·综述·

大数据驱动的罕见事件非均衡数据分析方法研究进展

周江杰[1]，王予童[2]，冯　天[1]，孟祥龙[2]，梁宝生[1]，王胜锋[2]

1. 北京大学公共卫生学院生物统计系（北京 100191）
2. 北京大学公共卫生学院流行病与卫生统计系（北京 100191）

【摘要】罕见事件在各学科领域广泛存在，例如疫苗和药品的罕见不良反应、临床罕见疾病以及发生概率很小的临床结局等。这类事件的研究之所以受到广泛关注，是因为其发生通常会带来难以估量的严重后果。在大数据场景下，已涌现出包括基于抽样、类别加权、集成学习以及深度学习的众多罕见事件分析方法。本文系统总结了当前罕见事件分析方法的研究进展，介绍其基本原理以及适用场景，通过分析现有方法的优缺点，梳理归纳罕见事件研究的挑战，探索相关领域潜在的研究方向，为研究者提供参考。

【关键词】罕见事件；非均衡数据；数据驱动；深度学习

【中图分类号】TP 181；R 195.1　　【文献标识码】A

Research progress on big-data-driven analysis strategies for imbalanced data of rare events

ZHOU Jiangjie[1], WANG Yutong[2], FENG Tian[1], MENG Xianglong[2], LIANG Baosheng[1], WANG Shengfeng[2]

1. Department of Biostatistics, School of Public Health, Peking University Health Science Center, Beijing 100191, China
2. Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing 100191, China
Corresponding authors: LIANG Baosheng, Email: liangbs@hsc.pku.edu.cn; WANG Shengfeng, Email: shengfeng1984@126.com

【Abstract】Rare events are widely prevalent in various disciplines, including rare adverse reactions to vaccines and drugs, clinical rare diseases, and low-probability clinical outcomes. The reason for research interest on such events is that their occurrence often brings incalculable and serious consequences. In the context of big data, numerous methods have emerged for rare event data analysis, including sampling based, category weighting, ensemble learning, and deep learning. This article systematically summarizes the research progress of current rare event data analysis methods, and introduces their basic principles and applicable scenarios. By analyzing the advantages and disadvantages of existing methods, the challenges of rare event research are sorted out and summarized, and potential research directions in related fields are explored to provide references for researchers.

DOI: 10.12173/j.issn.1005-0698.202411080
基金项目：“癌症、心脑血管、呼吸和代谢性疾病防治研究”国家科技重大专项（2023ZD0506600）；首都卫生发展科研专项项目（2024-1G-4251）
通信作者：梁宝生，博士，副研究员，博士研究生导师，Email：liangbs@hsc.pku.edu.cn
王胜锋，博士，研究员，博士研究生导师，Email：wangshengfeng@bjmu.edu.cn

【Keywords】Rare events; Imbalanced data; Data-driven; Deep learning

罕见事件（rare event）是指发生概率较低且在收集的数据集中阳性结局极为稀少的事件。罕见事件在各科学领域均广泛存在，如极端自然灾害、世纪金融危机、重大罕见公共事件、罕见药物不良反应等，虽然发生概率低但其破坏性严重或危害性强，对国家社会和人民生命财产安全和健康影响巨大[1-3]。预测罕见事件对预防与应对罕见事件至关重要。传统数据分析方法没有内置机制来处理罕见事件特有的类别不均衡问题，因为从样本整体看，罕见事件的信号显得非常微弱，另外罕见事件数据通常具有复杂的非线性关系或异常偏态分布，传统方法无法有效建模。因此导致对罕见事件的关注不足，无法捕捉罕见事件的关键特征，需要专门针对罕见事件的数据分析与研究的统计方法。

抽样技术和算法的发展、大数据技术的兴起，以及数据采集、存储和分析技术的进步，为罕见事件研究提供了新的思路、资源和工具。通过分析海量数据，建立罕见事件预测模型可以提升罕见事件预测和应对能力，识别潜在风险因素和预警信号，为预防和应对罕见事件提供数据支持。用大数据技术挖掘隐藏的模式和关联，有助于揭示罕见事件背后的复杂成因，推动罕见事件的认识、理解和进一步的研究。基于模型和数据分析的结果，还有助于合理分配资源，制定更有效的应急预案和响应策略，减少损失[4-6]。现有文献中报道的罕见事件数据分析方法多种多样，每种方法都有其独特的优势和局限性。本文对罕见事件的发生机制以及非均衡数据的处理和分析方法进行综述，分析比较现有方法的优缺点，归纳研究的难点和挑战，探索未来相关领域的潜在研究方向，为预防和应对罕见事件提供参考。

1 罕见事件的发生机制与分析难点

1.1 罕见事件的发生机制

罕见是相对概念，不同科学领域对罕见事件的发生原因及其发生率的定义有所不同。例如，临床医学中将罕见病定义为患病率＜ 1/500 000，或新生儿发病率＜ 1/10 000 的疾病[7]，而在机器学习领域则将在数据集中比例＜ 1% 的事件定义为极端罕见[8]。罕见事件的意义在于它们通常会对个体以及社会造成远大于常见事件的后果，因而早期识别、预测并对其进行干预具有重要作用。如自杀风险预测[9]、重症监护病房患者罕见事件预测[10]、手术后罕见并发症以及药品不良反应监测[11]等对于降低个体死亡风险提高生存质量有所帮助，而估计罕见病的发病率对于其预防、疾病负担研究、治疗药物研发决策都具有重要意义[8]。

罕见事件在数据层面常表现为样本的类别不平衡，在分类结局中表现为感兴趣的结局类别（如死亡、肿瘤复发等）在整体样本中出现的比例极低，因此机器学习领域常称为类不均衡数据或非均衡数据（imbalanced data）[12]。在生存结局中表现为感兴趣的结局事件右删失比例极高，删失状态指示变量取值 1 和 0 的类别不均衡。罕见事件与非均衡数据的概念既有联系又有区别，非均衡数据的概念更宽泛，因为罕见事件会导致非均衡数据，而非罕见事件情形也可能会出现非均衡数据。

如图 1 所示，常见的罕见事件来源包括：①自然发生，有些事件在自然界中发生的概率较其他事件低，如地震[13]、大洪水[14]或者极端气象事件[15]这样的自然灾害。医学领域的罕见病、遗传领域的罕见变异位点[16]也是自然发生事件。②人为干预或伦理限制，有的事件在自然条件下发生率较高，而引入人工干预后其发生率降低以至于成为罕见事件。例如 2021 年 6 月 30 日，世界卫生组织宣布中国正式获得无疟疾认证[17]，在中国因疟疾死亡事件成为罕见事件。干预也有来自伦理方面的限制，如自动驾驶领域的研究，在真实世界中收集危险情景下的驾驶数据是困难甚至不可行的[18]。③观测技术和成本的限制，在一些情况下，由于技术限制或者成本较高等原因，极难获取感兴趣的类别或者事件的样本，导致在数据层面样本量少，从而表现为罕见事件[19]。如航天员登陆月球的案例受实际成本和技术限制是罕见事件；抑郁症患者疾病进展早期的确诊，受技术限制确诊难度极高，实际观测到的样本极少。④抽样偏倚，同一事件或者疾病在不同阶段、不同地区的分布不同，由于抽样设计的不合理或者研究纳入排除标准设计的不合理，会导致获取数

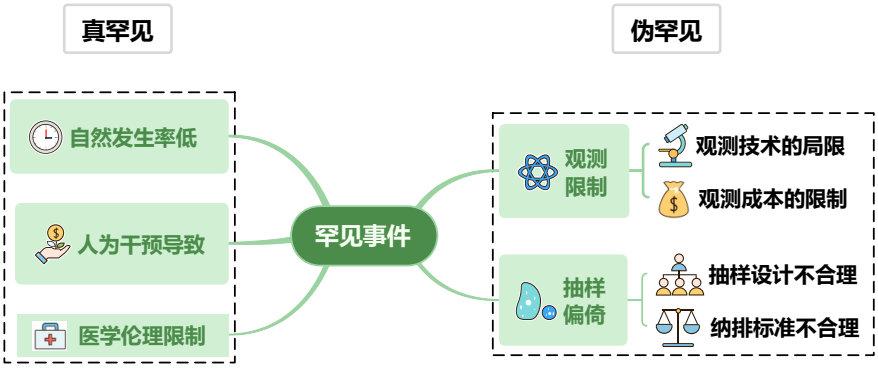

图1 罕见事件来源及发生机制

Figure 1. Sources and mechanisms of rare events

据中阳性事件率偏低，出现“罕见”事件。例如，阿尔茨海默病的人群发病率较低，可诊断的发病率为3%~8%，而大部分患者的确诊时间在60岁以后，如果纳入研究的年龄定为＜50岁，则观测得到的阳性率会较低。

从发生机制上看，罕见事件可以简单划分为真罕见与伪罕见两大类，其中由于观测技术和成本的限制与抽样偏倚导致的样本层面观测到事件“罕见”，但真实场景下事件的发生并不罕见的情形称为伪罕见事件（图1）。不同发生机制的罕见事件，在分析方法的选择上略有差异，其发生机制对分析结果的解释具有关键作用。实际研究中，应该尽量避免伪罕见情形的发生。由于观测技术局限或观测成本导致的伪罕见问题有时是不可避免的，但是由抽样设计或纳排标准导致的设计偏倚出现的伪罕见情形应该是可以有效避免的。前文提到的非均衡数据是真罕见和伪罕见两种情形在数据层面的统称，后文介绍的罕见事件分析方法主要针对真罕见事件的非均衡数据。

1.2 罕见事件数据分析的研究难点

在不同领域的罕见事件的研究目的存在一定差异，大致可以分为两类：一类是估计罕见事件的发生率[20]，另一类是探究影响罕见事件发生的因素，进行建模分析，预测罕见事件发生的可能性[18]。

各领域研究者广泛关注罕见事件分析方法的进展，一是因为其重要研究意义和实用价值，二是因为该类数据分析普遍存在的难点和若干挑战。

（1）类别不均衡导致模型分类的偏好：在罕见事件数据集中，少数类比多数类有明显较少的样本数量。这种类别分布的偏倚使得机器学习算法难以学习到数据的模式并准确地分类少数类[21]，因为算法通常基于分类的准确率，即正、负分类样本的分类概率，作为其优化目标，而在准确率计算中正负类占有相同的权重，因此分类算法会更倾向于将样本预测为负类。

（2）精确估计发生率需要大样本：由于罕见事件发生概率较低，想要准确估计其发生概率通常需要很大的样本量。如使用：

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n} I(Y_i=1) \quad \text{公式（1）}$$

估计发生率p，那么估计的标准误为：

$$\mathrm{se}(\hat{p})=\sqrt{\frac{p(1-p)}{n}} \quad \text{公式（2）}$$

欲使$\hat{p}$的误差波动范围小于10%，需要控制变异系数：

$$\mathrm{se}(\hat{p})/p \le 0.1 \quad \text{公式（3）}$$

即样本量$n \geqslant 100/p$。若p=10^{-4}，则所需样本量为1 000 000。

（3）类别重叠：在某些情况下，正类和负类样本的特征空间可能存在重叠或相似性[22]，使得分类算法识别罕见事件更加困难。

（4）高维特征：在大数据场景下，罕见事件数据可能会包含多维多模态特征，包括数值、类别、文本、图像数据等模态数据[23-24]。多模态特征的处理需要特征对齐、降维和筛选技术的支持。在结局事件发生率较低时，基于有监督的多模态特征处理技术是个难点。

（5）多事件交互作用：单一罕见事件不仅与多个变量有关联，还可能与多个事件存在交互作用，例如罕见基因变异位点涉及多个位点之间的交互[13]，自动驾驶中复杂的环境信息包括天气对事故发生的影响[16]。这种多元复杂结局需要使

用更复杂的模型进行刻画，而复杂度越高的模型通常所需参数规模更大、复杂度更高，训练所需的有效样本量也越大，这对罕见事件本身有限的样本信息是需要权衡的矛盾。

2 大数据驱动的罕见事件建模分析研究方法

开发能够处理独特性和减轻罕见事件固有偏见和限制的有效方法和算法是至关重要的。罕见事件的非平衡数据分析挑战由来已久，传统建模方法对非均衡数据集的估计、推断和预测存在较大偏差，需要先使用特定的预处理技术来有效放大罕见事件的信号，并结合传统建模分析方法，才能实现罕见事件的有效推断和准确预测。在大数据场景下，研究人员有更多的数据资源可用，基于数据驱动的方式已经开发出多种技术来处理类不均衡数据和罕见事件。包括数据层面的分析方法，如过采样（over-sampling）和降采样（under-sampling），以及算法层面的方法，如成本敏感学习、集成学习等。近年来，基于深度学习方法也开始被应用于罕见事件分析领域，为罕见事件的分析提供了一种新的思路。以下简要介绍这些方法的基本原理及适用场景。

2.1 基于抽样的方法

基于重抽样的方法有两种基本思路：一种是过采样，即增加罕见类别的样本数，实现罕见事件信号的放大，从而使得各类别达到平衡；另一种是降采样，即通过去除多数类别中对类别预测作用有限的样本来平衡各类别数据。另外也有将两者结合起来的综合算法。具体方法包括以下几种。

2.1.1 基于自助法的重抽样

自助法（bootstrap）在统计学中是指在原始数据集中通过有放回的随机抽样来估计总体参数的分布的非参数统计方法。而在此处自助法的目的是通过从多数类中随机抽取并删除样本来减少多数类，或通过随机复制少数类样本来增加少数类，以解决类别不平衡问题。自助法的优点是对样本的特征分布没有要求，对于混合类型的特征（如离散型和连续性变量混合）或者高维特征均可使用。尽管基于自助法的重抽样已被广泛使用，但它也存在一些局限性。过采样的缺点是可能导致过拟合，因为模型反复从重复样本中进行学习。同时简单随机降采样也可能会损失大量对于模型训练有用的信息。

2.1.2 基于样本生成的过采样

自助法产生的样本具有重复性，于是有研究者考虑使用算法来基于原始样本合成新的样本，从而使得类别达到平衡。一种流行的算法是 SMOTE（synthetic minority oversampling technique）[25]。SMOTE 是一种插值算法：选取两个样本，再生成介于两者之间的样本。SMOTE 算法基于欧式距离进行插值，适用于低维、连续型数据。对于离散型数据，或者文本、图片数据这样具有复杂特征空间的数据，SMOTE 并不适用[26]。同时 SMOTE 也会有增加数据噪声的风险 [27]。基于上述不足，SMOTE 算法出现了诸多变体。例如自适应合成采样（adaptive synthetic sampling，ADASYN）[28]，它根据学习难度改变样本的抽样权重，具体来说 ADASYN 倾向于生成更多学习难度高的样本。部分 SMOTE 变体及其原理和适用场景见表 1。除了基于插值的算法以外，另一类算法如 GNUS（Gaussian noise up-sampling）则通过向少数类别样本添加随机噪声达到数据增强（data augmentation）的目的 [29]。少数类别数据增强有助于提升模型的泛化能力以及降低估计偏差。

过采样方法要求数据本身具有较大的变异性，否则过采样可能因引入噪声而降低预测模型的性能（预测准确率等）[30]。对于医学数据，两类算法均有应用，也有研究者指出 GNUS 数据增强方法可能更适用于临床数据 [30]。

2.1.3 基于算法的降采样

简单随机降采样可能导致有用信息的损失，从而降低统计效率，因此在实际应用中通常采用基于算法的降采样。降采样的目的是减少多数类的样本数量，以减少模型对多数类的偏好，提高对少数类的识别能力。其基本原理是去除数据中对类别预测作用不大的样本，例如噪声，同时保留那些难以分类需要重点学习的样本。与过采样相比，其在平衡各类别的同时可以去除一些噪声数据，从而提高数据质量，同时减少模型训练和预测所需的计算资源和时间 [31]。

ENN（edited nearest neighbors）[32] 是一种常见的降采样算法，其原理是首先计算每个样本的

表1 SMOTE算法的常见变体

Table 1. Variants of the SMOTE algorithm

算法名称	基本原理	适用场景
SMOTE-N[25]	通过引入类别后验距离替代欧式距离达到处理离散数据的目的	适用于分类变量
SMOTE-NC[25]	连续型特征采用欧式距离，分类变量采用连续特征标准差的中位数作为距离度量	同时适用于连续型和分类变量
SMOTE-TLNN-DEPSO	1. 数据集分为多数类和少数类样本 2. 应用SMOTE生成合成样本 3. 使用TLNN识别噪声和边界样本 4. 使用DEPSO优化样本子空间	适用于处理不平衡数据集，特别是在存在噪声和边界样本问题时
SVM-SMOTE[33]	1. 训练SVM以获得支持向量 2. 使用支持向量生成新样本 3. 对于每个支持向量，找到其K-nearest neighbors，并使用插值或外推法创建样本	适用于SVM模型，特别是在类别边界较为清晰的情况下
DeepSMOTE[26]	1. 基于SMOTE的过采样方法 2. 使用encoder/decoder框架 3. 通过惩罚项增强的专用损失函数	适用于高维数据

注：SMOTE-N. SMOTE for nominal features; SMOTE-NC. SMOTE-nominal continuous; SMOTE-TLNN-DEPSO. SMOTE using a two-layer nearest neighbor (TLNN) and the hybrid differential evolution with particle swarm optimization (DEPSO); SVM-SMOTE. support vector machine based SMOTE; DeepSMOTE. deep learning based SMOTE。

K近邻，如果一个样本的K近邻大多都是另一个类别的样本，那么该样本应归类为噪声而被去除。相似性多数派欠采样技术（similarity majority under sampling technique，SMUTE）[34]则通过计算多数类样本与少数类样本之间的相关系数，选择与少数类样本相似性最高的多数类样本进行删除，以减少多数类样本的数量，SMUTE能够有效地将多数类和少数类样本分离，增加对每个类别的识别能力。除了K近邻以及相关系数算法之外，也有一类采用聚类模型对多数样本进行分层，在每层之内进行降采样的方法。这里的聚类模型起到了计算样本相似性的作用。

降采样方法一般为启发式方法，对噪声的判断更多依赖于既往经验，因此实际使用时可能出现误判。另外此类方法对数据集的样本量有一定要求，需要保证有足够的少数类别样本。在医学领域，特别是临床试验数据中样本量一般较小，可能并不适用降采样方法[30]。

2.1.4 两种采样方式相结合的算法

过采样和降采样方法均有其特点和适用范围，目前也存在一些将两者优势相结合的采样方法。一种算法是首先通过SMOTE生成合成样本来增加少数类的样本数量，随后采用ENN等降采样方法减少多数类样本，以减少SMOTE生成新样本过程中可能引入的噪声，此方法被称为SMOTE-ENN[35]。除了ENN算法，此类算法的第二步降噪部分也可用LOF[36]、DBSCAN[27]等其他算法。对于第一步的过采样算法，也可以采用深度学习方法进行数据降维，再使用SMOTE对低维隐空间进行采样[37]。

2.2 基于模型类别加权以及惩罚函数的方法

常规的分类算法中各类别样本的权重都是相同的，因此模型关注的是总体分类误差，在罕见事件的情境下，由于罕见事件样本数少，因此在总体分类误差中所占的权重低，故而模型预测结果将更倾向于多数类别。代价敏感学习（cost-sensitive learning）[38]则是一种为罕见且更重要的类别赋予更高样本权重的学习方法。目前这种思路已经被应用在多种分类算法中，包括Logistic回归[39-40]、随机森林[41]和支持向量机（support vector machine，SVM）[41]等。其中的加权Logistic回归是通过在对数似然函数中加入类别权重实现。加权支持向量机则是在损失函数中加入类别权重。加权随机森林（weighted random forest，WRF）[41]方法则是在两个步骤引入类别权重：一个是在树生成过程中，使用加权Gini准则来进行节点划分，另一步是在进行类别预测时对类别投票结果进行类别加权。此类方法在现有软件包中（如Python v 3.13-64 bit软件包“imbalanced-learn”）已被广泛实现，易于使用，但这类方法的不足之处在于模型的类别权重依赖于手动选取，但也有研究者提出了动态选取最优类别权重的方法[41]。

对于罕见事件的非均衡数据来说，由于事件发生数可能少于模型参数的数目，因此模型可能出现过拟合问题[42]。另外一种情况是罕见事件可能会导致分类模型，例如 Logistic 回归模型可能出现参数估计不收敛的情况（因为事件发生率 $p \to 0$，导致 $\text{logit}(p) \to \infty$）[13]。这种情况下可以采用罚函数对原始似然函数添加惩罚项，例如 Ridge[43]、Lasso[44] 以及 elastic net[45] 等，从而使得极大似然估计倾向于选择参数更少的模型。例如有研究者在分析罕见事件的生存数据时，在 Cox 模型中加入了 Ridge 或 Lasso 惩罚项，从而减少了传统模型的预测均方误差[46]。另一类惩罚项是 Firth 惩罚，即

$$\frac{1}{2}\log\left(I(\beta)^{-1}\right) \quad \text{公式（4）}$$

其中 $I(\beta)$ 为 Fisher 信息，计算机模拟表明该类型的惩罚在罕见事件情境下表现比包括 Ridge 在内的其他类型惩罚更优[39]。

应用层面，已有研究将类别加权法应用到术后关节脱位预测中（事件发生率 1%），使用逆概率加权（inverse probability of treatment weighting，IPTW）的 SVM 算法灵敏度达到了 87%。另有研究将其用于药品不良反应的预测[47]。

2.3 基于集成学习的方法

降采样的一个不足之处在于其损失了一部分样本信息，从而降低了估计的统计学效率。有研究者提出可以利用集成学习的思路，在保留全部罕见样本的基础上，多次进行降采样生成多个平衡数据集，然后将这些分类模型在这些数据集上的结果进行汇总得到一个集成分类器，从而达到提高统计学效率的目的。Li 等[48] 提出了一种针对罕见事件的分布式 Logistic 回归模型，该模型通过上述降采样策略生成多个数据集分布到计算集群中，集群每个节点则采用 IPTW 的 Logistic 回归对参数进行估计，最后中心计算机通过取各节点参数的均值得到最后的估计结果。平衡随机森林（balanced random forest）[49] 则将森林中每棵决策树的样本子集进行类别平衡：对于随机森林中的每个迭代，从少数类中抽取一个 bootstrap 样本，再从多数类中随机抽取相同数量的样本构成平衡数据子集。

除了 bagging 集成方法外，也有研究者提出基于 boosting 的集成学习方法，例如 SMOTEBoost[50]：结合了合成少数过采样方法（SMOTE）和标准提升过程。在每次迭代中，仅对少数类样本进行 SMOTE 抽样，新创建的合成少数类示例在学习分类器后被丢弃，不在原始训练集中添加。另外 Meta 分析也可视为一种集成方法，其特点是将各数据集中的结果进行汇总从而得到汇总效应值。罕见事件 Meta 分析的难点是零事件的存在，传统的 Mantel-Haenszel 法处理此类问题需要进行连续性校正，可能引入偏倚，Peto 法不需要连续性校正但只能用于估计比值比[51]。零事件的一种特殊情况是双零事件，即试验组和对照组事件发生数均为 0，一项模拟研究结果[52] 表明，在各组样本数不均衡的情况下，广义估计方程等单阶段模型比基于连续性校正的逆方差异质性分析模型（inverse-variance heterogeneous model，IVhet）的两阶段方法表现更佳。最近也有研究者提出基于贝叶斯[53] 和置信分布[54] 的算法，但目前此领域尚未有公认的最佳算法，因此需要借助敏感性分析等手段对结果进行解读[51]。

2.4 基于深度学习的方法

深度学习（deep learning）[55] 是基于深度神经网络的一个机器学习分支，深度神经网络具有强大的表示能力并在计算机视觉、自然语言处理以及强化学习等领域都得到了广泛应用。近年来，基于深度学习的方法也开始应用于罕见事件的分析当中。基于深度学习的罕见事件分析可以分为两种类别，一种是基于数据的方法，另一种是基于算法的方法。

基于数据的方法尝试使用深度学习模型生成新的罕见类别样本，特别是高维样本，例如使用生成对抗神经网络（generative adversarial networks，GAN）[56] 以及变分自编码器（variational autoencoders，VAE）[57]。GAN 通过两个相互对抗的神经网络——生成器（generator）和判别器（discriminator）来训练生成模型，生成器尝试生成与真实数据相似的假数据而尝试区分真实数据和生成器生成的假数据。通过这种对抗过程，生成器能够不断改进，最终生成高质量的合成数据。GAN 目前在生成新的医学影像数据上得到了应用[58]，通过生成罕见类别样本使得数据达到类别均衡从而辅助分类模型的训练。GAN 的不足之处在于训练过程不稳定并且容易产生高度重复的图像。VAE 是一种隐变量混合模型，其由编码器（encoder）

和解码器（decoder）两部分组成，VAE 能通过解码器采样的方式直接生成新样本，也能通过编码器将样本转换为隐空间表示。Dablain 等[26]便利用 VAE 学习样本的隐空间表示，并在隐空间应用 SMOTE 插值算法进行上采样来生成新样本，创新地将 SMOTE 算法与 VAE 相结合。

基于算法的方法尝试在现有的深度学习框架上，对算法进行修改以更适用于分析罕见事件。例如 Xiu 等[59]则基于 VAE 的框架，将原始 VAE 中的正态分布先验改为基于极值分布的先验分布，并假设罕见事件对应于潜空间中的极端值，进而在预测新型冠状病毒肺炎死亡事件发生概率的情境中取得了良好结果（曲线下面积：0.883）。Hamaguchi 等[9]则利用 VAE 的隐空间表示能力，构建了一个双分支网络学习到将图像中的不变特征和样本特异特征解耦，最后利用不变特征进行罕见事件预测，大幅降低了罕见事件检测需要的样本量。Yang 等[60]则为强化学习中的深度 Q 网络设计了一种基于逆类别概率的奖励函数，从而在预测重症监护病房患者急性事件发生和急诊新型冠状病毒肺炎诊断两种场景上表现超越了普通深度 Q 网络（DDQN 和 DQN）。

3 结语

罕见事件能产生远大于其在总体中所占比例相当的影响，近 20 年来，不断有研究者提出或者改进用于罕见事件分析方法。本文从罕见事件的定义出发，讨论了罕见事件分析的必要性及其分析难点，并梳理总结了现有罕见事件的非均衡数据建模分析方法及其适用范围（方法汇总见表 2 和图 2）。简单说，罕见事件的非均衡数据处理主要有两种思路：第一种思路是从样本出发，通过生成或者减少某一类别的样本来平衡各类别样本数目。此方法在非均衡数据分析领域被广泛使用，其特点是通用性，即任何模型都可以不经修改直接使用平衡后的数据集进行预测。其不足在于：数据集平衡之后并不一定会对模型的预测性能有提升，甚至可能由于新产生的噪声或是降采样对样本量的减少造成模型的性能损失。近年来随着深度学习技术的发展，研究者利用多种深度学习技术提出新的样本生成算法或是利用深度学习改

表2 罕见事件分析方法总结

Table 2. Summary of rare event analysis methods

类别	方法	基本原理	优缺点及适用场景
基于重抽样的方法	自助法	有放回随机抽样	自助法简单可行，对于混合类型的特征（例如离散型和连续性变量混合）或者高维特征均可使用。
	样本生成的过采样（SMOTE、ADASYN等）	基于样本间距离的插值算法	采样过程不需要样本分布相关知识，适用于已有样本量较小的情形，可能生成噪声数据
	基于算法的降采样（ENN、SMUTE等）	去除数据中的噪声	不需要样本分布相关知识，但对数据集的样本量有一定要求，需要保证有足够的少数类别样本，可能增加参数估计方差
	过/降采样结合算法（SMOTE-ENN等）	多阶段采样：过采样增加少数类别样本数、降采样减少生成的噪声	结合了两种采样方式的优点
基于类别加权以及罚函数的方法	加权法	为样本添加类别权重（少数类别权重更高）	对现有算法的改动较少，容易实现。类别权重的选取仍存在挑战
	罚函数法	在损失函数或似然函数中添加惩罚项，减少过拟合	能够减少罕见事件带来的模型参数估计不稳定的情况，并且起到减少过拟合的作用
基于集成学习的方法	Bagging法（分布式Logistic回归、平衡随机森林等）	生成多个平衡数据集并分别拟合模型，最后将模型结果进行汇总	通过平均多个模型的结果，降低模型参数估计的方差
	Boosting法（SMOTEBoost等）	每轮迭代创建平衡数据集，随后提升分类误差较大的样本的权重	模型偏差较小，但训练过程无法并行实现，大型数据集训练时间长
	Meta分析法	通过随机效应模型等方式汇总干预方法的效应值	适用于合并不同研究的效应值（如比值比、相对危险度等）
基于深度学习的方法	深度学习过采样（GAN、VAE等）	通过深度学习方法学习样本的分布，再从分布中抽取新样本	完全掌握样本的分布，适用于高维数据，例如图像数据等，要求大样本量以训练模型
	深度学习模型改进	在现有的深度学习框架上，对算法进行修改以更适用于分析罕见事件	适用于高维数据，要求大样本量以训练模型

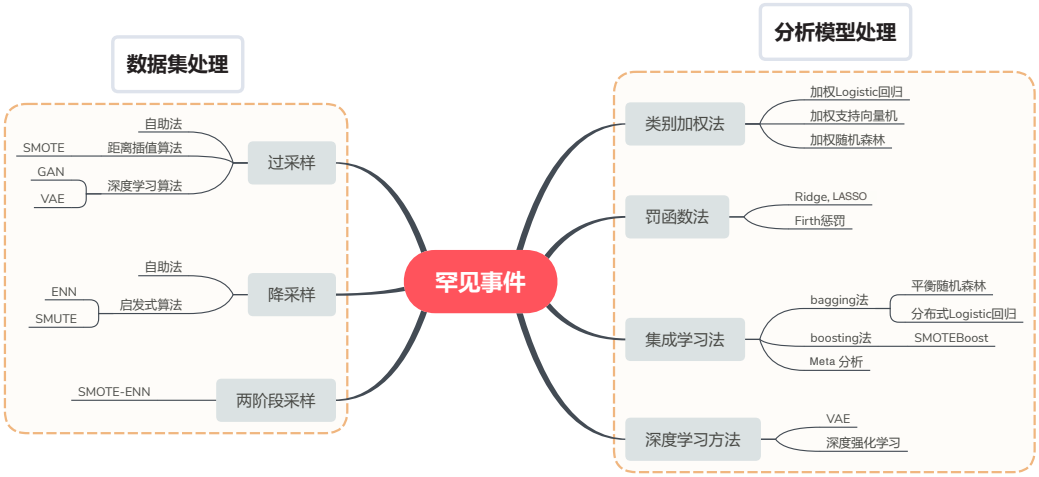

图2 罕见事件研究方法汇总

Figure 2. Summary of rare event analysis

善现有的成熟算法。使得这种分析思路能适用于当今大数据情境下，数据特征维度日益增加，传统低维算法不再适用的情形。第二种思路是改进现有的预测算法本身，使其能够基于现有的非均衡数据集提升预测性能。算法的改进思路有：调整类别权重以在损失函数中平衡各类别的影响、引入罚函数防止模型过拟合、将多个数据集的预测模型进行集成提升整体性能以及引入深度学习方法等。基于算法改进的思路不需要生成新的样本，因而可以加快分析的速度，另外针对特定模型的改进可能比通用的基于平衡数据集的方法更能保证预测性能的提升。其不足在于需要根据算法的不同设计专门的改进方式，调整类别权重的方法较为通用，但权重参数仍需要研究者选取。

尽管现有文献已经针对罕见事件的分类结局观测数据提出了大量的分析方法，但我们也注意到尚存在一些分析的局限性有待后续研究的解决。在事件交互层面，由于多个罕见事件或者罕见事件与其他事件之间可能存在的交互作用，如何识别并利用这种复杂的依赖关系成为分析的难点，目前针对此方向的研究仍然较少。在可解释性方面，随着深度学习算法在罕见事件分析领域的流行，其模型解释困难的特点也日益明显。尽管有研究尝试采用 SHAP [41] 等方法对模型的变量重要性进行解释，以及采用特征空间解耦 [52] 的方式对特征进行分组解释，但仍需要后续研究给出一个适用于罕见事件分析的通用框架。最后是数据方面，罕见事件数据由于其难以获取，通常单一数据集的样本量较小，如何建立一种有效的数据共享机制，提升整体的样本数是一个难点。目前此方向已经有了一些探索，例如医学领域的罕见病共享样本库等 [2]。除了数据总量的问题外，目前也缺乏利用多种模态数据进行分析的方法，现有研究通常只关注于某一种模态的分析。

基于以上难点，未来罕见事件的研究方向应当是由数据、智能以及协作共同驱动。一方面数据维度从低维向高维，从单模态数据扩展到多模态数据的分析。另一方面分析和预测方法更加智能化，多种深度学习甚至大模型方法将被引入其中。最后研究趋向于通过合作集成多领域、多地区知识，提高有效样本量和决策的稳健性。相信随着该领域的不断发展，跨学科合作和数智方法将为罕见事件预测带来变革性的进步。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 Barro RJ. Rare disasters and asset markets in the twentieth century[J]. QJ Econ, 2006, 121 (3): 823–866. DOI: 10.1162/qjec.121.3.823.

2 周颖，鲁云兰．严重或罕见的药物不良反应报告 8 例 [J]. 中国新药杂志，2002, 11(1): 92–94. [Zhou Y, Lu YL. 8 cases of severe or infrequent adverse drug reactions[J]. Chinese Journal of New Drugs, 2002, 11(1): 92–94.] DOI: 10.3321/j.issn:1003–3734.2002.01.023.

3 Li XT, Zhou J, Wang HS. Gaussian mixture models with rare events[J]. J Mach Learn Res, 2024, 25 (1): 1–40. DOI: 10.48550/arXiv.2405.16859.

4 贾玉春．浅谈大数据技术在安全生产管理中的创新应用 [J].

大数据与人工智能 , 2024, 5(2): 4–6. [Jia YC. Discussion on the innovative application of big data technology in safety production management[J].Big Data and Artificial Intelligence, 2024, 5(2): 4–6.] DOI: 10.12345/bdai.v5i2.16496.

5 Northcott R. Big data and prediction: four case studies[J]. Stud Hist Philos Sci Part A, 2020, 81: 96–104. DOI: 10.1016/j.shpsa.2019.09.002.

6 Chen W, Hu Y, Zhang X, et al. Causal risk factor discovery for severe acute kidney injury using electronic health records[J]. BMC Med Inform Decis Mak, 2018, 18 (Suppl 1): 13. DOI: 10.1186/s12911–018–0597–7.

7 Shyalika C, Wickramarachchi R, Sheth A. A comprehensive survey on rare event prediction[J]. ACM Comput Surv, 2024, 57(3): 1–39. DOI: 10.1145/3699955.

8 Chen W, Yang K, Yu Z, et al. A survey on imbalanced learning: latest research, applications and future directions[J]. Artif Intell Rev, 2024, 57(6): 137. DOI: 10.1007/s10462–024–10759–6.

9 Coley RY, Liao Q, Simon N, et al. Empirical evaluation of internal validation methods for prediction in large–scale clinical data with rare–event outcomes: a case study in suicide risk prediction[J]. BMC Med Res Methodol, 2023, 23(1): 33. DOI: 10.1186/s12874–023–01844–5.

10 Leisman DE. Rare events in the ICU: an emerging challenge in classification and prediction[J]. Crit Care Med, 2018, 46(3): 418–424. DOI: 10.1097/CCM.0000000000002943.

11 Oeding JF, Lu Y, Pareek A, et al. Understanding risk for early dislocation resulting in reoperation within 90 days of reverse total shoulder arthroplasty: extreme rare event detection through cost–sensitive machine learning[J]. J Shoulder Elbow Surg, 2023, 32(9): e437–e450. DOI: 10.1016/j.jse.2023.03.001.

12 郭丹 , 金晔 , 刘炜达 , 等 . 罕见病临床样本库的建立与应用[J]. 中国科学 : 生命科学 , 2024, 54(6): 1041–1049. [Guo D, Jin Y, Liu WD, et al. Development and application of rare diseases biobank[J]. Scientia Sinica(Vitae), 2024, 54(6): 1041–1049.] DOI: 10.1360/SSV–2023–0038.

13 Zhu C, Cotton F, Kawase H, et al. How well can we predict earthquake site response so far? Machine learning vs physics–based modeling[J]. Earthq Spectra, 2023, 39(1): 478–504. DOI: 10.1177/87552930221116399.

14 Anuar MAS, Rahman RZA, Soh AC, et al. A promising wavelet decomposition –NNARX model to predict flood: application to Kelantan river flood[J]. Int J Electr Comput, 2020, 11(2): 89–99. DOI: 10.32985/ijeces.11.2.4.

15 Sima RJ. Combining AI and analog forecasting to predict extreme weather[J/OL]. Eos, 101, [2025–02–13]. DOI: 10.1029/2020eo140896.

16 Wang X. Firth logistic regression for rare variant association tests[J]. Front Genet, 2014, 5: 187. DOI: 10.3389/fgene.2014.00187.

17 Xie Y, Yu Z. Mixture cure rate models with neural network estimated nonparametric components[J]. Comput Stat, 2021, 36(4): 2467–2489. DOI: 10.1007/s00180–021–01086–3.

18 Ding Y, Zou J, Fan Y, et al. A digital twin–based testing and data collection system for autonomous driving in extreme traffic scenarios[A]//The 6th International Conference on Video and Image Processing[C]. ICVIP, 2022: 101–109. DOI: 10.1145/3579109.3579127.

19 Hamaguchi R, Sakurada K, Nakamura R. Rare event detection using disentangled representation learning[A]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. IEEE Computer Society, 2019: 9319–9327. DOI: 10.1109/CVPR.2019.00955.

20 Cui T, Dolgov S, Scheichl R. Deep importance sampling using tensor trains with application to a priori and a posteriori rare events[J]. SIAM J Sci Comput, 2024, 46(1): C1–C29. DOI: 10.1137/23M1546981.

21 Olmuş H, Nazman E, Erbaş S. Comparison of penalized logistic regression models for rare event case[J]. Commun Stat–Simul C, 2019, 51(4): 1–13. DOI: 10.1080/03610918.2019.1676438.

22 Zhao L. Event prediction in the big data era: a systematic survey[J]. ACM Comput Surv, 2022, 54(5): 1–37. DOI: 10.1145/3450287.

23 Burns BL, Rhoads DD, Misra A. The use of machine learning for image analysis artificial intelligence in clinical microbiology[J]. J Clin Microbiol, 2023, 61(9): e0233621. DOI: 10.1128/jcm.02336–21.

24 Ashraf MT, Dey K, Mishra S. Identification of high–risk roadway segments for wrong–way driving crash using rare event modeling and data augmentation techniques[J]. Accid Anal Prev, 2023, 181: 106933. DOI: 10.1016/j.aap.2022.106933.

25 Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over–sampling technique[J]. J Artif Int Res, 2002, 16(1): 321–357. DOI: 10.1613/jair.953.

26 Dablain D, Krawczyk B, Chawla NV. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data[J]. IEEE Trans Neural Netw Learn Syst, 2023, 34(9): 6390–6404. DOI: 10.1109/TNNLS.2021.3136503.

27 Arafa A, El–Fishawy N, Badawy M, et al. RN–SMOTE: reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification[J]. J King Saud Univ–Com, 2022, 34(8): 5059–5074. DOI: 10.1016/j.jksuci.2022. 06.005.

28 He H, Bai Y, Garcia EA, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[A]//IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)[C]. IEEE, 2008: 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

29 Branco P, Torgo L, Ribeiro RP. Pre–processing approaches for imbalanced distributions in regression[J]. Neurocomputing, 2019, 343(8): 0925–2312. DOI: 10.1016/j.neucom.2018.11.100.

30 Beinecke J, Heider D. Gaussian noise up–sampling is better suited than SMOTE and ADASYN for clinical decision making[J]. BioData Min, 2021, 14(1): 49. DOI: 10.1186/s13040–021–00283–6.

31 Wang H. Logistic regression for massive data with rare events[A]// Proceedings of the 37th International Conference on Machine Learning[C]. PMLR, 2020: 9829–9836. https://doi.org/10.48550/arXiv.2006.00683.

32 Alejo R, Sotoca JM, Valdovinos RM, et al. Edited nearest neighbor rule for improving neural networks classifications[A]//Advances in Neural Networks–ISNN 2010[C]. Berlin, Heidelberg: Springer, 2010: 303–310. http://dx.doi.org/10.1007/978–3–642–13278–0_39.

33 Nguyen HM, Cooper EW, Kamei K. Borderline over–sampling for imbalanced data classification[J]. Int J Knowl Eng Soft Data Paradigm, 2011, 3(1): 4–21. DOI: 10.1504/IJKESDP.2011.039875.

34 Li J, Fong S, Hu S, et al. Rare event prediction using similarity majority under–sampling technique[A]//Soft Computing in Data Science, 6th International Conference[C]. Singapore: Springer, 2017: 23–39. DOI: 10.1007/978–981–10–7242–0_3.

35 Zhu Y, Hu Y, Liu Q, et al. A hybrid approach for predicting corporate financial risk: integrating SMOTE–ENN and NG Boost[J]. IEEE Access, 2023, 11: 111106–111125. DOI: 10.1109/ACCESS.2023.3323198.

36 Asniara, Maulidevia NU, Surendro K. SMOTE–LOF for noise identification in imbalanced data classification[J]. J King Saud Univ–Com, 2022, 34(6): 3413–3423. DOI: 10.1016/j.jksuci.2021.01.014.

37 Arafa A, El–Fishawy N, Badawy M, et al. RN–autoencoder: reduced noise autoencoder for classifying imbalanced cancer genomic data[J]. J Biol Eng, 2023, 17(1): 7. DOI: 10.1186/s13036–022–00319–3.

38 Thai–Nghe N, Gantner Z, Schmidt–Thieme L. Cost–sensitive learning methods for imbalanced data[A]//The 2010 International Joint Conference on Neural Networks[C]. IJCNN, 2010: 1–8. DOI: 10.1109/IJCNN.2010.5596486.

39 Puhr R, Heinze G, Nold M, et al. Firth's logistic regression with rare events: accurate effect estimates and predictions?[J]. Stat Med, 2017, 36(14): 2302–2317. DOI: 10.1002/sim.7273.

40 Maalouf M, Trafalis TB. Robust weighted kernel logistic regression in imbalanced and rare events data[J]. Comput Stat Data An, 2011, 55(1): 168–183. DOI: 10.1016/j.csda.2010.06.014.

41 He J, Cheng MX. Weighting methods for rare event identification from imbalanced datasets[J]. Front Big Data, 2021, 4: 1–11. DOI: 10.3389/fdata.2021.715320.

42 Rahman MS, Sultana M. Performance of firth–and logF–type penalized methods in risk prediction for small or sparse binary data[J]. BMC Med Res Methodol, 2017, 17(1): 33. DOI: 10.1186/s12874–017–0313–9.

43 Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics, 2000, 42(1): 80–86. https://doi.org/10.2307/1271436.

44 Tibshirani R. Regression shrinkage and selection via the Lasso[J]. J R Stat Soc B, 1996, 58(1): 267–288. DOI: 10.1111/j.2517–6161.1996.tb02080.x.

45 Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. J R Stat Soc B, 2005, 67(2): 301–320. DOI: 10.1111/j.1467–9868.2005.00503.x.

46 Ambler G, Seaman S, Omar RZ. An evaluation of penalized survival methods for developing prognostic models with rare events[J]. Stat Med, 2012, 31(11–12): 1150–1161. DOI: 10.1002/sim.4371.

47 Zhong S, Zhang J, Jiao J, et al. A machine learning case study to predict rare clinical event of interest: imbalanced data, interpretability, and practical considerations[J]. J Biopharm Stat, 2024, 11: 1–14. DOI: 10.1080/10543406.2024.2364722.

48 Li X, Zhu X, Wang H. Distributed Logistic regression for massive data with rare events[J]. Stat Sinica, 2024, 34(4): 2277–2300. DOI: 10.5705/ss.202022.0242.

49 Yağci AM, Aytekin T, Gürgen FS. Balanced random forest for imbalanced data streams[A]//24th Signal Processing and Communication Application Conference[C]. SIU, 2016: 1065–1068. DOI: 10.1109/SIU.2016.7495927.

50 Chawla NV, Lazarevic A, Hall LO, et al. SMOTEBoost: improving prediction of the minority class in boosting[A]//Knowledge Discovery in Databases: PKDD 2003[C]. Berlin, Heidelberg: Springer, 2003: 107–119. DOI: 10.1007/978–3–540–39804–2_12.

51 Efthimiou O. Practical guide to the Meta–analysis of rare events[J]. Evid Based Ment Health, 2018, 21(2): 72–76. DOI: 10.1136/eb–2018–102911.

52 Xu C, Furuya–Kanamori L, Lin L. Synthesis of evidence from zero–events studies: a comparison of one–stage framework methods[J]. Res Syn Meth, 2022, 13(2): 176–189. DOI: 10.1002/jrsm.1521.

53 Jansen K, Holling H. Rare events Meta–analysis using the Bayesian beta–binomial model[J]. Res Synth Methods, 2023, 14(6): 853–873. DOI: 10.1002/jrsm.1662.

54 Zabriskie B, Corcoran C, Senchaudhuri P. A comparison of confidence distribution approaches for rare event Meta–analysis[J]. Stat Med, 2021, 40(24): 5276–5297. DOI: 10.1002/sim.9125.

55 Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Cambridge, Mass: The MIT Press, 2016: 1–775.

56 Goodfellow I, Pouget–Abadie J, Mirza M, et al. Generative adversarial nets[A]//Proceedings of the 28th International Conference on Neural Information Processing Systems[C]. NIPS, 2014: 2672–2680. https://www.semanticscholar.org/paper/Generative–Adversarial–Nets–Goodfellow–Pouget–Abadie/86ee1835a56722b76564119437070782fc90eb19.

57 Kingma DP, Welling M. Auto–encoding variational Bayes[EB/OL]. (2022–12–10) [2025–02–12]. http://arxiv.org/abs/1312.6114.

58 Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review[J]. Med Image Anal, 2019, 58: 101552. DOI: 10.1016/j.media.2019.101552.

59 Xiu Z, Tao C, Gao M, et al. Variational disentanglement for rare event modeling[J]. Proc AAAI Conf Artif Intell, 2021, 35(12): 10469–10477. https://pubmed.ncbi.nlm.nih.gov/34888123/.

60 Yang J, El–Bouri R, O'Donoghue O, et al. Deep reinforcement learning for multi–class imbalanced training: applications in healthcare[J]. Mach Learn, 2024, 113(5): 2655–2674. DOI: 10.1007/s10994–023–06481–z.

收稿日期：2024 年 11 月 27 日 修回日期：2025 年 02 月 13 日
本文编辑：冼静怡 周璐敏