2019 年（第六届）全国大学生统计建模大赛

基于重采样与分类算法的永久性结肠造口危险因素探究

参赛单位：中南大学

参赛者姓名：国昕怡

李冰灵

邓牧野

目录

表格和插图清单

摘 要

目的： 研究结直肠癌手术治疗中需要进行永久性造口的概率，探究可能导致患者需做永久性造口的危险因素，为临床分析提供参考，降低还纳手术失败病人面临术后并发症的风险。

方法： 本文基于湘雅医院 2018-2019 年直肠癌手术 240 例病人的手术记录，剔除部分缺失及无效数据，最终得到 227 例有效样本点，其中永久性造口样本点 47 例，占比不足 20%，则样本点中多数类和少数类的比例差异较大，属于不平衡数据。一方面如果少数类样本数据过少会导致使用传统预测方法误差较大，另一方面传统的 SMOTE 算法只能对数值型数据进行重采样，而本文的数据集含有名义变量，故本文通过 Python3.7 软件，使用 SMOTE-NC 算法对数据进行平衡化处理。接着本文将数据集以 6:4 的比例分为训练集和测试集，并对训练集进行重采样，在原始训练集和重采样训练集上，分别建立基于 XGBoost 算法的分类模型，以及 L1 penalty Logistic 模型。

结果： 通过计算多个模型评价指标，发现 XGBoost 与 L1 penalty Logistic 在重采样之后，F-measure，G-mean,Recall 指标均有明显提升。可以认为使用重采样数据的模型在需要进行永久性造口的样本点预测上有了显著改进。考虑利用 XGBoost 的 F-score 得分排名靠前的变量进行变量筛选，以及 L1 penalty Logistic 给出的系数进行变量解释。本文认为（1）年龄（F-score:625，Logistic 系数：0.001952，两者均为重采样后结果，下文 Logistic 系数用 coef 代替），（2）肿瘤下缘距肛缘距离（F-score:462，coef: -0.09544），（3）腹腔是否给化疗药-0（即不给）（F-score:220，coef: 0.205214），（4）术前有无贫血-0（即非贫血）（F-score:205，coef: 0.021091）是更重要的变量，且年龄越大、肿瘤下缘距肛缘距离越近、腹腔未给化疗药、术前无贫血将会导致更高的需要永久性造口概率

结论： 在永久性结肠造口的研究中应当重点关注上述影响较大的指标，并以 XGBoost 等模型预测患者所需实施何种造口手术，减少还纳手术并发症的风险。

关键词：结肠回纳手术 SMOTE-NC XGBoost L1 penalty Logistic

一、绪论

（一）研究背景与意义

结肠直肠癌是全球最常见的三大癌症之一，患者死亡率仅居肺癌、乳腺癌之后，尤其在城市地区高发。Fanny 等人通过研究 1999-2016 年欧洲地区超过 1.4 亿人的数据发现，年龄在 20-29 岁的年轻人结肠癌发病增率长最为明显[14]。结肠造口术是直肠癌外科治疗的主要方式，肠造口即“人造肛门”，可分为永久性造口和临时性造口。做了临时性造口的患者在术后还需做结肠造口还纳手术，而结肠造口还纳手术并非是一个单纯的肠吻合手术，其术后并发症的发生率可高达 32.1%[6]，其中包括吻合口瘘，肠梗阻，术后感染，腹腔感染以及切口疝等，严重可危及患者生命。当患者还纳手术失败时，其临时性造口需要被改成永久性造口，这就致使该患者比一开始就做永久性造口手术后的病人要多面临造口还纳手术术后并发症的风险。因此，如果可在患者做结肠造口手术前就根据其相关身体指标以大概率预判出该患者是该做永久性造口还是可成功还纳的临时性造口，可降低那些可能还纳失败的病人面临还纳手术术后并发症的风险。故此，研究导致患者需要做永久性造口还是临时性造口这一区别的影响因素，以及利用模型对患者是否需要直接实施永久性造口手术给出预测概率参考将会是一个有价值的研究课题。

但在实际治疗过程中，直接做永久性造口的病例相较于临时性造口的病例，前者的样本点数目远少于后者。在统计学领域，一般把这种分布不均匀的数据集归结为不平衡数据集，在对不平衡数据集进行监督学习的时候，分类结果往往倾向于多数类，而忽视少数类，也就是传统的方法对样本点进行学习分类的时候，往往倾向于将样本点归为多数类(Negative 类)，而忽视少数类(Positive 类)。所以预测结果往往是总的分类准确率较高，而属于少数类样本点的正确分类效果很差，这就可能导致研究结果不具备实际意义，没有现实价值。

（二）文献综述

结肠造口手术是根据治疗的需要将近端结肠固定于腹壁，形成肠造口，粪便由此排出体外，故肠造口又称为“人造肛门”。造口可分为永久性和临时性两种。在 2017 年钟新强通过回顾性分析南昌大学第一附属医院自 2002 年至 2016 年

132 例结肠造口还纳术病人的临床资料后，得出结论：临时性结肠造口通常需要还纳，且还纳手术具有较高的术后并发症风险，其中切口感染是最常见的并发症[5]。 对于临时性造口，如若还纳手术失败或术后发生并发症，需要转做为永久性造口，对患者生理、心理都有一定影响，若能通过统计分析方法，在造口手术前根据患者的各项检查指标进行判断，选择合适的造口方式，具有重要的现实意义。

对于导致需要做永久性造口的危险因素，国内外的学者的研究方式主要基于假设检验的方式进行。如 2017 年 Xin Zhou 利用 PubMed、Embase 和 Cochrane Central Library 的数据库中的病例，运用 Meta-Analysis 方法，对影响永久性造口的危险因素进行了研究，认为较高的年龄、ASA 分级、并发症、手术方式、吻合口瘘、是否局部复发等是影响永久性造口的重要因素[16]。Wai Lun Law 等人针对 2000 年至 2014 年收集 314 例病例，利用 chi-squared test 和 Fisher’s exact test 以及 Mann - Whitney U test 进行分析，研究发现永久性造口患者是男性的概率大于女性，同时辅助放化疗等也是永久性造口的重要影响因素[12]。马得欣通过对 2008 年 6 月至 12 月实施 Miles 手术的 150 例患者的结肠造口资料进行实证分析后得出结论：直肠癌是消化系统常见的恶性肿瘤，其中 50%-60%需做永久性结肠造口手术[2]。2007 年 Marcel den Dulk 等人利用 COX Regression、Kaplan-Meier 等方法对造口还纳的逆转时间进行了研究，认为术前放射治疗明显降低了继发性造口逆转的可能性，但对原发性造口则没有影响；而年龄较大、继发性造口、术后并发症、复发等是造口逆转的主要影响因素[13]。Young Ah Kim 等人回归性分析了 2004 年 1 月至 2011 年 12 月 679 个直肠癌病例，其中临时性结肠造口 135 例，并得出结论：术后并发症如吻合口瘘、晚期原发疾病(第 4 期)、局部复发和共病是非逆转回肠造口的危险因素[10]。

（三）本文结构

本文将分为五章，第一章为绪论，第二章将对文章里涉及的主要知识做简单介绍，第三章将对 SMOTE-NC 重采样算法进行介绍，第四章介绍本文的主要模型：XGBoost 与 L1 penalty Logistic，第五章为实证分析，第六章为总结与展望。

二、预备知识

（一）结肠造口手术相关概念

1.结直肠癌

结直肠癌是常见的消化道肿瘤之一，其发病率在各肿瘤发病率中仅次于肺癌。据世界卫生组织国际癌症研究中心（International Agency for Research on Cancer, IARC）提供的资料显示，2012 年全世界约有 136 万余结肠癌新发病例，居恶性肿瘤第 3 位；死亡约 69 万例，居恶性肿瘤第 4 位[1]。结直肠癌以直肠癌和乙状结肠癌最为常见，其治疗可分为外科治疗放射治疗放疗综合治疗化学治疗免疫治疗等。目前对于结直肠癌患者，仍以外科手术为主要治疗方式。

2.结肠造口术

结肠造口术是指外科医生为了治疗如结直肠癌等肠道疾病而在腹壁上做一人为开口，并将一段肠管拉出开口外，翻转缝于腹壁的手术，目的是代替原来的会阴部肛门行使排便功能。这一人为开口被称为肠造口，也称“人造肛门”。

3.临时性造口

用于暂时通过造口将肠内容物排出体外，通过肠内容物的暂时性流转，可保护术后的远端吻合肠管得以休息和愈合，免受机械性损伤。

4.永久性造口

用于直肠以及全段或部分结肠切除术，这时肠道的延续性不能恢复，肠造口将永久用于代替肠内容物的输出。

5.结肠造口还纳术

结肠造口还纳术是在临时性造口手术后，待患者肠功能恢复，为解除患者的临时性造口所做的手术。如果原来的临时性造口手术比较顺利，还纳手术在造口手术几周后就可以进行，但复杂的造口其还纳时间一般要延迟到编年或 6 个月以上，其次还纳时间还取决于患者自身的身体情况。造口还纳手术有各种术后并发症，如伤口感染、裂开、吻合口瘘等，严重可危及患者生命。若患者还纳手术失败，临时性造口将被改成永久性造口。

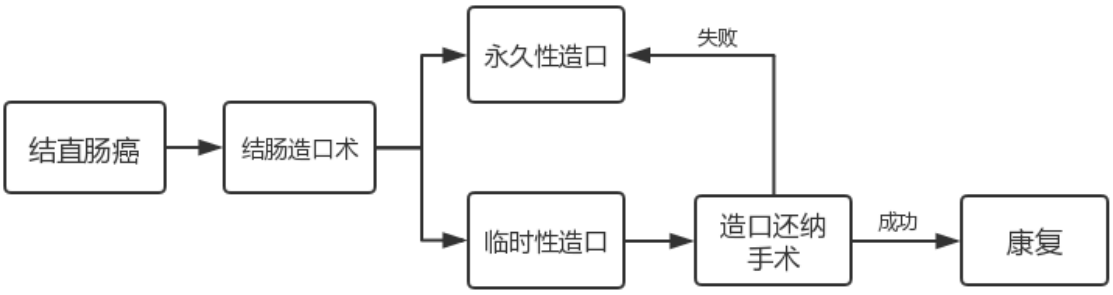

图 1　结直肠癌治疗简易流程图

（二）统计模型相关知识

1. 混淆矩阵

混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用 n 行 n 列的矩阵表示。其中，每一列代表了一种预测类别，每一行代表了数据的真实归属类别，矩阵中的每一个数值表示该数值所在行对应的真实类别被预测为所在列对应预测类的数据数目。

2. 模型评价指标

对于二分类问题，模型好坏的基础指标一般由一个混淆矩阵给出，如表所示：

表 1　混淆矩阵

	预测为负类	预测为正类
真实为负类	TN	FP
实际为正类	FN	TP

TN 表示真实为负类且被分为负类的数据个数；FP 表示真实为负类且被分为正类的数据个数；FN 表示实际为正类且被分为负类的数据个数；TP 表示实际为正类且被分为正类的数据个数。

由此，我们引入了如下指标：

① 误分率：反映在样本中有多少样本点被错误识别

$$Error = \frac{FP + FN}{TP + TN + FN + FP} \tag{2-1}$$

② 精确率：反映所有被预测为正类的样本点中实际为正类的样本点所占的比例

$$Precision = \frac{TP}{TP + FP} \tag{2-2}$$

③ 假正率：反映所有负类样本点中被预测为正类的样本点所占的比例，也记作FPR。

$$FPR = \frac{FP}{FP + TN} \tag{2-3}$$

④ 真正率：真正率即为召回率，反映所有正类样本点中被预测为正类的样本点所占的比例，也记作Recall，TPR。

$$TPR = \frac{TP}{TP + FN} \tag{2-4}$$

⑤ F-measure ：F-measure综合考虑Precision、Recall指标，即被预测为正类的样本点中有多少真实类别为正类，以及有多少真实正类样本点被预测为正类。其中β是用来调整Precision和Recall两个指标相对重要性的系数，若无特别说明，若无特别说明，一般β=1，本文中F-measure指标中的β=1

$$\frac{(1+\beta^2)\times recall \times precision}{\beta^2 \times recall + precision} \tag{2-5}$$

⑥ G-mean ：G-mean指标综合考虑TPR和FPR两种指标，原理与F-measure类似。

$$G-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{2-6}$$

⑦ AUC值

AUC（Area Under Curve）值是一种判别ROC曲线所对应模型好坏的指标，其值定义为ROC曲线下的面积，而ROC曲线名为接受者操作特性曲线（Receiver Operating Characteristic Curve），该曲线上各点反应了在不同判定标准下，对同一信号刺激的反应，是用来反应二值分类器分类效果的曲线。对于一个二分类问题，可将样本分为正类和负类，一般，正类对应少数类，标签为1，负类对应多数类，标签为0。而AUC值取值在0—1之间。对于一般的模型，ROC曲线一般在对角线上方，故AUC取值一般在0.5—1之间，越高越好。

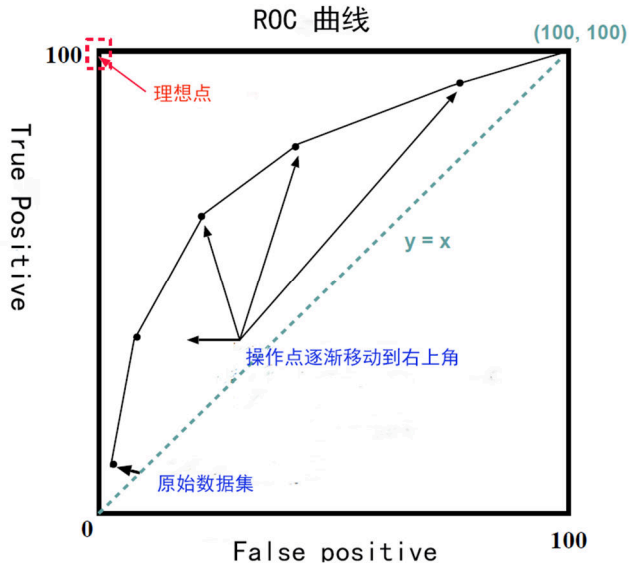

图 2　ROC 曲线

3. 不平衡数据集

数据不平衡是分类任务中典型存在的问题，是指在一个数据集中，每个类别下的样本数目差别很大。例如在一个二分类问题中，样本点共有 100 个，其中 80 样本点个属于类别 1,20 个属于类别 2，这便是一个不平衡数据集。常用的分类方法大多倾向于把样本数较少类别的样本错误地分到样本数较多的类别中，但在现实问题中，正确识别出小类样本往往更为重要，尤其在医学领域[4]。

三、SMOTE-NC 重采样

（一）SMOTE 算法介绍

SMOTE-NC 是 SMOTE 算法的改进，在讨论 SMOTE-NC 算法时候，SMOTE 算法是绕不开的。首先将给出 SMOTE 算法的原理介绍。

SMOTE 算法基于通过运用一定的方法进行“插值”来为少数类合成新的样本点的思想，是处理不平衡数据集较为常用的数据处理手段，在 2002 年由 Nitesh V. Chawl[8]等人提出，其原理是对于每个少数类样本点，计算其 K 近邻，在 K 近邻中随机选取 M 近邻，计算这 M 近邻与原始点的差值，通过线性插值的方法，合成新的少数类样本点，并不断重复以上过程，直至达到目标的样本之间的比例为止。具体步骤如下所示。

步骤 1：给定样本特征集合 $X=\{x_1,x_2,...,x_n\}$，其中多数类为 $N=\{x_1,x_2,...,x_{nnum}\}$，少数类记作 $P=\{x_1,x_2,...,x_{pnum}\}$，初始化给定一个 K 近邻的计算数 K（一般情况下默认 K=5）。

步骤 2：对于每个少数类的样本计算其 K 近邻，假设 x_i 的 K 近邻为，$\{x_{k1},x_{k2},...x_{kk}\}$，得到 K 近邻之后，跳转至下一步。

步骤 3：对于其 K 近邻，随机选择一个 M，且 M 满足条件 $M|0<M\le K$，对于其 K 近邻，再从中随机抽取 M 个样本点 $\{x'_{k1},x'_{k2},...x'_{kM}\}$。

步骤 4：对于得到的 M 个样本点，分别计算其与原始点之间的距离，得到距离 dif。

$$dif_{kj}=x'_{kj}-x_i \tag{3-1}$$

步骤 5:随机生成一个 0-1 之间的随机数 α,对于每一个计算出来的距离 dif，用 α 乘以 dif 并加至原始点 x_i 上，就得到了新的合成的样本点。

$$x_{new}=dif\times\alpha+x_i \tag{3-2}$$

步骤 6：重复以上步骤，直至达到设定的目标比例。

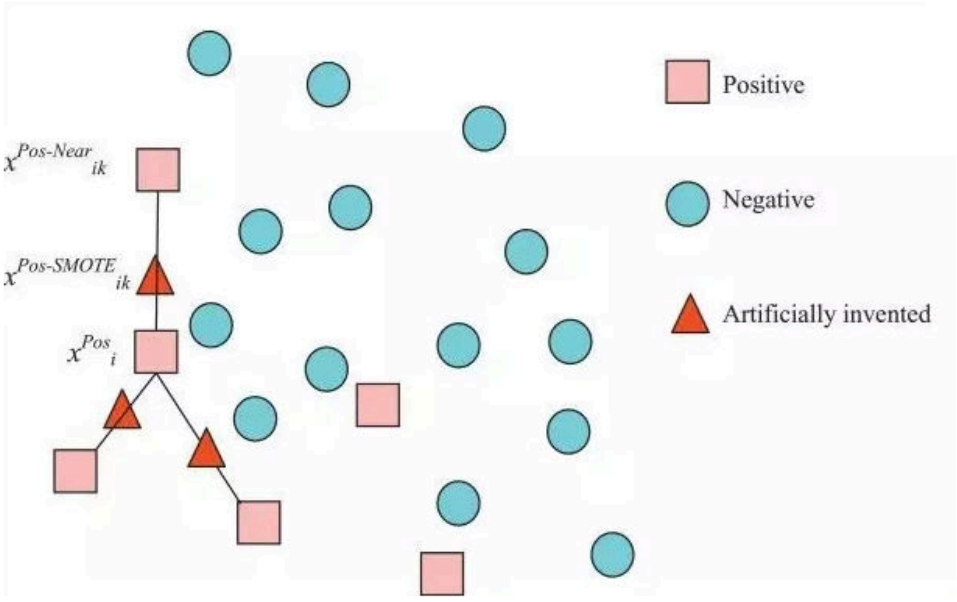

图 3　SMOTE 算法示意图

（二）SMOTE-NC 算法介绍

由上节 SMOTE 算法介绍我们可以发现，SMOTE 算法不适用于名义变量，特别是在计算距离时，传统的 SMOTE 算法失去了应对的能力，故 Nitesh V. Chawla 在文中也给出了一种基于 SMOTE 的适用于 Nominal & Continues 变量的重采样算法，故称之为 SMOTE-NC 算法。

SMOTE 算法基本思想与 SMOTE 算法完全一致，但由于引入了名义变量，故其在计算距离时候，并非简单的计算欧氏距离。

SMOTE-NC 算法创新之处描述如下：

1） 中值计算：计算少数类的所有连续型特征变量的标准差的中值，如果样本点的名义变量与其潜在的最近邻取值不同，那么在计算两点间的欧氏距离的时候，计算的中值将被考虑进欧式距离中。

2） 最近邻的计算：在连续特征空间中计算识别少数类样本与其他少数类样本的 K 近邻的时候，将包含上述计算的标准差的中值。

如给定两个的样本点 F1=1 2 3 A B C，以及 F2=4 5 6 A D E,F2 与 F1 之间的距离将会是 $[(4-1)^2+(6-2)^2+(5-3)^2+Med^2+Med^2]$。其中 Med 为标准差的中值。

四、XGBoost 与 L1 penalty Logistic 模型介绍

（一）L1 penaity Logistic 介绍

线性回归模型是一种流行的定量分析因变量与自变量之间相关关系的统计分析方法。然而在许多情况下，线性回归都会受到限制，如，当因变量是分类变量而不是连续变量时，线性回归就不适用。在定量分析分类变量时，常用的一种统计方法是对数线性模型（Log-linear model），Logistic 回归模型，是对数线性模型的一种特殊形式[11]。

现假设有一个理论上存在的连续随机变量 y_i^* 代表事件发生的可能性，其值域为 $(-\infty,\infty)$。当该变量的值跨越一个临界点 c 时，便导致事件发生了，即 $y_i=1$；否则 $y_i=0$，这里 y_i 是实际观测到的因变量取值。假设随机变量 y_i^* 和自变量 x_i 之间存在线性关系，即:

$$y_i^*=\alpha+\beta x_i+\varepsilon_i$$

于是，事件发生的概率为:

$$p(y_i=1|x_i)=p(\alpha+\beta x_i+\varepsilon_i>c)=p(\varepsilon_i>-\alpha-\beta x_i+c)$$

通常假设 ε_i 服从 Logistic 分布，根据 Logistic 分布的对称性，则有：

$$p(\varepsilon_i>-\alpha-\beta x_i+c)=p(\varepsilon_i\le\alpha+\beta x_i+c)=F(\alpha+\beta x_i+c)$$

其中，F 为 ε_i 的累积分布函数，即 Logistic 分布的累积分布函数。又不失一般性，可以假设 c=0，因此有：

$$p(y_i=1|x_i)=p(\varepsilon_i\le\alpha+\beta x_i)=\frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

无论 x_i 取任何值，事件发生的条件概率 $p(y_i=1|x_i)$ 的取值范围均在 0 至 1 之间。

若将事件发生的条件概率 $p(y_i=1|x_i)$ 记为 p_i，则 p_i 为第 i 个观测发生事件的条件概率，于是可得到如下 Logistic 回归模型：

$$p_i = \frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

这是一个由自变量x_i构成的非线性函数。

当有 k 个自变量时，Logistic 回归模型可以扩展如下：

$$p_i = \frac{1}{1+e^{-(\alpha+\sum_{m=1}^{k}\beta_m x_{mi})}} \quad (4\text{-}1)$$

这就是有 k 个自变量构造的 Logistic 函数。

接下来是对 Logistic 模型的参数估计，对于一般的 Logistic 模型来说，其损失函数的形式如下所示。

$$\sum_{i=1}^{n}\log(\exp(-y_i(X_i^T w+c))+1) \quad (4\text{-}2)$$

但是传统的 Logistic 模型对于变量维数过高的数据集处理效果很差，因为其并不具有选变量的功能，这将导致最后得出的模型系数可能出现没有意义或者方差过大的情形。故本文引入了带有一范数惩罚的 Logistic 模型，而带有一范数惩罚的 Logistic 模型能够很好的将部分参数系数严格的收束为 0，也就起到了选变量的功能。其损失函数如下所示。

$$\|w\|_1 + \sum_{i=1}^{n}\log(\exp(-y_i(X_i^T w+c))+1) \quad (4\text{-}3)$$

在线性回归中，对未知参数的估计采用的是最小二乘法，这一方法的原理是选择合适的参数估计值使得观测值与预测值之间的残差平方和达到最小。在线性回归中，极大似然估计与最小二乘估计得到的结果是一样的。与最小二乘法相比，极大似然估计法还可以用于非线性模型的参数估计。由于 Logistic 回归是非线性模型，因此，选择参数的极大似然估计值。

假设有 n 个观测作为样本，观测值$y_1, y_2, ..., y_n$，由于各观测是相互独立的，可以得到它们的联合概率分布，也是似然函数：

$$L(\theta) = \Pi p_i^{y_i}(1-p_i)^{1-y_i}$$

在得到带一范数惩罚的 Logistic 回归的似然函数后，采用坐标下降的方法即可迭代求解系数的最优解。

（二）XGBoost 介绍

Boosting 是一种基于弱分类器的监督学习算法,属于数据分类算法的一种。其思想是把多个弱分类器组合成一个强分类器。我们给它提供不同的训练集，它都能从数据中推断出一些新的东西。因此，每次都会得到不同的结果，将这些结果进行组合，得到一个好的最终模型。这样可以通过弱学习器的迭代达到优化模型的目的[7]。

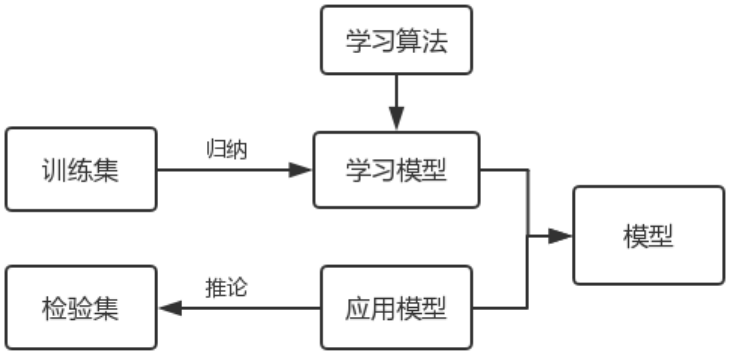

图 4　学习算法示意图

XGBoost 也是一种数据分类算法，属于一种迭代决策树算法[3]，在 boosting 基础上进行改进，所用到的树模型是 CART 回归树模型。

1.CART 回归树模型

对于一个训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，可以建立如下回归树模型[9]：

$$f(x)=\sum_{m=1}^{M}c_m I(x\in R_m)$$

其中 $R_m(m=1,\cdots,M)$ 表示回归树模型将输入空间划分成的第 m 个单元，总共有 M 个单元；c_m 代表每个单位的输出值。

预测误差可以用平方误差 $\sum_{x_i\in R_m}(y_i-f(x_i))^2$ 来表示，通过最小化平方误差得到每个单元上的最优输出值

一棵 CART 树包括两部分，回归树的整体结构 $f_k(x)$ 和各个叶子节点值 $w_q(x)$。其中，树的结构 $f_k(x)$ 是一个关于输入样本的函数，将样本映射到某一个叶子节点，k 表示第 K 棵树。而 $w_q(x)$ 表示序号为 q(x)的叶子节点值。容易得到：

$$f_k(x)=w_q(x) \tag{4-4}$$

XGBoost 包含大量的 CART 回归树，每棵树的复杂度定义为[15]：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2$$

其中 T 为叶子节点的个数，系数 γ 可以控制叶子结点的个数，$\|w\|$ 为叶子节点向量的模。λ 表示 L2 正则化系数，避免过拟合。

2.目标函数

XGBoost 对应的模型由多棵 CART 树组成，因此模型的目标函数可以写为：

$$L^{(t)}=\sum_i^n L(y_i,\hat{y}_i)+\sum_{k=1}^t\Omega(f_k)$$

即目标函数由两部分构成，第一部分是模型的训练误差，第二部分是正则化项，是 t 棵树的正则化项相加。

显然有

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t(x_i)$$

因此目标函数改写成：

$$L^{(t)}=\sum_i^n L(y_i,\hat{y}_t^{(t-1)}+f_t(x_i))+\sum_{k=1}^t\Omega(f_k)$$

$$=\sum_{i=1}^n L\left(y_i,\hat{y}_i^{(t-1)}+f_t\left(x_i\right)\right)+\sum_{k=1}^{t-1}\Omega\left(f_k\right)+\Omega\left(f_t\right)$$

由于前 t-1 棵树的复杂度之和为常数，为了简便将常数省略，目标函数化简为

$$L^{(t)}=\sum_i^n L(y_i,\hat{y}_t^{(t-1)}+f_t(x_i))+\Omega(f_t)$$

为了最小化目标函数，XGBoost 算法采用了在$f_t=0$处的泰勒二阶展开来近似，目标函数化为：

$$L^{(t)}=\sum_{i=1}^{n}\left[L\left(y_i,\hat{y}_i^{(t-1)}\right)+g_i f_t\left(x_i\right)+\frac{1}{2}h_i f_t^2\left(x_i\right)\right]+\Omega\left(f_t\right)$$

其中g_i为一阶导数，h_i为二阶导数：

$$g_i=\partial_{\hat{y}_i^{(t-1)}}L\left(y_i,\hat{y}_i^{(t-1)}\right)\quad h_i=\partial^2_{\hat{y}_i^{(t-1)}}L\left(y_i,\hat{y}_i^{(t-1)}\right)$$

因为$L\left(y_i,\hat{y}_i^{(t-1)}\right)$不影响目标函数的优化，将其省略，目标函数进一步化简为：

$$L^{(t)}=\sum_{i=1}^{n}\left[g_i f_t\left(x_i\right)+\frac{1}{2}h_i f_t^2\left(x_i\right)\right]+\Omega\left(f_t\right)$$

将（4-4）式代入有

$$L^{(t)}=\sum_{i=1}^{n}\left[g_i w_{q(x_i)}+\frac{1}{2}h_i w_{q(x_i)}^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$

$$=\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma T$$

$$=\sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}\left(H_j+\lambda\right)w_j^2\right]+\gamma T \qquad (4\text{-}5)$$

其中$\sum_{i\in I_j}g_i=G_j$，$\sum_{i\in I_j}h_i=H_j$。

式（4-5）可以看成是关于叶子节点的一元二次函数。最小化上式，得：

$$w_j^*=-\frac{G_j}{H_j+\lambda}$$

得到最终的目标函数：

$$L^*=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \qquad (4\text{-}6)$$

式（4-6）称为得分函数，它是衡量回归树模型拟合程度高低的标准，值越小，代表模型构建越好。

XGBoost 利用（4-6）式目标函数值作为评价函数。分裂后的目标函数值比单子叶子节点的目标函数的增益如下：

$$Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma$$

这里为了限制树生长过深，出现过拟合现象，添加阈值 γ，只有当增益大于该阈值才进行分裂。

3.近似算法

当样本数量非常大，某些特征取值过多时，如果遍历所有取值会花费很多时间，造成效率低下且容易出现过拟合。

XGBoost 想法是对特征进行分组，即找出分位点，将相邻两个分位点之间的样本分为一组，这样可以避免遍历所有取值，而只需要对各个分位点进行遍历，从而提高计算效率。该算法分为两种，全局近似和局部近似，前者是在建立模型前就计算出分位点，对样本进行事先划分，接下来每次分裂均为近似划分，而后者则是在某一次具体的分裂过程中采用近似算法。

五、基于重采样与分类算法的的模型表现分析

（一）数据介绍

本文的数据来源为湘雅医院 2018-2019 年直肠癌手术记录（本数据集只做研究，禁止外传），对名义变量数值化替换，剔除部分缺失数据以及无效数据，并且利用 Python 中的 category_encoders 包中的 one_hot 方法对名义变量进行编码，在数据预处理之后得到的样本点数目为 227 例，特征数目为 35（未编码前为 17），及编码后的变量对照表如表 2 所给出。完整的变量表由附件给出。

表 2　编码后变量对照表

编码后变量对照		
f0:主诊断-0	f12:有无腹部手术史-1	f24:ASA 分级-2
f1:主诊断-1	f13:有无吸烟史-0	f25:主要手术方式-0
f2:年龄	f14:有无吸烟史-1	f26:主要手术方式-1
f3:性别-0	f15:有无家族-0	f27:主要手术方式-2
f4:性别-1	f16:有无家族-1	f28:腹腔是否给化疗药-0
f5:术前有无转移-0	f17:术前有无新辅助放化疗-0	f29:腹腔是否给化疗药-1
f6:术前有无转移-1	f18:术前有无新辅助放化疗-1	f30:肿瘤位于腹膜反折-0
f7:术前有无高血压-0	f19:术前有无贫血-0	f31:肿瘤位于腹膜反折-1
f8:术前有无高血压-1	f20:术前有无贫血-1	f32:肿瘤位于腹膜反折-2
f9:术前有无糖尿病-0	f21: 肿瘤下缘距肛缘距离	f33:造口方式-0
f10:术前有无糖尿病-1	f22:ASA 分级-0	f34:造口方式-1
f11:有无腹部手术史-0	f23:ASA 分级-1	f35:造口方式-2

本文在重采样与非重采样的条件下，并在实验中，通过设定相同的随机数种子，按照训练集与测试集 6：4 的比例，利用重采样和未重采样的训练集分别构建 Logistic 模型与 XGBoost 模型共 4 个模型，并以测试集验证模型，本文中的所有结果均为测试集结果。

表 3　数据特征汇总表

特征	类别	特征	类别
主诊断	名义变量	年龄	数值变量

性别	名义变量	术前有无转移	名义变量
术前有无高血压	名义变量	术前有无糖尿病	名义变量
有无腹部手术史	名义变量	有无吸烟史	名义变量
有无家族史	名义变量	术前有无新辅助放化疗	名义变量
术前有无贫血	名义变量	肿瘤下缘距肛缘距离	数值变量
ASA 分级	名义变量	主要手术方式	名义变量
腹腔是否给化疗药	名义变量	肿瘤位于腹膜反折	名义变量
造口方式	名义变量		

（二）SMOTE-NC 重采样

在数据预处理之后，由于原始的数据集数据极度不平衡，故考虑对其进行 SMOTE-NC 算法重采样，原始数据中，永久性造口只有 47 例样本点，仅占所有所有样本点的 20%，数据分布不平衡，如果直接对其进行分类的话，对于少数类的分类将会偏差很大，是一个难以接受的情况。由于本文的数据集含有名义变量，而传统的 SMOTE 算法只能对数值型数据进行重采样，故本文使用 SMOTE-NC 算法对数据集进行重采样。并且将最终的少数类目标比例设置为 1：1。

表 4　原始数据样本点分布情况

类别	样本量	标签	比例
永久性造口	47	1	0.206
临时性造口	181	0	0.793
合计	228	\	1

并且本文将训练集和测试集整体按照 6：4 的比例分开，得到的未重采样的训练集中永久性造口数目为 26，临时性造口数目为 110，总样本点数目为 136，而重采样之后的样本点数目为 220，其中临时性造口与永久性造口均为 110 例。

表 5　原始数据样本点分布情况

类别	样本量	标签	比例
永久性造口（重采样）	110	1	0.5
临时性造口（重采样）	110	0	0.5
永久性造口（原始）	26	1	0.191
临时性造口（原始）	110	0	0.809

（三）XGBoost 算法分类及变量筛选

本文利用 Python3.7 对 XGBoost 算法进行编写，代码为调用 Python 的开源代码 xgboost 包，重要性评分由 plot_importance 包给出，本文给出 XGBoost 在重采样与未重采样的数据集上的分类结果与模型评价如图 5 和表 6 所示。表中加粗的为该指标的最优值。

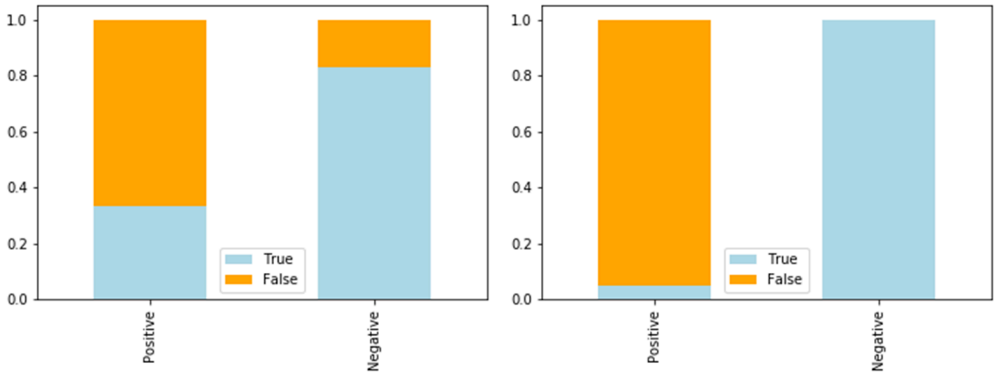

图 5　XGboost 分类结果可视化

表 6　XGBoost 模型评价

指标	重采样	未重采样
G-mean	**0.5255**	0.2182
F-measure	**0.3500**	0.0909
Recall	**0.3333**	0.0048
FPR	0.1714	**0.0000**
AUC	0.5782	**0.5809**

下面针对部分指标进行分析：

① G-mean 与 F-measure：针对 G-mean 与 F-measure 指标，重采样之后模型效果分别比未重采样提高了 140%和 288%，可以认为重采样之后的模型是远好于未重采样模型的整体分类效果的。

② FPR：FPR 反映了原本是多数类的样本被预测为少数类的比例，从中可见，重采样之后的算法不如原始算法，但这是符合逻辑的，并且也是在可以接受的范围内。

③ Recall：Recall 是原本是少数类的样本，有多大的比例被分类成了少数类样本，这是最值得关注的指标，因为算法的首要目标就是提高少数类样本的分类准

确率。重采样之后的模型效果从未重采样模型的 0.005 提高到了 0.333，从中可见重采样达到了预想的目的。

④ AUC：AUC 值是评价模型最重要的一个指标，而重采样之后的模型 AUC 值虽比未重采样略低，但几乎无差别。

通过上述实证分析，我们取 G-mean、F-measure 和 Recall 三个指标值的算数平均作为重采样算法对模型效果的提高程度，即重采样之后的模型效果从未重采样模型的 0.1046 提高到了 0.4029，可以发现，在进行重采样之后，模型相对于未进行重采样的模型有着显著的改进，下面给出未进行重采样和进行重采样之后的模型进行对比分析。

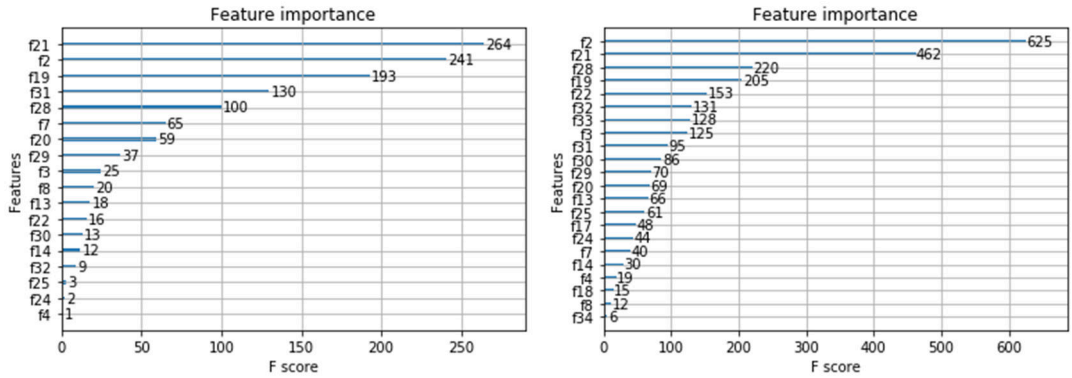

图 6　XGBoost 变量重要性排名

对于原始数据和重采样数据构建的 XGBoost 模型，筛选出来的最重要的五个变量分别如下表所示：

表 7　未重采样和重采样情况下的变量排名

排名	未重采样	重采样
1	肿瘤下缘距肛缘距离	年龄
2	年龄	肿瘤下缘距肛缘距离
3	术前有无贫血-0	腹腔是否给化疗药-0
4	肿瘤位于腹膜反折-1	术前有无贫血-0
5	腹腔是否给化疗药-0	ASA 分级-0

可以发现两者大致的排名趋势是一致的，但细节上有少许区别。

1.L1 penalty Logistic 算法分类及变量筛选

本文利用 Python3.7 对 L1 penalty Logistic 算法进行编写，代码为调用 Python 的开源代码 sklearn.linear_model 包，本文给出 L1 penalty Logistic

在重采样与未重采样的数据集上的分类结果与模型评价如图 7 和表 8 所示。表中加粗的为该指标的最优值。

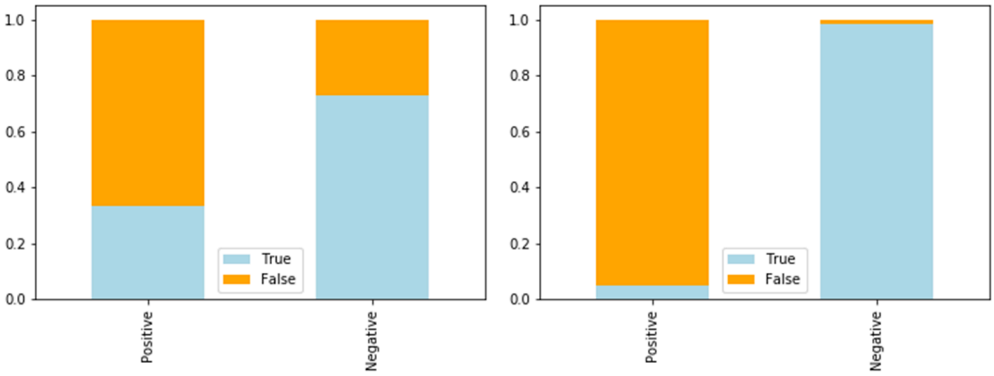

图 7　L1 penalty Logistic 分类结果可视化

表 8　L1 penalty Logistic 模型评价

指标	重采样	未重采样
G-mean	**0.4928**	0.2167
F-measure	**0.2979**	0.0869
Recall	**0.3333**	0.0476
FPR	0.2714	**0.0143**
AUC	0.4571	**0.5626**

具体的指标分析结果与 XGBoost 基本一致，重采样之后的模型相对于未重采样的模型更好，这里不再具体分析。

对于原始数据构建的 L1 penalty Logistic 模型，筛选出来系数不为零的变量及其系数如下表所示。

表 9　原始数据参数系数表

变量名	系数	变量名	系数
主诊断-1	-0.34216	年龄	0.001256
性别-0	-0.23251	术前有无转移-0	0.29204
术前有无转移-1	-0.26582	术前有无高血压-0	-0.36229
术前有无高血压-1	0.090531	术前有无糖尿病-1	-1.09777
有无吸烟史-1	-0.25235	有无家族史-1	-1.37825
术前有无新辅助放化疗-1	1.61042	术前有无贫血-0	0.095068
术前有无贫血-1	-1.793	肿瘤下缘距肛缘距离	0.000429
ASA 分级-0	-0.17902	主要手术方式-1	-0.57779

腹腔是否给化疗药-0	0.569447	腹腔是否给化疗药-1	-0.43109
肿瘤位于腹膜反折-1	0.181229	肿瘤位于腹膜反折-2	-1.35185
造口方式-0	1.549	造口方式-1	-0.02818

对于重采样数据构建的L1 penalty Logistic模型，筛选出来系数不为零的变量及其系数如下表所示：

表10　重采样数据参数系数表

变量名	系数	变量名	系数
年龄	0.001952	术前有无高血压-0	-0.56297
术前有无高血压-1	0.05458	有无吸烟史-0	0.054523
有无家族史-1	-0.1621	术前有无新辅助放化疗-1	-0.12402
术前有无贫血-0	0.021091	术前有无贫血-1	-1.2728
肿瘤下缘距肛缘距离	-0.09544	主要手术方式-2	0.743917
腹腔是否给化疗药-0	0.205214	腹腔是否给化疗药-1	-0.65165
肿瘤位于腹膜反折-1	0.708528	肿瘤位于腹膜反折-2	0.312779
造口方式-0	0.461258		

2.结论

本文结合XGBoost筛选结果，以及L1 penalty Logistic给出的参数系数，来共同分析可能的危险因素，以及可能的影响方向。通过分析上述四个模型给出的结果，本文认为以下变量是在严谨的医学实证分析中值得关注的因素。（仅分析XGBoost选出来的重要性前五的且L1 penalty Logistic给出的系数不为0的变量，其余变量由于其可能并不具有显著意义，以及篇幅限制，不作过多分析）

1) 年龄：在XGBoost模型构建的过程中，年龄变量对于分类结果的准确性起到了至关重要的作用，同时通过 Logistic 模型也可以发现，未重采样数据构建的模型中年龄系数为 0.001256，重采样之后为 0.001952，即两个模型都认为年龄对需要做永久性造口手术概率的影响是正向的，也就是年龄较大则更加推荐做永久性造口。这可能归结为年龄越大，身体的恢复能力也就越差，临时造口的恢复能力也就越不理想。

2) 肿瘤下缘距肛缘距离：首先，利用XGBoost模型构建，肿瘤下缘距肛缘距离被认为是对分类准确率影响较高的变量，但是对于肿瘤下缘距肛缘距离如何影响是否进行永久性造口，未重采样的Logistic与重采样之后的Logistic给出了不同的影响方向，前者系数为 0.00043，后者为-0.09544，但是模型

评价表明，重采样之后的 Logistic 模型更加合理，故本文认为可能肿瘤下缘距肛缘距离越近，越可能就进行永久性造口。

3) 腹腔是否给化疗药：腹腔是否给化疗药也是筛选出来的重要的变量。重采样之后腹腔给化疗药对应系数为-0.65165，可认为腹腔给化疗药将降低需要永久性造口的概率，即腹腔给化疗药能够在一定程度上促进治疗效果，让伤口能够更好的恢复，也就导致了需要永久性造口的概率降低。

4) 术前有无贫血：术前有无贫血同为筛选出来的重要影响因素，重采样之后术前无贫血对应系数为 0.0211，即术前贫血将会降低永久性造口的概率。

六、总结与展望

本文针对结肠直肠癌后续的可能手术处理方式进行了统计学上的研究，对于是否需要直接进行永久性造口手术进行模型构建，并且针对永久性造口样本点比例过少的情况，利用 SMOTE-NC 方法进行了重采样处理，这是当前学术界还几乎没有的处理方式。

同时本文并非直接对指标进行学术界流行的方差分析，而是通过构建分类模型，定量的给出哪些因素影响更高，以及给出影响因素的排名，并且结合 L1 penatly Logistic 模型进行系数的解释；而传统的方差分析，只能一刀切的给出是否显著影响，并不能做过多的解释。

尽管本文的研究并不能直接给出基于统计学意义上 p 值的、通过假设检验的具有显著影响的指标，这是未来需要继续研究的，但是本文所应用的方法，仍不失为对临床研究提供参考与研究方向上的建议。

参考文献

引文文献

[1] 李道娟，李倩，贺宇彤．结直肠癌流行病学趋势[J]．肿瘤防治研究，2015,42(3):305-310.

[2] 马得欣.150 例直肠癌患者术后结肠造口护理体会[J].中华结直肠疾病电子杂志,2012,1(02):31-33.

[3] 苏天培.基于 XGBoost 的糖尿病风险预测[J].科技视界,2019,(2):155-156.

[4] 王瑞． 针对类别不平衡和代价敏感分类问题的特征选择和分类算法[D].中国科学技术大学,2013.

[5] 钟新强．结肠造口还纳术 132 例临床分析[D].南昌大学,2017:1-15.

[6] 丁俊涛． 肠造口还纳手术临床分析[A]． 中国中西医结合学会.第十三届全国中西医结合大肠肛门病学术会议暨第三届国际结直肠外科论坛论文汇编[C].中国中西医结合学会:中国中西医结合学会,2009:2:119-120.

阅读型文献

[7] Anabel Gómez-Ríos, Julián Luengo, Herrera F . A Study on the Noise Label Influence in Boosting Algorithms: AdaBoost, GBM and XGBoost[J]. 2017:269-278.

[8] Chawla N V , Bowyer K W , Hall L O , et al. SMOTE: Synthetic Minority Over-sampling Technique[J].Journal of Artificial Intelligence Research, 2011, 16(1):321-357.

[9] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[J]. knowledge discovery and data mining, 2016: 785-794.

[10] Kim Y A , Lee G J , Park S W , et al. Multivariate Analysis of Risk Factors Associated With the Nonreversal Ileostomy Following Sphincter-Preserving Surgery for Rectal Cancer[J]. 2015:98-102.

[11] Liang Y , Liu C , Luan X Z , et al. Sparse logistic regression with a L1/2penalty for gene selection in cancer classification[J].BMC.Bioinformatics,2013,14(1):198.1-12.

[12] Mak J C K , Foo D C C , Wei R , et al. Sphincter-Preserving Surgery for Low Rectal Cancers: Incidence and Risk Factors for Permanent Stoma[J]. World Journal of Surgery, 2017:2912-2922.

[13] Marcel D D , Marije S , Peeters K C M J , et al. A multivariate analysis of limiting factors for stoma reversal in patients with rectal cancer entered into the total mesorectal excision (TME) trial: a retrospective study[J]. Lancet Oncology, 2007, 8(4):297-303.

[14] Vuik F E, Nieuwenburg S A, Bardou M, et al. Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years[J]. Gut, 2019:1-7.

[15] Zhang D, Qian L, Mao B, et al. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost[J]. IEEE Access, 2018: 21020-21031.

[16] Zhou X , Wang B , Li F , et al. Risk Factors Associated With Nonclosure of Defunctioning Stomas After Sphincter-Preserving Low Anterior Resection of Rectal Cancer[J]. Diseases of the Colon & Rectum, 2017, 60(5):544-554.

致谢

衷心感谢各位教授及专家在本论文的完成过程中提供的无私帮助与指导。感谢中南大学数学与统计学院刘心歌教授在论文选题、写作过程中始终对我们高标准、严要求，给予我们团队的热忱帮助和鼎力支持！

同时，还要衷心感谢中南大学湘雅医院日间手术中心在研究数据收集过程中给予我们的大力支持与无私帮助！