适用于含罕见事件海量数据的分布式逻辑回归

李学通[1],朱雪宁[2],和王汉生[1]

[1] 北京大学光华管理学院，中国北京；

[2] 复旦大学数据科学学院，中国上海。

arXiv:2304.02269v1 [stat.ME] 5 Apr 2023

摘要

大规模罕见事件数据在实践中十分常见。为应对海量罕见事件数据，我们提出了一种适用于分布式系统的逻辑回归新型分布式估计方法。在分布式框架下，我们面临两项挑战。第一项挑战是如何分配数据。对此，我们研究了两种不同的数据分配策略，即随机策略和复制策略。第二项挑战是如何选择合适类型的目标函数，以实现最佳渐近效率。随后，我们考察了欠抽样（US）和逆概率加权（IPW）两类目标函数。研究结果表明，对于含罕见事件的分布式逻辑回归，复制策略与 IPW 目标函数相结合是最佳方案。通过模拟研究和真实世界的瑞典交通标志数据集，展示了分布式方法在有限样本下的表现。

KEY 关键词：海量稀有事件数据；逻辑回归；分布式系统

1. 引言

二元回归中的海量稀有事件数据常见于科学研究领域及各类应用。从概念上说，稀有事件数据也称为类别不平衡数据，指正类中的样本数远少于负类中的样本数。例如，在在线搜索或推荐系统中，每天会产生数十亿次曝光。若将每次曝光视为一个样本，那么一次曝光产生点击的概率就非常低。因此，点击可视为稀有事件（Japkowicz et al., 2000; McMahan et al., 2013; Chen et al., 2016; Huang et al., 2020）。政治学领域的另一个例子是战争、否决、政变的发生，以及公民决定竞选公职；这些现象都曾被建模为稀有事件（King and Zeng, 2001; Owen, 2007; Neunhoeffer and Sternberg, 2019）。最后一个例子是高分辨率图像中的小目标检测，见图 1。假设将每个像素视为一个样本，并将其是否被边界框覆盖作为相应的响应变量。那么，小目标的边界框作为正样本，仅覆盖原始图像不到 1% 的区域（Zhu et al., 2016; Zhao et al., 2019; Chen et al., 2020）。其他重要的稀有事件数据例子还包括欺诈检测（Bolton and Hand, 2002; Hassan and Abraham, 2016）、药物发现（Zhu et al., 2006; Korkmaz, 2020）和罕见病诊断（Zhao et al., 2018; Zhuang et al., 2019）。如需全面综述，可参阅 Sun et al. (2009)、Haixiang et al. (2017) 和 Kaur et al. (2019)。

处理数据不平衡的一种常见方法是通过对负类欠采样（Drummond et al., 2003; Liu et al., 2008; Nguyen et al., 2012）或对正类过采样（Chawla et al., 2002; Han et al., 2005; Mathew et al., 2017）来平衡数据。现有文献大多关注用于分类的实用算法和方法，而缺乏统计理论保证。这些研究设计了采样策略



图 1：瑞典交通标志数据集中用于交通标志检测的一个示例。原始图像的尺寸为 960×1， 280×3。每个红色边界框用于标注包含交通标志的局部区域。被视为正实例的小目标的边界框仅覆盖原始图像的不到 1%。

或集成学习方法，以提高分类准确率（Krawczyk, 2016）。例如，Estabrooks et al. (2004) 通过实证研究探讨了不同重采样范式的有效组合，以提高分类准确率。Sun et al. (2007) 对 AdaBoost 算法进行了改造，以改进不平衡数据的分类。King and Zeng (2001) 研究了稀有事件数据中的逻辑回归，重点关注如何校正回归系数和概率估计中的偏差。Fithian and Hastie (2014) 利用逻辑回归模型的特殊结构，设计了一种新颖的局部病例-对照抽样方法。然而，这些理论研究基于一个常见假设，即事件发生的概率固定不变。从数学上描述稀有事件时，这或许并非最佳方式，因为这一假设意味着，随着总样本量趋于无穷大，稀有事件的数量也应以相同速率趋于无穷大。相反，对于稀有事件，更恰当的假设是，随着总样本量增加，正类比例应趋于零。

在这方面，Wang（2020）提出了一种新颖的理论框架，并对其进行了重新-

相应地，研究了所得估计量的统计性质。在其新颖的理论框架下，他证明了，全局最大似然估计量（GMLE）的收敛速率主要由正实例的数量而非总样本量决定。因此，GMLE 的收敛速率应显著慢于通常情形下的收敛速率。此外，Wang（2020）出人意料地发现，无论是欠采样方法还是过采样方法，都会导致参数估计中不必要的统计效率损失。那么，如何开发新的估计方法，从而获得统计有效的估计量，就成为一个至关重要的问题。值得注意的是，在本文余下部分，如果某个估计量与 GMLE 具有相同的渐近分布，我们就称其为统计有效的估计量。

值得一提的是，我们并不是最早研究大规模数据逻辑回归问题的研究者。既有文献已经取得了显著进展。一种可能的解决方案是子抽样。例如，Wang 等（2018）提出了一种子抽样方法，其设计依据是 Kiefer（1959）的 A-最优性准则。Wang（2019）进一步基于具有最优子抽样概率的子样本提出了更高效的估计量。Wang 等（2021）最近研究了具有不平衡二元响应的一般模型。另一种可能的解决方案是分布式计算，前提是能够使用并行计算系统。例如，Du 等（2018）提出了差分隐私方法，用于在多个参与方之间协作且准确地训练逻辑回归模型。Shi 等（2019）基于经典的 ADMM 算法（Boyd 等，2011）研究了分布式逻辑回归。Zuo 等（2021）提出了一种分布式子抽样程序，用于近似最大似然估计量。Wang 等（2016）针对线性 SVM 问题开发了一种代价敏感算法。尽管上述方法具有实用价值，但针对分布式

稀有事件数据分类问题及其严谨渐近理论的研究仍然非常有限。在缺乏坚实理论指导的情况下，我们无法为此提供统计效率高的估计量。这促使我们针对大规模稀有事件数据，开发一种具有坚实统计理论支持的新型分布式逻辑回归方法。

值得注意的是，为罕见事件下的逻辑回归开发一种分布式估计方法并非易事。我们至少面临以下两个具有挑战性的问题。第一个问题是分布式系统中数据在本地计算机上的分布。由于正样本总数远小于样本总量，在某些情况下，传统的纯随机数据分配策略可能并非最佳选择。例如，如果分配给本地机器的样本量很小，那么这种传统策略会导致每个分布式计算机节点上的正样本数量更少。这一过程会使从每台本地计算机获得的局部估计在统计上不准确，进而使最终合并得到的估计量在统计上效率低下。事实上，一种可能更好的选择是将所有正样本复制到每台本地计算机，再将负样本尽可能随机地分配给各台本地计算机。为方便起见，我们将传统的数据分配策略称为完全随机策略（RANDOM），将这一新策略称为复制策略（COPY）。因此，研究 RANDOM 和 COPY 两种策略下估计量的统计性质，便成为一个十分值得关注的问题。

第二个问题是目标函数的选择。如果采用 COPY 策略，每台本地计算机上的正负样本会变得更加均衡，从而使统计估计更加容易。然而，其副作用是，本地目标函数不再是全局对数似然函数的无偏估计。因此，尽管所得估计量仍然渐近正态，但其统计效率较低。这是 Wang（2020）发现的一个有趣现象。

为方便起见，我们将每台本地计算机上计算得到的这一估计量称为欠采样估计量。为解决这一问题，在每台本地计算机上提出了一种新型目标函数，该函数应当是全局目标函数的无偏估计。这自然引出了逆概率加权估计量（Fithian 和 Hastie，2014；Wang，2020）。随后，我们考虑如何获得分布式逻辑回归估计量。一种简单且常见的方法是对各本地计算机产生的估计量取平均。在文献中，这种方法被称为单轮（OS）方法（Zhang 等，2013；Rosenblatt 和 Nadler，2016；Chang 等，2017）。我们采用 OS 方法合并各本地的逆概率加权估计量，从而得到最终估计量，并将其称为逆概率加权估计量。

综上所述，我们旨在对现有文献做出以下重要贡献。首先，我们从理论上证明，传统的 RANDOM 分布式框架在许多情况下由于存在不可忽略的随机偏差项，无法在稀有事件数据上高效运行。其次，我们提出并严格研究了 COPY 策略。我们使用 US 类型的局部目标函数来构造 US 估计量。我们发现，若每个计算节点上的负例数量不足，US 估计量的偏差较低，但统计效率并不令人满意。最后，我们发现逆概率加权估计量在统计上比 US 估计量更有效，并且与 GMLE 具有相同的渐近行为。理论结果进一步通过大量数值研究得到验证。

本文其余部分安排如下。第 2 节介绍模型设定以及根据 Wang（2020）提出的三种重要基准估计方法。第 3 节介绍三种分布式估计方法及其渐近理论。第 4 节给出数值研究，并在此使用这三种分布式方法对瑞典交通标志数据进行应用分析。第 5 节对全文进行简要讨论并作出总结。所有技术细节均放在附录中。

2. L稀有事件数据的OGISTIC回归

2.1. 模型设定

假设总共有 N 个观测，其索引为 $1 \le i \le N$。第i个观测记为 $\big(X_i, Y_i\big)$，其中 $X_i \in \mathbb{R}^p$ 是一个 p维协变量， $Y_i \in \{0,1\}$ 是二元响应。假设 $\big(X_i, Y_i\big)$ 针对 $1 \le i \le N$独立生成，并将全部数据记为 $\mathcal{S}_F = \big\{\big(X_i, Y_i\big) : \ 1 \le i \le N\big\}$。令 $N_1 = \sum_{i=1}^{N} Y_i$ 为正例个数， $N_0 = N - N_1$ 为负例个数。为对其回归关系建模，考虑如下逻辑回归模型

$$P\big(Y_i = 1 \mid X_i\big) = p_i(\alpha, \beta) = \frac{e^{\alpha + X_i^\top \beta}}{1 + e^{\alpha + X_i^\top \beta}}, \tag{2.1}$$

其中 $\alpha \in \mathbb{R}$ 为截距， $\beta \in \mathbb{R}^p$ 为斜率参数。将 $\theta = (\alpha, \beta^\top)^\top \in \mathbb{R}^{p+1}$ 定义为完整参数向量，其真值由 $\theta^* = (\alpha^*, \beta^{*\top})^\top$给出。当 N发散至无穷时，若 θ^* 保持不变，正例数量将以 $O_p(N)$的速率发散。遵循Shao（2003），我们将 $O_p(\cdot)$定义如下。设$\{A_i\}$ 与 $\{B_i\}$ （其中 $1 \le i \le N$ ）为两个随机变量序列。则称$A_i = O_p(B_i)$，当且仅当对任意 $\varepsilon > 0$ 存在常数 $C_\varepsilon > 0$，使得$\sup_i P(\|A_i\| \ge C_\varepsilon \|B_i\|) < \varepsilon$。

在经典逻辑回归模型设定（2.1）下，现有理论表明，基于全数据的最大似然估计量（MLE） $\mathcal{S}_F$ 以 $O_p(N^{-1/2})$ 的速率收敛（Nelder and Wedderburn, 1972）。正如 Wang（2020）令人信服地论证的那样，这对于稀有事件数据的建模可能并非最佳选择。对于稀有事件数据，正例所占百分比极小。从统计角度看，更合适的做法是规定正响应率收敛到 0，随着

总样本量趋于无穷大。与此同时，我们希望协变量效应（由 β^*度量）保持不变，因为 X 的取值未知。否则，它在统计上无法被准确估计。因此，这表明我们应当用 α_N^*替换截距参数 α^*，且该参数应随着 $N \to \infty$发散至负无穷。具体而言，随着 $N \to \infty$，我们应当使 $\alpha_N^* \to -\infty$ 以适当的发散速率发散。然而，何为合理的发散速率需要更仔细的研究。在该假设下，随着$N \to \infty$，我们应当有 $P\big(Y_i = 1 \mid X_i\big) \approx e^{\alpha_N^* + X_i^\top \beta^*}$。于是我们有 $E(N_1) \approx N e^{\alpha_N^*} E\big(e^{X_i^\top \beta^*}\big)$。尽管正响应率（即 N_1/N）应随着 N 趋于无穷而收敛到 0，我们仍期望正例总数（即 N_1）发散至无穷。否则，我们无法一致地估计感兴趣的参数。这表明我们应当有

$$\alpha_N^* \to -\infty \quad \text{and} \quad \alpha_N^* + \log N \to \infty, \tag{2.2}$$

当 $N \to \infty$ 时。这成为所提出的理论框架中最重要的技术假设（Wang，2020）。

2.2. 相关方法

本小节根据 Wang（2020）介绍若干重要的基准估计方法。具体而言，我们将分别介绍全局最大似然估计、欠采样估计和逆概率加权似然估计。

全局最大似然估计

We 我们从使用 fu 的全局最大似然估计方法开始。 ll

数据。基于完整数据 $\mathcal{S}_F$ 的对数似然函数如下：

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \Big\{ Y_i \log p_i(\alpha_N, \beta) + (1 - Y_i) \log \big(1 - p_i(\alpha_N, \beta)\big) \Big\}, \tag{2.3}$$

其中 $p_i(\alpha_N, \beta) = e^{\alpha_N + X_i^\top \beta} / (1 + e^{\alpha_N + X_i^\top \beta})$。由此，我们可以得到 GMLE：

$\widehat{\theta}_{\mathrm{GMLE}} = \mathrm{argmax}_\theta \mathcal{L}(\theta)$。根据 Wang（2020）的定理 1，在适当条件下，GMLE $\widehat{\theta}_{\mathrm{GMLE}}$应当是 $\sqrt{N} e^{\alpha_N^*}$–一致且渐近正态的。这一结果表明，GMLE 的收敛速度完全由正例数量决定，这意味着额外增加大量负例所能带来的帮助有限。当负例总数过大，以至于难以在一台计算机上轻松处理时，这一结果尤为明显。

不过，我们应当指出，这绝不意味着大量负例对于提升效率毫无用处。大量理论和数值研究经验表明，即使收敛速度保持不变，更有效地利用负例仍可提升各种基准估计量的统计效率。然而，对于许多具有稀有事件的实际数据集，负例总数往往过大，以至于难以在一台计算机上轻松处理。在这种情况下，如何更有效地利用负例以提高估计效率，便成为一个备受关注的问题。

欠采样估计

在实践中，研究人员通常希望将所有正例纳入统计分析，因为正例较为罕见，因此很有价值（Drummond 等，2003；Liu 等，2008；Nguyen 等，2012）。随后，随机选取数量相同（或相近）的负例，以构建一个更加均衡的子样本。

随后，可以基于这个更加均衡的子样本估计感兴趣的参数。为方便起见，我们将这种常见做法称为欠抽样方法（Drummond 等，2003；Liu 等，2008；Nguyen 等，2012；Wang，2020）。这样一来，估计问题在计算上变得可行。从理论上讲，该问题可以表述如下。设 a_i 为二元指示变量，具有 $P(a_i = 1) = \pi$，并针对每个 i独立生成。这里， $a_i = 1$ 表示第 i 个实例被抽样， π 为抽样概率。因此，US 目标函数（2.1）变为

$$\mathcal{L}_{\mathrm{US}}(\theta) = \sum_{i=1}^{N} \Big\{ Y_i \log p_i\big(\alpha_N, \beta\big) + \big(1 - Y_i\big) a_i \log\big(1 - p_i(\alpha_N, \beta)\big) \Big\}. \tag{2.4}$$

为方便起见，我们称其为 US 目标函数。然后，我们得到一个 US 估计量：$\widehat{\theta}_{\mathrm{US}} = \mathrm{argmax}_\theta \mathcal{L}_{\mathrm{US}}(\theta)$。然而，Wang（2020）发现， θ_{US} 是针对 θ^* 的有偏估计量。因此，进一步得到去偏 US 估计量： $\widetilde{\theta}_{\mathrm{US}} = \widehat{\theta}_{\mathrm{US}} + (\log \pi, 0, \cdots, 0)^\top$。

比较（2.4）和（2.3）可知，唯一的区别在于对负样本的处理。对于（2.3），无论正样本还是负样本，都会使用全部样本。而对于（2.4），我们使用全部正样本；只有当相应的二元指示变量 $a_i = 1$ 时，才纳入负样本。这样一来，所有正样本都会被纳入，而使用的负样本数量则少得多。容易验证，这种形式在数学上等价于 Wang（2020）提出的形式。Wang（2020）严谨的理论分析表明，这种估计量仍然是 $\sqrt{N e^{\alpha_N^*}}$-相合的，并且渐近服从正态分布。然而，正如 Wang（2020）的定理 3 所示，如果正样本与负样本的比例不趋于零，US 估计量就无法达到与 GMLE 相同的效率。

———

US 估计量在统计上效率较低，关键原因在于（2.4）中的目标函数。通过欠抽样，得到的目标函数发生了实质性改变。直接后果是，它不再是总体对数似然函数的无偏估计量，因此导致 US 估计量效率低下。为解决这一问题，一种可行的方法是寻找总体对数似然函数的无偏估计量。这就引出了逆概率加权估计的如下目标函数（King 和 Zeng，2001；Fithian 和 Hastie，2014；Wang，2020）：

$$\mathcal{L}_{\mathrm{IPW}}(\theta)=\sum_{i=1}^{N}\Big\{Y_i\log p_i(\alpha_N,\beta)+(1-Y_i)a_i\log\big(1-p_i(\alpha_N,\beta)\big)/\pi\Big\}. \tag{2.5}$$

很容易验证，$E\{\mathcal{L}_{\mathrm{IPW}}(\theta)|\mathcal{S}_F\}=\mathcal{L}(\theta)$，这表明 $\mathcal{L}_{\mathrm{IPW}}(\theta)$ 是全局对数似然函数的无偏估计量。通过优化上述目标函数，可以得到逆概率加权估计量：$\theta_{\mathrm{IPW}}=\mathrm{argmax}_\theta\mathcal{L}_{\mathrm{IPW}}(\theta)$。Wang（2020）证明，逆概率加权类型估计量具有与 $\widehat{\theta}_{\mathrm{GMLE}}$ 相同的收敛速率 $O_p\big(1/\sqrt{Ne^{\alpha_N^*}}\big)$，但在统计上仍然是低效的。回顾一下，在本文中，我们将与 GMLE 具有相同渐近分布的估计量定义为统计有效估计量。

———

US 估计量和逆概率加权估计量的效率均不理想，这是可以理解的，因为这两种方法在估计时都只纳入了极少一部分负类样本。因此，应当有很大的可能性通过使用更多的负类样本（但不必使用全部负类样本）来提高统计效率。如果具备强大的分布式计算系统，这似乎尤其是一个很有前景的方向。借助分布式系统，我们应该能够多次计算各种局部估计量（例如 US 估计量和逆概率加权估计量）

。随后可以将它们聚合起来，形成一个更强大的估计量。然而，应计算哪种局部估计量，以及应如何组合这些估计量，才能使最终估计量达到与 GMLE 同等的效率，这些都是值得深入探讨的问题。因此，我们将在接下来的章节中系统地研究这些有趣的问题。

3. 分布式逻辑回归

3.1. 随机策略下的分布式极大似然估计

我们从最简单的分布式估计量开始，即采用随机策略得到的分布式极大似然估计量。为方便起见，将其称为限制极大似然估计（RMLE）。假设存在一个分布式计算系统，其中有 K 台本地计算机和一台中央计算机。分布式系统的典型架构如图 2 所示。本地计算机的索引为 $1 \leq k \leq K$。随后，RMLE 方法将完整数据 $\mathcal{S}_F$ 随机分配到各台本地计算机上，使每台计算机获得的数据量大致相等。记为 $\mathcal{S}_F = \mathcal{S}_+ \cup \mathcal{S}_-$，其中 $\mathcal{S}_+ = \{i: \ Y_i = 1\}$ 表示所有正例的集合，而 $\mathcal{S}_- = \{i: \ Y_i = 0\}$ 表示所有负例的集合。具体而言，令 $\mathcal{S}_k^R$ 为随机分配到第k台本地计算机的样本，其中 $\mathcal{S}_k^R = \mathcal{S}_{k+}^R \cup \mathcal{S}_{k-}^R$，$\mathcal{S}_{k+}^R = \{i: \ i \in \mathcal{S}_k^R, \ Y_i = 1\}$ 和$\mathcal{S}_{k-}^R = \{i: \ i \in \mathcal{S}_k^R, \ Y_i = 0\}$ 分别表示第k台本地计算机上的正例集合和负例集合。为方便起见，记为 $n_k = |\mathcal{S}_k^R|$。此外，令$n_{1k}^R = |\mathcal{S}_{k+}^R|$ 和 $n_{0k}^R = |\mathcal{S}_{k-}^R|$。数学上，若第 i 个观测被随机分配到第 k 台本地计算机，则记为 $a_i^{(k)} = 1$ 。于是，对于每个 i、$n_k = \sum_{i=1}^N a_i^{(k)}$ ，都有 $\sum_{k=1}^K a_i^{(k)} = 1$ 。我们还定义 $n = E\left(n_k\right) = N/K$。此外，还有$n_{1k} = \sum_{i=1}^N a_i^{(k)} Y_i$和 $n_{0k} = \sum_{i=1}^N a_i^{(k)} \left(1 - Y_i\right)$ 。

可以看出，采用 RANDOM 策略时，正例和负例都会随机分配到各台本地计算机上。因此，它们的相对

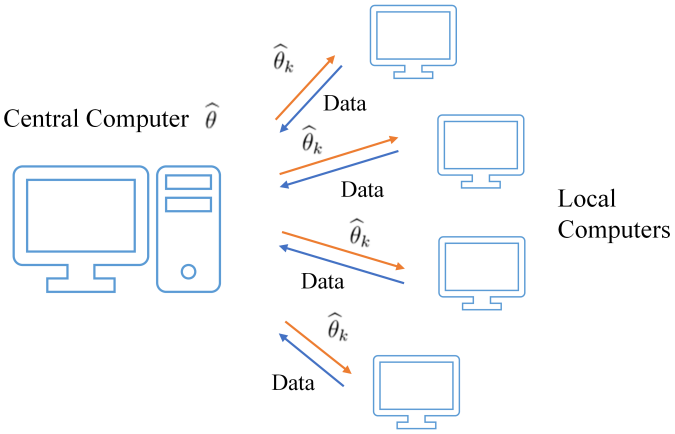

图 2：分布式系统示意图。

比例与完整数据集中的比例大致相同。也就是说，$n_{1k}/n_k \approx N_1/N$。这种方法的优点是，各台本地计算机上的数据分布与完整数据集保持一致。但其缺点是，分配到各台本地计算机上的正例数量会进一步减少，这可能导致所得估计量在统计上效率低下。具体而言，对于每台本地计算机，定义

$$\mathcal{L}_{\mathrm{R},k}(\theta) = \sum_{i=1}^{N} a_i^{(k)} \Big\{ Y_i \log p_i(\alpha_N, \beta) + (1 - Y_i) \log \big(1 - p_i(\alpha_N, \beta)\big) \Big\} \tag{3.1}$$

作为一个局部对数似然函数，其中包含 $P(a_i^{(k)} = 1) = 1/K$。然后，局部 MLE 计算为 $\widehat{\theta}_{\mathrm{RMLE},\,k} = \mathrm{argmax}_\theta \mathcal{L}_{\mathrm{R},\,k}(\theta)$。随后，每台本地计算机都应将该局部估计量报告给中央计算机。接下来，中央计算机汇总这些估计量，形成一个更强大的估计量。为实现这一目标，一种典型的汇总方案是单步策略（Zhang 等，2013；Chang 等，2017）。具体而言，最终估计量为 $\widehat{\theta}_{\mathrm{RMLE}} = \sum_{k=1}^{K} \widehat{\theta}_{\mathrm{RMLE},\,k}/K$。$\widehat{\theta}_{\mathrm{RMLE}}$ 的渐近分布见以下定理。

定理 1. 假设 *(C1)* $P(\|Z_i\| > M) \le 2\exp(-C_{\mathrm{Tail}} M^2)$ 满足 $Z_i = (1, X_i^\top)^\top \in \mathbb{R}^{p+1}$，其中 C_{Tail} 为某个正常数；*(C2)* $n \to \infty$ 当 $N \to \infty$ 时；*(C3)* $\log^2 N/(ne^{\alpha_N^*}) =$

$O(1)$以及 *(2.2)*。那么，我们有如下渐近表示式

$$\sqrt{Ne^{\alpha_N^*}}\big(\widehat{\theta}_{\mathrm{RMLE}}-\theta^*\big)=P_1+P_2/2+o_p\big(K/\sqrt{Ne^{\alpha_N^*}}\big),$$

其中 $P_1=-(Ne^{\alpha_N^*})^{1/2}K^{-1}\sum_{k=1}^{K}\ddot{\mathcal{L}}^{-1}{}_{\mathrm{R}\,k}(\theta^*)\dot{\mathcal{L}}_{\mathrm{R},k}(\theta^*)$ 以及 $P_2=(Ne^{\alpha_N^*})^{-1/2}KB(\theta^*)$。这里 $B(\theta^*)$ 是一个随机偏差项，满足 $C_{\min}\le E\{\|B(\theta^*)\|\}\le C_{\max}$，其中某些固定正常数为 $0<C_{\min}<C_{\max}<\infty$。

定理条件 (C1) 要求协变量分布的尾概率呈指数衰减（Zhang 和 Chen，2020）。定理条件 (C2) 意味着，每台本地计算机上的实例期望数量 $n=E(n_k)$ 应随着总样本量 $N\to\infty$ 趋于无穷大而发散。与此同时，我们有$E(n_{1k})=E(\sum_{i=1}^{N}a_i^{(k)}Y_i)=ne^{\alpha_N^*}E\{e^{X_i^\top\beta^*}/(1+e^{Z_i^\top\theta^*})\}$。根据条件 (C3)，我们要求本地计算机上的正例数量足够大。由定理 1 可知，$\sqrt{Ne^{\alpha_N^*}}\big(\widehat{\theta}_{\mathrm{RMLE}}-\theta^*\big)$ 可分解为三部分。第一部分是 P_1，其中 $P_1\to_d N(0,\Sigma^{*-1})$ 随着 $N\to\infty$。第二部分P_2是一个随机偏差项，其阶为 $K/(Ne^{\alpha_N^*})$； $B(\theta^*)$ 的解析表达式见附录 A.1。第三部分是相较于 P_2而言阶数更高、可忽略的项。若 K 足够小，即满足 $K/\sqrt{Ne^{\alpha_N^*}}\to 0$随着 $N\to\infty$，则应有 P_1 成为主导项。在这种情况下， θ_{RMLE} 与 Wang（2020）的 θ_{GMLE} 具有相同的渐近分布。否则，应由 $P_2/2$ 作为主导项。这会导致 θ_{RMLE} 的统计效率较低。

3.2. 采用无加权目标函数的欠采样

接下来，我们研究通过欠采样得到的分布式估计量的渐近性质。首先考察无权损失函数 (2.4) 所采用的 θ_{US}。为了获得

US 估计量，我们通过 COPY 策略将完整数据 $\mathcal{S}_F$ 分发到每台本地计算机。令 $\mathcal{S}_k^C$ 为 COPY 策略下分发给第 k 台本地计算机的样本。记 $\mathcal{S}_k^C = \mathcal{S}_{k+}^C \cup \mathcal{S}_{k-}^C$，其中 $\mathcal{S}_{k+}^C$ 和 $\mathcal{S}_{k-}^C$ 分别表示第 k台本地计算机上的正例和负例。对于 COPY 策略，当 $1 \le k \le K$ 时，我们有$\mathcal{S}_{k+}^C = \mathcal{S}_+$，这意味着所有本地计算机上的正例都相同。显然，COPY 策略的优点是分配给每台本地计算机的正例数远多于 RANDOM 方法。随后，将负例随机分配到各台本地计算机，使得 $\cup_k \mathcal{S}_{k-}^C = \mathcal{S}_-$，且对于任意 $k_1 \neq k_2$，都有 $\mathcal{S}_{k_1-}^C \cap \mathcal{S}_{k_2-}^C = \emptyset$。令 $n_{1k}^C = |\mathcal{S}_{k+}^C|$ 和 $n_{0k}^C = |\mathcal{S}_{k-}^C|$。我们通常要求 $n_{1k}^C = O_p(n_{0k}^C)$。换言之，分配给每台本地计算机的负例数不应远少于正例数，这在实践中也是最常见的情况。

随后，为每台本地计算机定义一个局部 MLE：$\widehat{\theta}_{\mathrm{US},\,k} = \mathrm{argmax}_\theta \mathcal{L}_{\mathrm{US},\,k}(\theta)$，其中

$$\mathcal{L}_{\mathrm{US},k}(\theta) = \sum_{i=1}^{N} \Big\{ Y_i \log p_i\big(\alpha_N, \beta\big) + \big(1 - Y_i\big) a_i^{(k)} \log\big(1 - p_i(\alpha_N, \beta)\big) \Big\}.$$

并回顾一下，$a_i^{(k)} = 1$ 如果第 i 个实例被分配给第 k 台本地计算机。于是，每个工作节点上的$\widehat{\theta}_{\mathrm{US},\,k}$ 等价于 Wang（2020）提出的欠采样估计量。完成局部估计后，每台本地计算机将局部估计量 $\widehat{\theta}_{\mathrm{US},\,k}$ 发送到中央计算机。类似地，采用 one-shot 策略可得到最终估计量 $\widehat{\theta}_{\mathrm{US}} = \sum_{k=1}^{K} \widehat{\theta}_{\mathrm{US},\,k}/K$。接下来，我们将在以下定理中分析 $\widehat{\theta}_{\mathrm{US}}$ 的渐近性质。

定理 2. 假设定理 *1* 中的条件同样成立。定义 $\flat = (\log K, 0, \cdots, 0)$以及 $\Sigma_2^* = E\big\{\big(1 + \gamma e^{X_i^\top \beta^*}\big)^{-1} e^{X_i^\top \beta^*} Z_i Z_i^\top\big\}$，其中 $\gamma = \lim_{N\to\infty} K e^{\alpha_N^*} \in [0, \infty)$。于是

具有如下渐近表示形式

$$\sqrt{Ne^{\alpha_N^*}}\big(\widehat{\theta}_{\mathrm{US}}-\theta^*-\flat\big)=\big(Ne^{\alpha_N^*}\big)^{-1/2}\Sigma_2^{*-1}K^{-1}\sum_{k=1}^{K}\dot{\mathcal{L}}_{\mathrm{US},k}\big(\theta^*+\flat\big)+o_p(1).$$

根据定理 2，我们知道 $\sqrt{Ne^{\alpha_N^*}}\big(\widehat{\theta}_{\mathrm{US}}-\theta^*-\flat\big)$ 可分解为两部分。对于第一部分，有 $\big(Ne^{\alpha_N^*}\big)^{-1/2}\Sigma_2^{*-1}K^{-1}\sum_{k=1}^{K}\dot{\mathcal{L}}_{\mathrm{US},k}\big(\theta^*+\flat\big)\to_d N\big(0,\Sigma_2^{*-1}\Sigma_1^*\Sigma_2^{*-1}\big)$，因为 $N\to\infty$，其中 $\Sigma_1^*=E\big\{(1+\gamma e^{X_i^\top\beta^*})^{-2}e^{X_i^\top\beta^*}Z_iZ_i^\top\big\}$。第二部分是一个高阶无穷小项。这里可以建立渐近正态性，因为 $\big(Ne^{\alpha_N^*}\big)^{-1/2}\Sigma_2^{*-1}K^{-1}\sum_{k=1}^{K}\dot{\mathcal{L}}_{\mathrm{US},k}\big(\theta^*+\flat\big)$ 可以写成一组经过仔细定义的独立随机变量之和；详情见附录 A.2 中定理 2 的步骤 4。因此，可以直接应用 Lindeberg-Feller 中心极限定理。由此，$\widetilde{\theta}_{\mathrm{US}}=\widehat{\theta}_{\mathrm{US}}-\flat$ 对 θ^*是 $\sqrt{Ne^{\alpha_N^*}}$-相合的。将这一结果与定理 1 的结果比较，我们发现了一些有趣的差异。首先，截距项必须包含一个额外的偏差校正项 $\flat$。这主要是由 US 设定下数据分布的扭曲所致。其次，我们发现，当 K 较大时，US 估计量的偏差小于 RMLE 估计量的偏差。这主要是因为，采用 COPY 策略计算的局部估计量的偏差小于采用 RANDOM 策略计算的局部估计量的偏差。

———

我们进一步讨论同时出现在 Σ_1^* 和 Σ_2^*中的常数 γ。正如 Wang（2020）所指出的，可以验证 $\gamma E(e^{X_i^\top\beta^*})\approx N_1/(N_0/K)$ 渐近成立，其中 N_0/K 表示每台本地计算机上的负类样本数。因此，$\gamma E(e^{X_i^\top\beta^*})$ 渐近地表示正类样本数与负类样本数之比。若 $\gamma=0$，则负类样本数占主导。因此，我们有 $\Sigma_2^{*-1}\Sigma_1^*\Sigma_2^{*-1}=\Sigma^{*-1}$。这意味着$\widehat{\ }$US 估计量与 GMLE θ_{GMLE}具有相同的渐近协方差矩阵。若 $0<\gamma<\infty$，则正类和负类样本数大致相当。这

意味着，与$\widehat{\ }$GMLE θ_{GMLE}相比，US 估计量在统计上效率较低。这一发现也与 Wang（2020）中的定理 2 一致。我们不考虑 $\gamma = \infty$，因为这意味着正类样本数远多于负类样本数。

3.3. 使用加权目标函数进行欠采样

第 3.1 节和第 3.2 节的分析表明，限制极大似然估计（RMLE）和 US 估计量都无法达到全局渐近效率。RMLE 未能达到这一效率，是因为分配给每台本地计算机的正样本数量过少。US 估计量未能达到这一效率，是因为每台本地计算机使用的 US 目标函数并非全局目标函数的无偏估计。因此，我们受到启发，决定构造一个新的本地对数似然函数，使其成为全局对数似然函数的无偏估计。同时，每台本地机器都应使用所有正样本。为此，我们提出如下逆概率加权（IPW）估计量。具体而言，我们仍通过 COPY 策略将完整数据 $\mathcal{S}_F$ 分发给每台本地计算机。接下来，我们为每台本地计算机定义一个局部 MLE$\widehat{\ }$，其形式为 $\theta_{\mathrm{IPW},\ k}$ =argmax$_\theta \mathcal{L}_{\mathrm{IPW},\ k}(\theta)$，其中

$$\mathcal{L}_{\mathrm{IPW},k}(\theta) = \sum_{i=1}^{N} \Big\{ Y_i \log p_i\big(\alpha_N, \beta\big) + K\big(1 - Y_i\big) a_i^{(k)} \log\big(1 - p_i(\alpha_N, \beta)\big) \Big\}.$$

$\widehat{\ }$因此，在每个工作节点上，$\theta_{\mathrm{IPW},k}$ 可视为 Wang（2020）提出的欠抽样加权估计量，该估计量也在（2.5）中给出。可以立即验证，$E\{\mathcal{L}_{\mathrm{IPW},k}(\theta)|\mathcal{S}_F\} = \mathcal{L}\big(\theta\big)$，其中回顾一下，$\mathcal{S}_F = \big\{\big(X_i, Y_i\big) :\ 1 \le i \le N\big\}$ 表示$\widehat{\ }$完整数据。然后，每台本地计算机将此本地估计量 $\theta_{\mathrm{IPW},k}$ 发送给中央计算机。类似地，采用单次通信策略，我们得到最终估计量$\widehat{\theta}_{\mathrm{IPW}} = \sum_{k=1}^{K} \widehat{\theta}_{\mathrm{IPW},k}/K$。如前所述，$\mathcal{L}_{\mathrm{IPW},k}(\theta)$ 现在是$\widehat{\ }$全局对数似然函数的无偏估计量，我们期望 θ_{IPW} 达到相同的渐近

效率，与 GMLE 相同。为此，我们在以下定理中分析 θ_{IPW} 的渐近性质。

定理 3. 假设定理 *1* 中的条件相同，则我们得到如下渐近表示：

$$\sqrt{Ne^{\alpha_N^*}}\left(\widehat{\theta}_{\mathrm{IPW}}-\theta^*\right)=\left(Ne^{\alpha_N^*}\right)^{-1/2}\Sigma^{*-1}\dot{\mathcal{L}}\left(\theta^*\right)+o_p(1).$$

根据定理 3，我们知道 $\sqrt{Ne^{\alpha_N^*}}\left(\widehat{\theta}_{\mathrm{IPW}}-\theta^*\right)$ 可以分解为两部分。对于第一部分，有 $\left(Ne^{\alpha_N^*}\right)^{-1/2}\Sigma^{*-1}\dot{\mathcal{L}}\left(\theta^*\right)\to_d N\left(0,\ \Sigma^{*-1}\right)$，其中 $N\to\infty$。第二部分的阶为 $o_p(1)$，属于高阶可忽略项。因此，$\widehat{\theta}_{\mathrm{IPW}}$ 是关于 θ^* 的 $\sqrt{Ne^{\alpha_N^*}}$-相合估计量。将这一结果与 Wang (2020) 中 GMLE 的结果进行比较，我们发现 θ_{IPW} 与 GMLE 具有相同的渐近分布。将其与定理 2 中 US 估计量的结果进行比较，我们发现，US 估计量在本地计算机上使用 US 目标函数 (2.4)，会对正例赋予过高的权重。然而，逆概率加权估计量在本地计算机上使用逆概率加权目标函数 (2.5)，对正例和负例赋予相同的权重。这是逆概率加权估计量优于 US 估计量的关键原因。特别是，定理 2 中给出的 γ 并未涉及其中；它表示正例与负例的渐近比例。因此，与 US 估计量（或 Wang (2020) 中的欠采样估计量）相比，我们无需 $\gamma=0$ 即可达到全局效率。我们大量的数值研究也表明，θ_{IPW}具有更好的有限样本表现。总而言之，具有较大 K 的 RMLE 估计量 θ_{RMLE} 会受到显著偏差的影响。如果每个计算节点上分布的负例数量不足，去偏 US 估计量 $\widetilde{\theta}_{\mathrm{US}}$ 也会因渐近协方差较大而在统计上效率低下。逆概率加权估计量 θ_{IPW} 则是最具吸引力的估计量。

值得注意的是，Wang (2020) 研究的 US 估计量和逆概率加权估计量，都不同于我们工作中研究的相应估计量。具体而言，Wang (2020) 中的这两种估计量基于一个子样本，该子样本包含所有正例，但只包含一小部分负例。这样可以节省大量计算成本。在这种情况下，Wang (2020) 发现 US 估计量比逆概率加权估计量更有效率。然而，我们工作中研究的 US 估计量和逆概率加权估计量均基于完整样本，只是以分布式方式进行计算。因此，对于我们的估计量，不仅所有正例都得到了充分利用，所有负例也同样如此。事实上，由于采用了 COPY 策略，所有正例都会被不同的本地计算机重复使用。相比之下，Wang (2020) 只使用了很小一部分负例。这使得我们研究的 US 估计量和逆概率加权估计量的理论性质与 Wang (2020) 中的估计量大不相同。这也是造成两组估计量性能差异的关键原因。

4. 数值研究

4.1. 模拟研究

模型设置与性能度量

为展示所提出方法在有限样本下的表现，本节进行若干项模拟研究。使用标准逻辑回归模型 (2.1) 生成包含协变量 $Z_i = (1, X_i^\top)^\top \in \mathbb{R}^5$ 的完整数据。这里，协变量 X_is 从 $N(0, \Sigma)$ 生成，其中 $\Sigma = (\sigma_{ij})$ 和 $\sigma_{ij} = 0.2^{|i-j|}$。总样本量为 $N = 10^4$, $10_5, 5 \times 10^5$ 和 10^6。对于固定的 N，我们设定 $\alpha_N^* = -0.45 \log N$ 和 $\beta^* = (1, 1, 1, 1)^\top$。这样做可以使 $P(Y = 1) \to 0$ 和 $E\big(N_1\big) \to \infty$ 作为 $N \to \infty$。接下来，我们将本地计算机的数量（即 K）设为

两种不同的情况。对于情况 1，在四种不同的样本量下，我们分别设定 $K = 17$、36、63、81。可以验证，对于 COPY 策略，每台本地计算机上的正实例数量约为负实例数量的 1.5 倍。相比之下，对于情况 2，我们相应地设定 $K = 2$、3、4、5。这样做后，COPY 策略分配给每台本地计算机的负实例数量应远大于正实例数量。接下来，考虑两种不同的数据分配策略（即 RANDOM 和 COPY）。然后 $\widehat{\ }$ $\widetilde{\ }$ $\widehat{\ }$为每台本地机器 k 得到三个局部估计量 $\theta_{\mathrm{RMLE}、k}$、 $\theta_{\mathrm{US}、k}$ 和 $\theta_{\mathrm{IPW}、k}$。$\widetilde{\ }$ $\widehat{\ }$这里，$\theta_{\mathrm{US}、k}$ 和 $\theta_{\mathrm{IPW}、k}$ 可视为$\widehat{\ }$ $\widetilde{\ }$ $\widehat{\ }$(Wang, 2020) 提出的欠抽样估计量。这使得我们在中央计算机上得到组合估计量 θ_{RMLE}、θ_{US} 和 θ_{IPW}。对于 US 方法，我们使用去偏估计量 θ_{US} $\widehat{\ }$ $\widehat{\ }$（而非 θ_{US}）作为最终估计量。为便于比较，也计算 GMLE θ_{GMLE}。为确保评估可靠，每项实验均随机重复进行 $M = 500$ 次。令 $\widehat{\theta}^{(m)} = (\theta_j^{(m)}: 1 \le j \le p+1)^\top$ 表示在第 m 次重复中得到的一个特定估计量$\widehat{\ }$ $\widehat{\ }$（例如，对于 $k = 1$ 或 θ_{RMLE}，$\theta_{\mathrm{RMLE}、k}$）。为评估估计精度，我们计算均方根误差（RMSE）：RMSE $= (p+1)^{-1} \sum_{j=1}^{p+1} \{M^{-1} \sum_{m=1}^{M} (\widehat{\theta}_j^{(m)} - \theta_j^*)^2\}^{1/2}$。然后根据其理论公式数值$\widehat{\ }$计算 $\widehat{\theta}_{\mathrm{GMLE}}$ 的 RMSE。此外，通过 BIAS$= (p+1)^{-1} \sum_{j=1}^{p+1} |\bar{\theta}_j - \theta_j^*|$ 估计 θ 的绝对偏差，其中 $\bar{\theta}_j = M^{-1} \sum_{m=1}^{M} \widehat{\theta}_j^{(m)}$。通过 SE $= (p+1)^{-1} \sum_{j=1}^{p+1} \{M^{-1} \sum_{m=1}^{M} (\widehat{\theta}_j^{(m)} - \bar{\theta}_j)^2\}^{1/2}$ 估计 $\widehat{\theta}$ 的标准误（SE）。

模拟结果

详细结果见图 3。这里我们研究局部估计量和分布式估计量。我们考虑了关于本地机器数量的两种不同情况（即情况 1 和情况 2）。由此得到四种组合，分别显示在不同的子图中。图 3 的纵轴表示

均方根误差的对数值。横轴表示总样本量，同样采用对数刻度。首先，左上子图展示了情况 1 下局部估计量的结果。在这种情况下，所研究的所有估计量的效率都远低于 GMLE，因为各种估计量的 log(RMSE) 值远大于 GMLE 的值。这是因为其他估计量（即 $\theta_{\mathrm{RMLE},k}$、 $\theta_{\mathrm{US},k}$ 和 $\theta_{\mathrm{IPW},k}$）都是局部估计量。这里 $\theta_{\mathrm{US},k} = \theta_{\mathrm{US},k} - \flat$ 是偏差校正估计量，其参数为$\flat = (\log K、0、\cdots、0)$。这些估计量所用的样本量远小于全局估计量所用的样本量。因此，它们的效率预计低于全局估计量。然而，在所有局部估计量中，我们发现 $\theta_{\mathrm{RMLE},k}$ 的表现始终最差。这是预期的，因为 RMLE 估计量所用的正例数量远少于其他局部估计量。相比之下，我们发现$\theta_{\mathrm{US},k}$ 的表现优于 $\theta_{\mathrm{IPW},k}$。这些观察结果与 Wang（2020）的结论一致。

图 3 右上方的面板展示了 CASE 2 中局部估计量的结果。与左上方面板相比，我们发现 GMLE 仍然是最佳估计量。然而，在所有局部估计量中，性能差异明显更小。这是因为在这种情况下，分配给本地机器的负样本数量足够多。因此，所有局部估计量的性能都趋近于全局估计量，其相对差异也随之消失。

左下图展示了情况 1中分布式估计量的 log(RMSE) 值。我们发现，与局部估计量（例如 $\theta_{\mathrm{US,k}}$ 和$\theta_{\mathrm{IPW,k}}$ 由（Wang，2020）提出的估计量）相比，分布式估计量（例如 θ_{IPW}）的表现有所改善，尤其是在欠采样的负实例数量不足时。例如，在左上图中，当 $N = 10^4$ 时，$\theta_{\mathrm{US,k}}$ 的均方根误差为 0.094，而$\theta_{\mathrm{IPW,k}}$ 的均方根误差为 0.121。作为比较， $\widehat{\theta}_{\mathrm{IPW}}$ 的均方根误差为

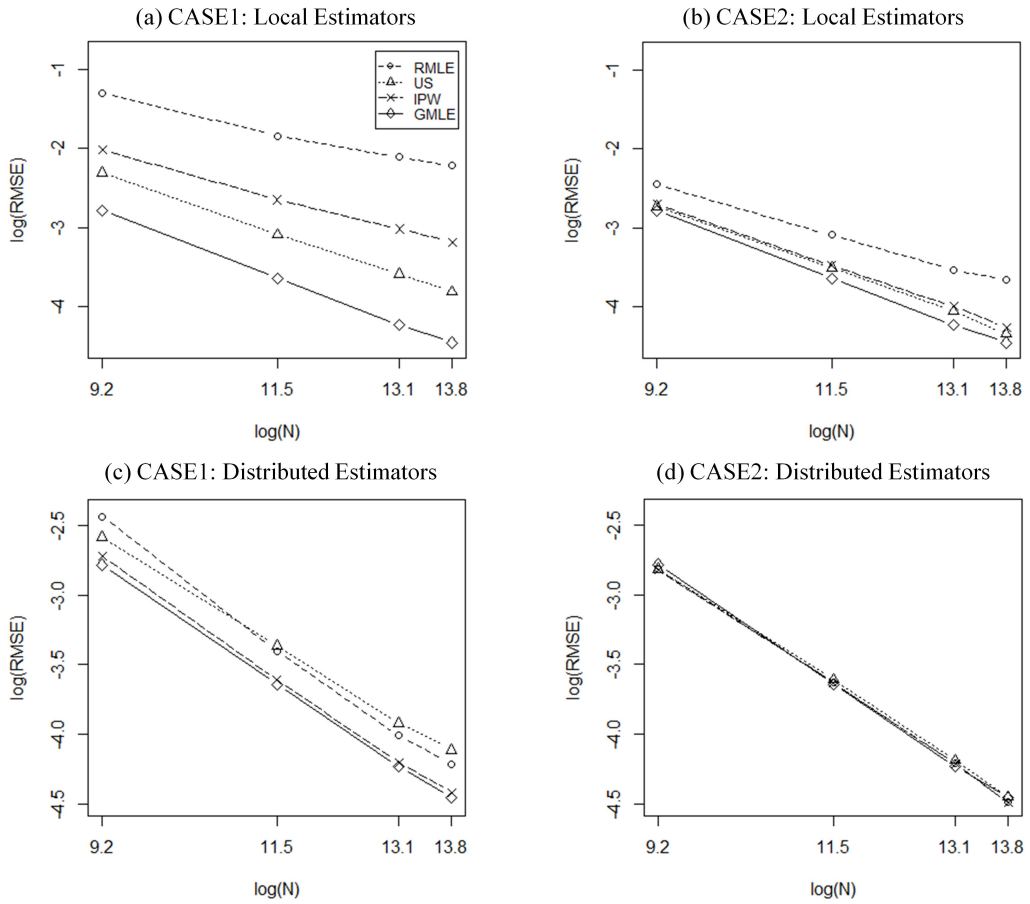

图3：局部估计量与分布式估计量的均方根误差（对数尺度）。横轴以对数尺度呈现总样本量 N 。上方各面板展示局部估计量。下方各面板展示分布式估计量。左侧各面板展示正例数量约为负例数量1.5倍的情形。右侧各面板展示负例数量远大于正例数量的情形。

0.065，与 GMLE 的均方根误差（即 0.061）接近。这表明，逆概率加权估计量对正负实例的比例不太敏感。在所有~分布式估计量中，我们发现去偏 US 估计量 θ_{US} 似乎是最差的估计量，因为其对应的 log(RMSE) 值始终最大。^相比之下，逆概率加权估计量 θ_{IPW} 表现最佳。随着分配给每台本地机器的负实例数量增加，不同分布式估计量之间的相对差异逐渐消失。右下图的结果体现了这一点。有关图 3(c) 的更多详细结果见表~1。根据表 1，我们发现，情况 1中的 RMLE 估计量 θ_{RMLE} 表现出一种

由于 K 相对较大，因此偏差较大。同时，去偏 US估计量$\sim\theta_{\mathrm{US}}$ 的标准误值较高。这些观察结果与所提出的定理1–2的理论发现一致。

表 1：CASE 1 下分布式估计量的模拟结果。

		RMLE			US			IPW		
N	$\sqrt{Ne^{\alpha_N^*}}$	BIAS	SE	RMSE	BIAS	SE	RMSE	BIAS	SE	RMSE
10^4	13	0.055	0.067	0.088	0.005	0.075	0.076	0.018	0.063	0.066
10^5	24	0.019	0.027	0.033	0.001	0.035	0.035	0.006	0.026	0.027
5×10^5	37	0.010	0.015	0.018	0.000	0.020	0.020	0.003	0.015	0.015
10^6	45	0.008	0.012	0.015	0.001	0.016	0.016	0.003	0.012	0.012

4.2. 瑞典交通标志数据分析

数据处理

为便于说明，我们给出一个有趣的真实数据示例。本研究使用的数据集是瑞典交通标志（STS）数据集，可从 *https://www.cvl.isy.liu.se/research/datasets/traffic-signs-dataset/* 公开获取。该数据集共包含 1,970 张经过标注的图像，其中标注了各种交通标志，并用边界框标出；图示见图 1。我们的目标是自动检测图 1 中的交通标志。为可靠评估，我们将全部数据随机分为两部分。第一部分包含 1,576 张图像（约占全部数据的 80%），用于训练；其余 394 张图像（约占全部数据的 20%）用于测试。这项任务包含两个重要步骤；参见 Girshick 等人（2014）；Girshick（2015）。第一步，需要在没有边界框信息的情况下，从输入图像中自动检测出一个足够紧致且包含交通标志的局部区域。第二步，需要将局部区域中检测到的交通标志分类为不同类别（例如，禁令标志、指示标志、警告标志和强制标志）。在本研究中，我们

重点关注第一步。随后，我们将说明如何把这项任务转化为一个逻辑回归问题。该问题的样本量较大，并且可以通过我们提出的方法以分布式方式高效求解。

具体而言，STS 数据集中的每张图像分辨率都相对较高；见图 4(a)。从数学上看，每张图像都可以表示为一个大小为$960 \times 1,280 \times 3$的张量；见图 4(b)。接下来，我们将预训练的 VGG16 模型应用于图像（Simonyan 和 Zisserman，2014）。VGG16 是一种经典的卷积神经网络模型，共有 13 个卷积层。最后两个全连接层被移除。随后，可以从最后一个卷积层提取一个大小为 $30\times40\times512$的特征图；见图 4(c)。这可以看作一张分辨率为 30×40、但共有 512 个通道的新“图像”。然后，我们将该特征图中的每个像素视为一个样本。因此，每张图像总共可以生成 $30 \times 40 = 1{,}200$ 个像素样本。对于每个像素样本，都可以构造一个维度为 512 的特征向量。因此，在这种情况下，我们有 $p = 512$ 。第 i 张图像记为 $\mathbb{X}_{i,k_1,k_2} \in \mathbb{R}^{512}$，其中 $1 \leq i \leq N$、$1 \leq k_1 \leq 30$ 和 $1 \leq k_2 \leq 40$；见图 4(d)。于是，总样本量为 $N = 1,970 \times 1,200 = 2{,}364{,}000$。

接下来，我们详细介绍响应 $\mathbb{Y}_{i,k_1,k_2} \in \{0,1\}$ 的构造方法。定义 $W_i = (W_{i,k_1,k_2})$ 为一个维度为 $960 \times 1{,}280$ 的二值矩阵，并且 $W_{i,k_1,k_2} \in \{0,1\}$。对于给定的图像及其边界框信息，定义 $W_{i,k_1,k_2} = 1$ ：若第 (k_1, k_2) 个像素位于边界框区域内；否则 $W_{i,k_1,k_2} = 0$ ；参见图 4(e) 和 (f)。随后，我们将 W_i 矩阵划分为一个由大小相同的块组成的 30×40 分块矩阵；参见图 4(f)。具体而言，将该分块矩阵记为 $\mathbb{W}_i = (\mathbb{W}_{i,k_1,k_2})$，其中 $\mathbb{W}_{i,k_1,k_2} \in \mathbb{R}^{32\times32}$，$1 \leq k_1 \leq 30$ 和 $1 \leq k_2 \leq 40$。接下来，计算分块矩阵 $\mathbb{W}_{i,k_1,k_2}$ 的平均值，并记为 μ_{i,k_1,k_2}。借助 TensorFlow 和 GPU，通过平均池化操作即可高效地以完全并行的方式完成此操作。定义 $\mathbb{Y}_{i,k_1,k_2} = I(\mu_{i,k_1,k_2} > 0.5)$。于是 $\mathbb{Y}_{i,k_1,k_2}$

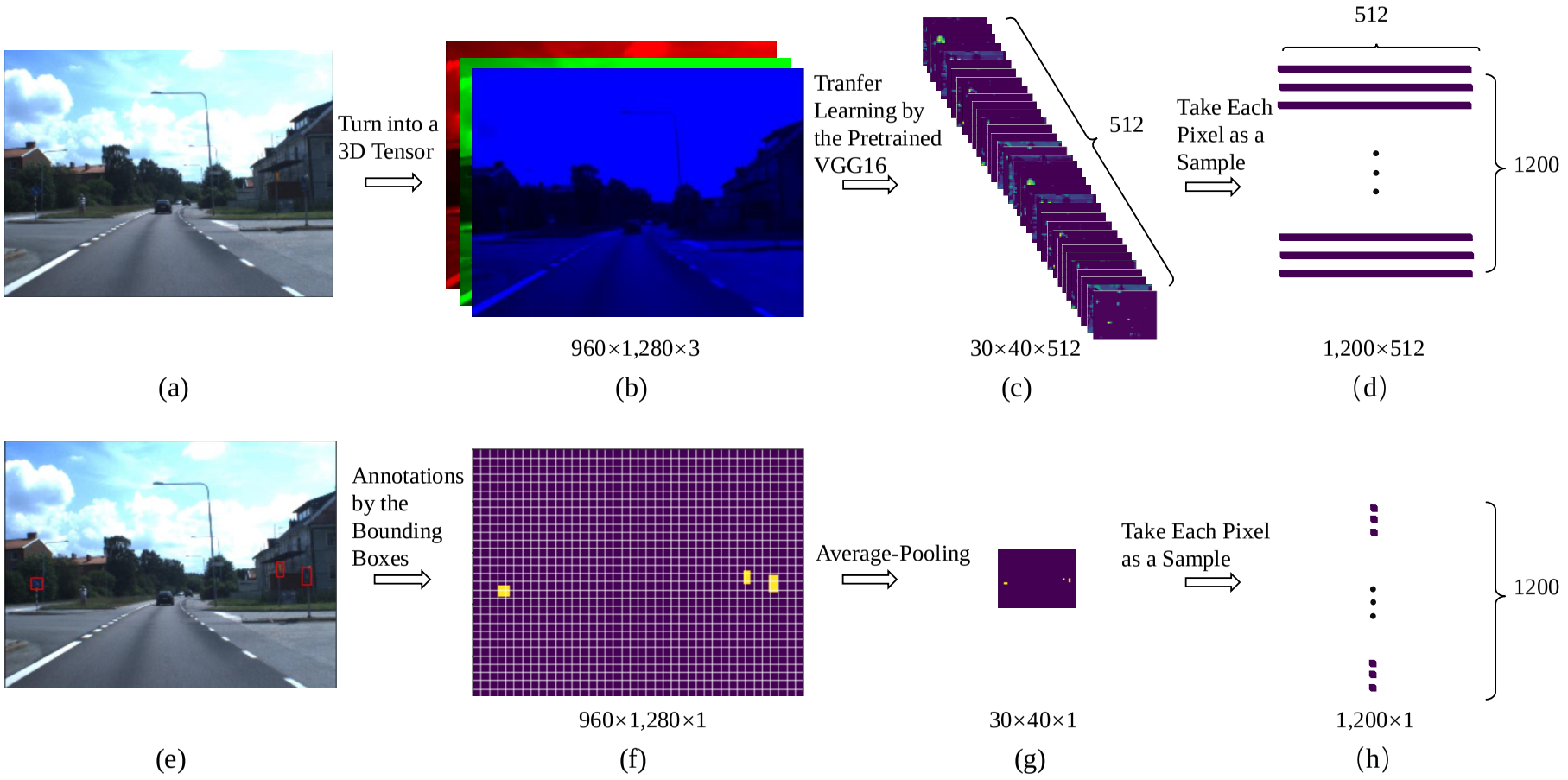

Figure 4: Illustration of the data preprocessing pipeline for one particular image. The top panel illustrates how the nonlinear features are generated. The bottom panel shows how the response is generated.

成为与 $\mathbb{X}_{i,\,k_1,\,k_2}$相关的二元响应变量；参见图 4(g) 和 (h)。它们都对应于原始图像中的同一区域。所有数据（8.59 GB，包括$\mathbb{X}_{i,\,k_1,\,k_2}$ 和 $\mathbb{Y}_{i,\,k_1,\,k_2}$）均存储在硬盘上。简单计算可知，$\mathbb{Y}_{i,\,k_1,\,k_2}$ 的样本均值为 0.225%，非常小。因此，我们可以将其视为罕见事件数据。

由于总样本量极大，因此需要进行分布式计算。为便于说明，我们将本地计算机的数量固定为 $K = 50$。由此，采用 RANDOM 策略时分配给每台本地机器的样本量约为 $N/K =$47, 280，采用 COPY 策略时约为 $N_1 + N_0/K = 52$、491。$\widehat{\quad}$ $\widetilde{\quad}$ $\widehat{\quad}$因此，基于训练数据计算了三个分布式估计量 θ_{RMLE}、 θ_{US} 和 θ_{IPW} $\widehat{\quad}$。为便于比较，还使用自行开发的牛顿-拉夫森型算法计算了 θ_{GMLE} 。如果在一台计算机上运行该算法，耗时会非常长；如果在分布式系统上运行，则由于牛顿-拉夫森型迭代，通信成本会非常高。简言之，此处开发该算法主要是为了进行理论比较。由于时间成本过高，无论是通信还是计算所致，该算法几乎无法用于实际应用。

性能结果

接下来，考虑测试数据中的第 i^* 张图像（$1 \le i^* \le N^*$），其中 $N^* = 394$ 表示用于测试的图像数量。对于第 i^* 张图像中的一个给定像素（k_1, k_2）$\widehat{\quad}$ $\widehat{\quad}$以及在训练数据上得到的一个特定估计量 θ（即 θ_{RMLE}），我们随后通过 $\widehat{p}_{i^*,k_1,k_2} = e^{\widehat{\theta}^\top \mathbb{X}_{i^*,k_1,k_2}} / (1 + e^{\widehat{\theta}^\top \mathbb{X}_{i^*,k_1,k_2}})$ 估计响应概率，并预测$\widehat{\mathbb{Y}}_{i^*,k_1,k_2} = I\big(\widehat{p}_{i^*,k_1,k_2} > c_{i^*}\big)$，其中 $c_{i^*} = \min\big\{\widehat{p}_{i^*,k_1,k_2} : \mathbb{Y}_{i^*,k_1,k_2} = 1,$ $1 \le k_1 \le 30, 1 \le k_2 \le 40\big\}$。这是 c_i^* 能够正确捕获所有正例的最大阈值。然而，为此付出的代价是产生了假阳性预测。

将测试数据中第 i^* 张图像的假阳性实例数定义为 $\mathrm{FP}^{i^*} = \sum_{k_1,\ k_2} I\big(\widehat{\mathbb{Y}}_{i^*,\ k_1,\ k_2} = 1\big) I\big(\mathbb{Y}_{i^*,\ k_1,\ k_2} = 0\big)$。随后计算其中位数，记为 FP^*。再将不同随机重复试验中的总体均值记为 FP。预测结果如图 5 所示。从图 5 可以看出，逆概率加权方法的 FP 值低至 1.88，远低于限制极大似然估计方法的 2.52 和 US 方法的 2.20。该值与 GMLE 方法的 1.88 相同。总之，在所有分布式估计量中，逆概率加权估计量表现最佳，其 FP 值最低，为 1.88。

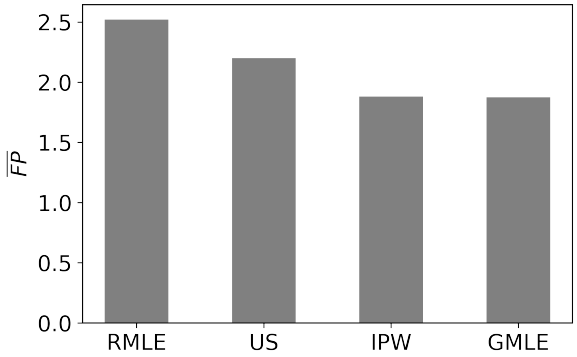

图 5：在测试数据上获得的 $\overline{\mathrm{FP}}$ 预测结果。

为进一步直观地了解预测准确性，我们在图 6 中展示了一些随机选取的预测结果。具体而言，图 6 的每一行都展示了测试数据中任意选取的一幅图像。第一列展示尺寸为 $960 \times 1{,}280$ 的原始输入图像。第二列展示 US 方法的预测结果。第三列展示限制极大似然估计方法的预测结果。第四列展示 GMLE 方法的预测结果。第五列展示逆概率加权方法的预测结果。最后一列展示真实标注区域。从图 6 可以看出，US 和限制极大似然估计方法的预测结果都非常嘈杂。GMLE 和逆概率加权方法的预测结果则好得多，而且两者非常接近。

5. 结语

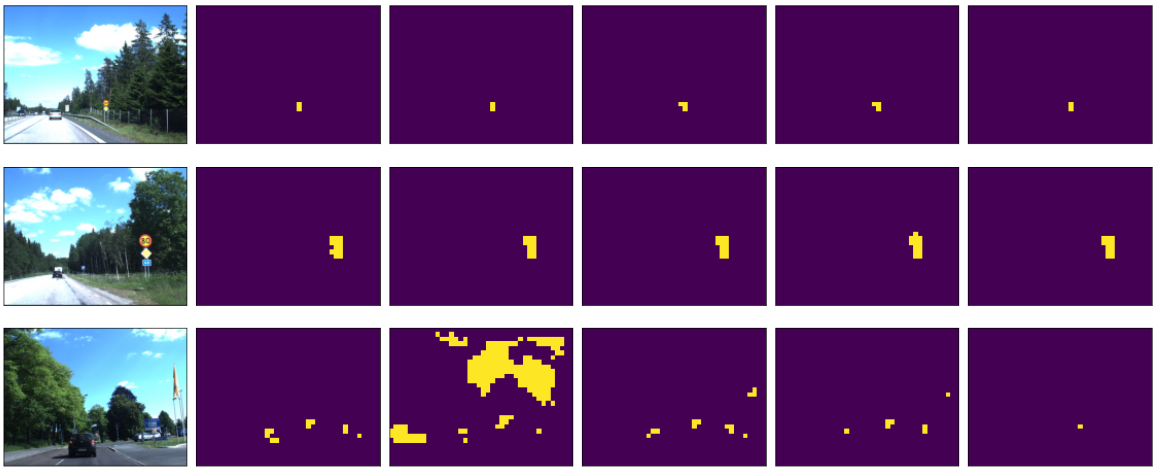

图 6：任意选取的若干测试示例，用于展示预测结果。每一行表示测试数据中任意选取的一张图像。第一列显示输入图像。第二列展示 US 方法。第三列展示限制极大似然估计方法。第四列展示逆概率加权方法。第五列展示 GMLE 方法。最后一列显示真实标注区域。

本研究探讨了大规模罕见事件数据的分布式逻辑回归问题。我们研究了两种不同的数据分布策略，分别是 RANDOM 和 COPY 策略。我们还研究了三种不同的估计量，分别是 $\hat{\theta}_{\mathrm{RMLE}}$、$\tilde{\theta}_{\mathrm{US}}$ 和 $\hat{\theta}_{\mathrm{IPW}}$。我们的结果表明，COPY 策略结合逆概率加权估计量的修正对数似然函数是最佳选择。由此得到的估计量在统计效率上可以与全局估计量相当。最后，我们讨论一些值得未来研究的主题。首先，本文聚焦于逻辑回归模型。未来的研究可以针对罕见事件数据，探讨更复杂、更一般的模型。其次，我们在分布式估计的最后一步采用了一次性策略。尽管这种策略在通信方面效率较高，但当数据非随机地分布在不同的本地机器上时，它未必是最佳选择（Zhu 等，2021）。在这种情况下，可以采用各种逆方差加权（IVW）方法（Lin 和 Xi，2011；Zhu 等，2021；Yu 等，2022）。IVW 的核心思想是对……进行加权

局部估计量的平均值。权重与 Hessian 矩阵的逆矩阵有关，而这些矩阵由本地计算机计算得出。如何将 IVW 思路与我们的 COPY 策略结合起来，用于分布式稀有事件数据分析，似乎是一个值得未来研究的有趣课题。最后，大型数据集中的协变量通常具有很高的维度。因此，如何基于这些分布式估计量进行特征选择或筛选，也值得考虑。

参考文献

Bolton, R. J. and Hand, D. J. (2002), “统计欺诈检测：综述，” 统计科学, 17, 235–255.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011), “通过乘子交替方向法进行分布式优化与统计学习，” 基础与趋势® 机器学习,3, 1–122.

Chang, X., Lin, S.-B., and Wang, Y.(2017), “分而治之的局部平均回归，” 电子统计学期刊, 11, 1326–1350.

Chawla, N. V., Bowyer, K. W., Hall, L. O., 和Kegelmeyer, W. P.（2002），《SMOTE：合成少数类过采样技术》，人工智能研究期刊，16，321–357。

Chen, G., Wang, H., Chen, K., Li, Z., Song, Z., Liu, Y., Chen, W., 和 Knoll, A.（2020），《小目标检测四大支柱综述：多尺度表示、上下文信息、超分辨率和区域提议》，*IEEE* 系统、人与控制论：系统汇刊。

Chen, J., Sun, B., Li, H., Lu, H., and Hua, X.-S. (2016), 《深度点击率预测：在...

展示广告，”收录于 第*24*届*ACM*国际多媒体会议论文集，第811–820页。

Drummond, C., Holte, R.C., 等（2003），《C4.5、类别不平衡与成本敏感性：为何欠采样优于过采样》，载于不平衡数据集学习研讨会 *II*，Citeseer，第 11 卷，第 1–8 页。

Du, W., Li, A. 和 Li, Q.（2018），《用于逻辑回归的隐私保护多方学习》，载于通信系统安全与隐私国际会议，Springer，第 549–568 页。

Estabrooks, A., Jo, T. 和 Japkowicz, N.（2004），《一种从不平衡数据集中学习的多重重采样方法》，计算智能，20，18–36。

Fithian, W. 和 Hastie, T.（2014），《局部病例-对照抽样：不平衡数据集中的高效子抽样》，统计年鉴，42，1693。

Girshick, R.（2015），“快速 R-CNN”，载于*IEEE* 国际计算机视觉会议论文集，第 1440–1448 页。

Girshick, R.、Donahue, J.、Darrell, T. 和Malik, J.（2014），“用于准确目标检测和语义分割的丰富特征层次结构”，载于 *IEEE* 计算机视觉与模式识别会议论文集，第 580–587 页。

Haixiang, G.、Yijing, L.、Shang, J.、Mingyun, G.、Yuanyue, H. 和 Bing, G.（2017），“从类别不平衡数据中学习：方法与应用综述”，《专家系统与应用》，73，220–239。

Han, H., Wang, W.-Y., 和 Mao, B.-H.（2005），《Borderline-SMOTE：一种用于不平衡数据集学习的新型过采样方法》，载于智能计算国际会议，Springer，第 878–887页。

Hassan, A. K. I. 和 Abraham, A.（2016），《使用不平衡数据分类对保险欺诈检测进行建模》，载于自然与生物启发计算进展，Springer，第117–127页。

Huang, J.-T., Sharma, A., Sun, S., Xia, L., Zhang, D., Pronin, P., Padmanabhan, J., Ottaviano, G. 和 Yang, L.（2020），《Facebook 搜索中的基于嵌入的检索》，载于第 *26* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 2553–2561 页。

Japkowicz, N. 等（2000），《从不平衡数据集中学习：各种策略的比较》，载于 *AAAI* 不平衡数据集学习研讨会，加州门洛帕克：AAAI Press，第 68 卷，第 10–15 页。

Kaur, H., Pannu, H. S. 和 Malhi, A. K.（2019），《机器学习中不平衡数据挑战的系统综述：应用与解决方案》，*ACM* 计算调查（*CSUR*），第 52 卷，第 1–36 页。

Kiefer, J.（1959），《最优实验设计》，《*Journal of the Royal Statistical Society: Series B (Methodological)*》，21，272–304。

King, G. 和 Zeng, L.（2001），《罕见事件数据的逻辑回归》，《*Political analysis*》，9，137–163。

Korkmaz, S.（2020），《基于深度学习的不平衡数据分类：药物发现》，《*Journal of chemical information and modeling*》，60，4180–4190。

Krawczyk, B.（2016），《从不平衡数据中学习：开放挑战与未来方向》，《*Progress in Artificial Intelligence*》，5，221–232。

Lin, N. 和 Xi,R.（2011），“聚合估计方程估计”，《统计学及其界面》，4，73–83。

Liu,X.-Y.、Wu, J. 和Zhou,Z.-H.（2008），“面向类别不平衡学习的探索性欠采样”，*IEEE*系统、人和控制论汇刊，*B* 部分（控制论），39，539–550。

Mathew, J.、Pang, C. K.、Luo, M. 和 Leong, W. H.（2017），“通过支持向量机核空间中的过采样对不平衡数据进行分类”，*IEEE* 神经网络与学习系统汇刊，29，4065–4076。

McMahan, H. B.、Holt, G.、Sculley, D.、Young, M.、Ebner, D.、Grady, J.、Nie, L.、Phillips, T.、Davydov, E.、Golovin, D. 等（2013），“广告点击预测：一线实践的视角”，载于第 *19* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 1222–1230 页。

Nelder, J. A. 和 Wedderburn, R. W.（1972），“广义线性模型”，英国皇家统计学会会刊：*A*辑（综合），135，370–384。

Neunhoeffer, M. 与 Sternberg, S.（2019），“交叉验证如何出错以及该如何应对”，政治分析，27，101–106。

Nguyen, H. M.、Cooper, E. W. 与 Kamei, K.（2012），“处理流数据类别不平衡的抽样技术比较研究”，载于第六届软计算与智能系统国际会议暨第十三届先进智能系统国际研讨会，IEEE，第1762–1767页。

Owen, A. B.（2007），“无限不平衡逻辑回归”。机器学习研究期刊，8。

Rosenblatt, J. D. 与 Nadler, B.（2016），“论分布式统计学习中平均法的最优性”，信息与推断：*IMA*期刊，5，379–404。

Shao, J. (2003),数理统计学, Springer Science & Business Media。

Shi, P.、Wang, P. 和 Zhang, H. (2019)，“分布式逻辑回归用于分离的大规模数据”，载于*CCF*大数据会议，Springer，第 285–296 页。

Simonyan, K. 和 Zisserman, A.（2014），“用于大规模图像识别的非常深的卷积网络”，*arXiv* 预印本 *arXiv:1409.1556*。

Sun, Y.、Kamel, M. S.、Wong, A. K. 和 Wang, Y.（2007），“用于不平衡数据分类的成本敏感提升算法”，模式识别，40，3358–3378。

Sun, Y.、Wong, A. K. 和 Kamel, M. S.（2009），“不平衡数据分类：综述”，国际模式识别与人工智能杂志，23，687–719。

Wang, H.（2019），“使用最优子样本对逻辑回归进行更高效的估计”，机器学习研究杂志，20。

—（2020），“针对罕见事件的大规模数据逻辑回归”，载于国际机器学习大会，PMLR，第9829–9836页。

Wang, H.、Gao, Y.、Shi, Y. 和 Wang, H.（2016），“一种面向大规模不平衡数据的快速分布式分类算法”，载于 *2016 IEEE* 第 *16* 届国际数据挖掘会议（*ICDM*），IEEE，第 1251–1256 页。

Wang, H.、Zhang, A. 和 Wang, C.（2021），“稀有事件数据的非均匀负采样与对数优势比校正”，载于 第三十五届神经信息处理系统会议。

Wang, H.、Zhu, R. 和 Ma, P.（2018），“大样本逻辑回归的最优子抽样”，美国统计学会会刊，113，829–844。

Yu, J., Wang, H., Ai, M. 和 Zhang, H.（2022），“适用于

海量数据的最大拟似然估计量的最优分布式子抽样”，美国统计学会会刊，117，265–276。

Zhang, H. 和 Chen, S. X.（2020），“统计推断的集中不等式”，*arXiv* 预印本 *arXiv:2011.02258*。

Zhang, Y.、Duchi, J. C. 和 Wainwright, M. J.（2013），“面向统计优化的通信高效算法”，《机器学习研究期刊》，14，3321–3363。

Zhao, Y.、Wong, Z. S.-Y. 和 Tsui, K. L.（2018），“面向罕见事件分类的不平衡医疗数据再平衡框架：形似音似混淆事件检测案例”，《医疗工程期刊》，2018。

Zhao, Z.-Q.、Zheng, P.、Xu, S.-t. 和 Wu, X.（2019），“基于深度学习的目标检测：综述”，《*IEEE* 神经网络与学习系统汇刊》，30，3212–3232。

Zhu, M., Su, W., and Chipman, H. A. (2006), “LAGO：一种计算效率高的统计检测方法，” *Technometrics*，48，193–205。

Zhu, X., Li, F., and Wang, H. (2021), “分布式系统的最小二乘近似，” 计算与图形统计学杂志，1–44。

Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., and Hu, S. (2016), “野外交通标志检测与分类，” 载于*IEEE*计算机视觉与模式识别会议论文集，第2110–2118页。

Zhuang, J., Cai, J., Wang, R., Zhang, J., 和 Zheng, W. (2019)，“Care：针对类别不平衡数据分类的病灶区域注意力机制”，载于*InternationalConference on Medical Imaging with Deep Learning*，PMLR，第588–597页。

Zuo, L., Zhang, H., Wang, H., 和 Sun, L. (2021), “基于分布式数据的大规模逻辑回归最优子样本选择”, *Computational Statistics*, 36, 2535–2562。