考虑边界稀疏样本的非平衡数据处理方法

汇报人：

2024-01-18

REPORTING

CATALOGUE

目录

- 引言
- 非平衡数据问题概述
- 边界稀疏样本特性分析
- 基于重采样技术的非平衡数据处理方法
- 基于代价敏感学习算法的非平衡数据处理方法

CATALOGUE

目录

- 基于集成学习算法的非平衡数据处理方法
- 实验设计与结果分析
- 总结与展望

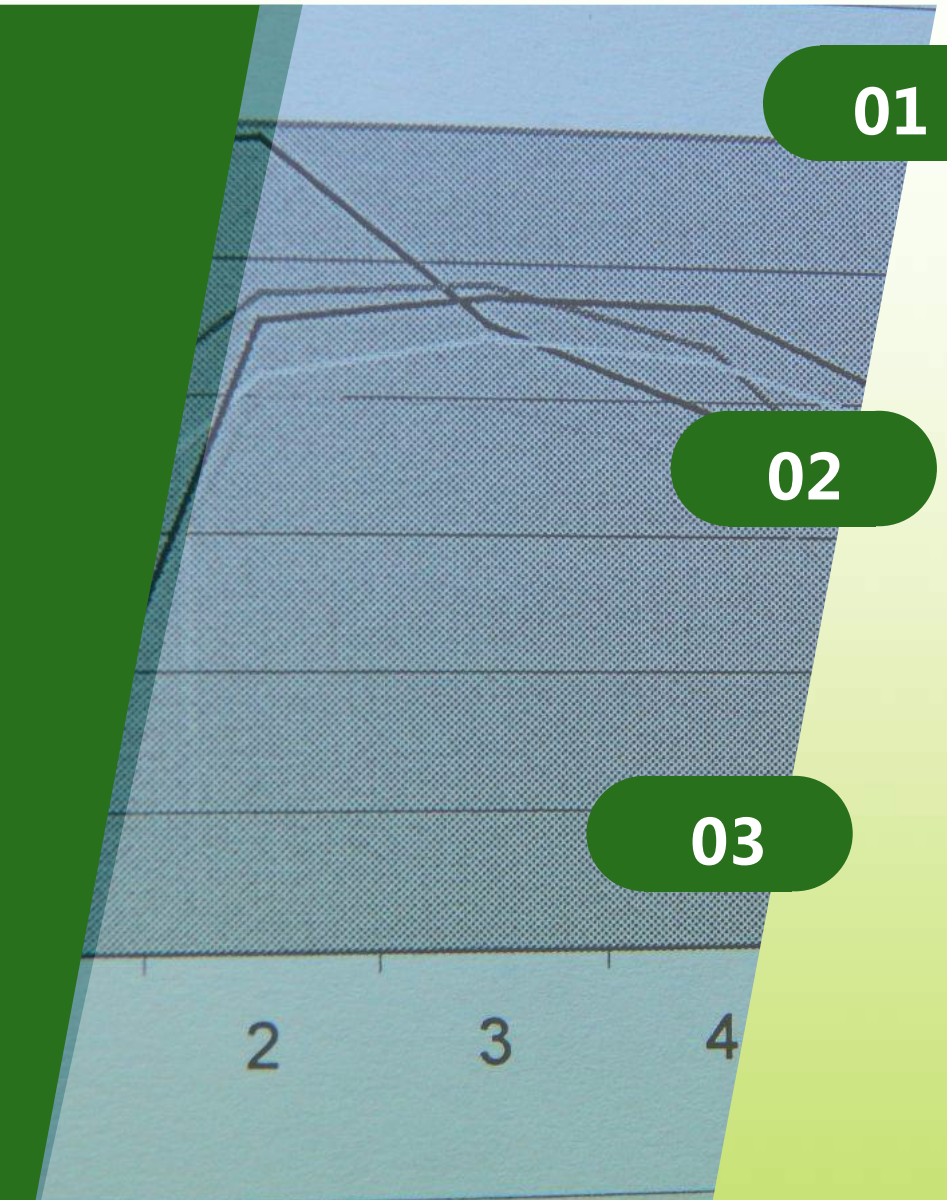
PART 01

引言

CATALOGUE

REPORTING

背景与意义



01

现实应用中的数据不平衡问题

在许多现实应用中，如医疗诊断、欺诈检测等，正常样本和异常样本的数量往往极不平衡，这给机器学习模型的训练和评估带来了挑战。

02

边界稀疏样本的重要性

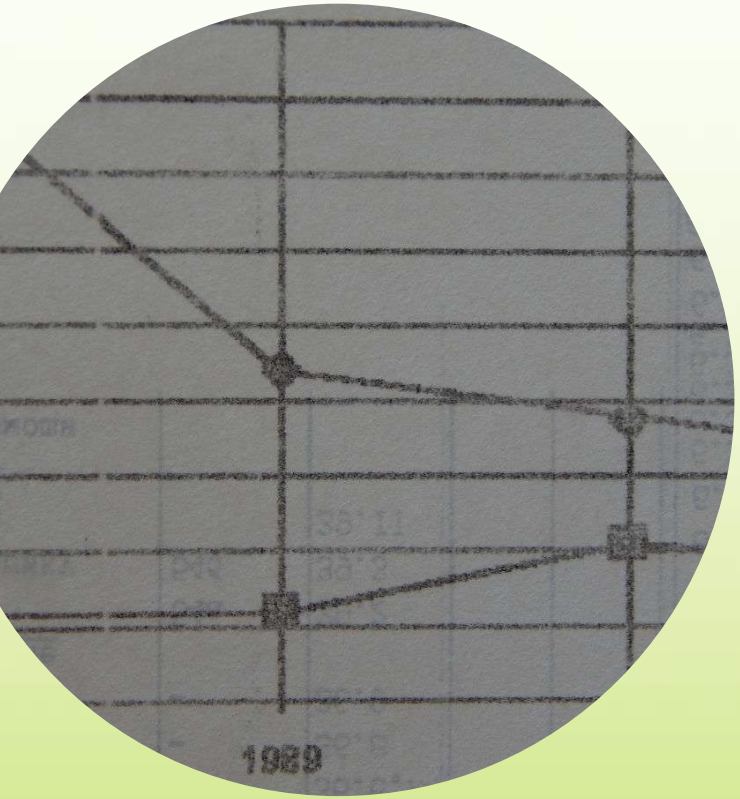
边界稀疏样本指的是那些位于类别边界附近且数量较少的样本。这些样本对于模型的分类性能至关重要，因为它们往往决定了模型的决策边界。

03

非平衡数据处理的意义

通过有效的非平衡数据处理方法，可以提高模型对少数类样本的识别能力，从而改善模型的整体性能，这对于实际应用具有重要意义。

国内外研究现状



过采样方法

通过增加少数类样本的数量来实现数据平衡，如SMOTE算法及其改进算法。

欠采样方法

通过减少多数类样本的数量来实现数据平衡，如随机欠采样、Tomek Links等。

代价敏感学习

通过为不同类别的样本设置不同的误分类代价来调整模型的训练过程。

集成学习方法

结合多种基学习器来提高模型对少数类样本的识别能力，如Bagging、Boosting等。

本文研究目的和内容

研究目的

本文旨在针对边界稀疏样本的非平衡数据问题，提出一种有效的处理方法，以提高机器学习模型的分类性能。

研究内容

首先，分析边界稀疏样本的特性及其对模型性能的影响；其次，提出一种基于合成样本和特征选择的非平衡数据处理方法；最后，在多个公开数据集上进行实验验证，并与现有方法进行对比分析。

PART 02

非平衡数据问题概述

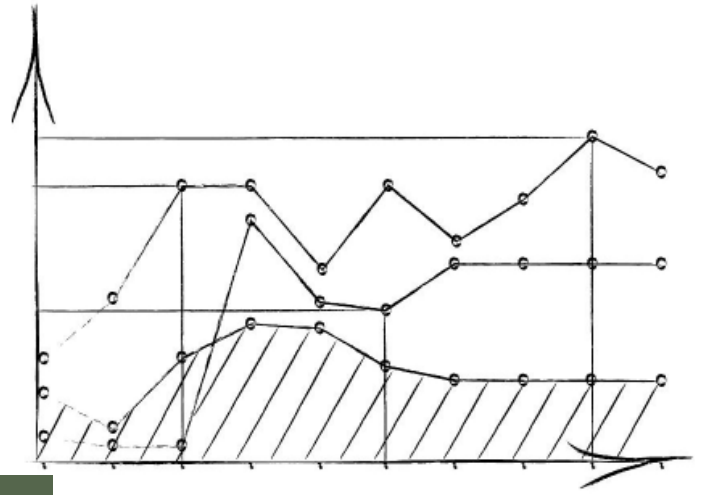
CATALOGUE

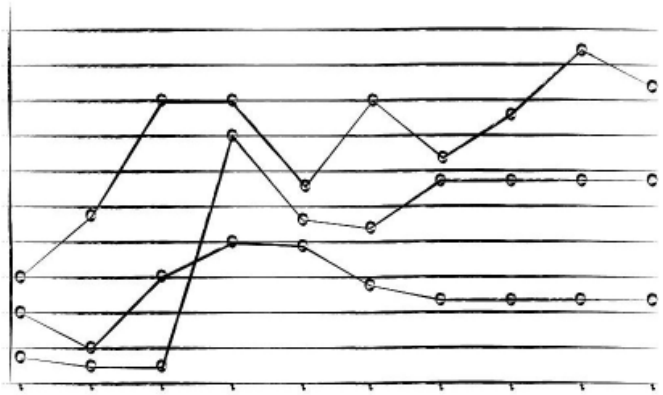
REPORTING

非平衡数据定义及分类

定义

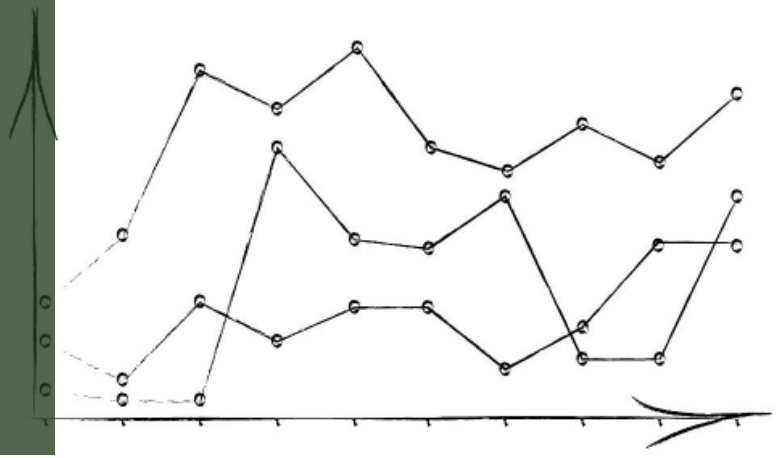
非平衡数据是指在分类问题中，不同类别的样本数量存在明显差异的数据集。

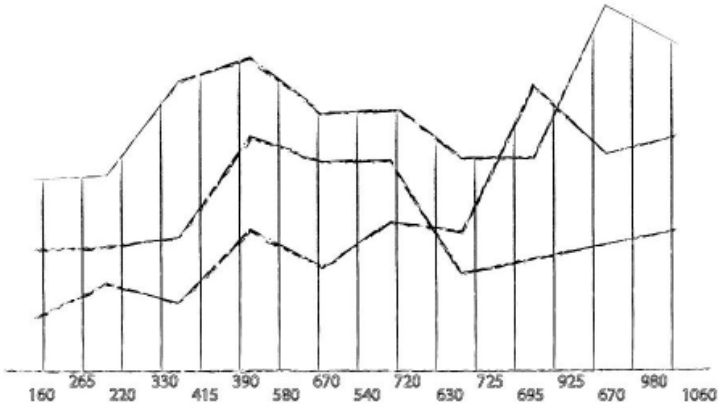




分类

根据样本数量差异的程度，非平衡数据可分为轻度非平衡、中度非平衡和重度非平衡三类。





非平衡数据对模型性能影响

准确率偏差

模型在训练过程中可能受到多数类样本的影响，导致对少数类样本的识别能力下降，从而降低了整体准确率。

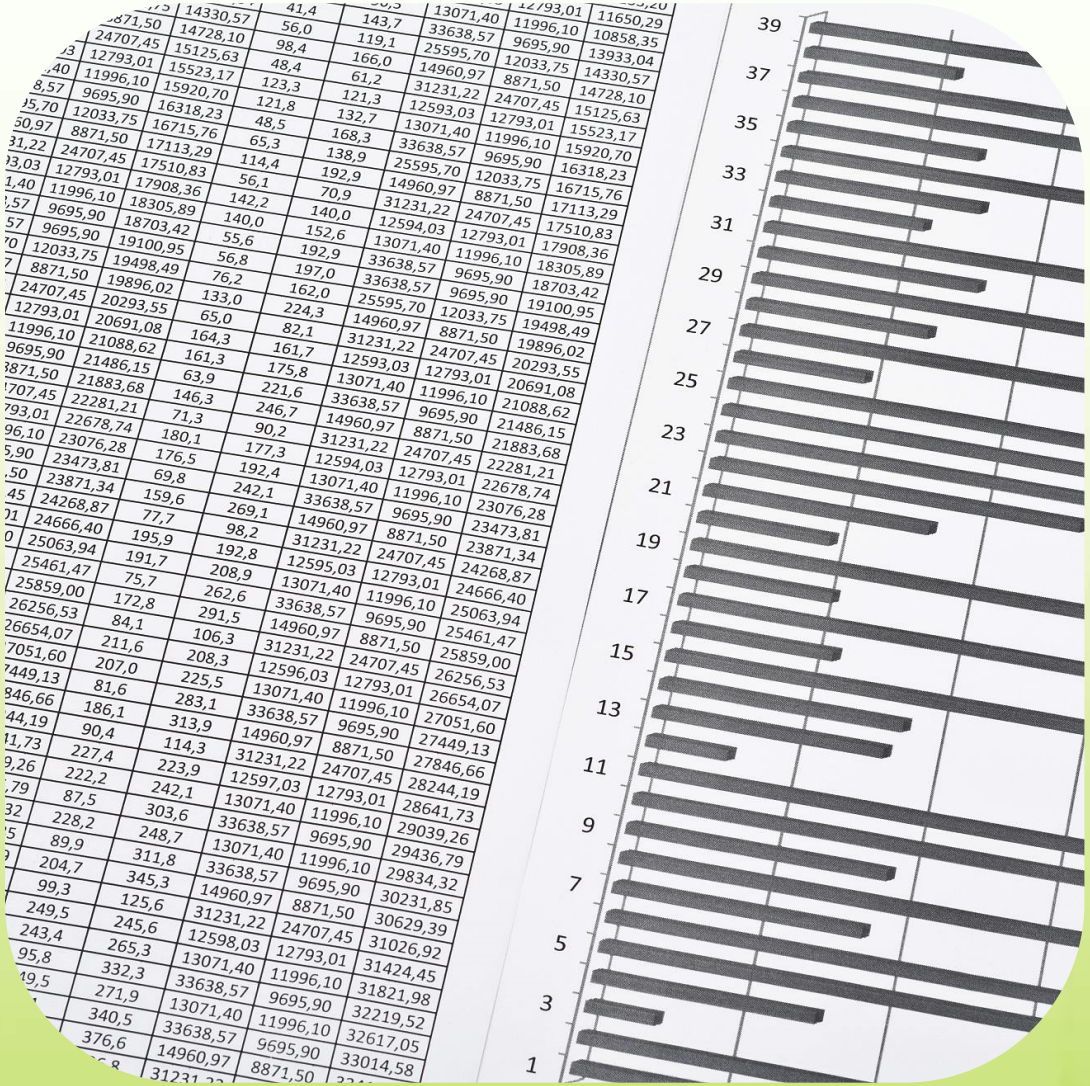
过拟合风险

当数据集严重不平衡时，模型可能过度拟合多数类样本的特征，而忽视少数类样本的重要信息。

泛化能力下降

非平衡数据可能导致模型在训练集上表现良好，但在测试集上性能不佳，即模型的泛化能力下降。

传统处理方法及其局限性



采样方法

包括过采样（增加少数类样本数量）和欠采样（减少多数类样本数量）。但过采样可能导致过拟合，欠采样则可能丢失重要信息。

代价敏感学习

通过为不同类别的样本分配不同的权重，使模型在训练过程中更加关注少数类样本。但权重的选择需要经验和实验调整。

集成学习方法

通过构建多个基分类器并结合它们的预测结果来提高整体性能。但集成学习的效果受到基分类器多样性和结合策略的影响。

PART 03

边界稀疏样本特性分析

CATALOGUE

REPORTING

边界稀疏样本定义及识别方法

边界稀疏样本定义

位于分类边界附近且数量较少的样本，对于分类器的性能具有重要影响。

识别方法

基于距离度量、密度估计等方法识别边界稀疏样本。

边界稀疏样本对分类器性能影响

降低分类器性能

由于边界稀疏样本数量较少，容易被分类器忽略，导致分类器性能下降。

增加模型过拟合风险

分类器在处理非平衡数据时，容易对多数类样本过拟合，忽略边界稀疏样本，进一步降低性能。

边界稀疏样本处理策略探讨

重采样策略

通过过采样少数类或欠采样多数类的方法平衡数据集，使分类器能够更好地关注边界稀疏样本。

特征选择策略

选择与分类任务相关的特征，降低特征维度，减少噪声干扰，提高分类器对边界稀疏样本的关注度。

集成学习策略

通过构建多个基分类器并结合它们的预测结果来提高整体性能，有效应对边界稀疏样本带来的挑战。

代价敏感学习策略

为不同类别的样本分配不同的误分类代价，使得分类器在处理非平衡数据时能够更多地关注边界稀疏样本。

PART 04

基于重采样技术的非平衡数据处理方法

CATALOGUE

REPORTING

过采样技术原理及实现方法

原理

过采样技术通过增加少数类样本的数量来实现数据平衡。它通过对少数类样本进行复制或者生成新的少数类样本来增加其数量，从而使得数据集中各类别的样本数量接近。

随机过采样

随机选择少数类样本进行复制，直到达到所需的样本数量。

SMOTE

通过对少数类样本及其近邻进行线性插值来生成新的少数类样本。

ADASYN

根据少数类样本的分布情况动态生成新的少数类样本，重点关注那些难以学习的样本。