

## 基于特征选择和 XGBoost 优化的术中低体温预测

曹立源<sup>1</sup>, 范勤勤<sup>1</sup>, 黄敬英<sup>2</sup>

(1. 上海海事大学物流研究中心, 上海 201306; 2. 浙江大学医学院附属邵逸夫医院麻醉恢复室, 杭州 310020)

**摘要:** 针对全麻手术患者术中低体温发生率高、影响因素复杂的问题, 提出了一种基于特征选择和 XGBoost 优化的术中低体温预测模型, 以更好辅助医生对全麻手术患者的临床诊断。首先, 利用随机森林(Random forest, RF)在处理高维数据集上的优势, 通过 RF 的袋外估计法进行特征选择。然后, 以极端梯度提升(XGBoost)为基础, 利用基于精英保留策略的遗传算法, 即 EGA 算法优化 XGBoost 超参数。最后, 根据最优参数训练预测模型, 并用于术中低体温预测。该模型组合利用 3 种算法优点, 提升模型泛化能力和预测精度。实验结果表明: 本文所提模型与逻辑回归、支持向量机等 7 种机器学习预测模型相比, 预测准确率、精确度、召回率、AUC 均有优势; 与现有其他预测模型相比均有提升。

**关键词:** 术中低体温预测; 特征选择; 极端梯度提升; 遗传算法

**中图分类号:** TP3      **文献标志码:** A

## Intraoperative Hypothermia Prediction Model Based on Feature Selection and XG-Boost Optimization

CAO Liyuan<sup>1</sup>, FAN Qinqin<sup>1</sup>, HUANG Jingying<sup>2</sup>

(1. Logistics Research Center, Shanghai Maritime University, Shanghai 201306, China; 2. Anesthesia Recovery Room, Shao Yifu Hospital Affiliated to Zhejiang University Medical College, Hangzhou 310020, China)

**Abstract:** In view of the high incidence of intraoperative hypothermia and complex influencing factors in patients undergoing anesthesia, a prediction model of intraoperative hypothermia based on feature selection and XGBoost optimization is proposed to better assist doctors in the clinical diagnosis of patients. Firstly, the random forest (RF) is used to deal with the high-dimensional data sets, and features are selected by the RF out-of-bag estimation. Then, XGBoost hyperparameters are optimized using the genetic algorithm based on elite retention strategy, i. e., EGA. Finally, the prediction is trained according to the optimal parameters and thus can be used to predict intraoperative hypothermia. This model combines the advantages of three algorithms to improve model generalization ability and prediction accuracy. The experimental result shows that the proposed model performs better other seven machine learning classification prediction models such as logistic regression, support vector machine, and so on in prediction accuracy, precision, recall and AUC, and overcomes the three representative hyperparameter tuning methods.

**Key words:** intraoperative hypothermia prediction; feature selection; XGBoost; genetic algorithm

## 引言

术中低体温是指体温在手术中任何时点均小于 $36\text{ }^{\circ}\text{C}$ ,是手术患者常见并发症<sup>[1]</sup>。国内外研究显示,术中有 $50\%\sim 70\%$ 的病人会出现低体温<sup>[2]</sup>,低体温会严重影响手术质量和患者术后康复<sup>[3-4]</sup>。全麻手术患者由于麻醉剂对机体温度调节等影响,常发生术中低体温问题<sup>[5]</sup>。

目前,许多学者对全麻手术患者术中低体温问题进行了深入研究,主要集中在3个方面:(1)术中低体温影响因素的研究。文献[6]采用回顾性研究方法,研究结果显示麻醉时间、手术时间、术中输血量等15个因素影响明显。文献[7]采用一元线性回归,实证了术前主动加温对术中低体温影响具有统计学意义。(2)术中低体温影响因素评价指标的研究。文献[8]采用专家咨询与层次分析相结合的方法,构建了一个由2个一级指标和21个二级指标组成的术中低体温影响因素评价指标体系。文献[9]采用多元线性回归,建立术中低体温预测因子评分方程,发现手术等级因子权重最大。(3)术中低体温预测模型的研究。文献[10]采用 Logistic 回归,构建腹腔镜术中低体温预测模型,结果显示模型拟合度中等偏上。文献[11]采用单因素分析、多因素 Logistic 回归,建立术中低体温预测模型,研究发现模型预测精度好于多元线性模型。近年来,机器学习在医疗诊断相关领域的研究增多。杨晶东等<sup>[12]</sup>提出一种集成学习机制下的鼻炎辅助诊断模型,研究表明该模型对多数类和少数类的分类精度较高。Chung 等<sup>[13]</sup>建立一个基于支持向量机的糖尿病诊断模型,给出了糖尿病特征变量筛选及诊断模型。张春富等<sup>[14]</sup>提出一个基于 GA\_XGboost 的糖尿病风险预测模型,结果显示预测精度优于支持向量机算法。岳鹏等<sup>[15]</sup>使用 Stacking 方法集成多种异质分类器,结果表明识别准确率高于单一分类器模型。Zhong 等<sup>[16]</sup>采用 GA-CG-SVM 方法,解决乳腺癌多元回归预测模型过拟合问题。雷雪梅等<sup>[17]</sup>建立一个改进的 XGBoost 模型,对高血压菜谱进行识别,结果显示用遗传算法优化模型参数的效果好于网格搜索法。以上研究取得了很好进展,但仍存在一定不足:回顾性分析法不能估算术中低体温发生率;一元线性回归对术中低体温的解释力不足;多元线性回归易忽略交互效应;Logistic 回归易产生过拟合;机器学习应用于术中低体温方面的研究较少,其应用性能有待进一步提高。

为此,本文提出一种基于特征选择和 XGBoost 优化的术中低体温预测模型。首先利用随机森林(Random forest, RF)的袋外估计法进行特征选择,然后利用基于精英保留策略的遗传算法,即 EGA 算法优化 XGBoost 超参数,最后根据最优参数训练预测模型并用于术中低体温预测。本文的主要贡献在于利用随机森林、遗传算法和极端梯度提升算法进行组合,结合3种算法优点,尝试机器学习在医疗诊断领域的新拓展;探索基于 Bagging 的集成学习算法 RF 和属于 Boosting 的集成学习算法 XGBoost 这两种不同集成学习算法的组合方法,提升算法模型的预测性能;所提模型对患者术中低体温的辅助诊疗具有较强适应性和效用性。实验中,采用 RF 袋外估计法获取特征重要性排序,得到低体温强相关特征,有效提高低体温的特征选择客观性和准确性;利用 EGA 对 XGBoost 参数进行调优,获得全局最优参数组合,模型性能提升,且 EGA 对 XGBoost 参数调优效果优于对 GBDT 和 SVM 参数调优效果,EGA 参数调优适用性有效性明显;EGA 作为参数调优方法,对 XGBoost 参数调优效果也优于网格搜索和随机搜索对 XGBoost 参数调优效果。这3项举措提升了模型预测精度,增强了模型应用于术中低体温预测的泛化性和有效性。

## 1 术中低体温预测模型

本文提出基于特征选择和 XGBoost 优化的术中低体温预测模型,其框架流程主要包括数据预处理、特征选择、超参数优化和模型评估4个步骤,如图1所示。

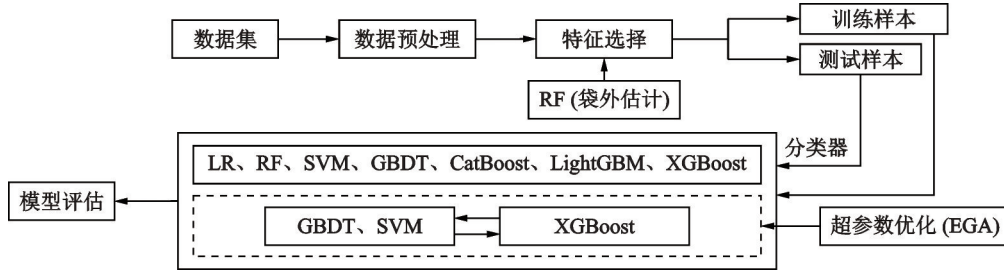


图1 术中低体温预测模型框架

Fig.1 A prediction model frame of intraoperative hypothermia

1.1 数据预处理

对全麻手术患者术中低体温原始数据进行数据清洗和数据转换预处理,包括:剔除记录不实数据,如腹部手术、食管手术的出血量为0 mL;剔除有缺失数据的病人数据,如手术麻醉信息系统中无体液出入量、BMI(体质指数)等记录;剔除逻辑核查有误数据等。同时,对数据进行归一化转换,去除数据的单位限制,将其转化为无量纲的纯数值,使数值化后的每个数值归一化到[0,1]区间,具体转化公式为

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

式中: $x_i, x_i^*$ 分别表示数据归一化前后的值, $x_{\min}, x_{\max}$ 分别表示样本数据中的最小值和最大值。

1.2 基于随机森林的特征选择

Breiman<sup>[18]</sup>在 Bagging 基础上提出了 RF 学习方法,RF 是一种基于集成学习思想,采用 Bootstrap 重抽样法从原始数据样本中抽取多个样本,利用节点随机分裂技术对每个样本进行决策树建模<sup>[19]</sup>,构建多棵决策树,并用投票法将多棵决策树的预测结果归总输出。

给定数据样本  $X$  和预测集  $Y$ ,在服从随机变量  $Y, X$  的分布的样本集中随机抽取  $n$  个样本作为训练集,从已训练过的样本集中随机抽取  $i$  个样本,每次随机抽样的结果为一棵决策树模型  $T_k(X, Y), k = 1, 2, 3, \dots, K$ 。其中, $X$  为自变量, $Y$  为因变量,则由  $k$  棵决策树组合形成的回归估计  $\bar{r}_n(X, D_n)$  为

$$\bar{r}_n(X, D_n) = E_Y[r_n(X, Y, D_n)] \tag{2}$$

$$r_n(X, Y, D_n) = \frac{\sum_{i=1}^n y_i I_{[X_i \in A_n(X, Y)]}}{\sum_{i=1}^n I_{[X_i \in A_n(X, Y)]}} I_{[E_N]}(X, Y) \tag{3}$$

式中: $D_n$  为数据集; $X(X_1, X_2, X_3, \dots, X_i)$  为  $D_n$  中被抽取的自变量随机参数; $E_Y$  为与  $X$  有关的期望函数; $I_{[\cdot]}$  为示性函数,当下标条件满足时  $I_{[\cdot]}=1$ , 否则  $I_{[\cdot]}=0$ ;  $y_i$  为在学习样本中与  $X_i$  相对应的因变量; $[X_i \in A_n(X, Y)]$  为包含一个  $X$  的随机单元分区; $I_{[E_N]}(X, Y)$  为判定条件,  $I_{[E_N]}(X, Y) = \left\{ \sum_{i=1}^n I_{[X_i \in A_n(X, Y)]} \neq 0 \right\}$ 。

联立式(2)和式(3)可得到 RF 对因变量  $Y$  的回归估计为

$$\bar{r}_n(X, D_n) = E_Y \left[ \frac{\sum_{i=1}^n y_i I_{[X_i \in A_n(X, Y)]}}{\sum_{i=1}^n I_{[X_i \in A_n(X, Y)]}} I_{[E_N]}(X, Y) \right] \tag{4}$$

式(4)表明,增加重抽样次数、决策树数量,改善期望 $E_Y$ ,模型的预测精度提升,同时不产生过拟合,这是因为RF在算法原理的随机抽样和集成学习上有优势。

假设给定一组决策树模型 $\{T_1(M), T_2(M), T_3(M), \dots, T_K(M)\}$ ,用上述方法可得训练数据集,由此进一步得到边缘函数

$$\text{mg}(X, Y) = \text{av}_k I(T_k(M) = Y) - \max_{j \neq Y} \text{av}_k I(T_k(M) = j) \quad (5)$$

式中: $X$ 为自变量; $Y$ 为因变量; $\text{av}$ 为平均值; $k$ 为决策树个数; $j$ 为 $Y$ 包含的不同类别。边缘函数代表正确与错误分类相差的幅度,与分类准确率呈正相关。进一步,泛化误差 $\text{PE}^*$ 可表示为

$$\text{PE}^* = P_{X, Y}(\text{mg}(X, Y) < 0) \quad (6)$$

式中 $P_{X, Y}$ 为概率分布。当决策树分类器数量 $k$ 不断增加时, $\text{PE}^*$ 收敛于

$$P_{X, Y}(P_\theta(T(X, \theta_k) = Y) - \max_{j \neq Y} P_\theta(T(X, \theta_k) = j) < 0) \quad (7)$$

式中: $P_\theta$ 为每棵决策树概率分布, $T(X, \theta_k)$ 满足大数定律, $\theta_k$ 为独立抽样且具有同分布的随机变量。由此可以说明在利用随机森林算法进行分类预测时,不会因为决策树个数 $k$ 的增加而产生过拟合问题。

RF的优势还体现在可以利用RF袋外估计法进行特征选择。RF是有放回抽样,在每一个初始训练集中,约有1/3的数据未被抽取,这些未被抽取数据被称为袋外数据(Out-of-bag, OOB)。基于RF袋外估计法可通过比较每个变量在加入噪声前后的模型OOB拟合误差来判定该变量的重要性及其对模型的影响程度,不存在偏向问题<sup>[20]</sup>,进而计算每个变量影响模型精度的下降率作为特征重要性度量指标,这就避免了简单地将特征用作分割属性的频率作为特征重要性度量指标所引致的无法准确反映特征间重要性的差异问题<sup>[21]</sup>。同时,对特征重要性由大到小排序,筛选出最优特征。基本过程是采用Bootstrap抽样法从原样本数据集中(大小为 $N$ )有放回地随机抽样 $N$ 次,构造一个大小同为 $N$ 的训练集,未出现在训练集里的样本数据为袋外数据;计算每一个特征重要性,并用袋外数据作为测试集,测试错误率记为袋外误差(error);计算所有基学习器的测试平均误差,以精度平均下降率(Mean descent altitude, MDA)作为特征重要性评价指标,按式(8)计算各个特征的MDA值<sup>[22]</sup>。

$$\text{MDA} = \frac{1}{n} \sum_{t=1}^n (\text{error}_t - \text{error}'_t) \quad (8)$$

式中: $n$ 为基学习器的总数; $\text{error}_t$ 为基学习器 $t$ 未加入噪声后的袋外误差, $\text{error}'_t$ 为基学习器 $t$ 加入噪声后的袋外误差。MDA指标下降越大,其所对应特征对预测结果影响越大,重要性越高,即可进行特征重要性排序,选择最优特征。

本文利用RF重抽样技术、泛化性和收敛性等优点,采用RF袋外估计法对高维度的术中低体温特征变量进行重要性排序和特征选择,并对特征选择的不同方案进行比较,进而以OOB数据的验证集上分类结果加予检验,筛选出最优特征集。

### 1.3 XGBoost超参数优化

XGBoost的基本思想是在梯度提升决策树(Gradient boosting decision tree, GBDT)基础上发展起来的,主要对GBDT目标函数进行正则化及二阶泰勒展开,在目标函数中加入正则项,使模型更为简单,防止过拟合,提升模型泛化性能<sup>[23]</sup>;同时利用一阶与二阶导数逼近目标函数,获取更多信息训练树模型<sup>[24]</sup>,提高算法的收敛速度和准确性。

假设一个具有 $n$ 个样本、 $m$ 个特征的数据集 $D = \{(X_i, y_i)\} (|D| = n, X_i \in \mathbf{R}^m, y_i \in \mathbf{R})$ ,其中: $|D|$ 表示集合 $D$ 的维度,即包含的样本数量; $X_i$ 为第 $i$ 个样本的特征向量; $y_i$ 为第 $i$ 个样本的标签即真实值,则由 $K$ 棵树集成的模型预测输出 $\hat{y}_i$ 为

$$\widehat{y}_i = \sum_{k=1}^K f_k(X_i) \quad f_k \in F \quad (9)$$

式中:  $F = \{f(X) = \omega_{q(X)} \mid q: \mathbf{R}^m \rightarrow T, \omega_{q(X)} \in \mathbf{R}^T\}$  为  $K$  棵树集成的函数空间,  $T$  为叶子节点个数;  $f_k$  对应一棵独立的树结构  $q$  与叶子权重  $\omega$ ,  $q$  为样本对应的叶子标签, 第  $i$  个叶子节点的分数即权重为  $\omega_i$ 。

为了学习  $K$  棵树的函数集合, 最小化下列正则化目标函数

$$L(\phi) = \sum_i l(y'_i, y_i) + \sum_k \Omega(f_k) \quad (10)$$

式中:  $\phi$  为  $K$  棵树的函数集合;  $l(y'_i, y_i) = (y_i - \widehat{y}_i)^2$  为损失函数项;  $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2$  为正则项, 用于对模型的复杂度进行惩罚, 其中  $\gamma$  和  $\lambda$  为正则化参数,  $\gamma$  控制叶子节点个数  $T$ ,  $\lambda$  控制叶子节点权重  $\omega$ , 防止过拟合。式(10)用于衡量预测值与标签值的差距。

为了快速优化目标, 要用二阶泰勒展开, 以更快更准地逼近优化目标函数。由于上述目标函数以函数作为参数, 因此采用加法学习方式<sup>[25]</sup>, 即在现有  $t-1$  棵树的基础上添加 1 棵最优化的 CART 即  $f_t$ , 使目标函数变为式(11)来优化目标。

$$L^{(t)} = \sum_{i=1}^n l(y_i, \widehat{y}_i^{(t-1)} + f_t(X_i)) + \Omega(f_t) \quad (11)$$

将式(11)进行二阶泰勒展开得到

$$L^{(t)} \approx \sum_{i=1}^n \left[ l(y_i, \widehat{y}_i^{(t-1)}) + g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] + \Omega(f_t) \quad (12)$$

$$g_i = \partial_{\widehat{y}_i^{(t-1)}} l(y_i, \widehat{y}_i^{(t-1)}) \quad (13)$$

$$h_i = \partial_{\widehat{y}_i^{(t-1)}}^2 l(y_i, \widehat{y}_i^{(t-1)}) \quad (14)$$

式中:  $g_i, h_i$  分别为一阶偏导和二阶偏导。因前面  $t-1$  棵树已训练完成, 故  $l(y_i, \widehat{y}_i^{(t-1)})$  为常数项, 移除该项得到第  $t$  次的简化目标函数为

$$\widetilde{L}^{(t)} = \sum_{i=1}^n \left( g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right) + \Omega(f_t) \quad (15)$$

定义  $I_j = \{i \mid g(X_i) = j\}$  为叶子节点  $j$  中的样本编号集合, 其中  $q(X_i)$  为  $X_i$  对应的叶子标签值。通过展开正则项  $\Omega$ , 可将式(15)写为

$$\widetilde{L}^{(t)} = \sum_{i=1}^n \left( g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (16)$$

式(16)是直接统计各个样本的值; 如式(16)变换成式(17), 就可把对样本的遍历转换为对树的叶子节点的遍历, 统计落在每个叶子节点中所有样本的一阶导数  $g_i$  和该叶子节点权重  $\omega$  的乘积, 以及二阶导数  $h_i$  和该叶子节点权重  $\omega$  的乘积, 使算法的收敛速度和准确性提升。

$$\widetilde{L}^{(t)} = \sum_{j=1}^T \left[ \left( \sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \quad (17)$$

式中  $\omega_j$  为叶子节点  $j$  的权重。

对于一个固定的树结构  $q(X)$ , 基于凸优化原则, 对  $\omega_j$  求导并令导函数等于 0, 得到叶子节点  $j$  的最优权重如式(18)所示, 相应的最优目标函数值如式(19)所示。

$$\omega_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (18)$$



$$\widetilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{\left( \sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (19)$$

式(19)用于评价一棵树模型的好坏,得分越小树模型越好。因此,如何优化模型超参数就成了关键。迭代次数过少,容易产生欠拟合;迭代次数过多,容易产生过拟合;学习率过小,梯度下降很慢;学习率过大,可能造成跨过最优值;模型中其他超参数,如树的深度、树的数量、最小叶子节点权值、L2正则化系数、叶子节点分裂时所需要最小损失减少量等都会对模型性能产生影响,因此本文将对XGBoost的学习率、树的深度、树的数量、最小叶子节点权值等6个超参数进行优化,以提高预测模型性能。

XGBoost超参数优化的目标为  $X^* = \arg \min_{X \in D} \widetilde{L}^{(t)}$ 。其中,  $D$  为  $X$  可选的数据集,  $L^{(t)}(q)$  为 XGBoost 的目标函数,  $X^*$  为使目标函数最优时  $D$  中的一组  $X$ 。

本文采用具有精英保留策略的遗传算法(EGA)优化XGBoost超参数。EGA的基本思想是把群体在进化中迄今出现的最好个体不进行遗传选择、交叉和变异操作而直接“复制”到下一代中,避免迄今出现的最好个体在遗传选择算子、交叉算子和变异算子操作中被破坏,从而使EGA朝着全局最优进化快速搜索到最优解。本文以实数和整数混合编码的方式来表示个体的染色体,每条染色体代表XGBoost待优化的一组超参数;把适应度函数设置为5折交叉验证的精度,用于衡量每条染色体的适应度;在编码和搜索空间规定的范围内随机产生数量为  $q$  个的超参数组合,作为初始种群启动遗传进化程序;通过评价种群,选出当前种群中适应度最好个体,若当前种群中最好个体的适应度比迄今出现的最好个体的适应度值高,则以当前种群中最好个体作为新的迄今出现的最好个体,直接“复制”进入下一代的繁衍中;利用算法产生新一代种群,对父代个体进行选择、交叉、变异,计算子代种群中的各个个体适应度值;用迄今出现的最好个体替换子代中最差个体,进化寻优。当遗传进化迭代次数达到最大或者终止条件时,输出最优染色体个体,即为XGBoost的最优超参数组合。本文将EGA优化XGboost超参数与优化GBDT和SVM超参数的效果进行比较;还将EGA优化XGboost超参数与网格搜索和随机搜索优化XGboost超参数的效果进行对比,验证本文采用EGA优化XGBoost超参数的合理性、有效性。

#### 1.4 模型评估

本文把混淆矩阵作为术中低体温模型性能评价指标,如表1所示。表1中:TP(True positive)表示实际有低体温,预测也有低体温;FN(False negative)表示实际有低体温,但预测没有低体温;FP(False positive)表示实际没有低体温,但预测有低体温;TN(True negative)表示实际没有低体温,预测也没有低体温。基于混淆矩阵,结合医学中的

评价指标<sup>[26]</sup>,选用分类模型中常用的4个指标作为模型评估依据,即:准确率(Accuracy)、精确度(Precision)、召回率(Recall)和  $F_1$  分数( $F_1$ -score)作为预测模型性能评估依据。计算公式分别为

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) \quad (20)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (21)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (22)$$

$$F_1\text{-score} = 2(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (23)$$

表1 混淆矩阵

Table 1 Confusion matrix

实际类别	预测类别	
	True(有低体温)	False(没有低体温)
True(有低体温)	TP	FN
False(没有低体温)	FP	TN

式中:Accuracy是所有预测正确的样本数与总样本数的比值;Precision是所有预测为正样本的样本中,真实为正样本的比例;Recall是预测为正样本的数量占有所有正样本总数的比例; $F_1$ -score是精确度和召回率的调和平均数。

同时,把ROC曲线下的面积AUC作为测量评价模型性能优劣的一个指标,用以上5个指标作为本文所提的术中低体温预测模型性能评估依据。

## 2 实验结果分析

### 2.1 数据集

本文以某市三甲医院2020年1月至2021年2月的全麻手术患者术中低体温病例数据为数据源,经过对数据进行预处理,删去较多缺失异常值的样本,最终入组294例。其中,纳入的病人中术中发生低体温的病人198例,未发生低体温的病人96例,低体温发生率为67%。

### 2.2 特征选择

通过数据预处理筛选出的全麻手术患者术中低体温样本数据294个,样本特征变量分别为全麻手术患者的年龄、性别、身体质量指数(BMI)、术前基础体温、麻醉时间、术中输液量、术中出血量、是否冲洗、ASA评级(ASA为美国麻醉医师协会分级)等15个特征为自变量特征,如表2所示。

随机森林生成,其主要步骤如下:

(1)在原始294个样本数据集中,有放回地随机抽样294次(每次抽1个,然后放回继续)组成训练子集,训练一棵决策树。未被抽到的样本则组成一个相应的OOB。

(2)决策树的每个节点处从15个特征中随机挑选 $m(m < 15)$ 个特征,按照节点不纯度最小原则进行分裂。

(3)每棵决策树尽最大程度生长,在生长过程中每个节点都按步骤2来分裂,一直到每个节点无法继续分裂为止。在决策树形成过程中不需要剪枝过程。

(4)按上述第1步到第3步构建随机森林,一共进行80次采样,组成80个训练子集,生成80棵决策树。而每一次未被抽到的样本则组成了相应的80个OOB。

在随机森林的构建过程中,利用OOB数据输入到相应的每个决策树分类器,根据多数投票法,得到随机森林模型分类结果的精度评价。当进行50次采样、生成50棵决策树时,模型分类效果最优,分类精确率为0.8006,如图2所示,故将决策树数量设为50。

特征重要性度量及特征选择方案。利用生成的随

表2 自变量特征赋值及特征重要性  
Table 2 Independent variable feature assignment and feature importance

自变量特征	赋值方式	特征重要性/%
术前体温	原值输入	26.54
BMI	原值输入	14.01
麻醉时间	原值输入	13.27
年龄	原值输入	12.95
术中输液量	原值输入	8.71
术中出血量	原值输入	8.39
是否冲洗	否=0, 是=1	5.19
胸科手术	否=0, 是=1	1.95
性别	女=0, 男=1	1.92
骨科手术	否=0, 是=1	1.72
普通外科手术	否=0, 是=1	1.53
泌尿外科手术	否=0, 是=1	1.23
ASA评级	评级 I=1; II=2; III=3	1.13
肛肠外科手术	否=0, 是=1	0.83
妇科手术	否=0, 是=1	0.61

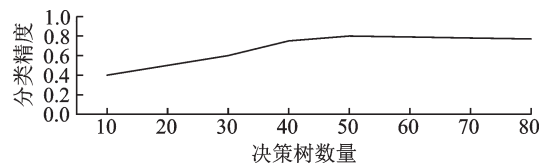


图2 模型分类精度随决策树数量变化

Fig.2 Model classification accuracy varies with decision trees

机森林、OOB误差估计法,计算15个特征变量进行重要性度量,重复进行10次计算,取每个特征重要性的10次计算平均值,得到各个特征变量对模型的重要性大小排序,如表2所示;按特征重要性由大到小排序,删除重要性最小的特征,对剩下的特征子集进行随机森林分类,重复上述步骤,得到OOB得分(分类精度)随特征个数 $m \in [1, 15]$ 变化曲线,如图3所示。由图3可知,当 $m$ 为7时,OOB得分最高。

把这7个特征的选择方案与大于7个特征、小于7个特征的方案进行比较,当 $m$ 为4、7和10时,OOB得分分别为0.695 3、0.802 4和0.790 5。从4个特征到7个特征时OOB得分迅速上升,说明增选这3个特征对模型重要性影响较大,之后再增选特征时OOB得分趋于缓慢下降,即 $m$ 为7个特征时是最优特征选择,这也印证了RF算法及袋外估计法的收敛性和泛化性优点。

特征最优选择及分类效果。为验证最优分类精度时的特征个数,基于特征重要性排序和特征选择方案比较结果,利用OOB数据集建立混淆矩阵,检测评价分类结果。表3展示了特征个数分别为4、7、12时采用随机森林算法进行分类检测时的效果比较。

由表3可知,当特征个数为7时分类效果最优,因此选取特征重要性排名前7个特征,即:术前体温、体质指数、麻醉时间、年龄、术中输液量、术中出血量和是否冲洗作为术中低体温预测模型的输入变量较为合理有效。

### 2.3 超参数优化

将特征选择后的数据集,按7:3比例随机划分为训练集和测试集,以筛选出的7个最优特征变量作为XGBoost模型输入变量,选用5折交叉验证,基于EGA优化XGBoost超参数。主要流程如下:

- (1)设置XGBoost超参数范围;
- (2)将遗传算法的适应度函数设为XGBoost在训练集上5折交叉验证的精度;
- (3)利用EGA优化XGBoost超参数,找出一个超参数组合使目标函数 $\widetilde{L}^{(t)}$ 达到最优值;
- (4)将最优超参数组合输入XGBoost算法进行训练,计算XGBoost预测精度;
- (5)若达到最大迭代次数,则输出XGBoost超参数最优值,否则返回步骤3。

本文设定的XGBoost超参数取值范围为(其他超参数保持默认设定值):树的深度(max\_depth), (3, 7);树的数量(n\_estimators), (0, 300);学习率(learning\_rate), (0.01, 0.3);最小叶子节点权重(min\_child\_weight), (0, 3);L2正则化系数(reg\_lambda), (0, 2);叶子节点分裂时所需要的最小损失减少量(gamma), (0, 4)。

根据遗传算法的特点,结合XGboost超参数范围,GA参数设置为:种群规模100,进化的最大代数100,进化停滞判断阈值 $1 \times 10^{-4}$ 、允许进化停滞代数20。经过24次迭代,遗传算法收敛并搜索到最优适应度函数值为0.829 2,如图4所示。此时输出XGBoost模型的最优超参数,如表4所示。

在上述GA参数设置下,用EGA优化主流分类器GBDT和SVM超参数,优化过程如图5、6所示。

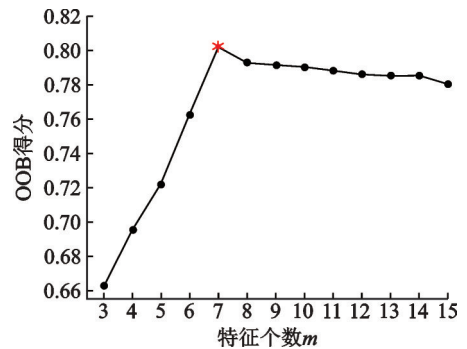


图3 前 $m$ 个特征的OOB得分结果

Fig.3 The first  $m$  features for OOB score

表3 不同特征数量采用RF算法分类效果比较

Table 3 RF classification effects of different feature numbers

特征个数	准确率	精确率	召回率	$F_1$
4	0.695 3	0.710 8	0.785 9	0.754 6
7	0.802 4	0.811 5	0.882 3	0.850 1
12	0.790 5	0.806 2	0.873 1	0.841 8



将优化后的3大分类器在验证集上交叉验证结果进行比较,如表4所示(表中括号内的数字为各分类器默认参数下结果)。可以看出,优化后的各个分类器的模型性能总体均有明显提升,如准确率指标,EGA-XGBoost提升了7.31%,EGA-GBDT和EGA-SVM也分别提升了4.87%和2.92%;但在具体指标上也有差异,EGA-GBDT和EGA-SVM各有4个指标提升、1个指标略有下降,说明EGA优化主流分类器超参数具有明显的适用性和有效性。EGA-XGBoost的5个指标均有明显提升,说明本文采用EGA优化XGBoost超参数的效果更佳。

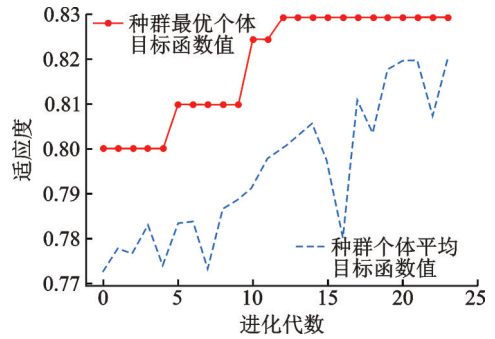


图4 EGA-XGBoost优化过程图  
Fig.4 Optimization process of EGA-XGBoost model

表4 EGA优化主流分类器超参数及交叉验证结果  
Table 4 Optimization of hyperparameters and results

	超参数	max_depth	n_estimators	learning_rate	min_child_weight	reg_lambda	gamma
XGBoost	最优值	5	146	0.201 4	0.825 6	1.024 1	2.996 4
	交叉验证	准确率		精确度	召回率	$F_1$	AUC
		0.829 2 (0.756 1)		0.858 5 (0.815 6)	0.892 1 (0.833 9)	0.873 2 (0.842 1)	0.865 4 (0.841 1)
	超参数	max_depth	n_estimators	learning_rate	max_features	subsample	min_leaf
GBDT	最优值	6	60	0.062 0	5	0.488 0	2
	交叉验证	准确率		精确度	召回率	$F_1$	AUC
		0.790 2 (0.741 5)		0.834 4 (0.818 9)	0.870 1 (0.848 7)	0.848 8 (0.840 1)	0.818 8 (0.832 0)
	超参数	惩罚参数 $c$			核函数参数 $g$		
SVM	最优值	219.351 1			0.003 9		
	交叉验证	准确率	精确度	召回率	$F_1$	AUC	
		0.795 1 (0.765 9)	0.799 8 (0.767 1)	0.934 7 (0.942 3)	0.860 2 (0.844 9)	0.843 3 (0.832 7)	

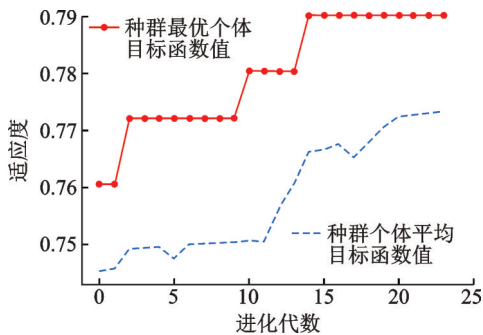


图5 EGA-GBDT优化过程图  
Fig.5 Optimization process of EGA-GBDT model

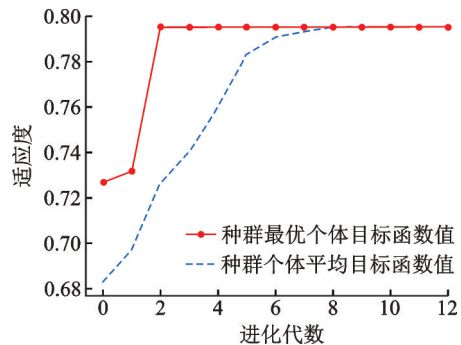


图6 EGA-SVM优化过程图  
Fig.6 Optimization process of EGA-SVM model

### 2.4 模型性能分析

经 EGA 优化 XGBoost 超参数后,使用训练集训练 EGA-XGBoost,模型在训练集上的准确率为 91.71%。将训练后的模型用于测试集进行术中低体温预测,测试集中,实际发生术中低体温 63 例,发生率为 70.79%,模型判断发生 60 例,误判 3 例,召回率为 0.952 4;实际未发生术中低体温 26 例,模型判断未发生 17 例,误判 9 例,精确度为 0.869 6;准确率为 0.865 2;ROC 曲线下面积为 0.894 7,如图 7、8 所示,模型各项性能指标表现良好。

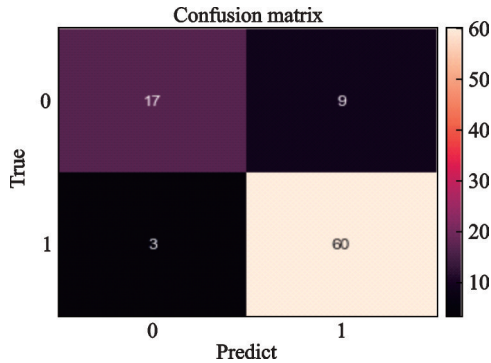


图 7 EGA-XGBoost测试集混淆矩阵

Fig.7 Confusion matrix of test set EGA-XGBoost

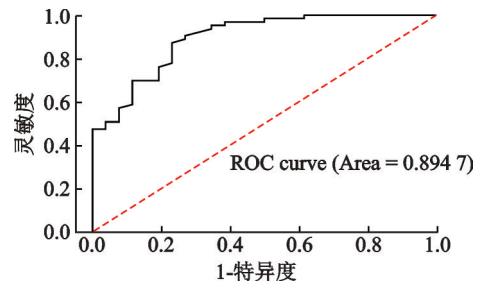


图 8 模型在测试集的 ROC 曲线下面积

Fig.8 Area under the ROC curve of the model

与其他主流分类器模型性能对比。使用 SVM、LR、RF、GBDT、XGBoost、LightGBM 和 CatBoost 这 7 种主流分类器(不进行上述超参数优化),通过与 EGA-XGBoost 相同的训练集训练、测试集测试,测试结果如表 5 所示。由表 5 可知,EGA-XGBoost 的准确率、精确度、召回率、 $F_1$  和 AUC 值均处于领先,本文所提的 EGA-XGBoost 模型性能明显占优。

表 5 模型在测试集上的测试结果对比

Table 5 Comparison of test results of model on the test

模型	准确率	精确度	召回率	$F_1$	AUC
SVM(RBF)	0.775 3	0.786 7	0.936 5	0.855 1	0.865 7
LR	0.797 7	0.800 0	0.952 4	0.869 6	0.890 7
RF	0.820 2	0.840 6	0.920 6	0.878 8	0.849 2
GBDT	0.831 5	0.863 6	0.904 8	0.883 7	0.848 6
XGBoost	0.831 5	0.863 6	0.904 8	0.883 7	0.833 3
LightGBM	0.797 8	0.846 2	0.873 0	0.859 4	0.819 3
CatBoost	0.831 5	0.852 9	0.920 6	0.885 5	0.880 3
EGA-XGBoost	0.865 2	0.869 6	0.952 4	0.909 1	0.894 7

与超参数优化的其他方法对比。网格搜索也是超参数调优的常用方法,其思想是穷举搜索,寻找待选参数最佳值,其缺点是耗时长。随机搜索调优方法是一种全局优化算法,只要随机的次数够多,就能找到最优或近优参数,其缺点是随机结果可能不一致。EGA、网格搜索和随机搜索对 XGBoost 超参数调优及预测结果如表 6 所示。

表6 XGBoost超参数调优方法及结果对比

Table 6 Comparison of XGBoost hyperparameter tuning optimization

方法	准确率	召回率	运行时间/s	超参数列表
EGA 调优	0.865 2	0.952 4	162	(5,146,0.201 4,0.825 6,1.024 1,2.996 4)
网格搜索调优	0.853 9	0.936 5	1 209	(6,50,0.1,2,1.5,0.3)
随机搜索调优	0.842 7	0.936 5	137	(3,167,0.195 6,0.224 5,1.754,2.521 4)

表6中超参数列表的顺序为:树的深度、树的数量、学习率、最小叶子节点权值、L2正则化系数和叶子节点分裂时所需要的最小损失减少量。从表6可知,这3大常用的超参数调优方法,都使模型精度提升,印证了XGBoost目标函数最优化的算法原理,不论选用哪个调参方法,若超参数优化,则模型性能提升,目标函数最优化。比较而言,EGA调优效果总体较好;网格搜索的准确率高于随机搜索,但很耗时,接近随机搜索的10倍;随机搜索耗时最少但精度略低,说明本文选用EGA优化XGBoost超参数较为合理有效。

与现有的其他预测模型对比,文献[9]基于多元线性回归建立术中低体温预测模型,AUC为0.771,泛化性不够好;文献[11]基于Logistic回归建立术中低体温预测模型,预测准确率为81%,灵敏度为0.639,AUC为0.814,模型性能不高。因此本文所提模型性能占优。

### 3 结束语

本文提出了一种基于特征选择和XGBoost优化的全麻手术患者术中低体温预测模型。针对术中低体温特征属性及数据维度高的问题,本文模型通过RF袋外估计法选择最优特征,采用EGA优化XGBoost超参数。实验表明本文模型与LR、SVM等7种预测模型相比具有更高的精度和稳定性;与现有其他预测模型相比具有更高精度,更适用于患者术中低体温的辅助诊疗,给出的7个特征重要性排序更具针对性和效用性,为医护人员对全麻手术患者采取有效性预防措施提供了更好参考。下一步将结合病人术后低体温数据的采集和分析,深入研究术中和术后低体温的内在联系及其关键影响因素,提出更为有效的预防性对策建议。

### 参考文献:

- [1] 郭莉. 手术室护理实践指南[M]. 北京:人民卫生出版社,2020: 105-107.
- [2] ROWLEY B, KERR M, VAN P J, et al. Perioperative warming in surgical patients:A comparison of interventions[J]. *Clin Nurs Res*, 2015, 24(4): 432-441.
- [3] PEYJEN Y, HUGH A C, NINA K, et al. Impact of postoperative hypothermia on outcomes in coronary artery bypass surgery patients[J]. *J Cardiothorac Vasc Anesth*, 2017, 31(4): 1257-1261.
- [4] RULLI F, STEFANI M, TORBA M, et al. Intraoperative continuous intestinal loop warming technique a prospective randomised trial[J]. *Ann Ital Chir*, 2017, 88: 237-241.
- [5] 李娟, 隋丽娟, 于英, 等. 术中低温危害及防治研究进展[J]. *当代医学*, 2017, 23(16): 199-200.  
LI Juan, SUI Lijuan, YU Ying, et al. Research progress of intraoperative hypothermia hazards and prevention[J]. *Contemporary Medicine*, 2017, 23(16):199-200.
- [6] 张曼曼, 段红香, 周晓阳, 等. 降低神经外科全麻手术患者术中低体温发生率的PDCA实践[J]. *护理学报*, 2021, 28(12): 10-15.  
ZHANG Manman, DUAN Hongxiang, ZHOU Xiaoyang, et al. PDCA practice to reduce the incidence of intraoperative hypothermia in patients undergoing general anesthesia in neurosurgery[J]. *Journal of Nursing*, 2021, 28(12):10-15.

- [7] 阮丽诗,袁浩斌.术前预热式主动加温措施对控制病人核心体温及术中出血量效果的Meta分析[J].护理研究,2021,35(8):1330-1337.  
RUAN Lishi, YUAN Haobin. Meta-analysis of the effect of preoperative preheating on the control of patients' core body temperature and intraoperative blood loss[J]. CHINESE NURSING RESEARCH, 2021, 35(8):1330-1337.
- [8] 黄一乐,胡文娟.手术患者术中低体温危险因素评价指标体系的构建[J].中国实用护理杂志,2016,32(35):2727-2730.  
HUANG Yile, HU Wenjuan. Construction of risk factors evaluation index system for intraoperative hypothermia in patients[J]. Chinese Journal of Practical Nursing, 2016, 32(35):2727-2730.
- [9] YI J, ZHAN L, LEI Y, et al. Establishment and validation of prediction equation to estimate of intraoperative hypothermia in patients receiving general anesthesia[J].SciRep,2017,7(1):1379-1392.
- [10] 普鹰,张莹,汤佳骏,等.腹腔镜手术患者术中低体温预测模型的构建及应用[J].中华护理杂志,2019,54(9):1308-1312.  
PU Ying, ZHANG Ying, TANG Jiajun, et al. The construction and application of low temperature prediction model in laparoscopic surgery[J]. Chinese Journal of Nursing, 2019, 54(9):1308-1312.
- [11] 史卓颖,张海伟,杜祥飞.全身麻醉病人术中低体温发生预测模型的建立[J].护理学报,2021,35(2):246-249.  
SHI Zhuoying, ZHANG Haiwei, DU Xiangfei. Establishment of a predictive model for hypothermia during general anesthesia [J]. Journal of Nursing, 2021, 35(2):246-249.
- [12] 杨晶东,孟一飞,荀菡基,等.集成学习机制下的鼻炎辅助诊断模型[J].数据采集与处理,2021,36(4):684-696.  
YANG Jingdong, MENG Yifei, XUN Rongji, et al. An assisted diagnosis model of rhinitis based on integrated learning mechanism[J]. Journal of Data Acquisition and Processing, 2021, 36(4):684-696.
- [13] CHUNG J W, KIM W J, CHOI S B, et al. Screening for pre-diabetes using support vector machine model[C]//Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Washington DC, USA: IEEE Press, 2014: 2472-2475.
- [14] 张春富,王松,吴亚东,等.基于GA\_XGBoost模型的糖尿病风险预测[J].计算机工程,2020,46(3):315-320.  
ZHANG Chunfu, WANG Song, WU Yadong, et al. Diabetes risk prediction based on GA\_XGBoost model[J]. Computer Engineering, 2020, 46(3):315-320.
- [15] 岳鹏,侯凌燕,杨大利,等.基于XGBoost特征选择的疾病诊断XLC-Stacking方法[J].计算机工程与应用,2020,56(17):136-141.  
YUE Peng, HOU Lingyan, YANG Dali, et al. Feature selection for disease diagnosis XLC-Stacking methods[J]. Computer Engineering and Applications, 2020, 56(17):136-141.
- [16] ZHONG L, MA C Y, ZHANG H, et al. A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA- CG- SVM method[J].Computers in Biology & Medicine, 2011, 41(11): 1006-1013.
- [17] 雷雪梅,谢依彤.用于高血压菜谱识别的基于遗传算法的改进XGBoost模型[J].计算机科学,2018,45(6):476-481.  
LEI Xuemei, XIE Yitong. Improved XGBoost model based on Genetic Algorithm for hypertension recipe recognition[J]. Computer Science, 2018, 45(6):476-481.
- [18] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [19] 李雪柔,陈飞燕,林爱文,等.基于随机森林回归的茶园扩张驱动机制分析[J].生态与农村环境学报,2020,36(1):44-52.  
LI Xuerou, CHEN Feiyan, LIN Aiwen, et al. Driving mechanism analysis of tea plantation expansion based on random forest regression[J]. Journal of Ecology and Rural Environment, 2020, 36(1):44-52.
- [20] GENUER R, POGGI JM, TULEAU-MALOT C. VSURF: Variable selection using random forests[J]. Pattern Recognition Letters, 2016, 31(14): 2225-2236.
- [21] BELGIU M, LUCIAN D. Random forest in remote sensing: A review of applications and future directions[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24-31.
- [22] 靳志杰,霍志红,许昌,等.基于特征选择和XGBoost的风电机组故障诊断[J].可再生能源,2021,39(3):353-358.  
JIN Zhijie, HUO Zhihong, XU Chang, et al. Wind Turbine fault diagnosis based on Feature selection and XGBoost[J]. Renewable Energy, 2021, 39(3):353-358.

- [23] LI S, ZHANG X. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm[J]. *Neural Computing and Applications*, 2020, 32: 1971-1979.
- [24] 肖跃雷, 张云娇. 基于特征选择和超参数优化的恐怖袭击组织预测方法[J]. *计算机应用*, 2020, 40(8): 2262-2267.  
XIAO Yuelei, ZHANG Yunjiao. Prediction method of terrorist attack organization based on feature selection and hyperparameter optimization[J]. *Computer Application*, 2020, 40(8): 2262-2267.
- [25] 张又文, 冯斌, 陈页, 等. 基于遗传算法优化 XGBoost 的油浸式变压器故障诊断方法[J]. *电力自动化设备*, 2021, 41(2): 200-206.  
ZHANG Youwen, FENG Bin, CHEN Ye, et al. Fault diagnosis method of oil-immersed transformer optimized by XGBoost based on genetic algorithm[J]. *Power Automation Equipment*, 2021, 41(2): 200-206.
- [26] 严露培, 姚丽丽, 李跃荣, 等. 手术患者非计划性术中低体温风险预测模型研究进展[J]. *护理学报*, 2021, 28(6): 13-16.  
YAN Lupei, YAO Lili, LI Yuerong, et al. Research progress of risk prediction model for unplanned intraoperative hypothermia in surgical patients[J]. *Journal of Nursing*, 2021, 28(6): 13-16.

#### 作者简介:



曹立源(1996-),男,硕士研究生,研究方向:机器学习、数据挖掘等,E-mail:598284988@qq.com。



范勤勤(1986-),通信作者,男,副教授,博士生导师,研究方向:多目标优化、机器学习、进化计算等,E-mail:forever123fan@163.com。



黄敬英(1984-),女,副主任护师,研究方向:护理管理和教育、急危重症护理、围手术期护理等,E-mail:h.j.y@163.com。

(编辑:夏道家)