* **机器学习——2023秋季学期**



**Differentiable Architecture Search: An In-depth Analysis**

——关于DARTs（神经网络搜索算法）的阅读报告

|  |  |
| --- | --- |
| 课 程： | 机器学习-2023年秋季课程 |
| 数据集选择： | [DARTS: DIFFERENTIABLE ARCHITECTURE SEARCH](https://arxiv.org/abs/1806.09055) |
| 课 程 教 师： | **曹 鹏** |
| 班 级： | 人工智能2101班 |
| 学 院： | 计算机科学与工程学院 |
| 作 者： | **谢 山** |
| 学 号： | **20216404** |

从UCI网站中选取不均衡数据集，进行不均衡数据的实验分析，可从以下几个研究点（不需要全做，也不限于以下几个点）进行探索，并完成一个报告。  
1. 不均衡数据对传统分类器的影响  
2. 不同采样算法的效果对比及分析  
3. 针对SMOTE算法的问题，可有针对性地提出新的解决方案  
会根据报告的选题角度，实验的数量，分析的深度，创新性和报告写作的规范性，进行综合打分。

## 摘要

## 引言

在当今信息时代，机器学习在处理各种应用中发挥着关键作用，尤其是在分类问题中。然而，面对现实中的不均衡数据集，传统分类器可能面临着严重的性能挑战。我们的研究旨在深入探究不均衡数据对传统分类器的影响，并比较不同采样算法的效果，以提高分类器在这一背景下的性能。

* 1. 研究背景

数据不均衡是指在数据集中各个类别的样本分布不平衡，这在实际问题中是非常常见的。例如，在信用卡欺诈检测中，正常交易相对于欺诈交易的比例可能极其不均衡。这种不均衡性可能导致传统分类器对多数类别过度拟合，而对少数类别的识别能力较弱。

* 1. 研究问题和目标

在这一背景下，我们的研究关注于以下几个问题：不均衡数据对经典分类器的性能产生何种影响？在不同采样算法的支持下，分类器的性能是否得以改善？通过实验，我们将选取信用卡欺诈数据集，并使用决策树、K-最近邻、支持向量机和逻辑回归这四个经典分类器，从而深入分析这些问题。

* 1. 数据集和分类器的选择

我们选择信用卡欺诈数据集作为我们实验的基准，因为它代表了一个典型的不均衡数据场景。同时，我们选取了决策树、K-最近邻、支持向量机和逻辑回归这四个分类器，以确保我们的研究具有一定的广泛性和可比性。

* 1. 采样算法的选取

🡨------🡪

通过这个研究，我们期望能够为处理不均衡数据集的机器学习任务提供实用的指导，并为选择合适的采样算法提供有力的支持。这对于提高分类器在真实应用中的可靠性和效果具有重要意义。

## 实验和方法

* 1. 数据集

这一数据集涵盖了欧洲信用卡持卡人在2013年9月进行的交易记录，其中详细呈现了为期两天的交易情况。数据集中包含492笔欺诈交易，总共记录了284,807笔交易。值得注意的是，该数据集的正类别（欺诈交易）在所有交易中仅占0.172%。

该数据集主要包含数值型输入变量，这些变量是通过主成分分析（PCA）进行转换得到的结果。为了保护用户隐私，数据集中的特征V1、V2、……、V28表示通过PCA获得的主成分，而未经PCA转换的特征则包括"Time"和"Amount"。其中，"Time"特征表示每笔交易与数据集中第一笔交易之间经过的秒数，而"Amount"特征则表示交易金额。最后，特征"Class"作为因变量，其取值为1表示欺诈交易，取值为0表示非欺诈交易。

* 1. 评价指标

# 采用AUPRC，为什么不选取混淆矩阵和准确率

* 1. 实验方法
  2. 分类器设置
     1. Decision Tree—参数设置
     2. KNN
     3. SVM
     4. LR
  3. 采样方法选取
     1. 过采样
        1. 随机过采样
        2. SMOTE
     2. 欠采样
        1. 随机欠采样
        2. Tomek links
        3. Cluster Centroids
     3. 混合采样
        1. SMOTE+Tomek links

## 实验结果

## 结论

## 参考文献