浙江理工大学学报,第51卷,第4期,2024年7月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2024.04.012

基于变换AUC的颈动脉斑块稳定性预测模型

李翔宇[1a],杨建萍[1b],吴　炯[2]

(1.浙江理工大学,a.计算机科学与技术学院;b.理学院,杭州 310018;2.浙江大学医学院附属浙江医院,杭州 310000)

摘　要:为了准确识别颈动脉斑块的重要生物标志物,在改进生物标志物所包含信息量的度量方法的基础上,通过向前逐步回归建立了基于变换AUC(Transformed area under curve)的颈动脉斑块稳定性预测模型。首先,在ROC(Receiver operating characteristic)空间提出变换AUC,并给出该指标在双正态分布模型和自由分布模型下的估计方法;然后,使用R统计软件,对比分析变换AUC与AUC等常用评估指标对非传统生物标志物的评估性能;最后,基于浙江医院提供的影像数据,利用变换AUC度量生物标志物的信息量,使用向前逐步回归筛选模型的方法建立高精度的颈动脉斑块稳定性预测模型。研究结果表明,该颈动脉斑块稳定性预测模型的AUC值达到0.9以上,能够准确识别斑块的稳定性,为临床医师对患者进行个性化诊疗提供更精准的参考依据。

关键词:变换AUC;向前逐步回归;斑块稳定性;生物标志物;模型筛选

中图分类号:TP181;R445.2　　**文献标志码**:A　　**文章编号**:1673-3851(2024)04-0529-08

引文格式:李翔宇,杨建萍,吴炯. 基于变换AUC的颈动脉斑块稳定性预测模型[J]. 浙江理工大学学报(自然科学),2024,51(4):529-536.

Reference Format: LI Xiangyu, YANG Jianping, WU Jiong. A prediction model of carotid plaque stability based on transformed AUC[J]. Journal of Zhejiang Sci-Tech University,2024,51(4):529-536.

A prediction model of carotid plaque stability based on transformed AUC

LI Xiangyu[1a], *YANG Jianping*[1b], *WU Jiong*[2]

(1a. School of Computer Science and Technology; 1b. School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2. Affiliated Zhejiang Hospital, Zhejiang University School of Medicine, Hangzhou 310000, China)

Abstract: To precisely identify critical biomarkers of carotid plaques, a model for predicting carotid plaque stability based on the transformed area under curve (transformed AUC) using forward regression was built on the basis that the method for quantifying the information content within biomarkers was improved. Firstly, transformed AUC was introduced in the receiver operating characteristic (ROC) space, and the estimation methods were provided under the binormal distribution model and free distribution model, respectively. Then, R statistical software was used to compare and analyze the evaluation performance of transformed AUC index with common evaluation indices such as AUC for non-traditional biomarkers. Finally, a carotid plaque stability prediction model with high-accuracy was built by using transformed AUC to measure the information of biomarkers and stepwise forward regression based on image data provided by Zhejiang Hospital. These research findings illustrate that the AUC value of the carotid plaque stability prediction model is above 0.9, indicating the model can accurately identify the plaque stability and provide more precise reference to clinicians for personalized diagnosis and treatment decisions.

Key words: transformed AUC; forward stepwise regression; plaque stability; biomarker; model selection

收稿日期:2023-11-16　　网络出版日期:2024-04-12

基金项目:国家自然科学基金项目(12071436);浙江省基础公益类项目(GF22H096743)

作者简介:李翔宇(1999—　),男,山东临沂人,硕士研究生,主要从事大数据分析、应用统计方面的研究。

通信作者:吴　炯,E-mail:wujiong0118@aliyun.com

0 引 言

脑卒中的发生已成为威胁人类生命安全的第二大原因[1]。医学研究表明,20%～30%的脑卒中是由不稳定的颈动脉斑块引起[2]:不稳定的颈动脉斑块会导致颈动脉狭窄或阻塞,使脑部血供减少,进而引起缺血性脑卒中。因此,如何准确有效地评估颈动脉斑块的稳定性,对颈动脉斑块患者实施针对性诊治,已成为医学研究中一个亟待解决的问题。

提取颈动脉斑块的重要生物标志物对于评估斑块稳定性至关重要。随着影像学技术的飞速发展,利用颈动脉斑块的影像数据,使用 AUC(Area under curve)、F1 分数(F1 score)等常用评估指标度量生物标志物所包含的信息量,已成为提取颈动脉斑块重要生物标志物的主要方法[3]。Saba 等[4]利用颈动脉斑块的 CT 测量值,使用二分类模型和 AUC,发现斑块密度变化值是评估颈动脉斑块稳定性的一个重要生物标志物。李杨等[5]利用颈动脉斑块的 CTA 测量值,使用假设检验和 AUC,发现斑块的体积是评估颈动脉斑块稳定性的另一个重要生物标志物。Lu 等[6-7]利用颈动脉斑块的 MRI 测量值,使用 logistic 回归模型和 F1 分数,发现颈动脉斑块的形状和位置是评估颈动脉斑块稳定性的重要生物标志物。目前,基于不同的常用评估指标已得到一些重要的颈动脉斑块生物标志物,但在实际应用中,仅仅使用这些重要生物标志物识别颈动脉斑块的稳定性,精度仍比较低。

AUC、F1 分数和 Cohen′s kappa 统计系数等是对 ROC 曲线(Receiver operating characteristic curve)使用描述性统计分析方法得到的一些评估指标[8-10]。ROC 曲线是一个二维图表,它以敏感度和特异性为变量,描述生物标志物在不同决策阈值预测二元疾病结果的操作特征,ROC 曲线所在的二维空间可称作 ROC 空间[11-13]。若 ROC 曲线完全位于单位正方形内的 45°对角线上方,称 ROC 曲线是有效的[14-15],此时相应的 AUC 等评估指标能准确地度量生物标志物所包含的信息量,且能够精确地评估生物标志物的重要性。然而,最近的一些医学病理研究表明,医学统计中常用的 AUC 等评估指标存在着某些缺陷,无法精确地评估某些生物标志物的重要性。如 Bantis 等[16]将肺表面活性蛋白 B(ProSFTPB)作为肺癌标志物进行了临床病理研究,发现 ProSFTPB 是肺癌的重要生物标志物;但是该生物标志物的 ROC 曲线是 S 型的,在医学统计分析中,其 AUC 等常用评估指标的值都很低,不能认为是肺癌的重要生物标志物。因此,为了能够准确评估生物标志物的重要性,提高医学分析时的准确性,必须改进目前医学统计中评估生物标志物重要性的方法。

本文首先在 ROC 空间中提出一种新的评估指标,即变换 AUC(Transformed AUC),并提出其在双正态分布模型和自由分布模型下的两种估计方法;然后,利用 R 软件对变换 AUC 在实际应用中的性能与 AUC 等常用的评估指标进行对比分析;最后,利用 ITK-SNAP 软件对浙江医院提供的颈动脉斑块的 MRI 影像数据进行分割和特征提取,使用变换 AUC 度量颈动脉斑块的生物标志物所包含的信息量,筛选出重要生物标志物,并结合向前逐步回归,建立评估颈动脉斑块稳定性的最优模型。本文提出的新指标和筛选模型的新方法能够有效应用于实际决策者筛选重要生物标志物,且能够提高医学分析时的准确性。

1 变换 AUC 及估计

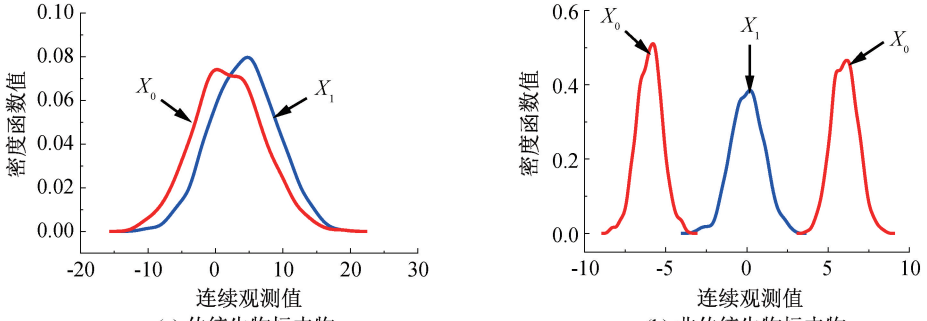
1.1 变换 AUC

在医学统计研究领域,一般生物标志物分为传统生物标志物和非传统生物标志物。假设 $X_0=[X|D=0]$和 $X_1=[X|D=1]$分别是患病和健康受试者的生物标志物的连续测试值,它们的密度函数分别为 $f(x)$和 $g(x)$。此生物标志物患病总体和健康总体之间的密度函数曲线和 ROC 曲线存在两种情况,示例分别如图 1 和图 2 所示。在生物标志物的统计分析中,常把具有有效 ROC 曲线的生物标志物称为传统的生物标志物,而不满足此条件的称为非传统生物标志物。使用 AUC 以及一些常用的评估指标通常不能有效地度量非传统生物标志物的信息量。例如,图 1 中的非传统生物标志物具有很高的信息量,是一个重要的生物标志物;但是它的 ROC 曲线不是有效的,对应的 AUC 值接近于 0.5,因此在医学统计分析时,不能认为该生物标志物是重要生物标志物。

本文针对非传统生物标志物 ROC 曲线的特征,提出了变换 ROC 曲线(Transformed receiver operating characteristic curve, TROC)和变换 AUC。假设 X_0 和 X_1 的分布函数分别为 $F(x)$和 $G(x)$,令 $u=F(x)$或 $u=G(x)$,对任意的 $u\in(0,1)$,称曲线

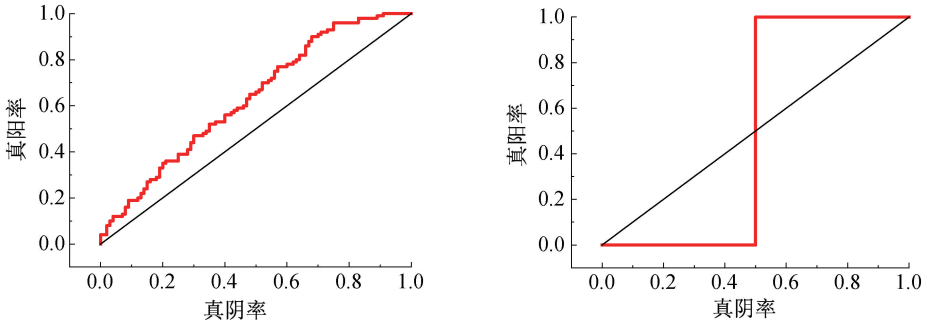
$$T_{\mathrm{ROC}}(u)=\begin{cases}G(F^{-1}(u)), & F^{-1}(u)\geqslant G^{-1}(u);\\ F(G^{-1}(u)), & F^{-1}(u)<G^{-1}(u)\end{cases}$$

为变换 ROC 曲线;称变换的 ROC 曲线与坐标轴所围的面积为变换 AUC,记为 A_1。



(a) 传统生物标志物　　(b) 非传统生物标志物

图 1　两类生物标志物的密度函数曲线示例



(a) 传统生物标志物　　(b) 非传统生物标志物

图 2　两类生物标志物的 ROC 曲线示例

显然,当 $F(x)$没有下穿 $G(x)$时,变化的 ROC 曲线与原始的 ROC 曲线形状相同,呈现凹形,此时生物标志物的变换 AUC 值等于 AUC 值。当 $F(x)$下穿 $G(x)$时,原始 ROC 曲线通常与单位正方形中的 45°对角线存在多个交点。此时将 45°线下方的 ROC 曲线部分对称翻转到 45°线上方,得到变换 ROC 曲线。计算可知,生物标志物的变换 AUC 值大于 AUC 值。因此,变换 AUC 可认为是改进的 AUC。

对于任意的连续递增非负有界函数 $\psi(x)$,有如下的非单调变换

$$h(x)=\begin{cases}\psi(x), & F(x)\leqslant G(x);\\ -\psi(x), & F(x)>G(x)\end{cases}\tag{1}$$

可以使生物标志物的变换 AUC 的值等于对它实施了非单调变换后的 AUC 的值,即 $A_1=P(h(X_0)>h(X_1))$,本文称此非单调变换为 H 变换。

1.2　双正态分布模型下的变换 AUC 参数估计

在医学生物标志物的诊断识别中,很多生物标志物是连续且服从双正态分布的。为了使变换 AUC 能更好地应用于医学中重要生物标志物的识别,本文首先提出了一种在双正态分布模型下的变换 AUC 参数估计。

假设某一生物标志物在患病受试者中的测试值 $X_0\sim N(\mu_0,\sigma_0^2)$,在健康受试者中的测试值 $X_1\sim N(\mu_1,\sigma_1^2)$。此生物标志物的变换 AUC 计算公式可表示为:

$$A_1=\begin{cases}\displaystyle\int_{-\infty}^{x_0}\Phi\left(\frac{x-\mu_1}{\sigma_1}\right)\mathrm{d}\Phi\left(\frac{x-\mu_0}{\sigma_0}\right)+\int_{x_0}^{\infty}\Phi\left(\frac{x-\mu_0}{\sigma_0}\right)\mathrm{d}\Phi\left(\frac{x-\mu_1}{\sigma_1}\right), & \sigma_0<\sigma_1;\\ \displaystyle\int_{-\infty}^{x_0}\Phi\left(\frac{x-\mu_0}{\sigma_0}\right)\mathrm{d}\Phi\left(\frac{x-\mu_1}{\sigma_1}\right)+\int_{x_0}^{\infty}\Phi\left(\frac{x-\mu_1}{\sigma_1}\right)\mathrm{d}\Phi\left(\frac{x-\mu_0}{\sigma_0}\right), & \sigma_0>\sigma_1;\\ \displaystyle 1-\Phi\left(\frac{\mu_1-\mu_0}{\sqrt{\sigma_0^2+\sigma_1^2}}\right), & \sigma_0=\sigma_1,\mu_0>\mu_1;\\ \displaystyle\Phi\left(\frac{\mu_1-\mu_0}{\sqrt{\sigma_0^2+\sigma_1^2}}\right), & \sigma_0=\sigma_1,\mu_0\leqslant\mu_1\end{cases}\tag{2}$$

其中:$x_0=(\sigma_0\mu_1-\sigma_1\mu_0)/(\sigma_0-\sigma_1)$为分布函数$F(x)$和$G(x)$的交点,$\Phi(\cdot)$为标准正态分布函数。

假设$X_{01},\cdots,X_{0m}$是来自于总体X_0样本容量为m的样本;$X_{11},\cdots,X_{1n}$是来自于总体X_1样本容量为n的样本。令$\hat{\mu}_0=m^{-1}\sum_{i=1}^{m}X_{0i}$,$\hat{\mu}_1=n^{-1}\sum_{j=1}^{n}X_{1j}$,$\hat{\sigma}_0^2=m^{-1}\sum_{i=1}^{m}(X_{0i}-\hat{\mu}_0)^2$,$\hat{\sigma}2_1=n^{-1}\sum_{j=1}^{n}(X_{1j}-\hat{\mu}_1)^2$。本文采用$\hat{x}_0=(\hat{\sigma}_0\hat{\mu}_1-\hat{\sigma}_1\hat{\mu}_0)/(\hat{\sigma}_0-\hat{\sigma}_1)$来估计分布函数$F(x)$和$G(x)$的交点$x_0$,构建双正态分布总体下变换 AUC A_1 的估计值$\hat{A}_1$:

$$\hat{A}_1=\begin{cases}\int_{-\infty}^{\hat{x}_0}\Phi\left(\frac{x-\hat{\mu}_1}{\hat{\sigma}_1}\right)\mathrm{d}\Phi\left(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0}\right)+\int_{\hat{x}_0}^{\infty}\Phi\left(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0}\right)\mathrm{d}\Phi\left(\frac{x-\hat{\mu}_1}{\hat{\sigma}_1}\right), & \hat{\sigma}_0<\hat{\sigma}_1;\\ \int_{-\infty}^{\hat{x}_0}\Phi\left(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0}\right)\mathrm{d}\Phi\left(\frac{x-\hat{\mu}_1}{\hat{\sigma}_1}\right)+\int_{\hat{x}_0}^{\infty}\Phi\left(\frac{x-\hat{\mu}_1}{\hat{\sigma}_1}\right)\mathrm{d}\Phi\left(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0}\right), & \hat{\sigma}_0>\hat{\sigma}_1;\\ 1-\Phi\left(\frac{\hat{\mu}_1-\hat{\mu}_0}{\sqrt{\hat{\sigma}_0^2+\hat{\sigma}_1^2}}\right), & \hat{\sigma}_0=\hat{\sigma}_1,\hat{\mu}_0>\hat{\mu}_1;\\ \Phi\left(\frac{\hat{\mu}_1-\hat{\mu}_0}{\sqrt{\hat{\sigma}_0^2+\hat{\sigma}_1^2}}\right), & \hat{\sigma}_0=\hat{\sigma}_1,\hat{\mu}_0\leqslant\hat{\mu}_1\end{cases}\tag{3}$$

因为$\hat{\mu}_0$、$\hat{\mu}_1$、$\hat{\sigma}_0^2$、$\hat{\sigma}_1^2$为μ_0、μ_1、σ_0^2、σ_1^2的极大似然估计,具有渐近相合正态性,而A_1是μ_0、μ_1、σ_0、σ_1的连续可微函数,因此$\hat{A}_1$是A_1的极大似然估计,也具有较高的估计效率。

1.3 自由分布模型下变换 AUC 的非参数估计

考虑到在实际应用中,也存在着连续但并不服从双正态分布的生物标志物。因此,本文还提供了自由分布模型下基于 H 变换的变换 AUC 非参数估计。满足 H 变换条件的函数有很多,本文选取了一个特殊的 H 变换函数,即:

$$h(x)=2\Phi(x-\mu_0)I\{F(x)\leqslant G(x)\}-\Phi(x-\mu_0)\tag{4}$$

其中:$I\{\cdot\}$为示性函数,$\mathrm{E}[X_0]=\mu_0$。

假设某一生物标志物在患病受试者中的测试值为X_0,在健康受试者中的测试值为X_1。$X_{01},\cdots,X_{0m}$是来自患病总体X_0样本容量为m的一个简单随机样本;$X_{11},\cdots,X_{1n}$来自健康总体X_1样本容量为n的一个简单随机样本。令

$$F_m(x)=m^{-1}\sum_{i=1}^{m}I(X_{0i}\leqslant x),$$

$$G_n(x)=n^{-1}\sum_{j=1}^{n}I(X_{1j}\leqslant x)\tag{5}$$

构建 H 变换h的非参数估计:

$$\hat{h}(x)=2\Phi(x-\hat{\mu}_0)\,I\{F_m(x)\leqslant G_n(x)\}-\Phi(x-\hat{\mu}_0)\tag{6}$$

变换 AUC 的非参数估计为:

$$\hat{A}^*=(mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}I(\hat{h}(X_{0i})>\hat{h}(X_{1j}))\tag{7}$$

由大样本理论可知,当样本量m、n足够大时,$\hat{A}^*$也具有相合渐近正态性,估计效率同样较高。

2 变换 AUC 评估性能的仿真分析

为了说明本文所提供的变换 AUC 能够准确地度量生物标志物所包含的信息量,本文设计了两个仿真实验,对比分析了变换 AUC 与医学生物统计中常用的生物标志物评估指标 AUC、F1 分数以及 Cohen's kappa 统计系数在评估方面的性能。

第一个仿真实验假设生物标志物服从正态分布,选取患病受试者生物标志物的测试值$X_0\sim N(2,10^2)$,健康受试者生物标志物的测试值$X_1\sim N(2.3,5^2)$。假设$X_{01},\cdots,X_{0m}$是来自患病总体X_0样本容量为m的一个简单随机样本;$X_{11},\cdots,X_{1n}$来自健康总体X_1样本容量为n的一个简单随机样本。设定样本容量$(m,n)=(50,50),(100,100),(200,200),(500,500),(520,500),(800,800),(830,800)$。变换 AUC 的值使用式(3)估计,采用经典的参数估计方法估计 AUC 的值,采用常用的计算方法计算 F1 分数和 Cohen's kappa 统计系数的值。在不同样本数量下均重复计算 10000 次,将计算结果的均值作为各指标值,实验结果如表 1 所示。

变换 AUC 和 AUC 的取值范围在 0 到 1 之间,取值越接近 0.5,表示其评估性能越差;F1 分数的取值范围在 0 到 1 之间,取值越接近 1,表示其评估性能越好;Cohen's kappa 统计系数的取值在-1到 1 之间,取值越大,表示其评估性能越好。根据表 1 的实验结果发现:用 AUC、F1 分数以及 Cohen's kappa 统计系数评估该生物标志物的识别能力,得到的结果都是低识别能力生物标志物;而变换 AUC

的估计值大于 0.6,说明该生物标志物具有一定的识别能力。显然这是一个非传统的生物标志物,因此,相较于 AUC 等常用的评估指标,变换 AUC 能更准确地度量非传统的生物标志物所包含的信息量,使用 AUC 等常用的评估指标筛选生物标志物时,该生物标志物极有可能被遗漏。

表 1　第一次仿真实验的变换 AUC 与常用评估指标的计算结果

评估指标	样本容量						
	(50,50)	(100,100)	(200,200)	(500,500)	(520,500)	(800,800)	(830,800)
变换 AUC 值	0.612	0.607	0.605	0.604	0.604	0.603	0.603
F1 分数	0.038	0.020	0.010	0.004	0.004	0.002	0.002
Cohen′s kappa	−0.981	−0.990	−0.995	−0.998	−0.998	−0.999	−0.999
AUC 值	0.511	0.511	0.510	0.511	0.511	0.510	0.511

第一个仿真实验的结果表明,在双正态总体模型下,变换 AUC 比 AUC 等常用的指标有更准确的评估能力。第二次仿真实验对非双正态总体模型下的变换 AUC 的评估性能进行分析。选取患病受试者生物标志物的测试值 $X_0 \sim N(2,4^2)$,健康受试者生物标志物的测试值 $X_1 \sim F(4,2)$。假设 $X_{01},\cdots,X_{0m}$ 是来自患病总体 X_0 样本容量为 m 的一个简单随机样本;$X_{11},\cdots,X_{1n}$ 来自健康总体 X_1 样本容量为 n 的一个简单随机样本。设定样本容量(m,n)=(50,50),(100,100),(200,200),(500,500),(520,500),(1000,1000),(1000,1200)。使用式(7)估计变换 AUC 的值,采用经典的非参数估计方法估计 AUC 的值。不同样本数量下均进行 10000 次重复计算,将计算结果的均值作为各指标值,实验结果如表 2 所示。

表 2　第二次仿真实验的变换 AUC 与常用评估指标的计算结果

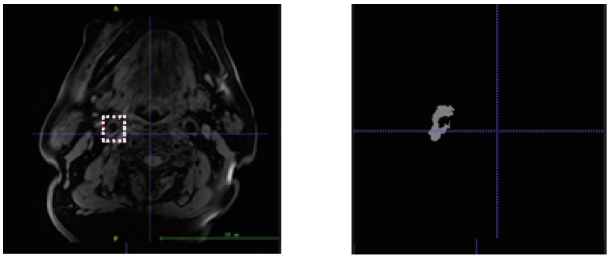
评估指标	样本容量						
	(50,50)	(100,100)	(200,200)	(500,500)	(520,500)	(1000,1000)	(1000,1200)
变换 AUC 值	0.580	0.594	0.601	0.605	0.605	0.606	0.607
F1 分数	0.003	0	0	0	0	0	0
Cohen′s kappa	−0.979	−0.990	−0.995	−0.997	−0.997	−0.998	−0.998
AUC 值	0.521	0.518	0.520	0.521	0.519	0.521	0.521

根据表 2 同样可以发现,相较于 AUC 等常用的评估指标,变换 AUC 能够更准确地度量非传统的生物标志物所包含的信息量。从这两个仿真实验结果可以发现,在实际应用中使用本文所提出的变换 AUC 评估生物标志物,可以防止重要生物标志物被遗漏。

3　颈动脉斑块稳定性评估的 logistic 回归模型

3.1　MRI 影像数据提取

本文使用的颈动脉斑块 MRI 影像数据由浙江医院提供。首先采用 ITK-SNAP 软件读取原始 MRI 影像数据,并进行图像分割处理;图像分割时,由于颈动脉斑块同周围组织的灰度相近,本文使用 ITK-SNAP 框选 ROI(Region of interest),通过调整阈值屏蔽灰度值过高或过低的部分,再在符合选定阈值范围内的区域添加参考点。然后,使用 ITK-SNAP 自动选择与参考点相连并且灰度相近的组织进行标注,在完成自动标注后,对有明显斑块的区域进行手动勾画,去掉明显没有斑块的部分。手动选取 ROI 及勾画的过程在临床医师的指导下进行,并经过检验与确认。图像 ROI 的选取及标注示例图像如图 3 所示。



(a) ROI选取　　(b) ROI标注

图 3　MRI 影像 ROI 的选取及软件的自动标注示例图像

颈动脉斑块 MRI 影像标注完成以后,本文利用 Python 中的 PyRadiomics 库对标注区域进行特征数据提取,共提取了 107 个生物标志物,包括三维形状特征(Shape-based)、一阶统计量(First-order statistics)、灰度共生矩阵(Gray-level co-ocurrence matrix)、灰度游程矩阵(Gray-level run-length matrix)、灰度区域大小矩阵(Gray-level size-zone

matrix)、灰阶相关矩阵(Gray-level dependence matrix)和邻域灰阶差分矩阵(Neighboring gray tone difference matrix),使用 $F_1,\cdots,F_{107}$ 表示;然后根据病患的病历信息,给每一个颈动脉斑块 MRI 影像添加了标签,其中稳定的颈动脉斑块影像用 $D=1$ 表示,不稳定用 $D=0$ 表示。共采集样本 141 个,其中:斑块稳定的有 35 个样本,不稳定的有 106 个样本。

3.2 MRI 影像特征的描述性统计分析

为了有效地提取颈动脉斑块稳定性的重要生物标志物,本文首先对每个生物标志物按类进行了描述性统计分析,主要包括计算各类生物标志物的均值、标准差和偏度,同时进行 Shapiro-Wilk 正态性检验,部分生物标志物的计算结果如表 3 所示。根据计算结果把生物标志物分为三类,分别是:第一类为类均值、方差具有较大的差异性且服从双正态分布,第二类为类均值、方差的差异性较小且服从双正态分布,第三类为不服从双正态分布。对第一类生物标志物,例如 F_6 等生物标志物,使用变换 AUC 度量这些生物标志物的信息量时,本文建议使用双正态分布模型下变换 AUC 的估计方法;对第二类生物标志物,例如 F_2 等生物标志物,这些生物标志物在进行初筛时应该去掉;对第三类生物标志物,例如 F_5、F_8 等生物标志物,建议使用自由分布模型下变换 AUC 的估计方法估计这些生物标志物的变换 AUC 值。

表 3 部分生物标志物的描述性统计分析计算结果

生物标志物	稳定组				不稳定组			
	均值	标准差	峰度	p 值	均值	标准差	峰度	p 值
F_1	0.743	0.139	2.891	0.877	0.602	0.144	2.410	0.051
F_2	0.509	0.166	1.964	0.228	0.432	0.130	3.999	0.077
F_3	6.683	2.389	1.948	<0.050	7.498	2.124	3.002	0.767
F_4	13.657	4.099	3.020	<0.050	18.048	4.799	2.718	0.324
F_5	14.566	9.690	24.034	<0.050	16.081	4.469	2.718	0.214
F_6	13.037	3.352	3.182	0.214	17.678	4.515	2.830	0.830
F_7	13.628	9.728	25.367	<0.050	13.401	3.276	2.137	<0.050
F_8	16.073	9.515	23.278	<0.050	19.357	4.334	3.012	0.959
F_{10}	9.872	2.625	2.547	0.188	10.443	2.336	3.697	<0.050
F_{13}	1.750	0.452	2.286	<0.050	1.631	0.438	2.037	<0.050
F_{38}	0.389	0.156	2.114	0.675	0.430	0.166	3.244	0.292
F_{42}	0.456	0.107	1.739	0.099	0.430	0.073	2.655	0.488
F_{43}	0.389	0.127	1.722	0.080	0.356	0.085	2.690	0.273
F_{57}	6.155	0.419	3.168	0.327	6.365	0.420	2.547	0.538
F_{100}	5.591	0.547	2.026	0.322	5.885	0.449	2.642	0.849

3.3 基于变换 AUC 的重要生物标志物的筛选

基于描述性统计分析,对具有区分能力的生物标志物,本文分别用 AUC、F1 分数、Cohen′s kappa 统计系数以及变换 AUC 评估它们的识别能力,部分结果如表 4 所示。从表 4 中可以看出,F_3、F_{85} 等生物标志物的变换 AUC 值和 AUC 值有明显不同,且 F1 分数均在 0 附近,Cohen′s kappa 统计系数都接近−1,若使用 AUC 等常用的评估指标来度量这些生物标志物的信息量,则这些生物标志物可能会被漏选;而这些生物标志物的变换 AUC 值显著大于 0.65,表明变换 AUC 能较好地度量这些生物标志物的信息量,可避免医学筛选时重要生物标志物被错失。因此,本文建议在一般的医学筛选生物标志物时,使用变换 AUC 评估生物标志物的信息量。

表 4 部分生物标志物的变换 AUC 值及其他常用评估指标值

生物标志物	变换 AUC 值	F1 分数	Cohen′s kappa	AUC 值
F_3	0.692	0.002	−0.983	0.610
F_{85}	0.687	0.027	−0.981	0.605
F_{40}	0.686	0.016	−0.982	0.580
F_{63}	0.684	0.036	−0.982	0.606
F_{74}	0.684	0.036	−0.982	0.602
F_{22}	0.680	0.007	−0.984	0.584
F_{27}	0.680	0.003	−0.984	0.599
F_{39}	0.678	0.042	−0.984	0.571
F_{62}	0.677	0.009	−0.985	0.581
F_{105}	0.676	0.011	−0.985	0.575

3.4 颈动脉稳定性预测模型的构建

本文根据变换 AUC 建立一个高精度的颈动脉

斑块稳定性预测模型,为了进一步说明在医学诊断识别中使用变换 AUC 的优势,本文还建立了基于 AUC 的颈动脉斑块稳定性预测模型,并进行了对比分析。

本文分别使用 AUC 和变换 AUC 度量生物标志的信息量,筛选出一些重要的生物标志物,建立变量池。具体过程如下:a)把 AUC 值大于 0.6 的生物标志物放入变量池Ⅰ,共 51 个生物标志物;b)筛选出变换 AUC 值大于 0.6 的生物标志物放入变量池Ⅱ,共 105 个生物标志物。

为了快速找到基于 AUC 的最优颈动脉斑块稳定性预测模型,本文将选用变量池Ⅰ,使用向前逐步回归筛选模型的方法,并使用 AIC(Akaike information criterion)值、AUC 值度量模型偏差值,以说明预测模型的精度。具体的建模过程如表 5 所示。

表 5　基于 AUC 和向前逐步回归筛选的预测模型

序号	模型	AIC 值	模型的 AUC 值
0(开始)	—	—	0
1	$+F_6$	131.47	0.798
2	$+F_1$	122.67	0.841
3	$+F_5$	117.62	0.864
4(结束)	$+F_{66}$	111.49	0.878

注:符号+表示向模型中添加变量。

对于变量池Ⅱ中的生物标志物 F,若其变换 AUC 值显著大于 AUC 值,则对它按式(4)进行 H 变换,变换后的生物标志物用 H_F 表示,将变换后的变量池Ⅱ记为变量池Ⅲ。为了快速找到基于变换 AUC 的最优颈动脉稳定性预测模型,本文使用变量池Ⅲ,采用了向前逐步回归筛选模型的方法,并使用 AIC(Akaike Information Criterion)值、AUC 值度量模型的精度。具体的建模过程如表 6 所示。

表 6　基于变换 AUC 和向前逐步回归筛选的预测模型

序号	模型	AIC 值	模型的 AUC 值
0(开始)	—	—	0
1	$+F_6$	131.470	0.798
2	$+H_{F20}$	116.250	0.851
3	$+H_{F10}$	105.540	0.901
4	$+H_{F66}$	94.490	0.920
5	$+H_{F41}$	82.260	0.943
6	$+H_{F22}$	66.310	0.970
7	$+F_1$	53.843	0.986
8	$+H_{F13}$	48.413	0.991
9	$+H_{F15}$	43.855	0.994
10(结束)	$+H_{F84}$	38.064	0.998

注:符号+表示向模型中添加变量。

由表 5 和表 6 发现,基于变换 AUC 的最优颈动脉斑块稳定性预测模型的精度显著高于基于 AUC 的最优颈动脉斑块稳定性预测模型,最重要的原因是:a)变量池Ⅱ中包含的有识别能力的生物标志物显著多于变量池Ⅰ,使用变换 AUC 评估生物标志物的识别能力能有效地防止重要生物标志物的遗漏;b)本文建议的非单调 H 变换能提高某些非传统的生物标志物的识别能力。因此,在实际应用中,基于变换 AUC 评估生物标志物的识别能力,对原始生物标志物进行非单调 H 变换,能提高医学诊断模型的精度。

4　结　论

为了准确评估颈动脉斑块的稳定性,本文在 ROC 空间提出了变换 AUC 及其在双正态模型下的参数估计方法和自由分布模型下的非参数估计方法,有效解决了常用的 AUC、F1 分数和 Cohen′s kappa 统计系数等评估指标不能准确度量非传统生物标志物的信息量的问题。实验研究表明,变换 AUC 能很好地度量非传统生物标志物的信息量,其值均在 0.6 以上,可以有效防止重要生物标志物的遗漏。此外,基于变换 AUC 建立的医学诊断预测模型,模型的 AUC 值达到了 0.9 以上,相较于传统的模型筛选方法,具有更高的精度。

本文提出了高精度颈动脉斑块稳定性预测模型及其应用过程,然而,对于变换 AUC 的估计问题,本文只提供了双正态模型下的参数估计方法和自由分布模型下的非参数估计方法,且这两种估计方法的准确性和可靠性尚未进行系统研究。此外,关于是否存在更有效的估计方法仍需进一步研究。

参考文献:

[1] Micari A, Nerla R, Vadalà G, et al. 2-year results of paclitaxel-coated balloons for long femoropopliteal artery disease: Evidence from the SFA-long study[J]. JACC Cardiovascular Interventions, 2017, 10(7): 728-734.

[2] Schmidt A, Piorkowski M, Görner H, et al. Drug-coated balloons for complex femoropopliteal lesions: 2-year results of a real-world registry[J]. JACC: Cardiovascular Interventions, 2016, 9(7): 715-724.

[3] 王雪利, 崔志新, 吕文君, 等. 颈动脉斑块的无创影像评价方法的研究进展[J]. 承德医学院学报, 2021, 38(2): 157-162.

[4] Saba L, Francone M, Bassareo P P, et al. CT attenuation analysis of carotid intraplaque hemorrhage [J]. AJNR American Journal of Neuroradiology, 2018, 39(1): 131-137.

[5] 李杨, 查云飞. CTA评价颈动脉斑块成分及体积与脑血管症状相关性[J]. CT理论与应用研究, 2016, 25(5): 601-607.

[6] Lu M M, Cui Y Y, Peng P, et al. Shape and location of carotid atherosclerotic plaque and intraplaque hemorrhage: A high-resolution magnetic resonance imaging study[J]. Journal of Atherosclerosis and Thrombosis, 2019, 26(8): 720-727.

[7] Lu M M, Yuan F, Zhang L C, et al. Segment-specific progression of carotid artery atherosclerosis: A magnetic resonance vessel wall imaging study[J]. Neuroradiology, 2020, 62(2): 211-220.

[8] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[9] 余昊,赵超群,杨建萍. 基于密度比模型的pAUC半参数估计方法及其应用[J/OL]. (2023-03-01)[2023-11-21]. http://kns.cnki.net/kcms/detail/33.1338.TS.20230331.0921.009.html.

[10] 赵超群,余昊,杨建萍. 正态总体决策曲线参数估计方法及其应用[J]. 浙江理工大学学报(自然科学), 2023, 49(3):379-387.

[11] 王彦光, 朱鸿斌, 徐维超. ROC曲线及其分析方法综述[J]. 广东工业大学学报, 2021, 38(1): 46-53.

[12] 王曼, 徐春燕, 施学忠. 医学论文中ROC曲线应用错误例析[J]. 编辑学报, 2019, 31(2): 159-161.

[13] 何小梅, 王林晓. Logistic模型和ROC曲线对替加环素致凝血异常的预测分析[J]. 中南药学, 2020, 18(9): 1577-1580.

[14] Zhou X H, Obuchowski N A, McClish D K. Statistical Methods in Diagnostic Medicine[M]. Hoboken: John Wiley & Sons, 2009:261-296.

[15] Zou K, Liu A, Bandos A, Ohno-Machado L, Rockette H. Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis[M]. Boca Raton: CRC Press, 2012:6-14.

[16] Bantis L E, Tsimikas J V, Chambers G R, et al. The length of the receiver operating characteristic curve and the two cutoff Youden index within a robust framework for discovery, evaluation, and cutoff estimation in biomarker studies involving improper receiver operating characteristic curves[J]. Statistics in Medicine, 2021, 40(7): 1767-1789.

(责任编辑:康 锋)