参赛队号：

2021年（第七届）全国大学生统计建模大赛

参赛学校	陆军军医大学
论文题目	基于脑出血患者院前指标的多种机器学习预测模型构建及比较研究*
参赛队员	王浩淼　曹若菲　林金欣
指导老师	伍亚舟

*本研究受国家自然基金项目（No. 81872716）资助

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

基于脑出血患者院前指标的多种机器学习预测模型构建及比较研究

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目 录

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表格和插图清单

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

摘 要

目的： 卒中是全球致残率、致死率最高的疾病之一，其中脑出血（Intracerebral hemorrhage，ICH）患者仅占卒中患者的20%，但致残、死亡人数与缺血性卒中无明显差异，由于病程短，发展迅速的特点，以农村地区为代表的经济欠发达地区，受限于交通以及基层医疗水平的欠缺，往往无法得到完整、可信的医疗检查从而不利于正确医疗干预的实施，给国家及患者家庭带来了巨大的精神与经济负担。目前ICH患者神经功能预后预测模型多基于传统方法如Logistic回归，主要关注院中指标对神经功能预后的影响，往往忽略院前指标的重要作用。ICH患者的救治时间窗短，须在有限时间内快速制定合理、有效的治疗方案，因此利用可快速获取的院前指标建立神经功能预后模型，阐明各指标变量在疾病发展的中发挥的作用，并利用该类指标变量建立治疗方案相关指南是未来研究的重要方向。

方法： 目前，国内基于有监督的机器学习方法对ICH患者神经功能预后分类模型相关研究较少，高性能预测模型缺乏，不利于脑出血相关临床实验的观察单位的纳入排除标准设置，极大程度阻碍了脑出血患者的前瞻性研究设计，导致相关疾病高质量循证医学证据进展迟缓。

结果： 本文首先使用Logistic回归模型、Lasso回归模型，对院前指标进行筛选，证明了院前指标相较于以往认为占据主导因素的院中指标，同样对ICH患者神经功能预后有重要影响，后利用多种新型机器学习方法如支持向量机、随机森林等建立神经功能预后预测模型，通过多模型间性能的比较研究，阐述了决策树类算法及其衍生算法随机森林在信息数据共享的大数据时代，在预测神经功能预后普遍不良的脑出血疾病的显著优势。

关键词：脑出血；院前指标；机器学习；比较研究

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一、前言

（一）研究背景与意义

2019 年全球疾病负担研究（the Global Burden of Disease (GBD) Study 2019）数据表明，脑卒中是导致全球死亡和致残的第二大原因,仅次于心脏病致死。其中出血性脑卒中占比 27.9%，约有 341 万例脑出血患者[1]。近年来，出血性脑卒中负担的增加主要源于人口老龄化和人口增长。《中国脑卒中防治报告 2018》[1]数据表明，2016 年中国出血性脑卒中发病率为 126.34/10 万，且患病率和发病率呈逐年上升趋势[2]。《中国卒中报告 2019》数据表明，2018 年中国卒中患者约有 300 万例，其中出血性脑卒中占比 14.9%，约有 44.7 万例。据统计，2018 年中国约有 2.1 万例脑出血患者死亡，占比 4.7%[3]。脑出血（intracerebral hemorrhage，ICH）是出血性脑卒中最主要、预后最差的类型,诊断标准为非创伤性因素导致脑内血管破裂,血液在脑实质内聚集,其症状突发,多见于情绪激动、用力过猛、劳累过度，患者常见症状包括头痛、恶心呕吐、意识障碍和肢体瘫痪等。在人口基数大、人口老龄化加重、国民饮食结构以及生活方式发生不断变化的国情背景下，ICH 的发病率逐年攀升，为家庭和国家造成了严重的经济负担。鉴于目前尚无有效手段改善 ICH 患者神经功能预后，分析、挖掘 ICH 患者脑出血的院前指标在疾病发展进程中的关键作用,以阐明各类指标变量与神经功能预后的相关性,对临床实验的设计以及有效干预措施的开发十分重要。

（二）研究现状及不足

临床上脑出血的诊断与评估中，就诊患者的病史与入院体征是必不可少的。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[4]其中病史采集包括：患者的年龄、性别、是否有高血压、糖尿病、冠心病和既往脑卒中等病史、是否吸烟、喝酒等等；一般体格检查和影像学检查包括：收缩压、舒张压、脑出血部位、是否破入脑室、出血量、是否发生脑疝等；常用的量表包括：格拉斯哥昏迷量表（Glasgow Coma Scale，GCS）、脑出血评分量表和美国国立卫生研究院卒中量表（National Institute Health Stroke Scale，NIHSS）。其中 GCS 是目前医学上应用最广的评估病人昏迷程度的指标，通过对病人的睁眼反应、语言反应和肢体运动等各种生理行为进行评估计分加和获得。NIHSS 是目前世界上广泛使用的对脑卒中评价的指标，比较全面地评价了脑卒中后的功能障碍。

根据以上检查，全面评估患者脑出血严重程度，采取相应的措施进行治疗，包括保守治疗和手术治疗。大多数患者均以内科保守治疗为主，而病情危重且有手术适应症者，可进行外科手术治疗。由于脑出血发病急、进展快以及大脑作为中枢神经系统高级部位，是生命机能的主要调节器等特点，脑出血患者早期死亡率较高，通过有效救治的患者往往也会出现情绪认知、运动功能、言语表达等功能障碍，常用改良 Rankin 量表（mRS）衡量卒中后患者神经功能恢复的状态。

临床治疗表明，不同患者的入院前情况及院内治疗措施等因素对脑出血患者致残程度和康复治疗有不同的影响。2013 年，有关研究者回顾性分析苏北人民医院 2008 年至 2011 年自发性脑出血青年患者的临床资料，采用逐步 Logistic 回归分析 ICH 患者预后良好与死亡的影响因素，结果表明 NIHSS 评分和 GCS 评分对预后判断具有较高价值，收缩压及血肿体积对预后也有较好的预测作用[5]。2018 年有关研究者回顾性分析南京中医药大学附属八一医院 2014 年至 2017 年收治的 134 例自发性脑出血患者临床资料，采用 t 检验和 χ^2 检验、多元 logistic 回归分析、Pearson 相关分析等方法，得出结论，年龄、入院收缩压、入院舒张

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

压、高血糖、高血压和高脂血症等病史以及 NIHSS 评分等指标在预后良好组与预后不良组之间具有统计学差异[6]。2021 年，有关研究者回顾性分析广东医科大学附属第一医院 2017 年至 2019 年间收治的 50 例老年患者和 50 例中青年患者，采用 t 检验和χ^2 检验探究不同年龄组 ICH 患者临床特征与预后，发现不同年龄组 ICH 病因不同，老年组主要病因在于高血压，中青年组主要病因为劳累和饮酒[7]。通过查阅文献，我们可以知道，以往国内大多数研究者通过回顾性分析脑出血患者入院前指标、院中治疗方法以及院内并发症等指标，采用假设检验如 t 检验和χ^2检验以及 Logistic 回归分析的传统统计方法探究影响 ICH 患者预后的独立危险因素。以往国内外研究者大多把性别、年龄、院前病史如糖尿病史和高血压史、院内药物使用史、院内血常规检查指标、院内影像学检查血肿体积、血肿位置、院内手术治疗方式等指标，用选定的机器学习方法，如 2021 年 Andrew N Hall 等人基于决策树的算法构建了用于预测 ICH 患者预后的高性能模型[8]。

（三）研究目的与内容

ICH 患者神经功能预后普遍较差，治疗方案拟定难度大，治疗效果低于预期，甚至出现不利影响，因此如何利用 ICH 患者病史资料以及入院体征，结合目前最新的机器学习分类模型构建神经功能预后预测模型是改善脑出血患者结局的重要问题。入院前指标，包含吸烟、喝酒、既往卒中史等既往病史资料等以及入院体征检查如 NIHSS、GCS 等量表评分，是制定治疗方案前 ICH 患者最易获取的指标变量，但是目前的研究者往往忽略院前指标的重要作用，故本文旨在探究 ICH 患者院前指标对患者神经功能预后的影响，并尝试建立一种性能较好的预测模型，为临床治疗提供一定的指导。

国内既往研究多采用传统的统计方法建模，不适用于复杂的临床情况以及临

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

床需求，本文采用有监督的分类机器学习方法，希望利用院前指标建立 ICH 患者神经功能预后的相关预测模型，并且比较研究基于各种机器学习方法建立的预测模型性能、适用场景以及优化方案，机器学习方法涵盖经典决策树、条件推断树、随机森林、支持向量机等五种机器学习方法。

本文技术路线如图 1 所示：

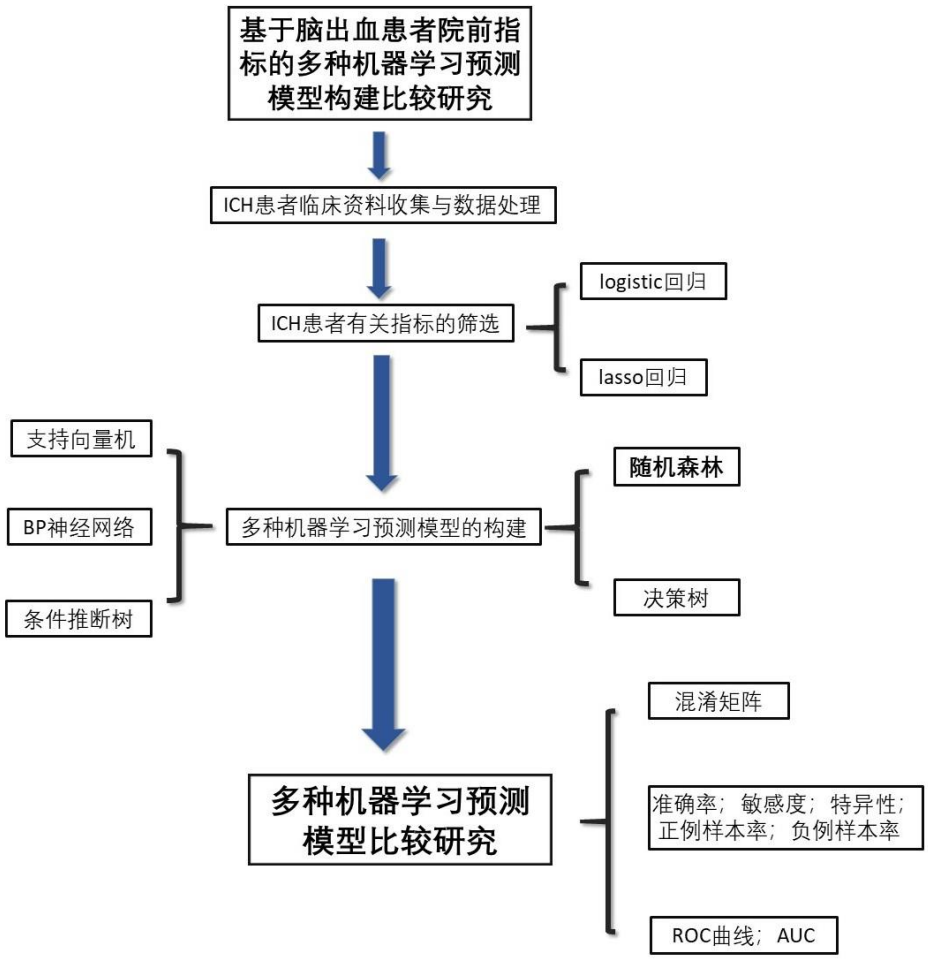


图 1　技术路线图

二、模型构建

本文首先以 ICH 患者院前指标及院中指标为自变量，神经功能预后情况为

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

结局变量，采用 Logistic 回归、lasso 回归进行变量筛选，探究各种指标对 ICH 患者神经功能预后是否有显著影响。后以 ICH 患者院前指标为自变量，神经功能预后情况为结局变量，训练支持向量机、BP 神经网络、决策树及随机森林等五个机器学习方法，以比较研究各类模型在基于院前指标预测 ICH 患者神经功能预后的性能。

（一）影响因素筛选方法

1.Logistic 回归（Logistic Regression）

Logistic 回归是处理二分类资料的标准方法。属于概率型非线性回归，多用于临床医学的鉴别诊断、评价治疗策略、分析疾病预后因素等。

设有一个应变量 Y 和 m 个自变量X_1，X_2，…，X_m，应变量 Y 是个二值变量时，取值为

$$Y=\begin{cases}1 & \text{阳性结果}\\ 0 & \text{阴性结果}\end{cases}$$

在一组自变量作用下阳性结果的发生概率记为$P=P(Y=1|X_1，X_2，\ …，X_m)$，简记为 P，则 Logistic 回归模型可以表示为

$$P=\frac{1}{1+\exp[-(\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)]} \qquad (2\text{-}1)$$

其中，α为常数项，$\beta_1,\beta_2,\ldots,\beta_m$为回归系数，将上述公式适当变换，得到 Logistic 模型的线性化表达：

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \qquad (2\text{-}2)$$

公式左端为阳性发生概率与阴性发生概率的比值再取自然对数，记作 logit(P)，Y 于是有

$$logit(P)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \qquad (2\text{-}3)$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

该公式称为 logit 模型。

记$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$，那么 Z 和 P 之间关系的 Logistic 曲线如下图 2 所示

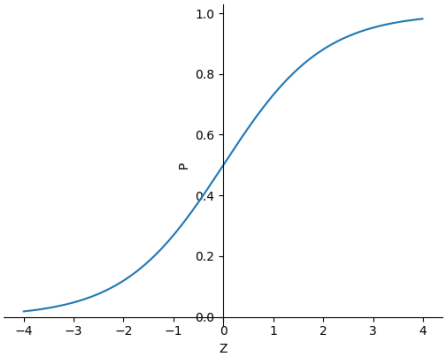


图 2　Logistic 曲线

当$Z \to +\infty$时，P 值渐近于 1，当$Z \to -\infty$时，P 值渐近于 0；P 值的变化在 0~1 的范围内变化，随 Z 值的增加或减小以点（0，0.5）为中心呈对称 S 形变化。该特点使得 Logistic 回归模型的这些点可以较好地用于生物反应资料。

Logistic 回归模型参数的意义：

回归系数$\beta_j (j = 1,2,\ldots,m)$表示的是因素$X_j$改变一个单位时 logit(P)的改变量，与衡量因素作用的比数比（odds ratio, OR,总体参数用 ψ 表示）有对应关系。如果对比某一因素的两组不同水平$X_{j(1)}$与$X_{j(0)}$的对应阳性结果，并假定其他因素的取值相同，则比数比的自然对数为

$$ln\psi_j = ln\left[\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\right] = logit(P_1) - logit(P_0) = \left(\alpha + \beta_j X_{j(1)} + \sum_{t\neq j}\beta_t X_t\right) - \left(\alpha + \beta_j X_{j(0)} + \sum_{t\neq j}\beta_t X_t\right) = \beta_j\left(X_{j(1)} - X_{j(0)}\right) \quad (2\text{-}4)$$

为$\psi_j = \exp\left[\beta_j\left(X_{j(1)} - X_{j(0)}\right)\right]$其中，$P_1$和$P_0$分别为$X_j$取值为$X_{j(1)}$和$X_{j(0)}$时的发病概率。特别地，若$X_j$赋值为

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$X_j = \begin{cases} 1 & 有 \\ 0 & 无 \end{cases}$$

则有和无X_j因素作用的两组的比数比为

$$\psi_j = \exp(\beta_j) \qquad (2\text{-}5)$$

当$\beta_j = 0$时，$\psi_j = 1$，说明X_j对阳性结果的发生不起作用；

当$\beta_j > 0$时，$\psi_j > 1$，说明X_j对阳性结果的起正向作用；

当$\beta_j < 0$时，$\psi_j < 1$，说明X_j对阳性结果的起负向作用。

一般而言，两组不同水平的$X_{1(1)}, X_{2(1)}, \ldots, X_{m(1)}$和$X_{1(0)}, X_{2(0)}, \ldots, X_{m(0)}$的综合比数比为

$$\psi = \exp\left[\sum_{j=1}^{m} \beta_j \left(X_{j(1)} - X_{j(0)}\right)\right] = \prod_{j=1}^{m} \exp\left[\beta_j \left(X_{j(1)} - X_{j(0)}\right)\right] = \psi_1 \psi_2 \ldots \psi_m \qquad (2\text{-}6)$$

由于ψ_j的值与常数项 α 无关，所以通常在因素重要性分析中将 α 看作无效参数。

变量筛选的基本思路：按事先规定的显著性水平，利用固定的算法把统计显著的变量筛选入模型，而将不显著的剔除在外。Logistic 逐步回归选用检验统计量为似然比统计量、Wald 统计量和记分统计量之一。大多统计软件使用似然比统计量。

一般对模型中的一部分回归系数是否为 0 作出检验，检验假设为

$$H_0: \beta_j = \beta_{j+1} = \cdots = \beta_{j+d} = 0$$

更经典的问题对一个回归系数作检验，检验假设为：

$$H_0: \beta_j = 0$$

似然比检验：基本思想为比较两种不同假设条件下的对数似然函数值，比较它们之间的差别大小。具体做法为拟合一个不包含准备检验因素在内的 Logistic 模型，求其对数似然函数值lnL_0，再将需要检验的因素放入模型中进行

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

配合，得到新的对数似然函数值lnL_1，假设前后两个模型分别含有自变量个数为 l 个和 p 个，计算似然比统计量 G 的公式为

$$G = 2(lnL_1 - lnL_0) \quad (2\text{-}7)$$

当样本含量较大时，在零假设下得到的 G 统计量近似服从χ^2分布，自由度 d=p-l。如果$G \geq \chi^2_{a,d}$时，表示新加入的 d 个自变量对回归有显著的贡献。

2.Lasso 回归(The Least Absolute Shrinkage and Selection Operator)

Lasso 回归模型于 1996 年由 Robert Tibshirani 提出，是一种在最小二乘法的基础上增加惩罚项对样本数据进行变量选择，减弱非显著性变量，筛选重要特征变量，从而实现数据降维的方法。

其基本思想为：通过设置惩罚项，使残差平方和在回归系数绝对值之和小于某值的情况下最小化，从而使某些回归系数压缩至零，将其所对应变量视作非显著性变量，筛选得到对应变量有关键影响的变量。[9]

Lasso 采用的是 L1 正则化：

$$\hat{\beta}(\lambda) = arg\,min\left(\frac{\|Y-X\beta\|_2^2}{n} + \lambda\|\beta\|_1\right) \quad (2\text{-}8)$$

$$\|Y - X\beta\|_2^2 = \sum_{i=1}^{n}(Y_i - (X\beta)_i)^2 \quad (2\text{-}9)$$

$$\|\beta\|_1 = \sum_{j=1}^{d}|\beta_j|, \quad \lambda \geq 0 \quad (2\text{-}10)$$

为一个惩罚项。复杂度的调整程度由 λ 控制，λ 越大，惩罚力度越大，获得的关键变量越少[9]。

（二）多种预测模型构建

1.BP 神经网络（Back Propagation Neural Network）：

1986 年由 Rumelhart 和 McCelland 在 Nature 发表论文 Learning

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

representations by back-propagating errors 提出的BP神经网络，是一种按误差逆传播算法训练的多层前馈神经网络，可解决非线性分类问题。其网络拓扑结构包括输入层、隐层（一层或多层）和输出层。

其基本思想是：将训练样本作为输入信号，沿网络拓扑结构正向分析，得到实际输出值。误差信号，即实际输出值与期望输出值的差值，沿网络拓扑结构反向回馈，通过不断修正调节网络权值，得到更接近期望输出信号的实际输出值。

实现步骤如下：

首先，输入训练样本 $\boldsymbol{X_k} = [x_{k1}, x_{k2}, \cdots, x_{kM}], (k = 1,2, \cdots, N)$，得到隐层各级神经元的输入信号$\boldsymbol{u}$和输出信号$\boldsymbol{v}$，$v_p^P(n) = y_{kp}(n), p = 1,2, \cdots, P$。然后，计算误差 $\boldsymbol{E}(n) = Y_k(n) - d_k$。迭代计算各级神经元的局部梯度 δ，公式为

$$\begin{aligned} \delta_p^P(n) &= y_p(n)\Big(1 - y_p(n)\Big)\Big(d_p(n) - y_p(n)\Big), p = 1,2, \cdots, P \\ \delta_j^J(n) &= f'\Big(u_j^J(n)\Big)\sum_{p=1}^{P}\delta_p^P(n)w_{jp}(n), j = 1,2, \cdots, J \\ \delta_i^I(n) &= f'\Big(u_i^I(n)\Big)\sum_{j=1}^{J}\delta_j^J(n)w_{ij}(n), i = 1,2, \cdots, I \end{aligned} \tag{2-11}$$

从而计算权值修正量Δw，公式为

$$\begin{aligned} &\Delta w_{jp}(n) = \eta\delta_p^{\mathrm{P}}(n)v_j^J(n) \quad w_{jp}(n+1) = w_{jp}(n) + \Delta w_{jp}(n) \quad j = 1,2, \cdots, J; p = 1,2, \cdots, P \\ &\Delta w_{ij}(n) = \eta\delta_j^J(n)v_i'(n) \quad w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n) \quad i = 1,2, \cdots, I; j = 1,2, \cdots, J \\ &\Delta w_{mi}(n) = \eta\delta_i^I(n)x_{km}(n) \quad w_{mi}(n+1) = w_{mi}(n) + \Delta w_{ml}(n) \quad m = 1,2, \cdots, M; i = 1,2, \cdots l \end{aligned} \tag{2-12}$$

直至学完所有训练样本$\boldsymbol{X_k}$。

2.SVM（Support Vector Machine）：

1995年Vapnik等人在统计学习理论SLT的基础上首次提出了一种新的有监督机器学习方法，即支持向量机的概念（Support Vector Machine，SVM）。该算法能够从理论上实现对不同类别间的最优分类，拥有较好的泛化能力。SVM是二分类算法，把具有多个属性的数据分为两类。

为进一步解决非线性问题，Vapnik 等人提出了一种通过将核技巧用于最大

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

间隔超平面来创建非线性分类器的方法。除了每一个点积都被一个非线性核函数代替之外，所得到的算法与线性样本分类在形式上是类似的。非线性分类器算法在转换成线性分类后的特征空间中拟合最大间隔超平面，实现分类。

实现步骤为：先给一组训练样本数据打上类别标签，让 SVM 模型使用这些打了类别标签的数据进行训练，训练后，给训练好的 SVM 模型新的无类别标签的数据，即输入测试集数据，SVM 模型就可以自动对这些新的数据分类，成为非概率二元线性分类器。

假设现在我们想要学习一个非线性分类规则，首先，给出一组非线性带标签的输入样本$\vec{x_i}$，$i=1,\dots,n$，及其对应的期望输出$y_i \in \{+1,-1\}$，$\vec{x_i}$对应于转换数据点的线性分类规则$\varphi\left(\vec{x_i}\right)$。然后，给出一个核函数$k$，满足$k\left(\vec{x_i},\vec{x_j}\right)=\varphi\left(\vec{x_i}\right)\cdot\varphi\left(\vec{x_j}\right)$。我们知道向量$\vec{w}$的分类在转换空间中满足

$$\vec{w}=\sum_{i=1}^{n} c_i y_i \varphi\left(\vec{x_i}\right) \tag{2-13}$$

其中c_i可以通过优化问题求解获得。对于所有i，在约束条件$\sum_{i=1}^{n} c_i y_i = 0$和$0 \le c_i \le \frac{1}{2n\lambda}$下

$$\begin{aligned} f(c_1 \dots c_n) &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i \left(\varphi\left(\vec{x_i}\right)\cdot\varphi\left(\vec{x_j}\right)\right) y_j c_j \\ &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i k\left(\vec{x_i},\vec{x_j}\right) y_j c_j \end{aligned} \tag{2-14}$$

达到最大化。其中系数c_i可以通过二次规划求解获得。同样，我们可以找到一些i值，使得$0<c_i<\frac{1}{2n\lambda}$，从而使得$\varphi\left(\vec{x_i}\right)$位于转换空间的间隔的边界上，然后用$i$值求解

$$\begin{aligned} b=\vec{w}\cdot\varphi\left(\vec{x_i}\right)-y_i &= \left[\sum_{j=1}^{n} c_j y_j \varphi\left(\vec{x_j}\right)\cdot\varphi\left(\vec{x_i}\right)\right]-y_i \\ &= \left[\sum_{j=1}^{n} c_j y_j k\left(\vec{x_j},\vec{x_i}\right)\right]-y_i \end{aligned} \tag{2-15}$$

最后，新的没有带标签的非线性数据点通过此模型计算，进而达到二元线性

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分类目的。

$$\vec{z} \mapsto \operatorname{sgn}\left(\vec{w} \cdot \varphi\left(\vec{z}\right) - b\right) = \operatorname{sgn}\left(\left[\sum_{i=1}^{n} c_i y_i k\left(\vec{x}_i, \vec{z}\right)\right] - b\right) \quad (2\text{-}16)$$

3.决策树(Decision Tree)

1966 年，概念学习系统（Concept Learning System，CLS）提出决策树算法的概念。1979 年，J.R Quinlan 基于 CLS 提出 ID3 算法扩大了决策树模型的适用范围。1993 年，J.R Quinlan 在 ID3 算法基础上提出 C4.5 算法再次使得该算法能够有效地解决诸多实际问题。此外，ID3 还发展出另一种分类回归决策树算法（CART）。决策树由特征节点和分类叶节点组成，从根节点到一个特定的叶节点的路径构成一个具体的输出类认知表示（Yi）j，其层次结构画成的图形像一棵树，因此称为决策树。从根节点到任意叶节点的路径都可以生成一条唯一的 if-then 规则:

$$\begin{aligned} &\text{if } (x)_1 < w_1, \text{ then } \underline{Y}_1; \\ &\text{if } (x)_1 \geqslant w_1 \text{ and } (x)_2 < w_2 \text{ , then } \underline{Y}_1 \text{ ;} \\ &\text{if } (x)_1 \geqslant w_1 \text{ and } (x)_2 \geqslant w_2, \text{ then } \underline{Y}_2 \end{aligned} \quad (2\text{-}17)$$

路径上的每一个内部节点都是规则的条件，路径上的叶节点就是该规则的结论。

构建决策树的基本思想是以信息熵为度量，构造一棵熵值下降最快的根节点，即先判断对结果影响大的条件，到叶节点处时熵值将为零，即做出决策。按照变量选择的不同的标准可以将决策树类模型划分为经典决策树与基于条件推断的条件推断树。

决策树算法是最常用的分类算法之一，具有良好的解释性，是一种可将分类过程与结果可视化的分类算法。但通过上述基本思想生成的决策树可能由于训练集合的噪音和异常值而产生许多不必要的分支，这导致了决策树算法可能过拟合。为避免生成过于复杂的决策树，应裁掉一些子树和叶节点，称为决策树的剪枝。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.随机森林(Random Forest）

随机森林是一种常用的机器学习算法，是 bagging 算法的扩展变体，它以决策树为基本单元并将多个决策树的输出结合起来得到一个结果，属于集成学习[11]。当一个新样本需要归类时，它的结果不是仅仅取决于某一棵决策树的结果，而是让森林里每一棵决策树都产生一个结果，看看哪一类得到结果多，选出结果最多的那类作为输出。我们可以知道构建随机森林实质就是构建多棵决策树。随机森林里面的每一棵决策树都是没有关联的。随机森林可以同时处理分类和回归问题，具有易用性和灵活性。

模型构建实现步骤：

① 随机抽取数据：数据集总量为 N，随机有放回地从中抽取数据，构成子数据集，作为子决策树的训练集。

② 特征的随机选取：样本存在 M 各指标变量，每次分类都随机地从中选择 m 个（满足 m 小于 M），再从这 m 个变量中选择最佳的分类标准。从而保证各决策树之间的低相关性，提升了模型的效能[9]

③每棵树都尽量最大程度地生长，无剪枝过程

④将生成的多棵树决策树组成随机森林，对于分类问题，采用简单投票原则决定最后输出结果；对于回归问题，由多棵决策树预测值的平均数构成最终的预测结果[12]。

（三）模型评价指标

1.混淆矩阵（Confusion Matrix）

混淆矩阵是用于评价模型分类能力的 N×N 矩阵。在本文中由于结局变量为二分类变量，因此采用 2×2 矩阵，将实际类别与预测类别进行比较，以此评价

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

模型的分类预测效果。

	实际为正样本	实际为负样本
预测为正样本	TP	FP
预测为负样本	FN	TN

混淆矩阵包含四部分信息：真正类（True positive，TP)，表明实际是正样本预测成正样本的样本数。假正类（False positive，FP)，表明实际是负样本预测成正样本的样本数。假负类（False negative，FN)，表明实际是正样本预测成负样本的样本数。真负类（True negative，TN)，表明实际是负样本预测成负样本的样本数。

然而，混淆矩阵得到的统计结果是数量，不能直接由此判断模型性能优劣，因此一般计算由混淆矩阵衍生的另外五个评价指标，计算公式如下：

$$准确率 = \frac{TP+TN}{TP+FN+FP+TN} \qquad 正例命中率 = \frac{TP}{TP+FP}$$

$$敏感度 = \frac{TP}{TP+FN} \qquad 负例命中率 = \frac{TN}{TN+FN}$$

$$特异度 = \frac{TN}{FP+TN}$$

模型预测效果评判中，最常用的一个统计量是准确率（Accuracy,ACC)，指被正确分类样本单元所占比重，即分类器是否总能正确划分样本单元的概率。敏感度（Sensitivity）指成功预测正样本的概率，衡量分类器对正样本的识别能力。同样地，特异度（Specificity）指成功预测负样本的概率，衡量分类器对负样本的识别能力。正例命中率（Positive Predictive Value）或精确度（Precision)指预测正确的样本与被预测为正样本之比。负例命中率(Negative Predictive Value）指预测正确的样本与被预测为负样本之比。

其中准确率适用于多分类评估，将多分类问题转换成二分类问题进行求解，将关注的类化为一类，其他所有类化为一类。正例命中率和负例命中率是二分类

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

指标。分析指标时，我们希望这两个评价指标都越高越好，然而事实上这两者在某种情况下是矛盾的，正例命中率高时，负例命中率低，本文所构建的模型也表现出该特性。因此在不同应用条件下需要根据实际需求判断哪个评价指标更重要。

2.受试者工作特征曲线（Receiver Operating Characteristic Curve，ROC）

ROC 曲线是以敏感度为纵坐标，特异度为横坐标绘制的曲线。其绘制过程：对所有样本按预测概率排序，以每个样本的预测概率为阈值，计算相应的正例命中率和负例命中率，以线段连接即可。ROC 简单直观，通过图示即可直观观察模型的准确性，ROC 曲线越靠近左上角，表示模型准确性越好。本文将各分类器模型的 ROC 曲线绘制在同一坐标中，方便直观判断各模型的预测性能。

AUC（Area Under ROC Curve）：是另一个评价二分类模型的指标，其定义为 ROC 曲线与横坐标（特异度）之间的面积之和。 AUC＞0.5 表示该算法模型有训练效果，AUC 值越大表示模型分类性能越好，AUC=0.5 表示该算法模型没有效果，AUC＜0.5 表示出现标签标注错误等情况。

三、数据采集及处理

（一）数据来源

采用回顾性研究的方法，依据预设的纳入排除标准，收集 2000 年至 2004 年的 ICH 患者人口学资料、疾病发病情况、既往病史、个人史、入院生命体征以及治疗方式等进行评估后入组。鉴于近 20 年间脑出血患者的诊治手段并未出现实质性进展，致死率及致残率均未显著改变，同时该研究主要聚焦于对比各类机器学习模型应用于 ICH 患者基于院前指标预后评估的适用性，因此该回顾性数据能够满足本研究需求。结局变量按照 mRS＜3 划分为预后良好组，mRS≥3 划分为预后不良组。分析指标包括名义变量：性别、有无高血压、有无糖尿病、有无冠

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

心病、是否吸烟、是否饮酒、既往卒中史、是否发生脑疝，数值变量：年龄、收缩压、舒张压、入院是 GCS 量表评分、入院时 NIHSS 量表评分等 13 个指标变量。剔除有缺省值的观察单位，并按照表 1 方式对义变量进行数值替换。本文在使用同一组数据，将数据集随机的划分为数据量满足 7:3 的训练集与测试集，在设定相同的 seed 值(=95453)的情况下，建立多种预测模型以增强实验的可重复性。

纳排入除标准：①纳入标准：年龄≥18 岁，经影像学检查确诊的脑出血患者；②排除标准：患者为由脑外伤、脑肿瘤、脑动脉瘤和动静脉畸形等疾病所引起的继发性脑出血，或严重心肝肺肾等重要脏器功能不全者。

诊断标准：①吸烟：依据 1997 年 WHO 制定的吸烟标准，诊断连续或累计吸烟 6 个月及以上为吸烟患者；②饮酒：每日监控 ICH 患者的饮酒量以及酒品的酒精度数，计算日均值，计算每日酒精摄入量（依照 Sacanilla 公式，酒精摄入量为日饮酒量与酒精度数乘积的 80%）＞30g 且持续时长超过一年患者诊断为饮酒。

表 1　编码变量对照表

变量	编码	变量	编码	变量	编码
性别男	0	脑疝	0	无气管切开	1
性别女	1	无脑疝	1	上消化道出血	0
有糖尿病	0	手术治疗	0	上消化道不出血	1
无糖尿病	1	保守治疗	1	再出血	0
有冠心病	0	出血部位	0	无再出血	1
无冠心病	1	出血部位	1	脑梗死	0
吸烟	0	出血部位	2	无脑梗死	1
不吸烟	1	出血部位	3	颅内感染	0
饮酒	0	出血部位	4	无颅内感染	1

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

不饮酒	1	脑积水	0	癫痫	0
有既往卒中史	0	无脑积水	1	无癫痫	1
无既往卒中史	1	肺部感染	0	神经功能预后良好	0
突入脑室	0	无肺部感染	1	神经功能预后不良	1
无突入脑室	1	气管切开	0		

注：出血部位-0：脑室出血； 出血部位-1：额叶、颞叶、枕叶、岛叶； 出血部位-2：基底节区、丘脑、内囊；出血部位-3：小脑； 出血部位-4：脑干（脑干＞小脑＞丘脑）、脑桥、中脑（出血部位编号数值赋予是基于 ICH 患者神经功能预后经验结论）。

（二）数据分析

本文调用 R x64 4.0.5 中的开源包如 e1071 包、randomForest 包、glmnet 包、rpart 包、nnet 包等构建五种基于机器学习的模型，包括支持向量机(Support Vector Machine, SVM)、经典决策树 (Decision Tree)、随机森林(Random Forest)、BP(back propagation)神经网络以及条件推断树（Confidence Inference Tree）。本文在 seed 设置为 95453 的情况下，通过分析模型测试集混淆矩阵计算准确率、敏感度、特异度、正例命中率、负例命中率等指标综合分析、阐述各种基于机器学习的分类模型的性能、适用场景等差异。

1.支持向量机（Support Vector Machine, SVM）

本文在 seed 值设置为 95453 的条件下，利用 R x64 4.0.5 版本开源 e1071 包进行模型建立。数据中包含年龄、性别等达 17 个指标变量，因此调用 svm()函数并选择采用径向基函数（Radial Basis Function，RBF）将样本中的观察对象指标投射至高维空间，对分割的数据集进行训练。其中 gamma 和 cost 是影响

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

模型性能的主要参数，前者控制分割超平面的形状，其确定与样本所使用的向量数相关。后者代表犯错的成本，其值的变化会影响分类边界的复杂程度，cost 过大可能出现过拟合现象，使模型对新数据的预测效能减低，相反地，cost 过小分类边界更光滑但是容易产生欠拟合的情况发生。不同的参数组合可能出现性能的优化，为确定最佳参数组合可通过设立多个组合进行模型建立，后经验证发现表现最佳组合。本文改变参数组合尝试进行模型优化，预测性能未见显著提升，故模型建立中的 gamma 与 cost 均采用 svm()函数的默认值，即 gamma 为预测变量个数的倒数（=1）、cost 默认值为 1。

2.决策树（Decision Tree）

本文在 seed 值设置为 95453 的条件下，利用 R x64 4.0.5 版本开源 rpart 包与 party 包分构建经典决策树模型及其一种重要的变体模型条件推断树。数据中结局变量系神经功能预后良好与不良，属于二元输出变量，可以使用决策树模型进行分类预测。

① 经典决策树

通过不断地选择目前最佳变量实现样本观察单位的二分类，尽可能实现两类变量纯度最大化，即可实现对新样本的预测，但是该算法通常会出现树的大小及分支数过大，发生过拟合的情况，因此在模型建立的过程中，需要使用复杂度参数（complex parameter，cp）作为惩罚决策树分支数的指标，同时计算通过基于训练样本的 10 折交叉验证误差，调用 plotcp()函数可视化该结果见附图 1，在保证交叉验证误差满足最小交叉验证误差一个标准差范围内，复杂度参数最小的树即最佳的决策树。由附图 1、 图 3 可知本文建立的决策树的最佳树为由入院时 NIHSS 评分、是否脑疝、治疗方式以及收缩压五个节点进行的四次变量分割。

② 条件推断树

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

条件推断树与传统决策树具有相同的思想，通过基于显著性检验的变量分割，对预测对象进行所有可能的二元分割方式，从中选取最显著的分割后，在分割子群中重复该流程至所有分割无显著性差异或经达到最小节点不可继续分割。图 4 变量分割路径提示入院时 NHISS、脑疝以及治疗方式是预后的重要分类指标，这一结果与前文 Logistic 回归变量筛选结果接近。

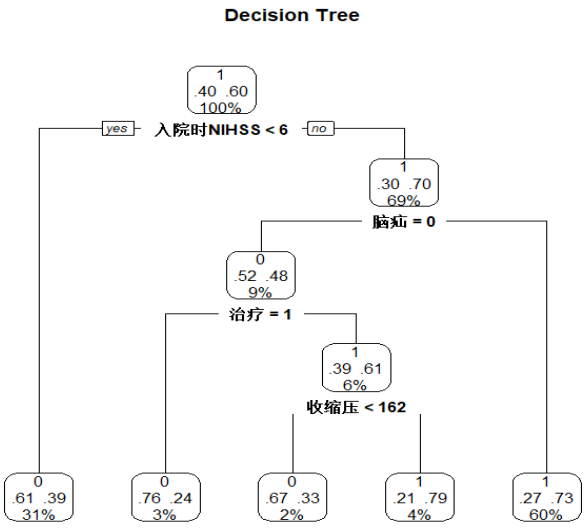


图 3　经典决策树变量分割途径

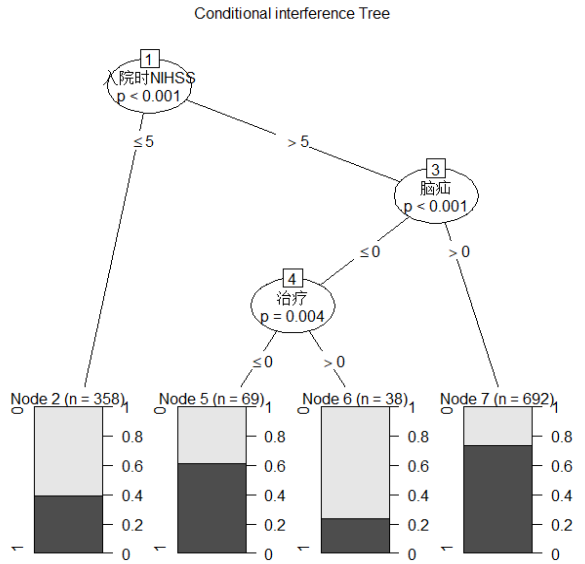


图 4　条件推断树变量分割途径

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3.随机森林（Random Forest）

本文在 seed 值设置为 95453 的条件下，利用 R x64 4.0.5 版本开源 randomForest 包进行随机森林模型构建。模型录入的院前数据包含既往病史、入院体征等不同属性的资料，通过随机森林进行样本单元和变量进行抽样，建立大量的决策树预测模型，其中每棵决策树均针对某一特定指标分类表现出良好性能，所有决策树依次对预测对象指标进行分类并记录，取该指标分类众数类别即为随机森林所预测的该指标的分类归属，并在子群中不断重复至分类完成。为获得最佳参数组合，每个节点变量抽取数设置为（指标个数-1），计算袋外预测（out-of-bag，OOB）误差，由图 5 可知该参数取值 4 时误差最低，效果最佳。后固定变量抽取值为 4，依次计算决策树最小数目 1 至最大数目 1000 误差图 6，提示决策树取值 1000 有误判率趋于稳定且最低。提示在变量抽取值为 4，生成 1000 棵决策树的情况同时兼顾了训练成本和准确率，附录图 2 显示随机森林模型的分类可视化结果。

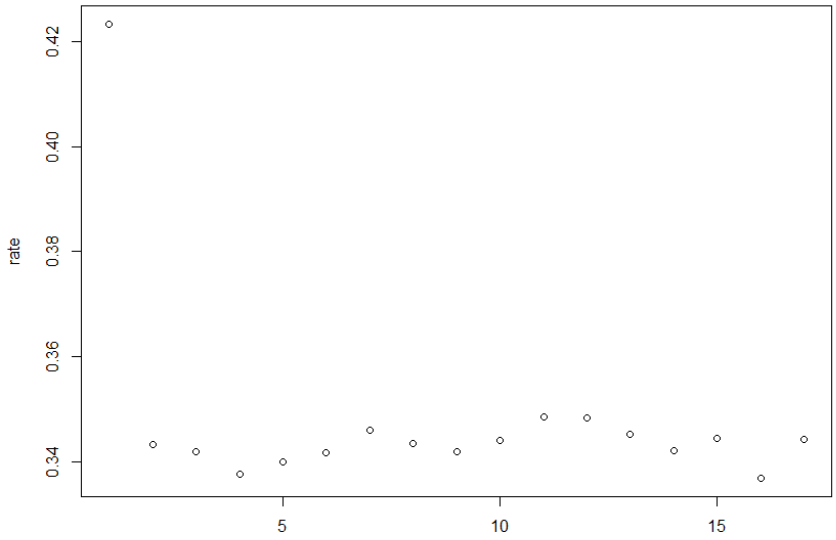


图 5　节点变量与 OOB 误差关系图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

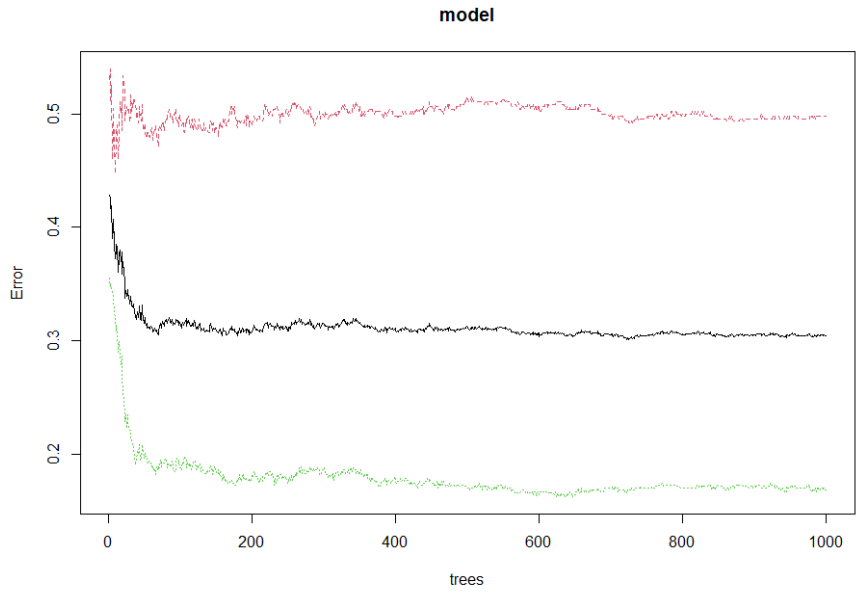


图 6　决策树数目 OOB 误差关系图

4.BP 神经网络（Back Propagation　Neural Network）

本文在 seed 值设置为 95453 的条件下，利用 R x64 4.0.5 版本开源 nnet 包实现 BP 神经网络的实现。模型通过输入层录入预测指标，经各条边路对结果进行预测，逆向回溯错误预判边路获取信息用于修正模型。模型性能依赖于合适的参数组合，且设计参数众多难以实现高性能模型搭建。

四、模型结果及分析

（一）影响因素筛选结果

1.Logistic 回归

本文在 seed 值设置为 95453 的条件下，利用 R x64 4.0.5 版本对院前指标及院中指标进行模型建立以及重要影响因素筛选。首先纳入全部指标，包括院前、院中指标建立模型 1（附表 1，2），依据 0.5 为神经功能预后良好与不良的分类

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

阈值，认为预测值小于 0.5 归类为神经功能预后良好，否则归类至神经功能预后不良，表 2 中 Z 检验提示脑疝、入院时 NIHSS 以及 GCS 评分等四项院前指标变量与肺部感染、上消化道出血、再出血、脑梗死等四项院中指标可能为脑出血患者神经功能预后的关键影响因素，目前的神经功能预后预测模型构建往往集中于治疗后的院中指标，Logistic 回归分析结果显示院前指标与院中指标同样处于重要的地位，利用院前指标对于构建脑出血患者神经功能预后模型，能够极大程度上指导临床治疗方案的制订，具有重要的研究价值。纳入 17 项院前指标建立模型 2，Z 检验提示高血压、脑疝、入院时 NIHSS、GCS 评分是脑出血患者神经功能预后的关键影响因素。为进一步验证变量筛选是否合理，纳入该四项指标再次建立模型 3，并对模型 2 与模型 3 进行显著水平 α 为 0.05 的方差分析，计算得 P-Value 为 0.5894＞α，表明其余指标变量对神经功能预后无显著影响。依据专业知识，怀疑 GCS 量表评分与 NIHSS 评分可能存在共线性，故采用方差膨胀系数(variance inflation factor，VIF)对两变量进行检验发现，两指标变量确存在该问题。在模型 3 的基础上排除 GCS 评分再次建立模型 4，并对模型 3 与模型 4 进行显著水平 α 为 0.05 的方差分析，计算得 P-Value 为 0.002762＜α，认为模型 4 具有更好的预测效果，高血压、是否脑疝、入院时 NIHSS 三个指标变量为脑出血患者神经功能预后的关键院前指标。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2　院前及院中指标 Z 检验

指标变量名称	P 值	Z 值
脑疝	0.000074	3.963
入院时 NIHSS	0.000000	6.307
GCS 评分	0.000794	3.355
肺部感染	0.000000	-6.807
上消化道出血	0.000122	3.842
再出血	0.030502	2.164
脑梗死	0.000098	3.895

2. Lasso 回归

Logistic 回归分析显示指标变量确存在共线性的问题，复杂临床指标往往也存在潜在的共线性问题，使用 Lasso 回归解决指标变量共线性对预测模型建议具有重要意义。本文在 seed 值设置为 95453 的条件下，利用 R x64 4.0.5 版本开源 glmnet 包进行模型建立。Lasso 回归过程中不断更迭模型，通过 Dev%解释基于当前变量的残差比例，类似于线性模型的决定系数，该值越接近 1 提示模型性能越强，基于前文 Logistic 回归模型变量筛选结果结合图 7、8 及附录表 3，选择平均最小错分误差对应的 λ 值，筛选出以下四个变量指标如下表 3。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

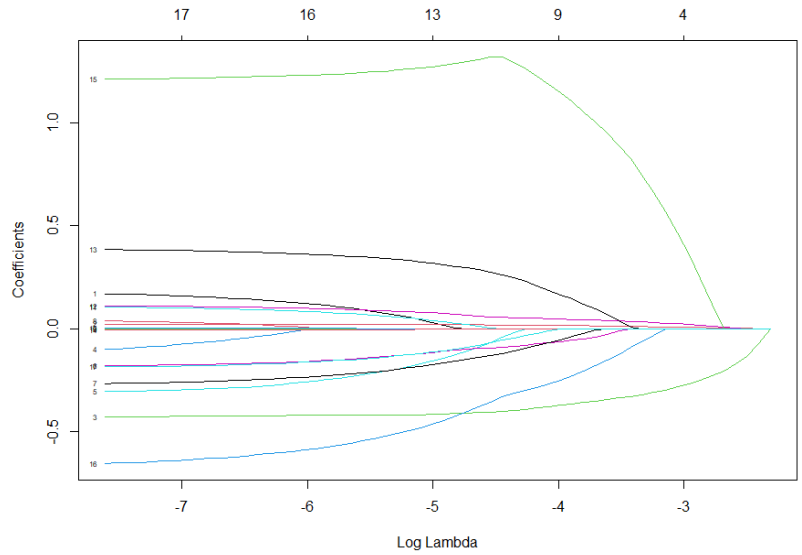


图 7　Lasso 回归系数与 λ 值的变化关系图

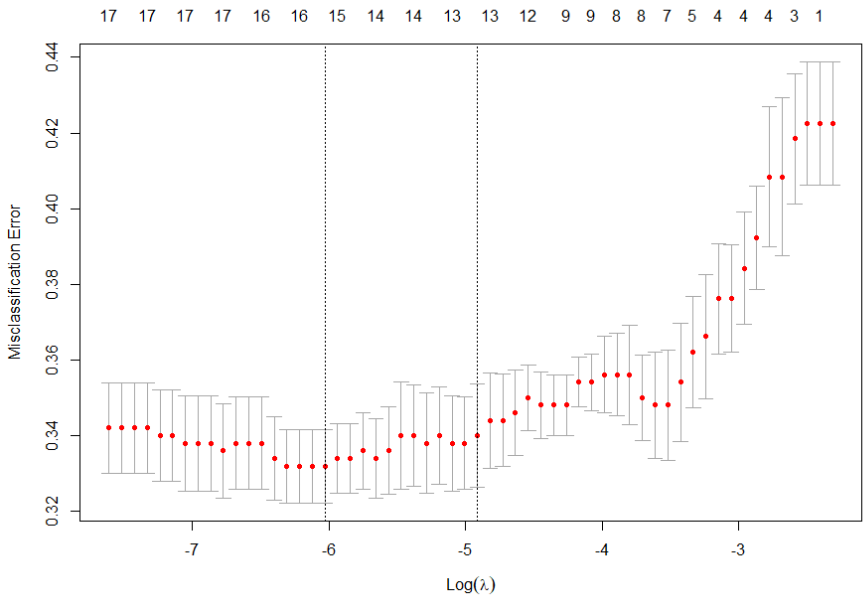


图 8　交叉验证 λ 和错分误差图关系图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 3　Lasso 回归系数

指标变量名称	Lasso 回归系数
脑疝	0.574591137
入院时 NIHSS	0.027054078
有高血压	-0.002382282
治疗方式	-0.299321073

通过 Logistic 回归与 Lasso 回归对 17 个院前指标变量与脑出血患者神经功能预后进行分析，筛选所得影响因素基本一致。

（二）多种机器学习方法预测结果及分析

1. 多种机器学习方法预测结果

表 4 至 9 为各模型混淆矩阵结果。图 9　为各模型预测准确率可视化结果。图 10 为各模型 ROC 曲线汇总图，各模型 ROC 曲线见附录图 3。表 10 为各模型其余评价指标结果。

表 4　支持向量机混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	121	66
预测为神经功能预后不良	89	221

表 5　逻辑回归混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	88	40
预测为神经功能预后不良	122	247

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 6　BP 神经网络混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	112	62
预测为神经功能预后不良	98	225

表 7　随机森林混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	109	46
预测为神经功能预后不良	101	241

表 8　条件推断树混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	125	56
预测为神经功能预后不良	185	231

表 9　经典决策树混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	129	59
预测为神经功能预后不良	81	228

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

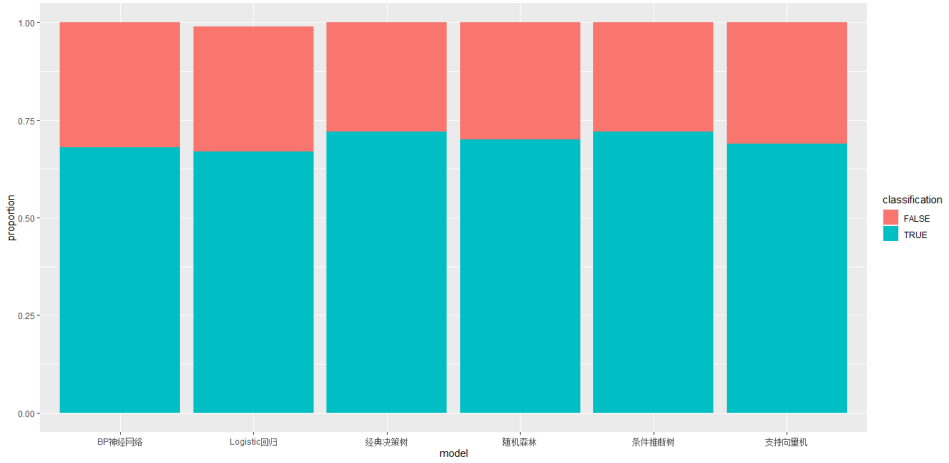


图 9　预测结果类型构成

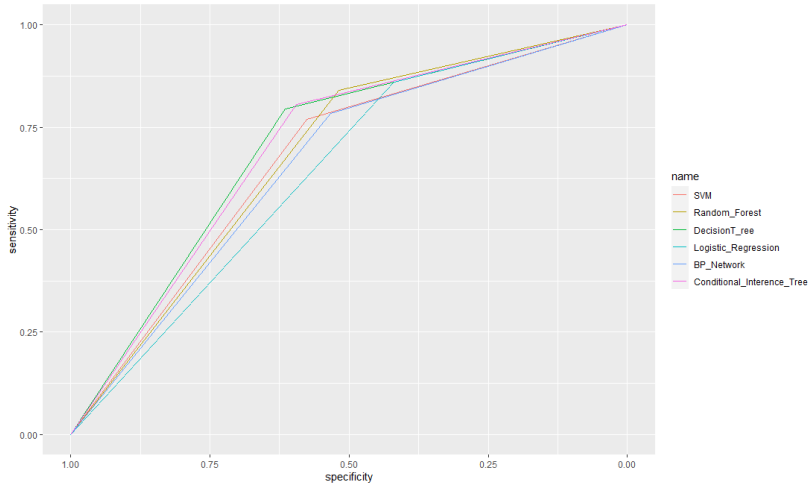


图 10　各模型 ROC 曲线

表 10　模型分类性能评价指标

	准确率	敏感度	特异度	正例命中率	负例命中率	AUC
Logistic 回归	0.67	0.42	0.86	0.69	0.67	0.640
支持向量机	0.69	0.58	0.77	0.65	0.71	0.673
BP 神经网络	0.68	0.53	0.78	0.64	0.70	0.659
随机森林	0.70	0.52	0.84	0.70	0.70	0.679
条件推断树	0.72	0.60	0.80	0.69	0.73	0.700
经典决策树	0.72	0.61	0.79	0.69	0.74	0.704

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2.结果分析

① 模型分析

ICH 患者神经功能预后多较差，如何发现神经功能预后结局良好的患者是预测模型的核心内容。提前发现 ICH 神经功能预后良好的患者，收集其治疗前的各项指标，有利于进一步挖掘影响神经功能预后的关键影响因素。敏感度以及正例命中率两个指标协同反映模型正确预测神经功能预后良好患者的能力，据表 10，五种基于机器学习算法模型中基于决策树基本分类思想的经典决策树、条件推断树以及随机森林相较于基于传统分析方法的 Logistic 回归分类性能更优越，Logistic 回归模型的预测能力高度依赖高质量数据，在大数据时代难以实现数据的快速、高效的清洗，传统模型势必逐渐淘汰。基于新型机器学习方法构建的模型中经典决策树与其变体条件推断树两个模型在预测 ICH 患者神经功能预后表现高度一致，尽管基于大量决策树构建的随机森林模型敏感度表现略低于前两者，良好的特异度表现仍然提示随机森林模型未来广阔的应用前景，尤其是在指标变量属性划分复杂的医学领域。同时，相较于 BP 神经网络等机器学习算法，随机森林表现出良好的移植性，不依赖于复杂的参数组合优化，对于异质性较大的数据可以通过增加决策树等方法消除或减小极端情况的偏倚，并且具有良好的解释性，可以采用决策树算法拆分随机森林，单独地解释某个指标变量对 ICH 患者神经功能预后的影响，能够更好地适应不断更新的海量数据。

② 影响 ICH 患者神经功能预后因素分析

Logistic 回归、Lasso 回归结果均支持入院时 NIHSS、是否已发生脑疝以及是否合并高血压是三个重要的影响因素，图 11 显示随机森林基于 OOB 误差及 gini 值的重要性测度散点图、图 3 显示经典决策树变量分割途径、图 4 显示条件推断树变量分割途径均支持入院时 NIHSS 系影响 ICH 患者神经功能预后最重

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

要的因素，尽管高血压并非被所有模型视作关键影响因素，但其某一侧面的表现形式，如收缩压仍然被视作相关影响因素，提示如何采用一种新的函数处理血压数据构建一种性能更强的统计量对于 ICH 患者神经功能预后具有重要的意义。图 5 所示条件推断树变量分割途径提示我们脑疝对于影响 ICH 患者神经功能的神经功能预后可能是通过两条途径，第一是脑疝直接造成的生理性损伤，第二是脑疝是否发生是医生决定是否进行手术治疗的主要判断依据，从而影响治疗方案的设计对 ICH 患者神经功能预后产生影响。因此，何种情况下的脑疝发生手术治疗具有更大的临床获益的相关研究也具有重要意义。

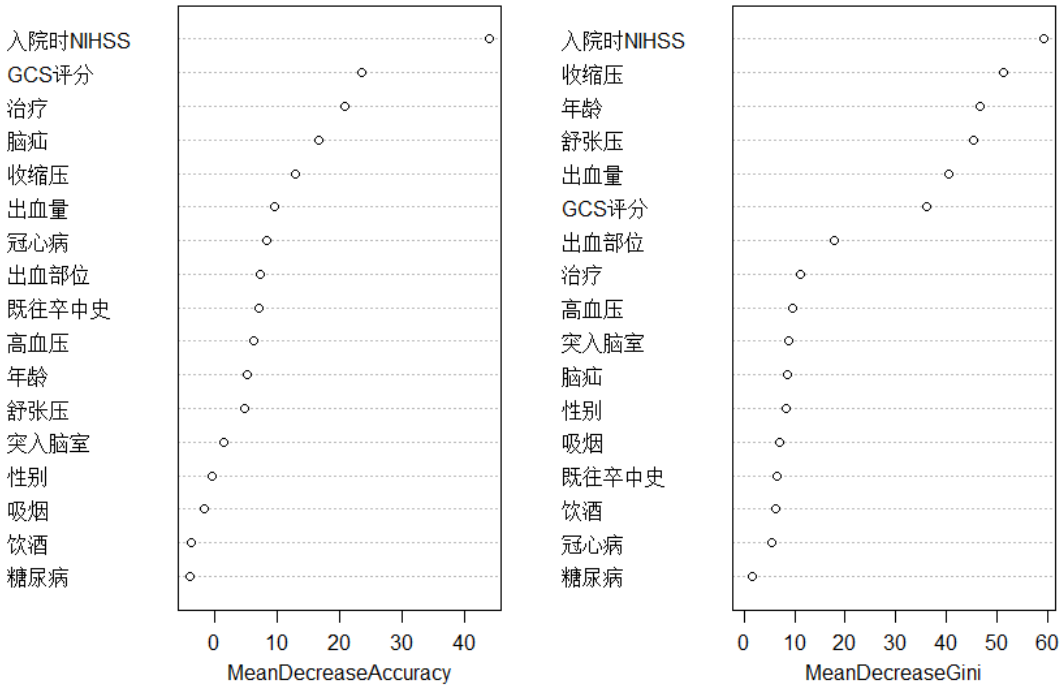


图 11　输入变量重要性测度散点图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

五、总结与展望

（一）主要结论

（1）ICH 患者以既往病史与入院体体征评估为主院前指标对于神经功能预后评估具有重要意义，其中入院时 NIHSS 评分以及脑出血是否合并脑疝是最重要的两个关键影响因素。

（2）有监督的机器学习方法相较于传统方法整体预测性能好，对于数据质量要求低，尤其在预测神经功能预后良好 ICH 患者具有明显优势，更符合临床的预测需求，利于后续随机对照实验的统计与专业设计。

（3）决策树类算法以及基于此思想的随机森林模型，不依赖于复杂的参数优化以及特定训练集分割，拥有较好的泛化性、鲁棒性以及移植性，能够在原理层面规避异常数据造成的模型过拟合或欠拟合而影响模型性能。更重要的是，该类模型能够给出基于重要性评分的指标变量排序，能够直观表现各变量的重要程度，同时可以进行变量拆分与组合通过单棵决策树可视化某特定变量在临床疾病发展所发挥的作用。

（二）实际应用

（1）结合医院 HIS 系统，为基层医院提供 ICH 患者脑出血预后给出预测，利于医生快速制订医疗干预方案，如是否进行手术以及采取何种手术方式，并据此判断是否需要提升患者护理等级以及尽早进行康复训练，尽可能提升患者神经功能预后，减轻家庭与国家经济压力。

（2）该研究为中枢神经系统损伤相关疾病提供了结合大数据、新型统计方法与临床医学的边缘融合的研究新范式。例如该类模型可以推广至重大灾难救治场

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

景或军队战场救治场景，提供一种可以被现场救援人员或基层战士等非医疗群体轻松使用的“黑箱模型”，决策哪些患者急需接受紧急救治，合理利用有限的医疗资源实现伤员获益最大化，保障人民的生命健康。

（3）ICH 患者发病急、临床情况复杂、病情变化迅速，难以建立基于大量人群的队列进行前瞻性研究，推动循证医学体系的建立。利用决策树类算法，在无院中干预的情况下以神经功能预后对入院患者进行分类，能够提高研究对象的同质性，增加统计的可信性推动脑出血诊治体系的规范化。

（4）决策树类及其衍生模型随机森林具有良好的解释性，基于该类算法构建的机器学习模型能够给出各类指标的重要性排名，并且就疾病发生发展情况给出决策路径图。据此，临床工作者能够就决策树给节点设计相应的干预研究，探索脑出血疾病的临床变化规律以及其后的病理生理规律，加深医务工作者对该疾病的认知，为救治该类患者提供基础理论基石。

（三）特色与创新

（1）国内多采用 Logistic 回归等传统统计方法对 ICH 患者神经功能预后影响因素的进行研究，本文通过采用 ICH 患者院前指标，运用决策树、随机森林、BP 神经网络等机器学习方法建立神经功能预后分类模型，提升了模型预测性能与临床需求相适应，扩大机器学习分类算法在脑出血相关疾病的研究中的应用，并对多种机器学习方法进行了比较研究，阐述了不同模型间的区别。

（2）国内针对脑出血神经功能的预后多局限于院中指标，难以阐明院前指标对疾病发展的重要影响，本文采用包括既往病史、入院体征检查如入院时 NIHSS、是否合并脑疝等第一手院前指标进行模型建立，能够更好地指导临床治疗方案的拟定。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（四）不足与展望

（1）各模型在均能一定程度上正确预测 ICH 患者神经功能预后，但模型分类的准确率均低于预期水平，并且通过多次实验验证发现训练集的分割显著影响模型性能，提示我们选择训练集的组成形式与模型参数优化具有同样重要的作用，由于随机森林模型 OOB 误差分析的特性可使用全部数据集而避免进行训练集的分割，其余几种模型则可通过 k 折交叉验证选取最佳训练集样本单元组合提高预测性能。

（2）已有文献报道建立针对 ICH 患者神经功能预后，准确率达 83%的随机森林模型[13]，经对比发现，其模型纳入的指标变量代表性更强、与神经功能预后相关性的临床证据等级更高，提示随机森林模型的建立后应当不断应用于临床实践，挖掘关键影响因素并纳入模型不断迭代模型以实现模型性能的优化与完善。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1]. Roth GA, Mensah GA, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study[J]. J Am Coll Cardiol. 2020 Dec 22;76(25):2982-3021.

[2]. 王陇德,刘建民,杨弋,彭斌,王伊龙.我国脑卒中防治仍面临巨大挑战——《中国脑卒中防治报告 2018》概要[J].中国循环杂志,2019,34(02):105-119.

[3]. 王拥军,李子孝,谷鸿秋,翟屹,姜勇,赵性泉,王伊龙,杨昕,王春娟,孟霞,李昊,刘丽萍,荆京,吴静,徐安定,董强,David Wang,赵继宗.中国卒中报告2019（中文版）[J].中国卒中杂志,2020,15(10):1037-1043.

[4]. 朱遂强;刘鸣;崔丽英;连立飞;张苏明. 中国脑出血诊治指南(2019). 中华神经科杂志. 2019. 052(12)：994-1005

[5]. 陈兰兰,万琪,陈蓓蕾,张娴娴,叶青,张扬威,李晓波.青年自发性脑出血神经功能预后的相关因素分析[J].中华急诊医学杂志,2013,22(09):1016-1020.

[6]. 李鸣,梁冲,吴鹤鸣,顾振兴.入院血糖水平与自发性脑出血患者近期神经功能预后的相关性研究[J].东南国防医药,2018,20(06):592-595.

[7]. 陈昶春,柯志通,钟晖东.不同年龄组自发性脑出血的临床特征及神经功能预后分析[J].中国医学工程,2021,29(05):67-69.

[8]. Hall AN, Weaver B, Liotta E, Maas MB, Faigle R, Mroczek DK, Naidech AM. Identifying Modifiable Predictors of Patient Outcomes After Intracerebral Hemorrhage with Machine Learning. Neurocrit Care. 2021 Feb;34(1):73-84.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[9]. 许国泽, 基于影像组学的乳腺癌腋窝淋巴结转移诊断研究, 2020, 暨南大学. 第 70 页.

[10]. Robert, T., Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996. 58(1).

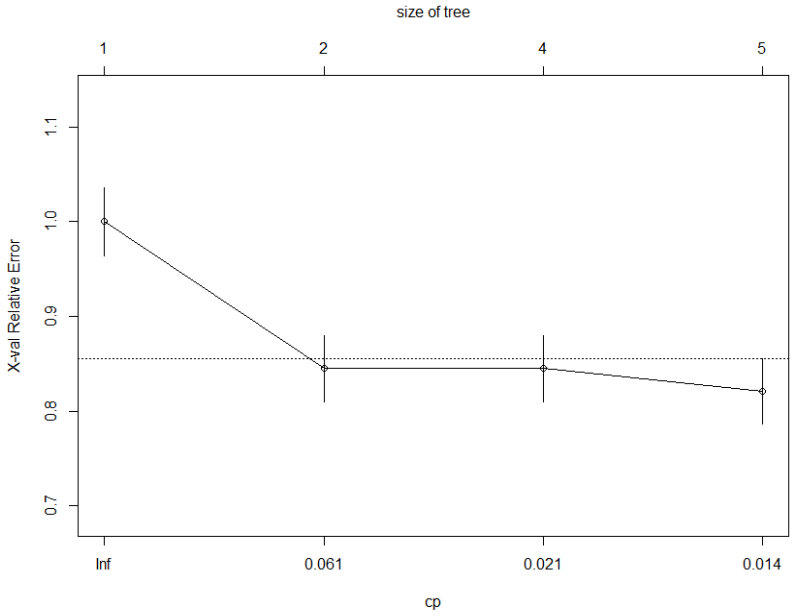
[11]. Leo, B., Bagging predictors. Machine Learning, 1996. 24(2).

[12]. 董师师,黄哲学. 随机森林理论浅析[J]. 集成技术,2013,01:1-7.

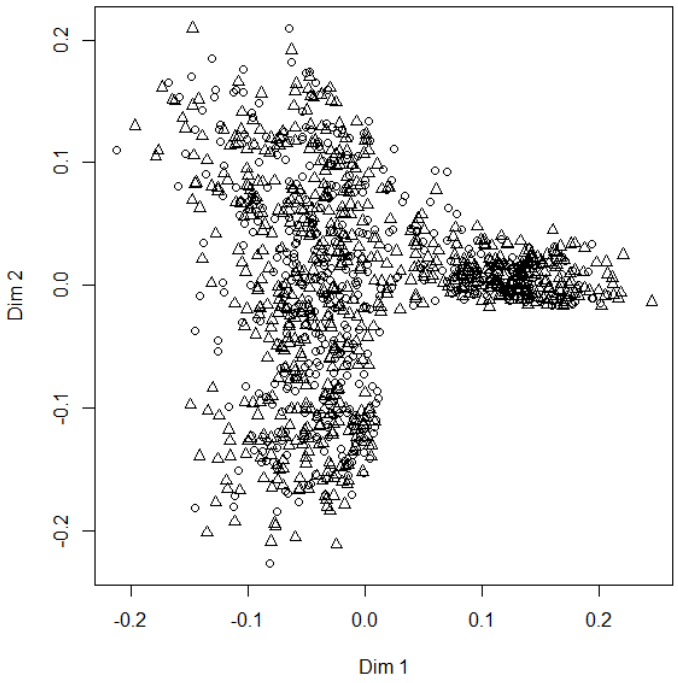
[13]. Wang HL, Hsu WY, Lee MH, Weng HH, Chang SW, Yang JT, Tsai YH. Automatic Machine-Learning-Based Outcome Prediction in Patients With Primary Intracerebral Hemorrhage. Front Neurol. 2019 Aug 21;10:910.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

附录

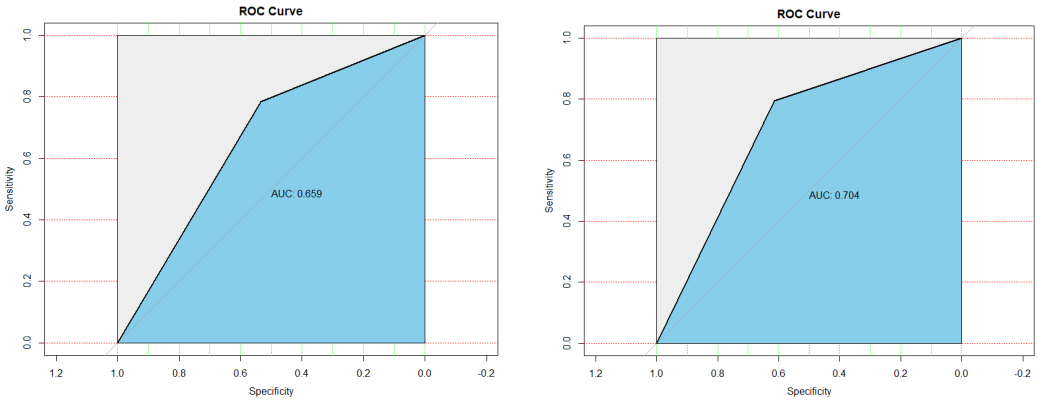


附图 1　交叉验证与复杂度参数关系图



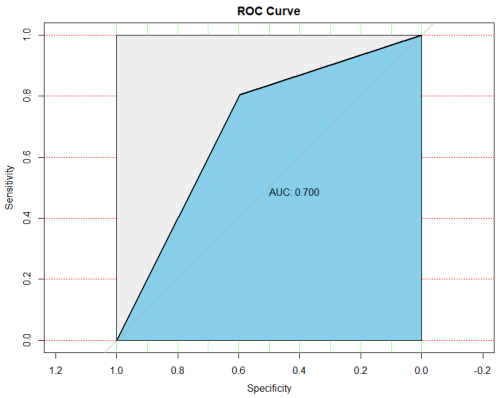
附图 2　随机森林可视化

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

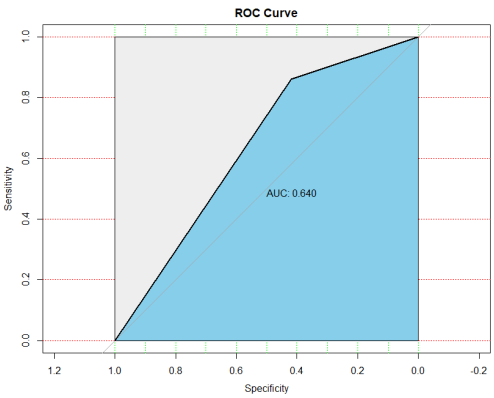


(A)

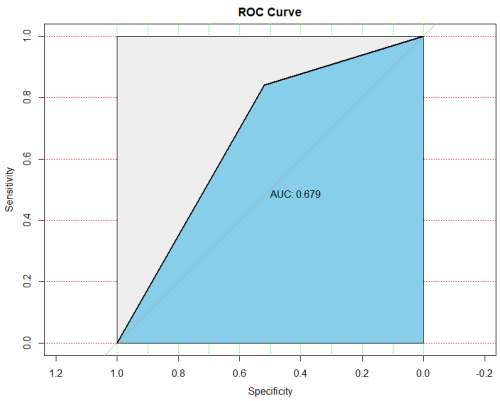
(B)



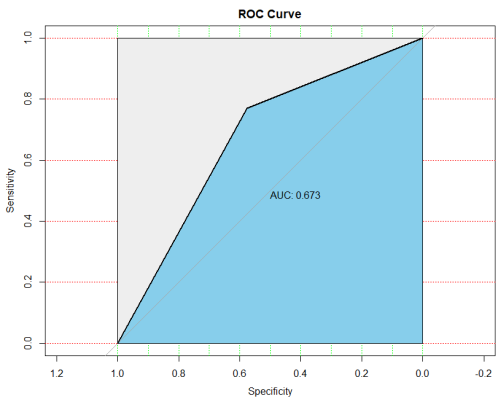
(C)



(D)



(E)



(F)

附图 3　各模型 ROC 曲线

注：图表 3（A-F）分别为 BP 神经网络、经典决策树、条件推断树、Logistic 回归、随机森林和支持向量机的 ROC 曲线。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

附表 1　Logistic 回归模型 1

变量	Logistic 回归系数	变量	Logistic 回归系数
性别	0.217169	出血量	-0.003451
年龄	0.003900	脑疝	1.113315
高血压	-0.220460	治疗	-0.169857
糖尿病	0.073999	出血部位	-0.144702
冠心病	0.419752	脑积水	0.200572
吸烟	0.152625	肺部感染	-0.961721
饮酒	-0.260316	气管切开	0.001924
既往卒中史	-0.247278	上消化道出血	1.021255
收缩压	0.002248	再出血	0.462209
舒张压	0.002274	脑梗死	1.432644
GSC 评分	0.134431	颅内感染	0.306856
入院时 NIHSS	0.113561	癫痫	0.171691
突入脑室	0.155444		

附表 2　Logistic 回归模型 1 混淆矩阵

	实际为神经功能预后良好	实际为神经功能预后不良
预测为神经功能预后良好	97	80
预测为神经功能预后不良	95	225

附表 3　lasso 回归系数

编号	Df	%Dev	Lambda
1	0	0.00	0.099360
2	1	0.50	0.090530
3	2	1.01	0.082490
4	3	1.76	0.075160
5	3	2.37	0.068480
6	4	3.37	0.062400
7	4	4.28	0.056860
8	4	5.04	0.051810
9	4	5.68	0.047200
10	4	6.22	0.043010
11	5	6.76	0.039190
12	5	7.21	0.035710
13	6	7.61	0.032540
14	7	8.04	0.029640

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

15	7	8.41	0.027010
16	8	8.72	0.024610
17	8	9.00	0.022430
18	8	9.23	0.020430
19	8	9.43	0.018620
20	9	9.62	0.016960
21	9	9.79	0.015460
22	9	9.94	0.014080
23	10	10.07	0.012830
24	11	10.20	0.011690
25	12	10.36	0.010650
26	12	10.54	0.009707
27	12	10.69	0.008845
28	13	10.82	0.008059
29	13	10.94	0.007343
30	13	11.04	0.006691
31	13	11.12	0.006096
32	14	11.19	0.005555
33	14	11.25	0.005061
34	14	11.31	0.004612
35	14	11.38	0.004202
36	14	11.41	0.003829
37	14	11.44	0.003489
38	14	11.46	0.003179
39	14	11.48	0.002896
40	15	11.49	0.002639
41	16	11.51	0.002405
42	16	11.52	0.002191
43	16	11.53	0.001996
44	16	11.53	0.001819
45	16	11.54	0.001657
46	16	11.55	0.000150
47	17	11.55	0.001376
48	17	11.56	0.001254
49	17	11.56	0.001142
50	17	11.56	0.001041
51	17	11.57	0.000948
52	17	11.57	0.000864
53	17	11.57	0.000787
54	17	11.57	0.000717
55	17	11.57	0.000654
56	17	11.57	0.000596

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

57	17	11.57	0.000543
58	17	11.57	0.000495

附表 3 输出结果的每一行代表一个模型，Df 为自由度，表示非零线性模型拟合系数的个数；%Dev 代表由模型解释的残差比例，对于线性模型其相当于决定系数 R2 值，其值介于 0 和 1 之间，越接近于 1 说明模型的拟合效果越好。Lambda 即 λ 值。随着 λ 值的逐渐缩小，非零线性模型拟合系数的个数逐渐增多，%Dev 逐渐增大，%Dev 最大值也仅为 11.57%，表明拟合效果不佳。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致谢

本研究受到国家自然基金项目（No. 81872716）资助

衷心感谢指导老师在本次比赛，论文选题、数据分析、论文撰写中提供的无私指导，面对种种客观困难的情况下，始终给予我们安慰与鼓励。

同时，还要感谢大学医院神经外科团队在研究数据的收集与处理提供的大力帮助。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net