# 势函数聚类的优化下采样SVM分类方法

汇报人：

2024-01-13

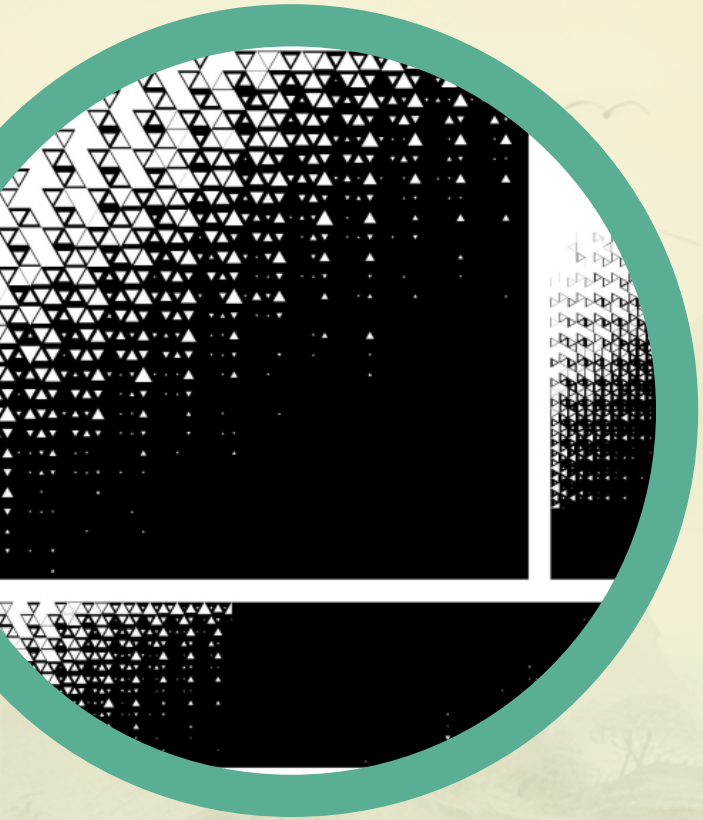
# 目录


- 引言
- 势函数聚类算法
- 优化下采样策略
- SVM分类器构建与训练
- 实验设计与结果分析
- 总结与展望

# 01

# 引言

# 研究背景与意义



## 大数据时代下的分类问题

随着互联网和物联网技术的快速发展，数据量呈现爆炸式增长，如何有效地处理和分析这些数据，并从中提取有用的信息，成为当前研究的热点问题。

## 不平衡数据分类的挑战

在实际应用中，很多数据集存在类别不平衡的问题，即某一类别的样本数量远远大于其他类别。传统的分类算法在处理这类问题时往往效果不佳，因此需要研究专门针对不平衡数据的分类方法。

## 势函数聚类与SVM的结合

势函数聚类是一种基于数据点之间相似度的聚类方法，能够有效地处理大规模数据集。支持向量机（SVM）是一种广泛应用的分类算法，具有优秀的泛化性能。将势函数聚类与SVM相结合，有望提高不平衡数据分类的准确性和效率。

# 国内外研究现状及发展趋势

## 国内外研究现状

目前，国内外学者已经提出了一些基于采样、代价敏感学习等策略的不平衡数据分类方法。其中，下采样方法通过减少多数类样本的数量来平衡数据集，但可能导致重要信息的丢失。因此，如何在下采样过程中保留关键信息，成为当前研究的重点。

## 发展趋势

随着深度学习等技术的不断发展，未来不平衡数据分类方法将更加注重模型的自适应能力和可解释性。同时，针对特定领域和应用场景的不平衡数据分类方法也将得到更多关注。

# 研究内容与创新点

- 研究内容：本研究旨在提出一种基于势函数聚类的优化下采样SVM分类方法。首先，利用势函数聚类对多数类样本进行聚类，并根据聚类结果选择性地删除部分样本，以实现下采样。然后，将处理后的数据集输入到SVM分类器中进行训练和预测。

# 研究内容与创新点

01 创新点：本研究的创新点主要体现在以下几个方面

02 1. 结合势函数聚类和SVM的优点，提出了一种新的不平衡数据分类方法。

03 2. 在下采样过程中引入聚类思想，能够更准确地识别并保留关键信息。

04 3. 通过实验验证了所提方法在不平衡数据分类中的有效性和优越性。

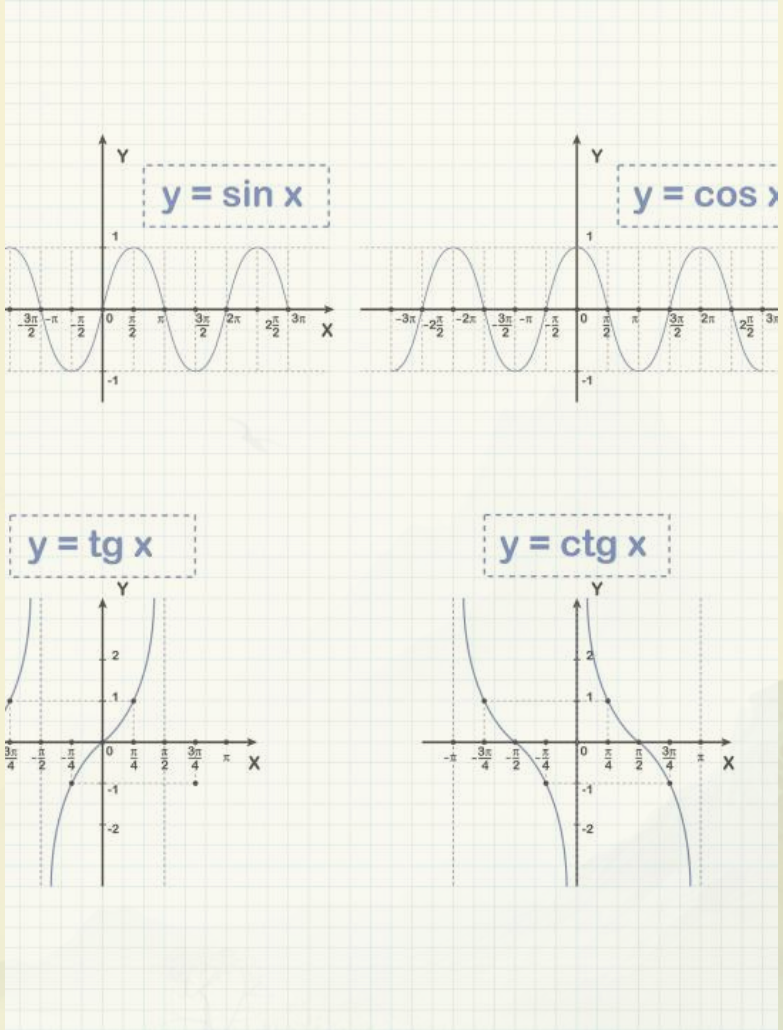
# 02

# 势函数聚类算法

# 势函数定义与性质

## 势函数定义

势函数是用于描述数据点之间相互作用力的函数，通常与数据点之间的距离有关。在聚类算法中，势函数用于衡量数据点之间的相似度或亲密度。




## 势函数性质

势函数具有非负性、对称性和可加性。非负性表示数据点之间的相互作用力总是大于等于0；对称性表示两个数据点之间的相互作用力是相等的；可加性表示多个数据点之间的相互作用力可以相互叠加。

# 聚类算法原理及流程

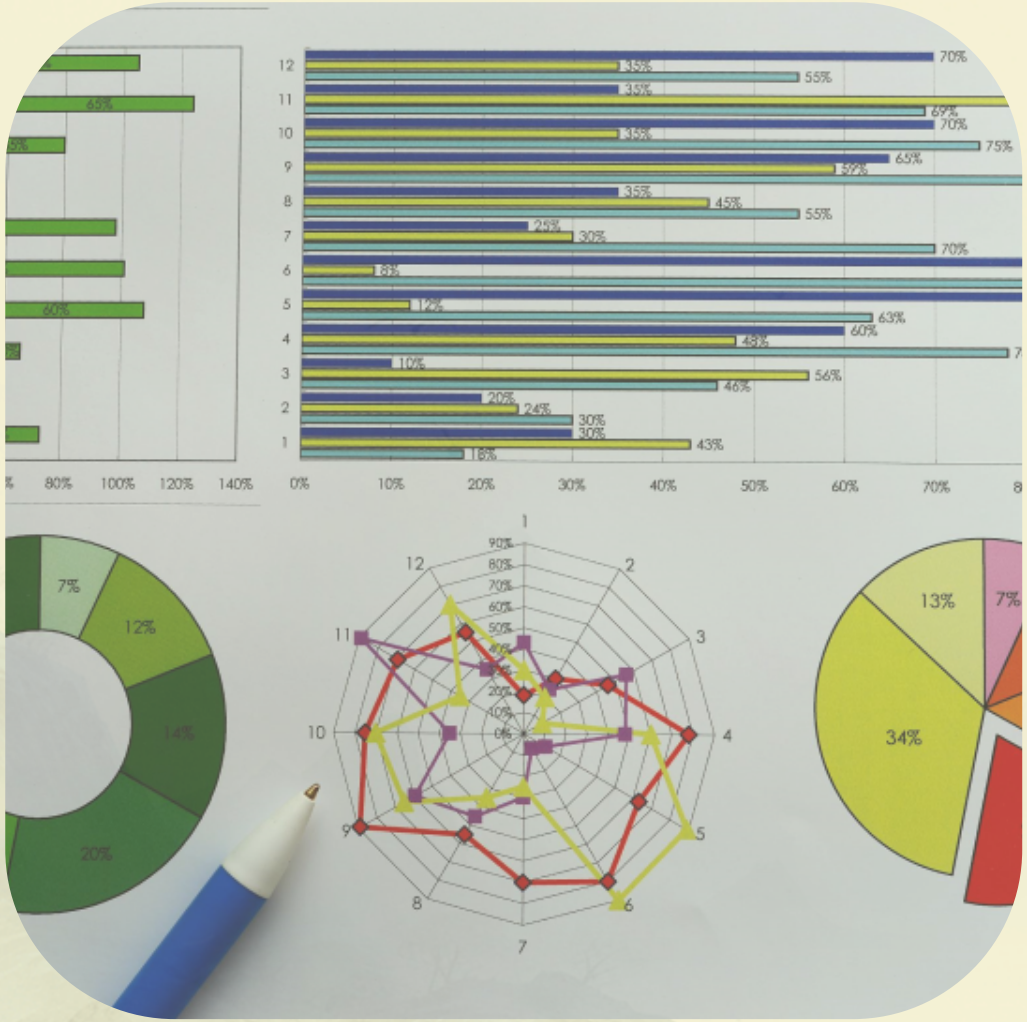
## 聚类算法原理

势函数聚类算法基于数据点之间的势函数值进行聚类。算法通过计算数据点之间的势函数值，将数据点划分为不同的簇，使得同一簇内的数据点相似度高，不同簇之间的数据点相似度低。

## 聚类算法流程

首先，初始化聚类中心或选择代表性的数据点作为聚类中心；然后，计算每个数据点与聚类中心之间的势函数值，并根据势函数值将数据点划分到相应的簇中；接着，更新聚类中心，重新计算数据点与新的聚类中心之间的势函数值，并进行数据点的重新划分；重复以上步骤，直到达到收敛条件或达到最大迭代次数。

# 聚类效果评价指标



## 内部评价指标

内部评价指标主要基于聚类结果本身的信息来评价聚类的效果，如轮廓系数、Calinski-Harabasz指数和Davies-Bouldin指数等。这些指标通过计算簇内紧凑度和簇间分离度来评估聚类的效果。

## 外部评价指标

外部评价指标需要真实的类别标签信息来评价聚类的效果，如调整兰德系数、调整互信息和标准化互信息等。这些指标通过比较聚类结果与真实类别标签的一致性来评估聚类的准确性。

# 03

# 优化下采样策略

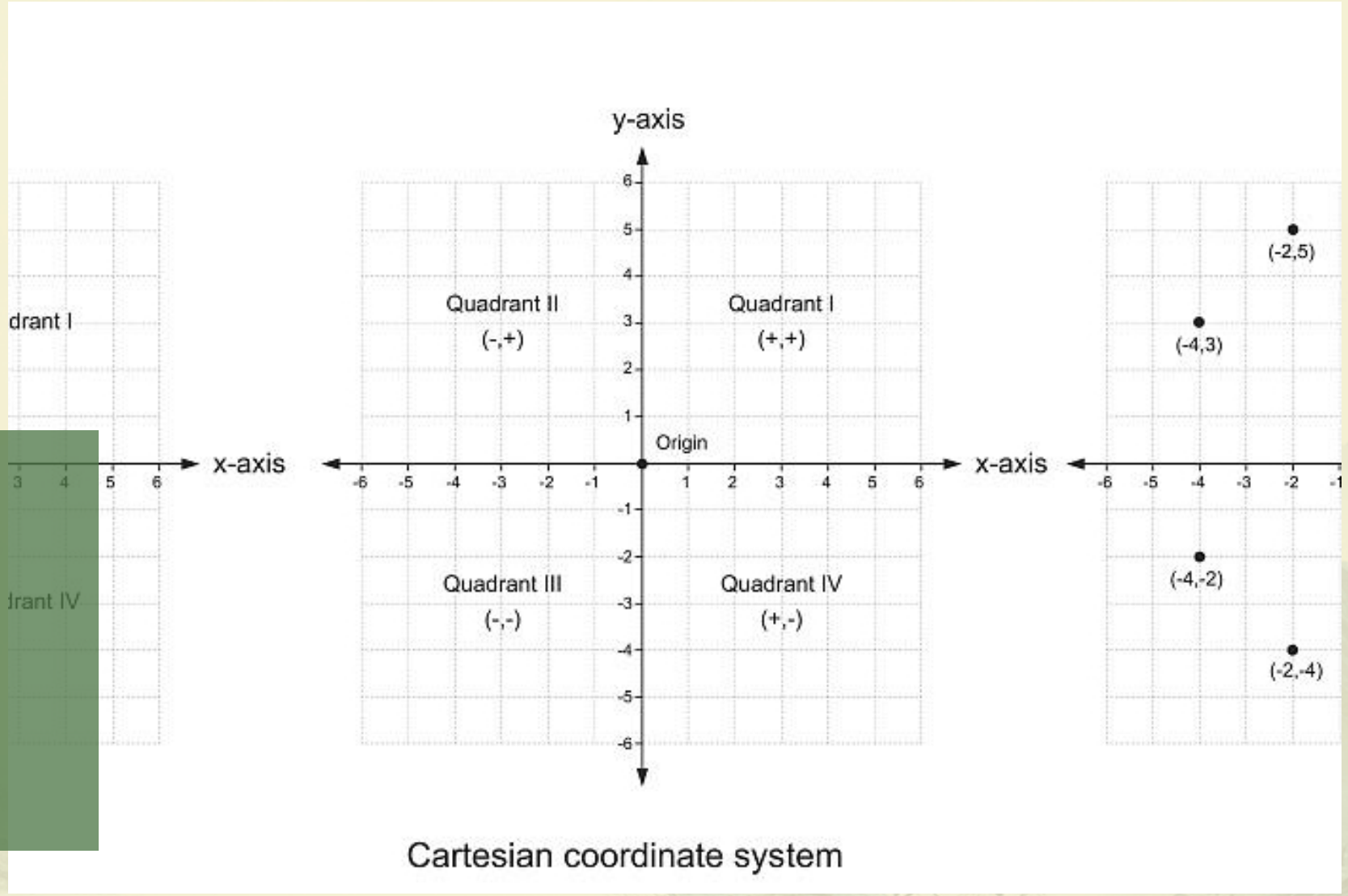
# 下采样方法概述

## 下采样定义

下采样是一种处理不平衡数据集的方法，通过减少多数类样本来平衡数据集中各类别的样本数量。

## 传统下采样方法

随机下采样和启发式下采样是两种常见的传统下采样方法，前者随机选择多数类样本进行删除，后者则根据某些启发式规则来选择要删除的样本。



Cartesian coordinate system