通过模型复杂度驱动的类别比例调节，提升类别不平衡数据的临床预测建模：阿片类药物过量预测的实证研究

Yinan Liu[1], Xinyu Dong[1], Weimin Lyu[1], 硕士，Richard N. Rosenthal[1], 医学博士，Rachel Wong[1],医学博士，Tengfei Ma[2], 博士，Fusheng Wang[1], 博士[1]石溪大学，美国纽约州石溪[2]IBM Research，美国纽约州约克敦高地

摘要

类别不平衡问题广泛存在于医学领域，并严重降低临床预测模型的性能。大多数缓解这一问题的技术都会重新平衡类别比例，而且通常假设重新平衡后的比例应由原始数据决定，而不考虑所用的模型。本研究挑战了这一主流假设，提出将最优类别比例与模型复杂度关联起来，从而针对每个模型调节类别比例。我们针对阿片类药物过量预测问题开展的实验凸显了调节类别比例所带来的性能提升。严谨的回归分析也证实了所提出理论框架的优势，并表明控制模型复杂度的超参数与最优类别比例之间存在统计显著的相关性。

引言

类别不平衡问题在医学机器学习中很常见，因为“正类”患者通常只占总人群的一小部分。例如，在 Health Facts 数据库中，阿片类药物过量的正例率低于 1%，阿片类药物使用障碍约为 2%，急性肾损伤则介于 3% 至 20% 之间。[22, 23, 24] 在这些情况下，我们可以使用 F1 分数、精确率和召回率等指标来评估模型性能，因为分类错误可能会造成误导。例如，始终预测为负类的分类器，其分类错误率已经很低。优化 F1 分数通常通过调整输入分布来实现，而不改变机器学习算法的内部结构 10。比如，为了“促使”分类器更加关注正例，我们可以对正例误分类施加更高的惩罚（下文称为“代价敏感学习”），或在训练数据中增加正例的数量（下文称为“重采样技术”）。SMOTE、ADASYN 和 GAN 系列方法等技术[3, 5, 19, 20]基于类似原理。对于代价敏感学习和重采样技术，我们只需调整一个超参数，即正类权重标量（PWS）：在代价敏感学习中，我们使用该标量放大每个正类实例对应的代价项；在重采样技术中，则用它提高正类实例被抽样到的概率。关于如何确定这一超参数的研究 [3][8, 9, 12, 14, 17, 21]认为，PWS 应当仅由数据结构决定。令 ! 表示负类实例与正类实例的比率。将类别频率的倒数 ! 或类别频率倒数的平方根 "! 作为 PWS，是两种常见做法。以阿片类药物滥用检测问题为例，其中正类实例仅占 1%，按照 "! 规则，我们会在训练集中将正类实例复制 10 次；按照 ! 规则，则复制 100 次。

本文提出以下假设：*H1.* 不同误差指标之间的权衡：PWS 只能在 F1 分数与其他误差指标之间实现权衡，并不能产生严格更优的模型。*H2. F1* 分数优化中的偏差与方差之间的权衡：更大的 PWS 通常会提高样本内 F1 分数，减少有效样本，并降低测试性能。

验证/我们的方法论

我们首先开展一项小规模合成研究，以突出说明调整输入分布或成本函数只能带来一种权衡，例如以牺牲其他指标为代价提高 F1 分数，而不会产生严格更优的模型。随后，我们通过在四个广泛使用的模型系列中，使用不同的超参数训练大量模型，分析 PWS 在预测阿片类药物过量中的作用；再通过回归分析理解最优 PWS 与模型复杂度之间的关系，并验证我们的假设。

理论与实践意义

从理论上看，我们的研究发现与当前用于解决类别不平衡问题的机器学习技术截然不同。F1 分数综合衡量精确率和召回率，在数据不平衡时更受青睐 [31]。通过将最优类别比例与模型

复杂度联系起来，我们阐明了优化 F1 分数时的方差-偏差权衡。从实践角度看，我们的研究发现表明，纳入正类权重标量（PWS）后，大多数现成模型只需对超参数搜索进行近乎零成本的调整，就能显著提高 F1 分数。在此之前，医学机器学习研究更多依赖成熟的机器学习库，而非设计新的专用模型（主要是出于成本考虑）；此外，超参数搜索通常受可用计算资源限制，也往往忽略 PWS，从而导致模型并非最优。图 1 显示，对训练数据重新加权并不能产生更好的模型。

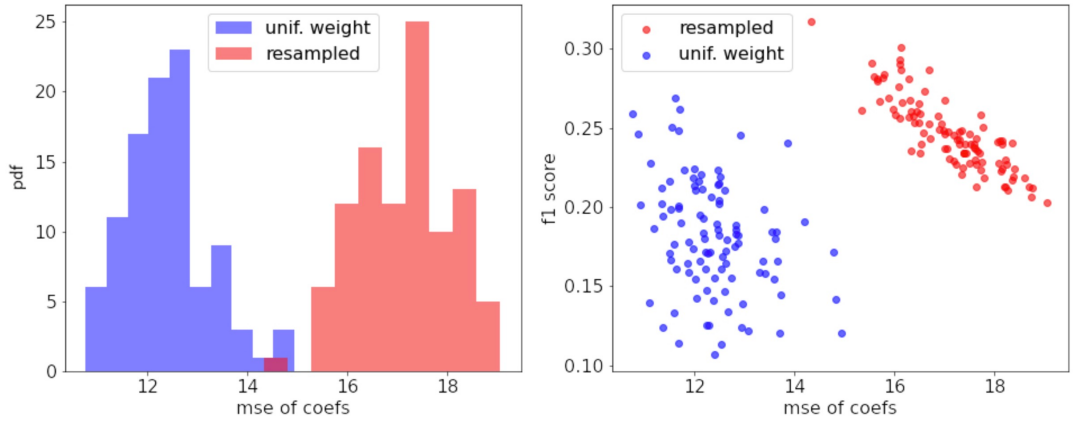

图 1。重新加权训练数据只会带来权衡，并不会产生更好的模型。对于模型，PWS 只会带来不同的权衡（本例中 F1 分数更高而均方误差更差），并不会产生严格更优的预测。我们使用逻辑线性模型，在 #$_0$ （蓝色）和 #$_1$ （红色）条件下生成合成训练数据。从 #$_0$ 中抽取的数据只有 5% 为正例，而从 #$_1$ 中抽取的数据约有 50% 为正例。随后，我们使用这些数据集拟合逻辑模型。左：估计系数与真实值相比的 $₂-误差直方图显示，使用 #$_1$ （更平衡的数据）下的数据的模型表现更差。右：估计系数的 F1 分数与均方误差散点图显示，使用平衡数据的模型均方误差更差，但 F1 分数有所提高。

初步

问题表述

我们考虑一个使用 %-维特征的二分类问题。我们的目标是拟合一个函数

$$y \sim f(x),$$

其中 & ∈{0, 1}，且为 $x \in R^d$。我们将) (·) 视为常用的机器学习黑箱，例如梯度提升决策树 [4, 11, 13, 15, 16] 或神经网络[7], , 并将 $T = \{(x_i, \ \&i)\}_{i \le n}$ 视为训练数据集。类别不平衡问题是指正标签的比例显著小于负标签的比例；在这种情况下，通常使用更稳健的指标（例如 F1 分数）来替代均方误差（MSE）或分类错误，以评估模型性能。我们将 F1 分数定义为：

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + TN)},$$

其中，TP 是真阳性的数量，FP 是假阳性的数量，FN 是假阴性的数量。需要注意的是，精确率和召回率也可以用于定义 F1 分数 [31]。

训练及解决类别不平衡问题. 许多工具和算法通常沿用现有的机器学习架构，并调整成本函数或训练数据的分布，以促使分类器减少对正例的误判并提高 F1 分数。具体而言，我们令 $(&i, &?) 表示机器学习模型中使用的损失函数，例如，$(&, &@) = (& −&@)² 表示均方误差，$(&, &@) = −(& · $:C(&@) + (1 −&) · $:C(1 −&@)) 表示二元交叉熵损失。在不存在类别不平衡问题的情况下，通常我们会最小化

$$\sum_{(x_i, y_i) \in T} l\,(y_i, f(x_i)) = \sum_{(x_i, y_i) \in T} l\,(y_i, \hat{y}_i).$$

(1)

成本敏感学习和重采样常用于解决类别不平衡问题，两者的目标都是提高正例误判的成本，并且在某些条件下等价 CITE。成本敏感学习会提高正例的成本，即新的成本函数变为

$$\sum_{(x_i,y_i)\in T_P} \gamma \cdot l(y_i, f(x_i)) + \sum_{(x_i,y_i)\in T_N} l(y_i, f(x_i)), \quad (2)$$

其中，$T_P = \{(x_i, \&i) \in T | \&_i = 1\}$ 是 5 中正例的子集，$T_N = T - T_P$ 是负类实例的子集，而 G > 1 是 PWS。重采样会将*Eq.1*中使用的训练集替换为 #（即，我们优化$\sum \$_{(x_i,y_i)\in D} (\&i,)(x_i))$，其中 # 中的每个元素都从 5 中抽样，使得当 $\&_i = 0$ 时，$(x_i, \&i)$ 被抽取的概率为 J；当 $\&_i = 1$ 时，被抽取的概率为 G · J。这里，G 为 PWS，J 为归一化常数，使所有概率质量之和为 1。

方法论与实验设计

本节介绍用于验证*H1*和*H2*的实验设置。我们首先考察一个合成数据集，以获得直观认识。然后，我们说明基于真实数据针对阿片类药物过量问题所开展的实验。

误差指标权衡的直观理解

如前所述，我们先从一项小规模合成研究入手，并使用逻辑回归模型生成的合成数据集，来理解 PWS 与模型复杂度之间的关系。回想一下，在逻辑模型中，

$$Pr[y_i = 1 \mid x_i] = \sigma(\omega^{\intercal} x_i + \omega_0). \quad (3)$$

在标准的收敛分析中，我们假设观测到来自分布 # 的独立同分布 样本（x_i，&*i*），并求得一个可优化似然函数的估计器（N?, N?₀）。最大似然估计器（MLE）（N?，N?₀）是无偏的，其方差通常与 x_i 的协方差矩阵有关（即 MLE [18] 的渐近正态性）。

我们考虑两个分布 $\#_0$ 和 $\#_1$，其中 $\#_0$ 是会导致 & 不平衡的原始分布，而 $\#_1$ 则调整了 x 的分布，使正类 & 的比例更大。我们考虑使用相同数量的数据训练两个模型；在这两个分布下构造的 MLE 都渐近正态，即二者均值相同且与真实值一致，但方差不同。由于两个估计器使用相同数量的样本进行训练，其中一个估计器的质量不太可能占据优势；也就是说，当数据分布从 $\#_0$ 变为 #₁时，置信区间（集）或等距轮廓只是发生旋转，并未收缩。

举例来说，令 % = 100 （特征数量）和 O = 120 （观测值数量），并使用 式*3* 中指定的模型生成训练数据。我们将 N_0 设为一个绝对值适当的负数，使得当 $Q[x] = 0$时，正类实例所占比例较小。对于 #₀，我们假设 x 服从标准正态分布。对于 #₁，我们按照 $\#_0$ 抽取足够大的训练数据，然后删除其中很大一部分负类实例，以使正类实例和负类实例大致平衡。我们指出，$\#_1$ 模拟了标准的过采样技巧。

对于 $\#_0$ 和 $\#_1$，我们分别生成 100 组数据，并考察估计器的分布。参见图1，我们发现，如果以估计器的均方误差（图1左侧）衡量，来自 $\#_1$ 的估计器不如来自 $\#_0$ 的估计器，但在 #₁下，F1 分数更高。在图1（右侧）中，我们看到，$\#_1$ 下模型的散点从右下方向左上方移动，这表明存在权衡，即以牺牲估计器准确度为代价来提高 F1 分数。

这项合成研究的结果帮助我们提出了*H2*：较大的 PWS 会将置信集旋转至更有利于 F1 分数的方向，但也会降低有效样本量。以 PWS → ∞ 为例。在这里，我们实际上只关注正例，训练集规模也显著更小。因此，我们预期会出现一种权衡：随着模型复杂度增加，我们首先需要增大 PWS；但到了一定程度，PWS 就应停止增大并缩小，以免因有效样本不足而导致置信集的体积过度扩大。这也意味着，PWS 应针对每个模型进行优化，而不是设为一个固定值。

数据集

我们采用了来自 Cerner 的 Health Facts[22] 数据。Health Facts 是美国一个大型电子健康记录数据库，包含来自 600 多家参与临床机构的近 6900 万名患者的信息。该数据库以患者的就诊记录为结构基础，其中包含诊断、医疗操作、患者人口统计信息、药物剂量和给药信息、生命体征、实验室检查结果、手术病例信息、医疗系统属性及其他临床观察。对于这项阿片类药物过量研究，我们从 Health Facts 中提取了所有开具阿片类药物处方的患者。提取后，剩余 5,231,614 名患者，其中 44,774 名为阳性病例。阳性率不足 1%。另请参见图 2，了解我们的数据收集流程。

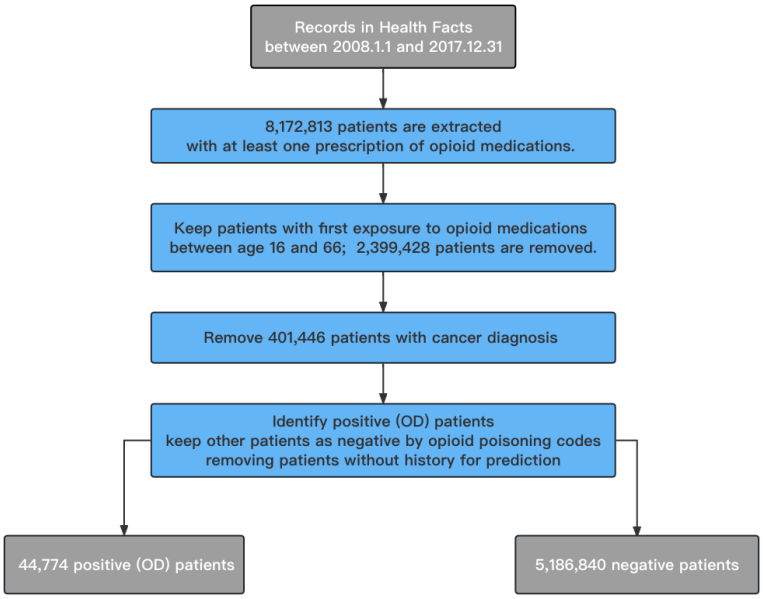

图2。收集和生成实验数据的流程图

为从数据集中构建特征输入，我们遵循预定义流程[22] 提取了 1185 个特征，包括 414 个诊断代码特征、394 个实验室检查特征、3 个人口统计特征、227 个临床事件特征和 147 个药物特征。基于构建的特征矩阵，我们采用 BERT 模型，并通过掩码语言模型和下一句预测这两种经典的预训练方式对其进行训练[33]。训练后，每位患者输入的输出将作为 50 维的患者表征。我们使用这些表征开展研究。

模型

我们选取了四种广泛应用于应用机器学习研究的机器学习模型。这些模型在公开基准测试中通常优于更复杂和／或定制的模型[2, 11, 15]。其中两种基于提升分类树（LightGBM 和 CatBoost），另外两种基于深度学习。我们训练了 8,000 个 LightGBM 模型、2,000 个 CatBoost模型和 16,000 个神经网络模型，共计 26,000 个模型。对于每个模型，我们在 1 到 150 之间搜索，以找到最优 PWS。表 1 列出了模型、超参数及其取值。

表 1。所考察的四种机器学习模型及其超参数和值。

模型名称	**超参数**	**超参数的取值**
LightGBM	数量 估计器（树）	100、500、1000、2000、5000, 8000, 10000
	树的最大深度	4, 5, 7
	最大数量 树的叶节点	100, 300, 500, 1000
	学习率	0.1、0.05、0.01
CatBoost	数量 估计器（树）	100, 500, 1000, 2000, 5000, 8000、10000
	树的最大深度	4、5、 7
	学习率	0.1、0.05、0.01

MLP （两者均加权 成本和 重采样)	隐藏层的宽度 层	50, 100
	层数	1, 2
	缩小 初始化	1, 0.1
	批归一化	不，只在输入之后，以及每个 线性-ReLU 层
	学习率	10^{-1}，10^{-2}，10^{78}，10^{79}
	动量	0.1, 0.9

LightGBM [11]

LightGBM 模型采用了一套高效的启发式学习规则，同时保持了相当的计算效率。它实现了梯度提升回归树的一种变体，并逐步构建由基于树的基学习器组成的集合。每个基学习器都是一棵决策树，其深度和叶子数量由超参数控制。训练新的基学习器后，模型会根据新基学习器的估计值对响应变量进行残差化调整。

CatBoost [15]

CatBoost 模型实现了梯度提升决策树的另一种变体。它在训练过程中整合了更多新型启发式方法和统计估计技术。这些新技术妥善解决了基于提升的模型中所谓的“目标泄漏”问题。

神经网络 加权代价函数

我们考虑具有一个或两个隐藏层的多层感知机。我们通过成本函数接入 PWS，并优化公式 *2*。中的成本函数。

神经网络重采样

我们使用与上述相同的一组模型架构。我们通过重采样实现 PWS，即在随机梯度批次中，正类实例被采样的概率是负类实例的 G 倍，其中 G 为 PWS。

模型复杂度

接下来，我们将介绍可能与模型复杂度相关的超参数，这些模型将用于分析模型复杂度与最优 PWS 之间的关系。*(i)* LightGBM 和 CatBoost：叶子数量、树深度和学习率都可能与这些模型的复杂度相关。叶子数量也与 LightGBM 的模型复杂度相关（CatBoost 无法调整此超参数）。*(ii)* 神经网络：每层宽度和层数与模型复杂度的关系最为密切。

一般而言，判断某个超参数是否与模型复杂度相关，并不总是显而易见的。*(i)*我们认为，学习率、批次大小、动量或随机初始化的可学习参数的尺度并不属于神经网络的模型复杂度，尽管有证据表明它们可能有助于拟合正则化[32]。*(ii)*虽然使用加权代价函数的神经网络与采用重采样的神经网络具有相同的架构，但它们拟合训练数据的容量似乎不同，因此模型复杂度可能也不同。*(iii)*即使树/估计器的数量较少，CatBoost 仍能很好地拟合训练数据，这表明 CatBoost 的模型复杂度可能更难界定。下一节将说明，量化模型的复杂度确实是一项微妙的工作。

我们考虑具有一个或两个隐藏层的多层感知机。我们以两种方式训练这些神经网络。（*i*） 使用权重调整成本函数。我们通过成本函数引入 PWS，并优化公式 *2*。中的成本函数。（*ii*） 使用重采样。我们通过重采样引入 PWS，即在一个随机梯度批次中，正类实例被抽样的概率是负类实例的 G 倍，其中 G 为 PWS。

神经网络的模型复杂度。神经网络的规模（例如每层宽度和总层数）与模型复杂度的关联最为密切。实际情况则更为复杂。例如，尽管两类神经网络（采用权重调整成本的神经网络和采用重采样的神经网络）理论上最小化的是同一个效用函数，但采用重采样的神经网络似乎更擅长拟合训练数据，因此其模型复杂度可能更高。

实验结果

本节介绍实验结果和回归分析。我们首先考察通过为四类具有不同超参数的模型调整 PWS 所获得的性能提升。我们的基线模型假设 PWS 是一个常数，取值为 10 或 100。表2列出了结果。

表2。 不同 PWS 下模型的性能。我们考察了 LightGBM（lgbm）、CatBoost、采用加权代价的多层感知机（mlp.wcost）以及采用重采样的多层感知机（mlp.wsample）。我们还考虑了三种确定 PWS 的方法，包括网格搜索（后缀为“.opt”的行）、使用线性规则将其设为常数 100（后缀为“w.100”的行），以及使用平方根规则将其设为常数 10（后缀为“w.10”的行）。对于训练集和测试集，我们展示 F1 分数和误差；对于测试集，我们展示 ROC、精确率和召回率。对于使用网格搜索的方法，PWS 列表示最优 PWS；对于使用线性规则或平方根规则的方法，则表示所使用的常数。

模型	f1_测试集	PWS	ROC_测试集	精确率_测试集	召回率_测试集	误差_测试集	f1_训练集	误差_训练集
lgbm.opt	**53.1%**	**25**	95.1%	57.8%	49.2%	1.5%	49.1%	1.5%
基线：lgbm.fixed.w	51.5%	10	95.0%	57.3%	47.1%	1.4%	53.2%	1.3%
基线：lgbm.fixed.w	51.3%	100	95.1%	59.3%	45.4%	6.5%	48.6%	6.5%
catboost.opt	**52.1%**	**35**	94.4%	58.6%	47.1%	2.8%	57.2%	2.5%
基线：cb.fixed.w	51.0%	10	95.1%	57.5%	45.9%	1.4%	52.2%	1.2%
基线：cb.fixed.w	50.4%	100	93.8%	56.9%	45.3%	4.4%	53.7%	4.1%
mlp.wcost.opt	**50.8%**	**70**	86.9%	55.4%	47.2%	4.9%	48.7%	4.8%
基线：mlp.wcost.10	50.3%	10	82.4%	51.2%	49.8%	2.2%	48.8%	2.2%
基线：mlp.wcost.100	50.4%	100	90.3%	54.0%	47.4%	7.8%	48.7%	7.8%
mlp.wsample.opt	**50.8%**	**57**	86.4%	52.1%	49.6%	5.6%	49.1%	5.5%
基线：mlp.wsample.10	50.5%	10	79.3%	56.4%	45.8%	1.4%	50.2%	1.4%
基线：mlp.wsample.100	50.3%	100	90.8%	53.3%	47.9%	8.7%	48.7%	8.6%

我们有三项主要研究发现。第一，仅调整 PWS 就能提升所有情境下的 F1 分数。由于模型的表现往往会在约 50% [6] 时趋于平稳，lgbm.opt 达到 53.1% 的 F1 分数，意味着显著提升。第二，依赖于模型的最优 PWS 与 H2 一致。第三，这两项性能指标之间的权衡（F1 分数提升，但分类错误增加）也证实了 *H1*。

我们还绘制了四类模型的 F1 分数散点图。在图2中，每个点对应一个具有固定超参数集的模型，x 轴表示模型在训练数据上的 F1 分数，y 轴表示模型在测试数据上的 F1 分数。

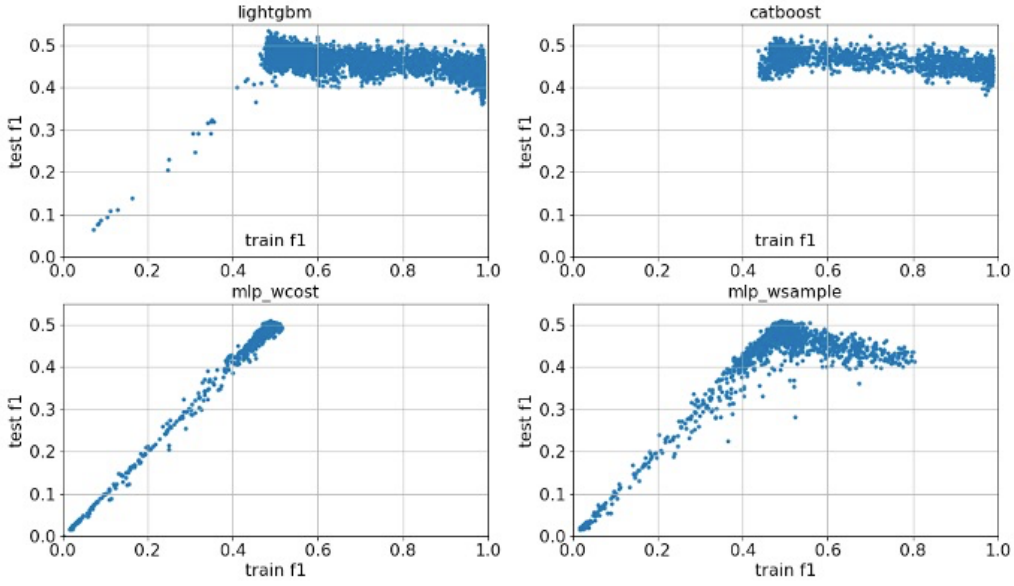

图 3。 LightGBM、CatBoost、采用加权代价的多层感知机（mlp_wcost）和采用重采样的多层感知机（mlp_wsample）的训练集与测试集 F1 分数散点图。

我们发现了三个主要结果。首先，四类模型都存在方差-偏差权衡现象。当模型的训练集 F1 分数提升时，测试集 F1 分数也会随之提升，直到训练集和测试集的 F1 分数都达到约 50%。此后，尽管

训练集 F1 分数仍可继续提升，但测试集 F1 分数要么保持不变（CatBoost模型），要么下降（采用重采样的 LightGBM 和多层感知机）。其次，尽管 LightGBM 和 CatBoost 都是梯度提升回归树的变体，但它们的训练行为存在显著差异：所有 CatBoost模型都能很好地拟合训练数据（F1 分数为 ≥ 0.4），而相当一部分 LightGBM 模型无法恰当地拟合训练数据（即 F1 分数为 ≤ 0.4）。第三，所有采用加权代价函数的多层感知机都无法拟合训练数据（分数为 ≤ 0.6），而采用重采样的多层感知机对训练数据的拟合效果更好。请注意，这两类神经网络使用相同的一组架构。所有这些发现都表明，通过超参数来表征模型复杂度极具挑战性。

线性回归分析

本节基于我们训练的 26,000 个模型，分析 PWS 与模型复杂度之间的关系。我们采用线性回归，因为它在估计模型系数和 X统计量时具有更稳健的医疗操作，尽管模型复杂度变量与 PWS 之间可能并非严格呈线性关系。我们首先分析 LightGBM，因为其统计特性广为人知，且所需的启发式方法较少，随后分析 CatBoost 和神经网络模型。

***Lightgbm*回归模型**

如前所述，与 LightGBM 复杂度相关的超参数包括估计器数量、学习率、每棵树的最大深度，以及每棵树的最大叶子数量。对于每组超参数，我们找出 Y 个最佳 PWS 值。然后构建 Y 个观测值，每个观测值都以模型超参数作为协变量，以最优 PWS 作为响应变量。接着，我们拟合以下模型：

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

其中，x_1 是估计器数量，x_2 是学习率，x_3 是最大树深度，x_4 是叶子数量，响应变量 & 是对应的最优 PWS。我们分别对训练数据和测试数据进行拟合（即，为训练数据选择训练 F1 分数最高的模型，为测试数据选择测试分数最高的模型）。每组超参数对应的 PWS 数量 Y 决定了估计准确度与估计质量之间的权衡。当 Y ＝1 时，我们只关注最优 PWS，从而获得最理想的响应变量，但观测值总数较少。当 Y 增大时，我们为每组超参数选取排名前 Y 的模型；尽管其中会选入一些非最佳模型，但观测值总数也会增加。我们考察了多个 Y 值，以提高分析的稳健性。

根据 *H2*，我们预计，随着模型复杂度增加，为了提升测试性能，PWS 最终需要缩小。此外，随着所用模型变得更加复杂，训练数据的最优 PWS 应当增大。具体而言，当我们使用简单模型时，决策边界集合相当僵化（即线性模型的决策边界是线性的）。在这种情况下，较大的 PWS 可能会使决策边界偏移到不理想的位置，从而损害训练数据和测试数据的 F1 分数。 使用更复杂的模型时，决策边界拥有更高的自由度，因此增大 PWS 可以提高训练集的 F1 分数。

表3列出了估计系数、与该系数相关的X统计量，以及每个协变量与响应变量之间的相关系数。X统计量（无论关系是否具有统计显著性）及其方向最为重要。当某个变量的X统计量绝对值至少为3时（对应J值 0.003），我们认为该变量具有统计显著性。

表3。 LightGBM回归分析

k	mode	计数	估计器数量 (!1) 贝塔系数-¹	t统计量	corr	学习率 (!2) 贝塔系数-²	t统计量	corr	最大深度 (!3) 贝塔系数-³	t统计量	corr	叶子节点数(!4) 贝塔系数-⁴	t统计量	corr
	test	252	-0.0027	-3.4665	-0.2138	-62.6178	-0.8179	-0.0505	-3.9539	-1.7608	-0.1086	0.0019	0.2269	0.0140
1	训练集	252	0.0062	10.5165	0.5177	384.4424	6.6113	0.3255	5.7409	3.3444	0.1646	-0.0026	-0.4086	-0.0201
	test	756	-0.0024	-5.4263	-0.1933	-18.9180	-0.4278	-0.0152	-3.5534	-2.7223	-0.0970	-0.0028	-0.5710	-0.0203
3	训练集	756	0.0044	13.5818	0.4128	342.6457	10.6497	0.3237	5.4452	5.7331	0.1743	-0.0029	-0.8238	-0.0250
	test	1260	-0.0021	-5.9845	-0.1660	-15.6023	-0.4524	-0.0125	-3.0111	-2.9575	-0.0820	-0.0003	-0.0739	-0.0021
5	训练集	1260	0.0040	15.5481	0.3781	307.4436	12.0934	0.2941	5.1650	6.8824	0.1674	0.0017	0.6099	0.0148

讨论。 我们发现，结果与我们基于 H2 提出的预期一致。首先，在训练数据的回归模型中，β_1 至β_4 均为正值，这表明随着模型复杂度的增加，PWS 会增加，以优化 F1 分数。我们发现，在测试数据的回归模型中，β_1 , β_2, 和β_4 为负值，这表明随着模型复杂度的增加，PWS 会降低，以优化 F1 分数。此外，我们发现β_3（与最大深度相关的系数）也呈现出类似的趋势，但其统计显著性较低。

***CatBoost* 回归模型**

我们模拟 LightGBM，但由于无法直接控制 CatBoost 超参数中的叶子数量，我们的回归变为：

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

其中，x_1 为估计器数量，x_2 为学习率，x_3 为最大树深度，& 为相应的 PWS。我们继续考察 Y = 1、3、5，并在训练集和测试集上运行模型。我们再次指出，衡量 CatBoost模型的复杂度可能比较微妙，因为即使 CatBoost模型仅由少量树（估计器）组成，也仍能很好地拟合训练数据。为解决这一问题，我们在回归分析中剔除了树数量较少的 CatBoost模型。表4列出了包含 2,000 棵或更多树的 CatBoost模型的结果。

表4。包含 2,000 棵或更多树的 CatBoost回归分析

k	mode	n	估计器数量_（!1）			学习 _率（!2）			最大_深度（!3）		
			贝塔系数."1	t统计量	corr	贝塔系数."2	t统计量	corr	贝塔系数."3	t统计量	corr
	test	36	0.0033	1.3943	0.2051	520.8121	2.6380	0.3881	-13.4259	-2.3037	-0.3389
1	训练集	36	0.0006	2.1468	0.2643	88.7907	3.8464	0.4736	2.5995	3.8147	0.4697
	test	108	0.0022	1.4830	0.1414	230.4582	1.8909	0.1803	-1.6923	-0.4704	-0.0448
3	训练集	108	0.0005	1.8348	0.1596	73.6998	3.2451	0.2822	2.5384	3.7862	0.3293
	test	180	0.0018	1.5067	0.1122	120.7273	1.2368	0.0921	-1.8289	-0.6347	-0.0473
5	训练集	80	0.0004	1.4496	0.1027	58.3983	2.8842	0.2044	2.1347	3.5715	0.2531

讨论。结果与LightGBM相似。我们发现，训练数据的β_i 大多为正。与学习率和最大深度相关的系数在训练集中显著，而在测试数据中，所有系数均不显著（与0相比，即我们无法拒绝所有系数均为0的原假设）。与 LightGBM不同，在测试集中，与估计器数量和最大深度相关的系数为正（尽管不显著），这表明CatBoost更善于防止过拟合。总体而言，这也反映出，对于采用高级统计技术的较新模型，衡量模型复杂度更加困难。

多层感知机回归模型

与基于提升的模型相比，多层感知机回归分析需要针对每组与复杂度相关的超参数考察更多模型。除了调整隐藏层的宽度和数量外，我们还需要调整许多其他超参数（即随机初始化尺度、是否进行批量归一化、学习率和动量），以优化训练性能。为解决这一问题，我们从回归分析中移除无关的超参数。

我们的最终回归模型仅包含两个协变量：

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

其中，x_1 为每个隐藏层的宽度，x_2 为隐藏层的数量。我们检查了 1,000 个模型，并增加 Y（即 Y = 100、200、300），以获得更稳健的响应变量。表 5 列出了结果。

表 5。多层感知机的回归分析

k	mode	n	隐藏 _size			n_层		
			贝塔系数."1	t统计量	corr	贝塔系数."2	t统计量	corr
	test	800	-0.0016	-0.3696	-0.0131	-3.1375	-0.9698	-0.0343
100	训练集	800	0.0324	8.8693	0.2720	38.1875	13.6954	0.4200
	test	1600	0.0010	0.3371	0.0084	-1.1750	-0.5111	-0.0128
200	训练集	1600	0.0174	5.9822	0.1457	16.1938	7.3153	0.1781
	test	2400	-0.0006	-0.2439	-0.0050	0.1083	0.0577	0.0012
300	训练集	2400	0.0079	3.3094	0.0673	5.3292	2.9400	0.0598

讨论。我们发现，多层感知机的结果与 LightGBM 和 CatBoost 相似。我们发现，在训练集中，β_1 和 β_2 为正且显著，而在测试集中，所有系数均不显著，这再次表明，两个相互竞争的因素（提高 F1 分数与降低方差误差）共同决定了 PWS 的最优值。

正如我们所指出的，由于宽度和隐藏层数量可能无法恰当地表征深度神经网络的复杂度，这或许就是我们在测试集上未发现 β_1 和 β_2 呈显著负相关的原因。因此，我们需要寻找其他协变量来表征模型复杂度。一个自然的统计量是模型的样本内表现，即模型对训练数据的拟合效果越好，其拟合能力就越强，也越可能具有较高的复杂度。因此，我们在回归模型中加入样本内 F1 表现，并进行估计：

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

其中，x_1 是每个隐藏层的宽度，x_2 是隐藏层的数量，x_3 是样本内 F1 分数。表 6 展示了结果。

表 6。具有样本协变量的多层感知机回归分析

			隐藏的 _size			n_层次			训练集 F1 分数_		
k	mode	n	贝塔系数."1	t统计量	corr	贝塔系数."2	t统计量	corr	贝塔系数."3	t统计量	corr
100	test	800	0.0087	2.2328	-0.0131	3.7541	1.2726	-0.0343	-2,469.1991	-13.8136	-0.4335
200	test	1600	0.0081	2.9607	0.0084	4.2574	2.0507	-0.0128	-2,859.9630	-19.9198	-0.4392
300	test	2400	0.0066	2.9997	-0.0050	5.8308	3.4562	0.0012	-3,022.3433	-25.1486	-0.4495

讨论。我们发现，β_3 对最优 PWS 有非常显著的负向影响。换言之，当模型对样本内数据的拟合效果更好时，需要较小的 PWS，才能在测试数据上取得更好的表现。这一发现证实了 H2。

结论

本研究重新审视了在重平衡过程中用于确定类别比例、以应对类别不平衡问题的广泛应用规则。我们提出将最优类别比例/正类权重标量（PWS）与用于解决类别不平衡问题的临床预测模型的复杂度联系起来。针对阿片类药物过量问题开展的实验表明，该框架显著提升了预测能力。此外，回归分析发现，控制模型复杂度的超参数与 PWS 之间存在统计显著的相关性。该理论框架可用于解决各类类别不平衡问题，并以近乎零的成本提升预测性能。

参考文献

1. Kayley Abell-Hart、Sina Rashidian、Dejun Teng、Richard N Rosenthal 和 Fusheng Wang。阿片类药物过量患者居住地远离治疗机构：纽约州服务不足人群的地理空间分析。《JMIR公共卫生与监测》，8(4):e32133，2022。 2. Casper Solheim Bojer 和 Jens Peder Meldgaard。Kaggle 预测竞赛：一个被忽视的学习机会。《国际预测学杂志》，37(2):587–603，2021。 3. Nitesh V Chawla、Kevin W Bowyer、Lawrence O Hall 和 W Philip Kegelmeyer。SMOTE：合成少数类过采样技术。《人工智能研究杂志》，16:321–357，2002。 4. Tianqi Chen、Tong He、Michael Benesty、Vadim Khotilovich、Yuan Tang、Hyunsu Cho、Kailong Chen、Rory Mitchell、Ignacio Cano、Tianyi Zhou 等。Xgboost：极端梯度提升。R 软件包版本 0.4-2，1(4):1–4，2015。 5. Yin Cui、Menglin Jia、Tsung-Yi Lin、Yang Song 和 Serge Belongie。基于样本有效数量的类别均衡损失。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9268–9277页，2019。 6. Xinyu Dong、Rachel Wong、Weimin Lyu、Kayley Abell-Hart、Jianyuan Deng、Yinan Liu、Janos G Hajagos、Richard N Rosenthal、Chao Chen 和 Fusheng Wang。用于可解释阿片类药物过量风险预测的集成式 LSTM-HeteroRGNN 模型。《医学人工智能》，135:102439，2023。 7. Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。《深度学习》。MIT 出版社，2016。 8. Haibo He、Yang Bai、Edwardo A Garcia 和 Shutao Li。ADASYN：用于不平衡学习的自适应合成采样方法。载于《2008年IEEE国际神经网络联合会议（IEEE计算智能世界大会）》，第1322–1328页。IEEE，2008。 9. Chen Huang、Yining Li、Chen Change Loy 和 Xiaoou Tang。为不平衡分类学习深度表征。载于《IEEE计算机视觉与模式识别会议论文集》，第5375–5384页，2016。 10. Nathalie Japkowicz 和 Shaju Stephen。类别不平衡问题：一项系统研究。《智能数据分析》，6(5):429–449，2002。 11. Guolin Ke、Qi Meng、Thomas Finley、Taifeng Wang、Wei Chen、Weidong Ma、Qiwei Ye 和 Tie-Yan Liu。LightGBM：一种高效的梯度提升决策树。《神经信息处理系统进展》，30，2017。 12. Dhruv Mahajan、Ross Girshick、Vignesh Ramanathan、Kaiming He、Manohar Paluri、Yixuan Li、Ashwin Bharambe 和 Laurens Van Der Maaten。探索弱监督预训练的极限。载于《欧洲计算机视觉会议论文集（ECCV）》，第181–196页，2018。

13. Llew Mason、Jonathan Baxter、Peter Bartlett 和 Marcus Frean。作为梯度下降的提升算法。《神经信息处理系统进展》，12，1999。 14. Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。词语与短语的分布式表征及其组合性。《神经信息处理系统进展》，26，2013 11 15 . . Liudmila Prokhorenkova、Gleb Gusev、Aleksandr Vorobev、Anna Veronika Dorogush 和 Andrey Gulin。CatBoost：具有类别特征的无偏提升。《神经信息处理系统进展》，31，2018。 16. Robert E Schapire 和 Yoav Freund。《提升：基础与算法》。《Kybernetes》，42(1):164–166，2013。 17. Yu-Xiong Wang、Deva Ramanan 和 Martial Hebert。学习对尾部进行建模。《神经信息处理系统进展》，30，2017。 18. Larry Wasserman。《统计学全书：统计推断简明课程》，第26卷。Springer，2004。 19. Hao Yang 和 Yun Zhou。IDA-GAN：一种新型不平衡数据增强生成对抗网络。载于《2020年第25届国际模式识别会议（ICPR）》，第8299–8305页。IEEE，2021。 20. Tingting Zhou、Wei Liu、Congyu Zhou 和 Leiting Chen。基于 GAN 的不平衡数据分类半监督方法。载于《2018年第4届国际信息管理会议（ICIM）》，第17–21页。IEEE，2018。 21. Yang Zou、Zhiding Yu、BVK Kumar 和 Jinsong Wang。通过类别均衡自训练进行语义分割的无监督域适应。载于《欧洲计算机视觉会议论文集（ECCV）》，第289 12 22 305页，2018– . . Dong X、Deng J、Hou W、Rashidian S、Rosenthal RN、Saltz M、Saltz JH、Wang F。基于时间深度学习，利用电子健康记录预测有阿片类药物处方患者的阿片类药物过量风险。《生物医学信息学杂志》。2021年4月1日；116:103725。 22. Dong X、Deng J、Rashidian S、Abell-Hart K、Hou W、Rosenthal RN、Saltz M、Saltz JH、Wang F。利用深度学习识别服用阿片类药物患者的阿片类药物使用障碍风险。《美国医学信息学会杂志》。2021年7月30日；28(8):1683-93。 23. Xu Z、Feng Y、Li Y、Srivastava A、Adekkanattu P、Ancker JS、Jiang G、Kiefer RC、Lee K、Pacheco JA、Rasmussen LV、Pathak J、Luo Y、Wang F。重症监护中急性肾损伤风险的预测建模：对类别不平衡问题的系统性调查。《AMIA联合峰会转化科学论文集》。2019年5月6日；2019:809-818。PMID：31259038；PMCID：PMC6568062。 24. X. -Y. Liu、J. Wu 和 Z. -H. Zhou，“类别不平衡学习的探索性欠采样”，载于《IEEE系统、人类与控制论汇刊B辑（控制论）》，第39卷，第2期，第539-550页，2009年4月，doi：10.1109/TSMCB.2008.2007853。 25. I. Fakhruzi，“采用装袋法的人工神经网络，以解决临床预测中的数据集不平衡问题”，《2018年国际信息与通信技术会议（ICOIACT）》，印度尼西亚日惹，2018，8350824 27 895 898，10 1109 2018 pp. - doi: . /ICOIACT. . . Sun S、Wang F、Rashidian S、Kurc T、Abell-Hart K、Hajagos J、Zhu W、Saltz M、Saltz J。利用循环自编码器和生成对抗网络生成纵向合成电子健康记录数据。载于《异构数据管理、多数据存储与医疗保健分析：VLDB研讨会论文集，Poly 2021和 DMAH 2021，线上会议，2021年8月20日，修订版精选论文》第7卷，2021年（第153-165页）。Springer International Publishing。 26. van den Goorbergh R、van Smeden M、Timmerman D、Van Calster B。类别不平衡校正对风险预测模型的危害：使用逻辑回归进行说明和模拟。《美国医学信息学会杂志》。2022年9月；29(9):1525-34。 27. Vandewiele G、Dehaene I、Kovács G、Sterckx L、Janssens O、Ongenae F、De Backere F、De Turck F、Roelens K、Decruyenaere J、Van Hoecke S。不平衡数据上的过度乐观预测结果：过采样应用中缺陷与益处的案例研究。《医学人工智能》。2021年1月1日；111:101987。 28. Florence C、Luo F、Rice K。2017年美国阿片类药物使用障碍和致命阿片类药物过量的经济负担。《药物与酒精依赖》。2021年1月1日；218:108350。doi：10.1016/j.drugalcdep.2020.108350。电子版发表于2020年10月27日。PMID：33121867；PMCID：PMC8091480。 29. Sasaki Y。F-measure 的真相。2007。网址：https://www.cs.odu. edu/mukka/cs795sum09dm/Lecturenotes /Day3/F-measure-YS-26Oct07. pdf [访问日期：2021-05-26]。 30. Wu L、Li J、Wang Y、Meng Q、Qin T、Chen W、Zhang M、Liu TY。R-drop：神经网络的正则化 dropout。《神经信息处理系统进展》。2021 905 33 6 34 10890Dec ; : - . Devlin J、Chang MW、Lee K、Toutanova K。BERT：用于语言理解的深度双向 Transformer 预训练。arXiv 预印本 arXiv:1810.04805。2018年10月