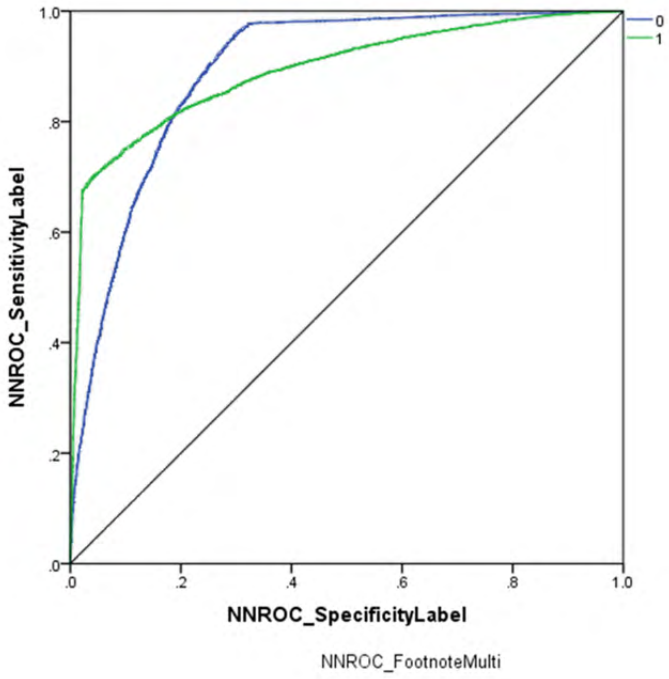

图 5　神经网络 ROC 曲线图

（二）基于随机森林的预测模型

1.随机森林方法介绍

随机森林本质上是一种特殊的 bagging（装袋算法）方法。首先，用 bootstrap 方法生成 m 个训练集，在后对每一训练集创造一棵决策树，在节点寻找特征进行分叉时，并非对全部的特征找到相应使得指标（例如信息增益）最大的。而在特征中随机抽取相应的部分特征，在随机抽取的特征中间寻找最优解，并将之用于节点处，从而进行分叉。其过程如图 6 所示：

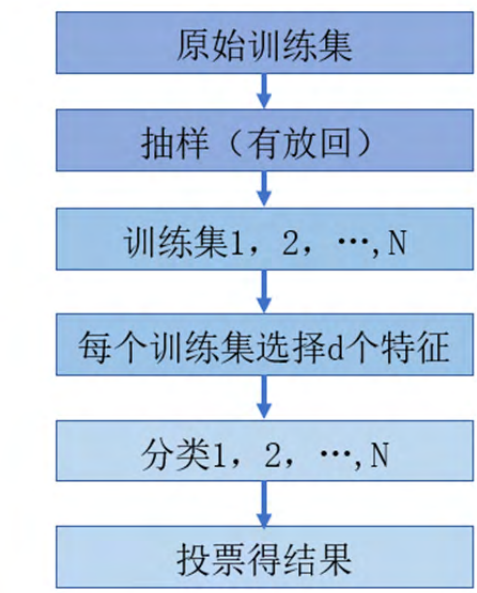

图 6　随机森林预测过程

其中信息熵为度量样本集合纯度的常用指标。令 $p_k, k=1,2,...,m$ 代表第 k 类样本在输入样本集合 D 中所占的比例，假设一共有 m 个类别，则该集合 D 的信息熵为：

$$\mathrm{Ent}(D)=-\sum_{k=1}^{m} p_k \log_2 p_k \tag{9}$$

信息熵 $\mathrm{Ent}(D)$ 的值越小，则代表 D 的纯度越高。

假定属性 a 有 V 种取值，因此输入数据集合 D，根据属性 a 可以将 D 划分为 V 个分支集合，那么根据属性 a 划分得到的信息增益为：

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\left|D_v\right|\left|D\right|\mathrm{Ent}\left(D_v\right) \tag{10}$$

其中，D_v 代表根据属性 a 的某个取值 a_v 得到的分支集合。

信息增益的值越大，就表明使用属性 a 进行划分带来的纯度提升的难度越大，故信息增益的选择属性问题可以根据此解决。

决策树则是用树的结构来构建分类模型，一个节点代表着对应的一个属性。预测结果则是依据属性进行划分，进入这个节点的子节点，直至叶子节点，一定的类别被每个叶子节点代表，分类从而实现。决策树结构图如图 7 所示。

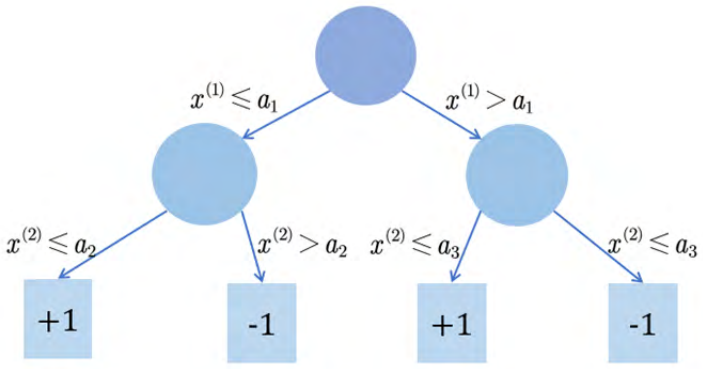

图 7　决策树结构图

2.随机森林模型的建立

根据袋外错误率选取最优参数 *mtry*（节点数），即用于二叉树的最佳变量个数在指定节点中。随机森林所具有的一个重要优点是，不必对其进行交叉验证来得到样本误差的一个无偏估计。可以直接在模型内部进行评估，即在生成的过程中就可以直接对误差建立一个无偏估计。

在对每棵树进行构建的时候，对训练集采用不同随机且有放回地抽取的样本。故对于每个树而言（假设对于第 k 棵树），没有参与第 k 棵树的生成的训练实例大概有1/3，它们称为第 k 棵树的袋外样本。

由于上述的采样特点，进行 OOB（袋外误差）估计在本文中就被允许的。首先对每个样本，通过 OOB 样本的树对其分类情况（大约1/3 的树）进行计算。进而该样本的分类结果，则以简单的多数投票结果得出。最后，随机森林的 OOB 误分率则用误分个数占样本总数的比率。

根据上述理论，本文采用 R 语言，在随机森林树数 1000 的条件下从 1 开始进行循环遍历节点数，计算每一个节点数下模型出现的袋外误差率，并根据所得结果绘制相应的散点图。结果如下图所示，当 *mtry* 的值为 4 时，袋外误差值最低，故选取 4 为最佳变量个数。

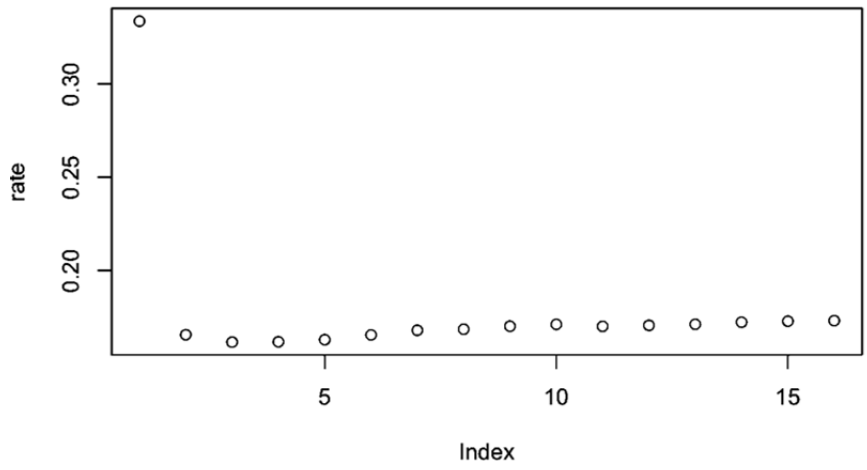

图 8　mtry 误差率图

选取最佳变量个数后，在 *mtry* 为 4 的基础上，接着选取最佳参数 *ntree*，即指定随机森林所包含的最佳决策树数目。将节点数定为 4，从 1 开始不断提高模型的树数至 1000，计算每个树数下模型预测的的误差率，并生成相应的误差图（图 8）。对小于 1000 的 *ntree* 树数进行逐个遍历，计算每个决策树数下的错误率，绘制模型误差率随着决策树数的变化图。如图 9 所示，随着 *ntree* 的增大到 200 左右时，模型错误率逐渐趋于稳定，模型错误率的变化率不再产生较大变化，故选取最佳 *ntree* 的参数值为 200。

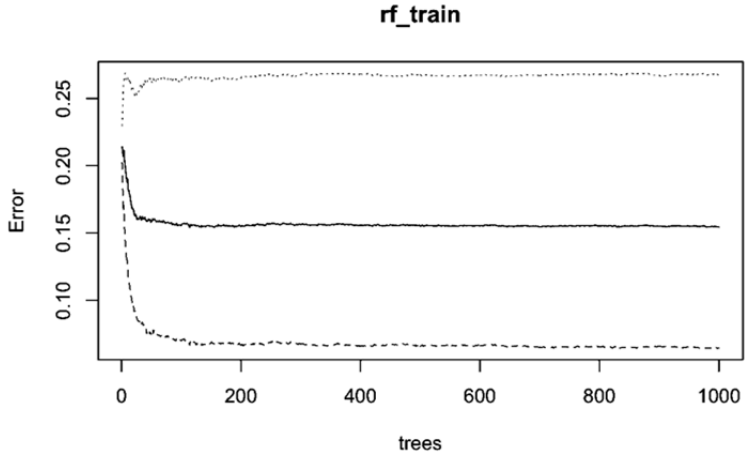

图 9　ntree 误差图

3.随机森林模型的检验

随机森林实际上是一种基于决策树的装袋算法模型。本文根据上述步骤，先用循环遍历的方法，根据袋外误差率的散点图，在*mtry*为 4 时，模型的袋外误差率最低，故选择*mtry*的参数值为 4。接着，在将*mtry*的值定为 4，从 1 开始不断提高随机森林中决策树的个数至 1000，绘制模型误差率随着决策树数的变化图，*ntree*在 200 左右时，模型错误率趋于稳定，故本文将*ntree*参数值定为 200。本文用 R 语言建立出*mtry*参数值为 4，*ntree*参数值为 200 的随机森林模型，并利用训练好的模型对测试集进行预测，预测准确度为 0.84。随机森林模型的生成过程，符合正常医务人员对患者进行诊断的过程，即建立判别规则，不断向下拓展，最终确定患者的疾病。同时生成的混淆矩阵如图 10 所示，可以看到模型犯第一类错误和第二类错误的概率。根据混淆矩阵，模型可以对大部分的样本进行精准的预测。这反映模型能够较好地对病人的冠心病的发病情况进行预测，对冠心病的初筛情况进行一定的辅助作用。

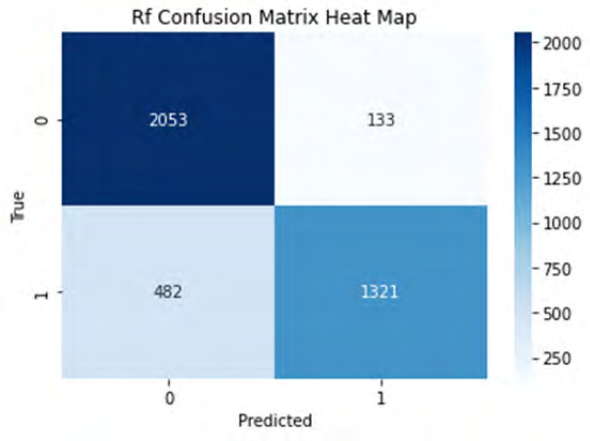

图 10　随机森林混淆矩阵热力图

（三）基于朴素贝叶斯的预测模型

在这一节，我们使用朴素贝叶斯模型，继续对前文所提及的住院患者特征信息数据建立冠心病预测模型。不同于之前两个机器学习算法的是：朴素贝叶

斯模型是一个建立在坚实概率论与数理统计学基础上的“标准”概率统计模型。下面我们首先对贝叶斯统计学做一个简短的介绍，接着将叙述高斯朴素贝叶斯的假定，最后建立高斯朴素贝叶斯模型并对它的效果进行评价。

1.朴素贝叶斯方法介绍

贝叶斯统计学开端于 1763 年的英国，托马斯·贝叶斯(1702-1761)在论文《机遇理论中一个问题的求解》中所描述的问题可以转化成如下叙述：假设有一个随机变量 X 服从二项分布，其中这个二项分布具有一个确定的、但仍未知的参数 θ（θ 表示一次试验中失败的概率），同时我们在总共 n 次试验中观察到了 r 次失败，那么 θ 落在两个定值（比如说 θ_1 和 θ_2）之间的概率是多少？

通过使用经典统计学的方法理论，可以得到有关参数 θ 的估计、置信区间和假设检验，但贝叶斯的思想不止于此，他试图借由 θ 的概率分布函数 $p(\theta)$ 表达出参数的不确定性。因此贝叶斯统计分析中决定性的一步就是把参数 θ 视作随机变量，$p(\theta)$ 称作 θ 的先验分布，而在给定观测后的 $p(\theta|y)$ 称作后验分布。沟通它们的是下面著名的贝叶斯公式：

设 $B_1, B_2, ..., B_n$ 是样本空间 Ω 的一个分割，即 $B_1, B_2, ..., B_n$ 互不相容，且 $\bigcup_{i=1}^{n} B_i = \Omega$，如果 $p(A) > 0, p(B_i) > 0, i = 1, 2, ..., n$，则

$$p(B_i \mid A) = \frac{p(B_i)\,p(A \mid B_i)}{\sum_{j=1}^{n} p(B_j)\,p(A \mid B_j)} \tag{11}$$

朴素贝叶斯是一种衡量标签和特征之间的概率关系的有监督学习算法，高斯朴素贝叶斯则通过假设各特征是独立服从正态分布，估计给定特征下各类别的条件概率。对于固定特征下的取值，高斯朴素贝叶斯有如下公式：

$$p(x_i \mid Y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{12}$$

2.朴素贝叶斯模型的建立

本文使用 Python 中的 scikit-learn 工具包构建高斯朴素贝叶斯模型。对于我们的具体问题，(11)式中的 B 是 *w.CHD*（是否患冠心病），$P(B=1)$ 以及 $P(B=0)$ 可以由观测值直接估计；A 是一个随机向量，其中分量（记作 A_k，$k=1,2,...,p$）是 *age*（年龄）、*w.brain*（是否患脑梗死）等随机变量（见第二章第一节变量选择的结果）。

又据高斯朴素贝叶斯假设，(11)式中 $P(A|B_i)$ 可以分解成 $\prod_{k=1}^{p} P(A_k|B_i)$。

而对于任意给定的 k，$P(A_k|B_i)$ 可以由(12)式给出（在离散场合下概率将由分布律求和得到）。这样，在给定特征信息的条件下，一名患者患冠心病的概率就被计算出来了，进一步如果这个概率大于 0.5，那么模型将会给出该名患者患冠心病的预测。

它在测试集上的分数是 0.839，这表明预测的准确率达到了 80%以上。下面给出布里尔分数的公式，它是一种衡量我们的概率距离真实标签结果的差异的指标。

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{n}(p_i - o_i)^2 \tag{13}$$

其中 N 是样本数量，p_i 为朴素贝叶斯预测出的概率，o_i 是样本所对应的真实结果，只能取到 0 或者 1，如果事件发生则赋值为 1，如果事件不发生则赋值为 0。布里尔分数的取值区间是从 0 到 1，分数越高表示预测结果越差，校准程度越差，因此布里尔分数越接近 0 越好。

经过计算，此模型的布里尔分数为 0.126，这是一个较低的分数。

3.朴素贝叶斯模型的检验

朴素贝叶斯是一个“不建模”的算法，故要由其他角度对所建立进行模型

刻画。

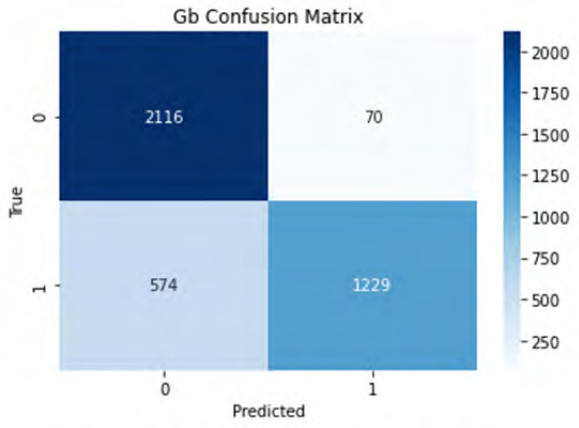

图 11　高斯朴素贝叶斯混淆矩阵热力图

由图 11 可知模型犯第一类错误和第二类错误的概率，即此模型将未患病判成患病的个案数仅有 70，而相应的将患病判成未患病的错误数很多。

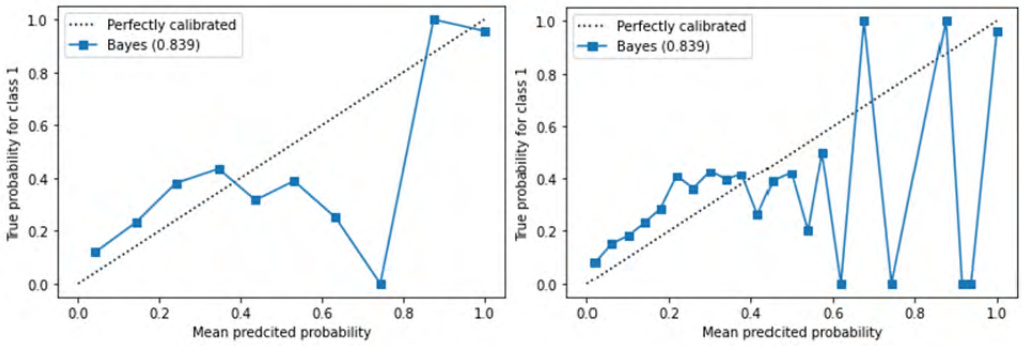

图 12　高斯朴素贝叶斯 10 箱、25 箱可靠性曲线

图 12 被称作概率预测的可靠性曲线。它的算法如下：给定一个总箱数 n，设箱数为 $i=1,2,...,n$，将所预测的概率从小至大排列并均匀分箱，在每一箱中计算它们的预测概率的均值（作为横坐标），对于第 i 箱，计算箱内观测为 1 的个数 / 箱内观测数（作为纵坐标）。所得折线越接近第一象限角平分线说明拟合的效果越好。据图 12 可知，此模型在判断较小概率 (<0.5) 上表现较好，在判断大概率 (>0.8) 表现较好，而对于较大概率 $(0.5<0.8)$ 表现欠佳。

模型所预测概率的分布如图 13 所示：

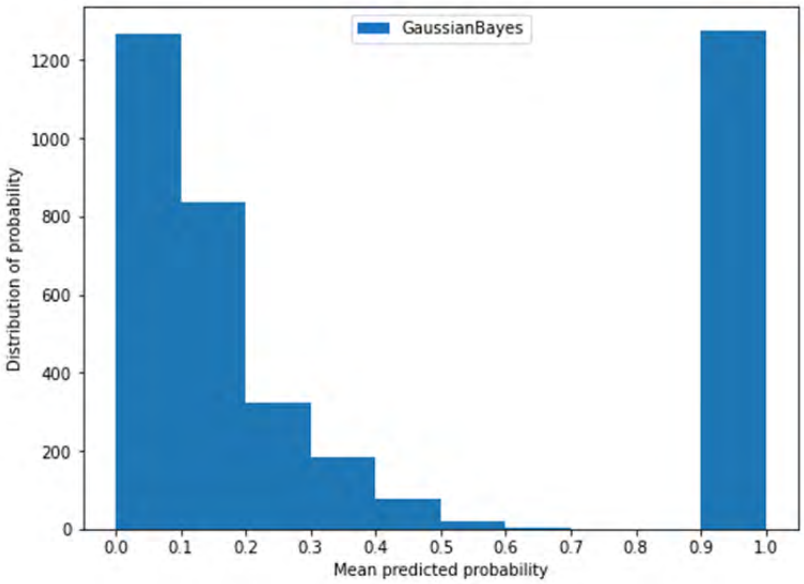

图 13　高斯朴素贝叶斯预测概率直方图

由上图我们看到，对于测试集上的数据，此模型给出了大量较小概率(< 0.5) 的判断，例如，模型给出了 1200 多个 0 到 0.1 之间的概率。这与图 1 展示的犯第二类错误较多存在一定的联系（在这里第二类错误就是把患冠心病判为不患冠心病)。同时对于图 2 所展现的问题：模型在较大概率上的判断不佳我们无需过于担心，因为它虽然在给出 0.5 到 0.8 概率时准确度很差，但是仅仅给出了很少数量的判断。

（四）集成学习下的 logistic 回归

如果随机地向几万人征询一个复杂问题的意见，然后融合他们的回答。在很多情形下，我们会发现这个回答比相关领域专家的回答还要好。同样地，如果将一组分类器的结果聚合，最终结果很可能也会比单个分类器的效果要好得多。而在机器学习领域，这种思想和技术被称作“ 集成学习 ”。

通过上述三种模型，我们已经训练好了一些分类器，包括一个神经网络分类器、一个随机森林分类器和一个朴素贝叶斯分类器。接着，我们将使用集成方法中的堆叠法（stacking）建立一个集成的、混合的分类模型来增强模型的分

类能力。

1.堆叠法

现在要实现聚合所有分类器给出的分类，可以使用简单的函数（比如硬投票）来实现，但我们还可以选择训练一个新的模型来实现这样的聚合，这就是堆叠法的基本思想。

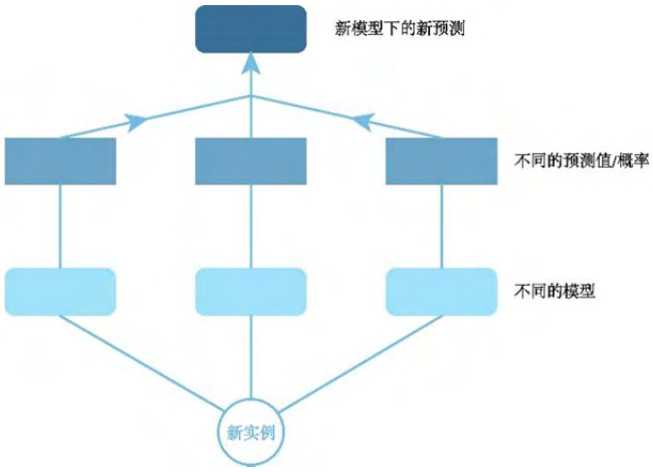

图 14　集成学习示意图

图 14 展示了这样一个过程：底部的三个不同模型给出了不同的预测值/概率，再经由顶部的新模型（最终的分类器，也称混合器）将所得到的预测值/概率作为输入，进行最终预测。

训练混合器经常使用留存集的方法，下面具体展示在模型在住院患者特征信息数据上的运行过程。首先将数据分成两个子集：训练集和测试集，其中训练集已经用以建立前文所述的三个模型（见图 15）。

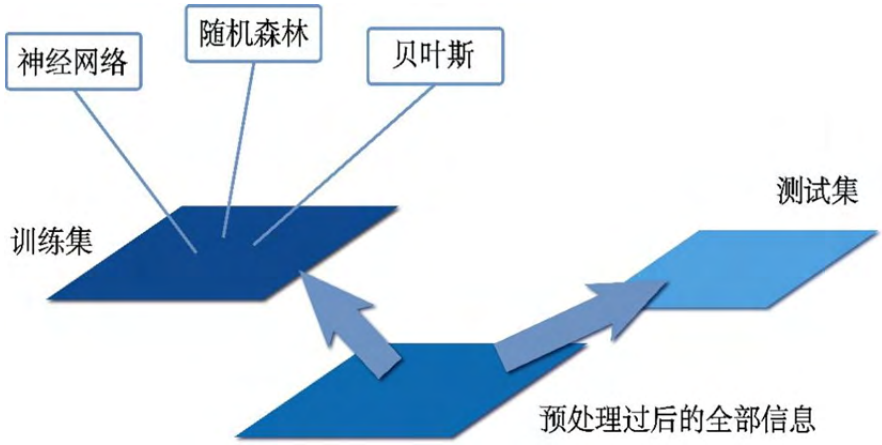

图 15　训练第一层

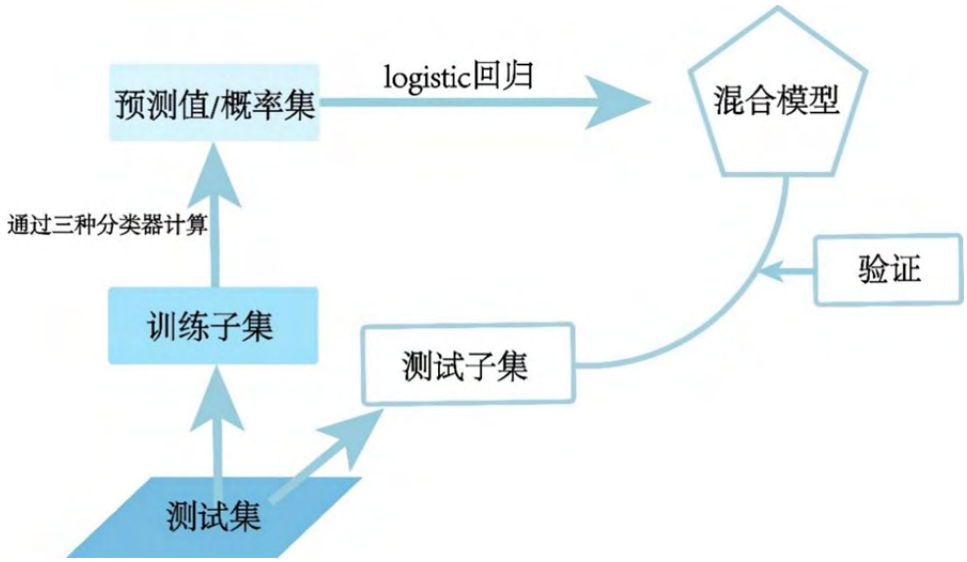

图 16　训练混合器

现在对于测试集上的数据，我们先选取一部分，使用已经建立好的三个模型对它们给出预测的概率，并把这些所得到的概率作为输入特征，构造一个新训练集（图 16 中的预测值/概率集），并保留目标值。在这个集合上训练混合器（logistic 回归分类器），让它根据第一层得到的概率来预测目标值。

2.logistic 回归

logistic 回归来源于线性回归，我们首先给出需要求解的线性回归方程：

$$z = \theta_0 + \theta_1 x_1 + ... + \theta_p x_p \tag{14}$$

令 $\theta^T = (\theta_0, \theta_1, ..., \theta_p)$，　$X = (x_0, x_1, ..., x_p)^T$，　上式就可以改写为

$$z = \theta^T * X \tag{15}$$

线性回归所要解决的问题，是构造一个函数来刻画输入和输出的线性关系（其中，输入通常是一个矩阵 **X**），而构造的关键则是找出(15)式中的θ^T，著名的最小二乘法就是用来求解线性回归中参数的数学方法。

通过(15)式中函数 z，线性回归使用输入的特征矩阵 **X** 来输出一组连续型的标签值 y，以达到预测连续型变量的目的。特别地，当 y 满足伯努利分布的离散型变量时，通过引入联系函数，将线性方程 z 变换为$g(z)$，且$g(z)$的值在(0,1)之间。最后设定一个阈值（通常是 0.5），当$g(z)$小于阈值时的标签为类别 0，当$g(z)$大于阈值时样本的标签为类别 1，这样就得到了一个分类模型。对于逻辑回归而言，这个联系函数选用为 sigmoid 函数：

$$g(z) = \frac{1}{1+e^{-z}} \tag{16}$$

将(15)代入(16)就得到：

$$g(z) = y(x) = \frac{1}{1+e^{-\theta^T X}} \tag{17}$$

$g(z)$就是我们逻辑回归所得到的函数值，由此可以导出某样本的标签值。注意到$y(x)$的取值在[0,1]之间，因此$y(x)$和$1-y(x)$之和为 1。事实上可以证明：

$$\ln\left(\frac{y(x)}{1-y(x)}\right) = \theta^T X \tag{18}$$

$\frac{y(x)}{1-y(x)}$可被视作形似几率，将其取对数的就得到(15)式中的线性回归 z。

事实上，我们是在对线性回归模型的预测结果取对数几率来让其的结果无限逼近 0 和 1。因此这个模型被称为“对数几率回归”（Logistic Regression）。逻辑回归的目的和线性回归是一致的：求解θ^T来构建一个尽可能拟合所输入的特征矩阵 **X** 的函数$y(x)$，并向预测函数中输入新的特征来得到对应的标签值。

在逻辑回归中，通常使用损失函数来测度已建立的模型拟合训练集时产生的信息损失，以此判断所求解到的参数θ^T的优劣。逻辑回归的损失函数由极大似然估计推导而来，具体可以写作：

$$J(\theta)=-\sum_{i=1}^{m}(y_i\times\log(y_\theta(x_i))+(1-y_i)\times\log(1-y_\theta(x_i))) \qquad (19)$$

其中，θ表示求解出的参数向量，m是样本个数，y_i是第i个样本上真实的标签，$y_\theta(x_i)$是第i个样本上，基于参数θ构建的模型的函数返回值，x_i是样本i不同特征的取值。优化的目标，是求解出使$J(\theta)$最小的θ取值。

我们指出：对逻辑回归中过拟合的控制，通过正则化来实现。在本节中我们使用的是L_2范数：

$$J(\theta)_{L_2}=C\times J(\theta)+\sqrt{\sum_{j=1}^{n}(\theta_j)^2} \qquad (20)$$

(20)式中的C是用来控制正则化程度的参数，n是方程中参数的个数，同时也是方程中特征的总数。在这里，参数θ_0是截距，它通常不参与正则化的。C越小，损失函数越小，模型对损失函数的惩罚越重，正则化程度越强，参数θ^T的范数会被压缩得越来越小。

3.模型的建立

这里依旧使用 Python 中的 scikit-learn 工具包来构建模型，求解最佳θ^T的方法本文选择的是坐标下降法（Coordinate Descent）,这是一个简单高效的非梯度优化算法。不同于沿着梯度最速下降的方向寻找函数最小值（梯度优化算法），坐标下降法按一定顺序沿着坐标轴的方向最小化目标函数值。

基于上一节的讨论，首先要确定一个合适的(20)式中的C，以期模型在测试集上能有一个较好的效果。

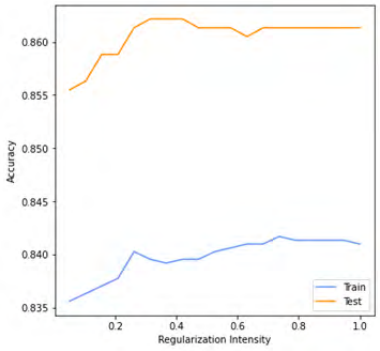

图 17　混合模型准确率随参数 C 变化曲线图

图 17 展示了所建模型分别在训练集和测试集上的分数，可以看到当 C 取 0.4 时模型在测试集上的表现最好。

经过五次迭代之后模型已经停止计算，所得各变量的斜率系数见下表：

表 4　混合模型各变量斜率系数

变量	贝叶斯所给概率	神经网络所给概率	随机森林所给概率
系数	1.17673225	1.24792519	3.24421645

此模型的在测试集上的准确率达到了 0.86382，图 18 给出了它判断两类别的 ROC 曲线和混淆矩阵。在医学中，纵坐标 *TPR* 被称为真阳性检出率，横坐标 *FPR* 被称为假阳性率。曲线右下方面积为 0.91，这说明模型判断效果较好。

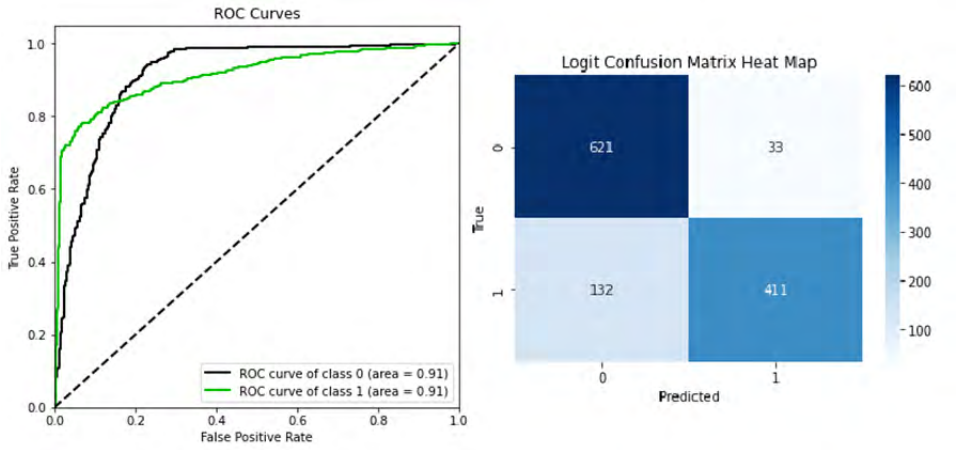

图 18　混合模型的 ROC 曲线与混淆矩阵

四、基于图像分类算法预测模型的建立

（一）心电图介绍

心电图一般是以时间长短和电压大小表现检测者心脏活动，用横坐标表示时间，纵坐标表示电压，再根据两者共同描绘出图形进行心脏病的诊断。通过心电图，可以对心脏状况进行观察，很快地对心脏病进行生理和病理性判断。此外还可以通过原始心电图帮助诊断情况复杂的心律失常。观察病人进行治疗用药后的冠心病恢复情况，判断药物疗效也可以用心电图。因此心电图对诊断冠心病有着十分重要的意义。本文通过对心电图的图像分类算法，实现对冠心病的进一步筛选。本文选取的心电图数据如下所示：

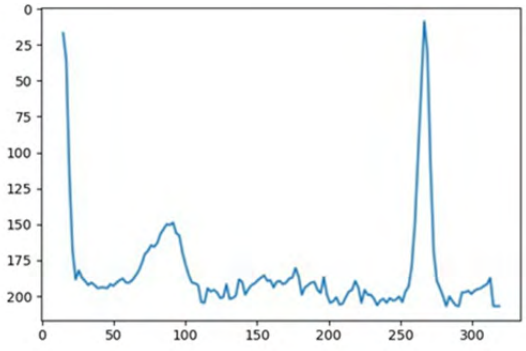

图 19　心电图数据示例图

（二）残差神经网络介绍

Resnet 网络允许将初始输入信息直接传输到后层。Resnet 开始处理神经网络深度增大时的随机梯度下降问题。精度将首先提高，达到饱和状态。当深度不断增加时，如果模型太深，预测精度将下降。这不是过度拟合造成的问题，而是因为训练集本身的误差的进一步增加。

ResNet 有较多旁路的支线将输入层和后面的层直接连接，使后面的层可以直接对残差进行学习，ResNet 和普通的卷积神经网络所具有的最大的区别在这

里。这种结构也被称为 skip connections 或 shortcut。

对传统的卷积层来说，总会存在信息丢失等问题。ResNet 在某种程度上解决了这个问题，通过将输入信息直接传到输出，从而达到简化学习目标和难度。

通过 shortcut connection 实现这个残差网络结构，网络的参数和计算量并不会因为此加法增加，同时但却可使模型的训练速度加快、训练效果提高，退化问题能够得到较好的解决。

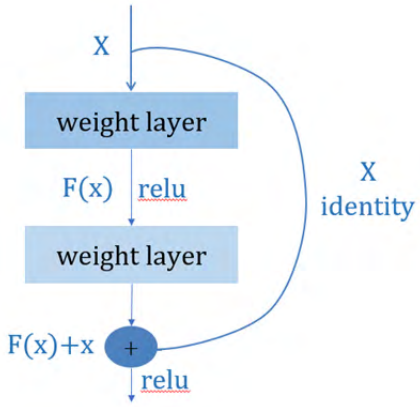

图 20　残差单元图

（三）基于 pytorch 的 resnet34 网络模型

ResNet 网络结构是在 VGG、GoogLeNet 网络后又一经典网络。VGG、GoogLeNet 网络逐渐加深网络结构会出现准确率下降的问题。其主要原因是网络层数加深会引起梯度消失与梯度爆炸。由微软研究院的何凯明等提出残差网络结构，在网络结构中引入了残差单元，如图中虚线框部分所示。残差单元多次堆叠进而构成了 ResNet18、34、50、101 的网络结构，层数远远增大，准确率也出现了较大的提升。本文选用了 ResNet34 网络结构，如图 21 所示。

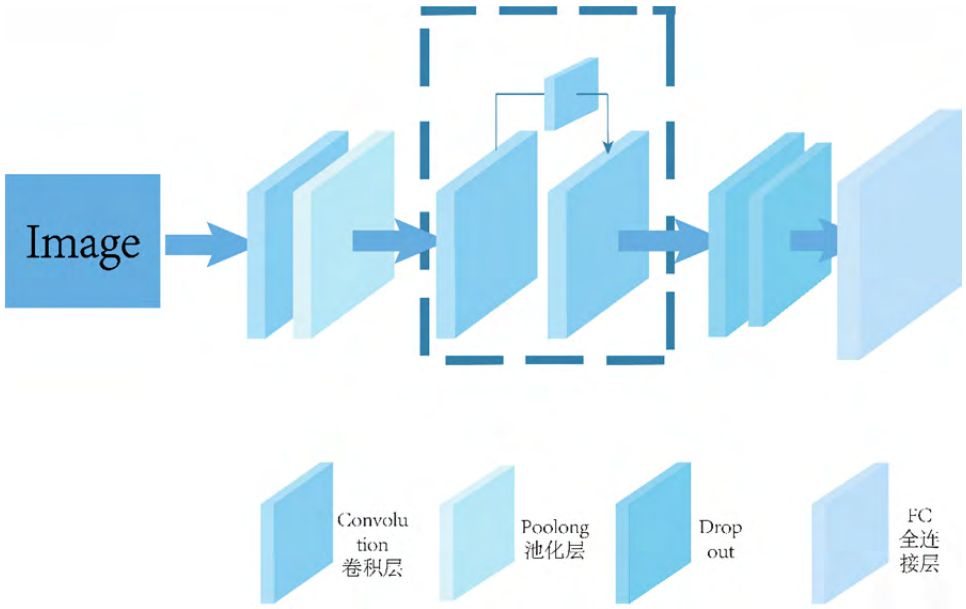

图 21　ResNet34 网络结构

本文将训练样本分为 0、1，即没有心脏病和有心脏病两类图像。将图片数据预处理后，本文用 ResNet34 残差神经网络模型对 3607 张分类训练集图片进行训练，经过训练得出网络模型训练权重。利用训练好的模型权重，再用 ResNet34 残差神经网络模型对余下的 110 张验证图片进行验证，得出 110 张训练集图片的预测结果。预测结果内容为 class（图像属于的分类）和 prob（在此分类下的概率）。其中一个预测样本的预测结果如下图所示，此心电图预测结果为没有患有心脏病的预测概率为 0.943。从心电图的样本图像中，图像的 ST 段有所抬高，考虑到心肌缺血，具有较大概率患有冠心病，这与图像分类模型预测结果相吻合。在对预测集进行的预测中，预测准确度达到 95%以上，且即使在预测错误的个案中，预测得出的 prob 也较低，能够为医务人员提供较为清晰准确的引导和支持，为进一步的检查提供帮助。

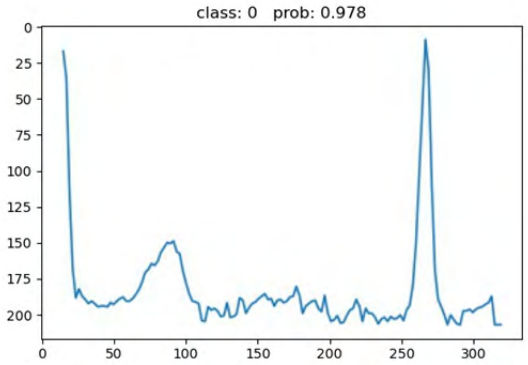

图 22　心电图预测结果图

五、主要结论和挑战

（一）主要结论

朴素贝叶斯模型起源于经典的数学理论，具有理论基础，且分类效率稳定。它对大量的训练和查询具有很高的速度。在使用大规模训练集时，每个项目的特征数相对较少。该算法对小规模数据有很好的处理能力，能够处理多种分类任务，适合于增量训练。它对缺失数据不敏感，算法相对简单，结果解释容易理解。

朴素贝叶斯模型需要预先计算先验概率，对输入数据的表达形式较为敏感。由于朴素贝叶斯模型使用了样本独立性假设，故若样本属性具有关联时，其效果较差。

随机森林算法由于采用集成算法的思想，其精度优于大多数单一算法，准确性较高。同时由于两个随机性的引入使随机森林具有不易陷入过拟合的优点。同时作为非线性分类模型，随机森林可以处理非线性数据，对数据集的适应力强，同时无需数据集规范化。

但是当随机森林的决策树的个数较多的时候，训练需要的时间和空间的开

销较大，在噪音较大的样本集上，随机森林容易陷入过拟合的问题。

非线性映射能力是 BP 神经网络的一个重要功能，它可以充分完成从输入到输出的映射，通过 3 层神经网接近任意精度的非线性连续函数。对于内部机制复杂的问题，BP 新轻网模型具有很强的非线性映射能力，BP 神经网具有较强的自学适应能力，在训练过程中，该模型独立提取输出数据和输出数据之间的规则，并将学习内容从网络的可独立应用于值。

虽然 BP 神经网络的精度较好，但也存在一些不足。BP 神经网络具有对初始的权重非常敏感的特性，极有可能存在使算法陷入局部极值的问题，权值收敛则到局部极小点，常常导致网络训练失败的发生。神经网络算法的收敛速度较慢，会出现“ 锯齿形现象 ”，导致算法低效。神经网络对样本依赖性较强，选取合适的训练集也是预测准确的关键。

在最后本文将上述三种模型进行集成学习。假设决策的差异性是各弱分类器间具有的，分类决策边界不同会出现，即在决策时很可能犯下有差异的错误。将它们通过某种算法结合后，就能得到更加正确的边界，整体的错误就得以减少，更好的分类效果由此实现。本文采用的 Stacking 算法分为两个阶段，首先多个基础的分类模型使用来预测分类。然后，一个全新的学习模块与其预测结果相结合，以此降低泛化误差。

本文分别利用朴素贝叶斯、随机森林和 BP 神经网络三种模型，对患者生理特征数据和患有冠心病情况进行预测。朴素贝叶斯模型，较为简单，对数据的训练和处理时具有较高的速度，且理论基础坚实。由于本文数据进行预处理，基本排除数据直接存在可能的关联性，故使用朴素贝叶斯算法得到了较好的预测效果。随机森林模型，相较于朴素贝叶斯模型，需要较大的时间和空间的开销，同时需要人工进行相关参数的确定。在确定了相关参数后，模型预测效果较好，对于处理本文中存在的非线性数据，具有较高的精度。BP 神经网络模型

是一个黑箱模型，对初始的权重非常敏感，本文选取模型中的默认初始权重，并且确定隐藏层数为 3，学习率为 0.1，阈值为 0.05 进行 BP 神经网络模型的建立。BP 神经网络模型收敛时运行速度较慢，预测出的结果具有较高的精度，对于本文数据存在的非线性数据，具有良好的泛化性。但是，患者生理特征数据和患有冠心病情况在朴素贝叶斯、随机森林和 BP 神经网络三种模型下的预测效果都并没有达到较好的精度。于是本文接着将上述三种模型预测得出非患有心脏病（即数据分类为 1）的概率作为自变量，是否患有心脏病作为因变量，构造 logistic 回归模型。在集成学习的改进后，预测的精度达到了 0.8638，模型准确度得到了显著的提高。

表 5　模型准确率对比表

模型	准确度
朴素贝叶斯	0.839
随机森林	0.84
神经网络	0.84
基于 logistics 模型的集成学习	0.8638

对于心电图图像分类问题，采用 ResNet34 网络结构，先对 3607 张图片进行预训练，接着采用训练好的预训练权重，对剩下的 100 张图片进行分类预测，得到的准确率较高。

由于冠心病早期症状不明显，患者自身很难察觉。同时，对于是否患有冠心病具有重要判别意义的心脏造影手术，不仅价格昂贵，同时对患者的身体也有一定的损伤。本研究力图在经济和健康的最小损失下，对人们冠心病的风险进行准确预警和预测。由于患者特征和心电图图像都是最简单，最易获取，且不需要高端的医疗设备下即可获取的数据，本文在这两种模型的基础上分别建立模型，从不同类型的数据进行预测。在上述模型的预测结果的准确度下，此研究对冠心病的进一步筛查具有重要的意义。

（二）研究中存在的挑战

本文研究中可能由于样本的选择出现“伯克松悖论”，这条悖论是医学统计中的一种偏倚，指两个通常独立的事物会在特定场合下关联起来，由此产生的相关性容易带来认知上的偏差，导致两个本来无关的变量之间体现出貌似强烈的相关关系。因为本文中的样本数据来源于医院，这是一个受限样本，所收集的信息基本上是已经患有与相关心血管疾病或身体已经感到不适的就诊人员，这便导致了个体被纳入研究样本的机会有所不同。

当前，我国社会主要矛盾转化为人民日益增长的美好生活需要和不平衡不充分的发展之间的矛盾，在卫生健康领域主要表现为群众对健康有了更高需求，要求看得上病、看得好病，看病更舒心、服务更体贴。诊断模型的每一次优化都有利于提供更高准确率的诊疗和更优质的健康资源，有助于我国实现更可持续、更高质量的发展。

参考文献

[1] 李方舟，苏小婷，孙润宸，等．全球冠心病预测模型准确性的系统评价[J]．中国胸心血管外科临床杂志，2021,28(03):288-298.

[2] 戴小毛．冠心病应用64排128层螺旋CT冠状动脉成像的诊断分析[J]．影像技术，2017,29(05):12-13.

[3] Dutta A, Batabyal T, Basu M, et al. An efficient convolutional neural network for coronary heart disease prediction[J]. Expert Systems With Applications, 2020,159.

[4] 李雨洁，郑锐龙，杨旭明．基于数据挖掘技术的冠心病诊断预测模型[J]．医学信息，2020,33(24):14-17.

[5] 赵金超，李仪，王冬，等．基于优化的随机森林心脏病预测算法[J]．青岛科技大学学报(自然科学版)，2021,42(02):112-118.

[6] 孙建州．贝叶斯统计学派开山鼻祖——托马斯·贝叶斯小传[J]．中国统计，2011(07):24-25.

[7] Lunn D, Jackson C, Best N, et al. The BUGS Book[M]. Taylor and Francis.

[8] 高惠璇．应用多元统计分析[M]．应用多元统计分析，2005.

[9] Johnson R A, Wichern D W. Applied Multivariate Statistical Analysis, 6/E[J]. Technometrics, 2005,47(4):517.

[10] Ge·Ron A. Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems[J]. 2017.

[11] 黄健，林进意，林秋萍．动态心电图联合颈动脉彩超诊断冠心病的临床

研究[J]. 世界复合医学, 2021,7(03):41-43.

[12] 张梦雨. 基于ResNet和注意力机制的花卉识别[J]. 计算机与现代化, 2021(04):61-67.

[13] 赵叶红, 杨卫民. 基于粗糙集和ResNet34网络算法的森林火灾预测研究[J]. 信息与电脑(理论版), 2020,32(20):61-63.

[14] 王跃, 王卫东, 赵蕾, 等. 基于迁移学习的胃镜图像自动识别多分类系统的研究[J]. 中国医疗设备, 2021,36(03):81-84.

[15] 闵宇航. 基于深度残差网络ResNet-50的绿萝状态识别系统[J]. 科学技术创新, 2021(08):92-93

[16] 宋碳. 基于ICU病人电子病历数据的死亡率预测分析的研究[D]. 黑龙江大学, 2019.

[17] 张惠玲. 基于股吧文本的主题挖掘及其股票投资应用[D]. 华南理工大学, 2018.

[18] 张吉刚, 梁娜. 基于灰色BP网络的企业财务绩效评价研究[J]. 中国商贸, 2013(26):84-85.

[19] 钟杰. 基于计算机视觉的危险车间人员检测及定位技术应用研究[D]. 南京理工大学, 2019.

致谢

在本次论文的撰写中，我们既运用了平时学习的专业知识，也学习到之前未曾接触过领域的知识，提高了学习能力和专业知识应用实践能力。同时，小组分工合作的方式让我们学会如何与其他成员沟通配合，使得每一位小组成员都能够发挥其优势。

感谢我们的指导老师 XXX 老师和 XXX 老师，他们在主题的拟定、方法的选择和论文的写作过程中都为我们提供了诸多建议与帮助，他们用扎实的专业知识是指引着我们学习的方向，他们严谨的治学精神不断激励着我们，使我们充满信心迎难而上，直面困难。

最后，再次感谢比赛过程中给予我们帮助的老师和同学们！