# 不平衡工作总结和展望

## 一、问题：

### 不平衡问题：

#### 类内不平衡

少数类中聚成不同的簇，且各个簇之间的样本数量差值较大

#### 类间不平衡

不同类别样本数量差距较大

#### 数据稀疏

绝对稀疏和相对稀疏

稀疏问题的定义，是否需要区分多数类和少数类，

针对少数类绝对稀疏，

#### 边界重叠问题

少数类和多数类的簇边界有重叠，重叠区域的样本需要重点考虑。

#### 噪声样本

某类样本中的部分样本单独出现在不同类别的簇中，很大可能会被作为噪声样本处理，除此外，少数类的抗真实噪声能力较差，

### 总结

1上述问题中，类间、类内不平衡，噪声样本，边界重叠，都是样本分布中出现的问题，

2 哪个问题对分类结果的影响程度最高？

就目前所阅读的文献而言，边界重叠问题研究较多；噪声样本问题研究较少；类间不平衡则是传统的不平衡问题，一直在研究中；类内不平衡的样本中，少数类样本可能会被作为噪声样本处理，从而演变成噪声样本问题；

### 高维不平衡问题：

定义：维度绝对高；维度和样本数量差距较大，相对高维；

## 二、衡量指标

### 2.1 数据衡量指标

1. 不平衡率，即多数类：少数类，

2. 广义不平衡率 利用样本周围的类标与本身的类标是否相同，计算这个均值，基于数据的分布描述

数据分布： 密度，

数量指标：不平衡率

数据分散程度：熵

分类结果：F1值，召回率，准确率，AUC，g-mean，

## 三 数据集

1. 数据集来源
2. 每个数据集中分别出现了哪些不平衡问题？
3. 如果现有数据集不能体现出某种特殊问题，可以考虑人工合成数据集，有哪些合成工具？