天善智能
TIANSHAN SOFT

第16章　不平衡数据问题及处理

《Python数据科学：技术详解与商业实践》

讲师：Ben

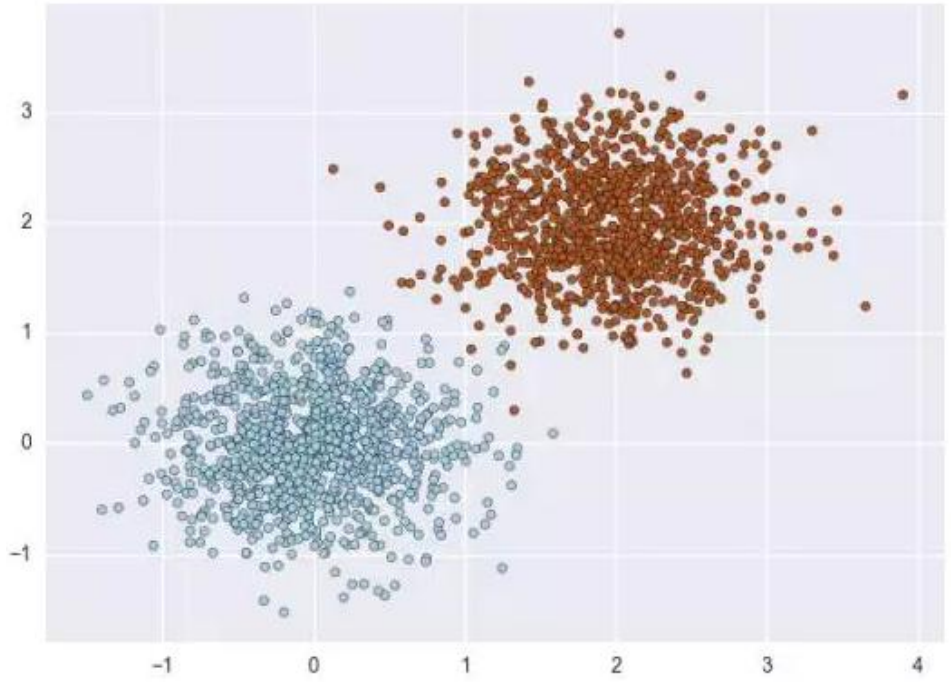
自我介绍

- 天善商业智能和大数据社区　　讲师 –Ben
- 天善社区 ID - Ben_Chang
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

主要内容

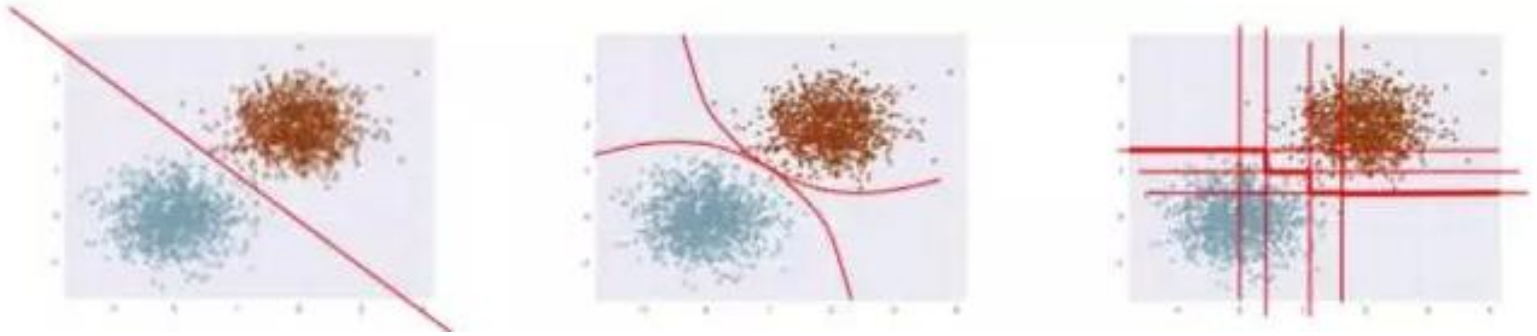
- 不平衡分类概述
- 欠采样法
- 过采样法
- 综合采样

以往学习的算法：



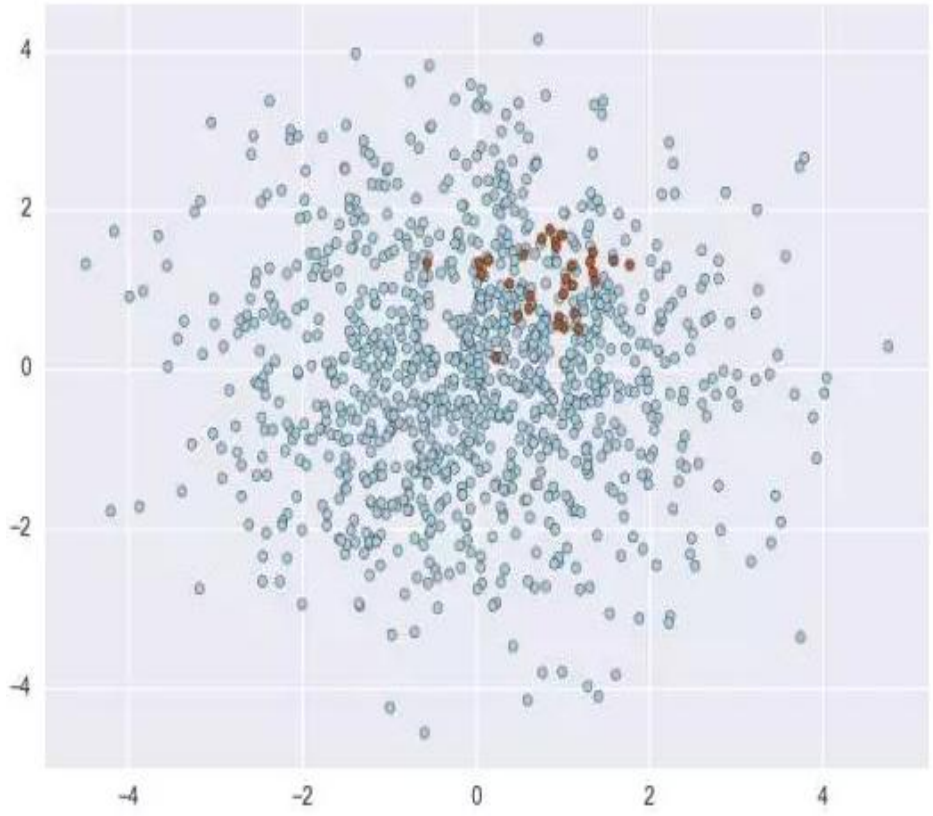
在以上章节中均认为数据是对称分布的，即正负样本的数量相当。这样可以把注意力集中特定算法上，而不被其它问题干扰。

分类算法的目标是尝试学习出一个能够分辨二者的分离器（分类器）。根据不同的数学、统计或几何假设，达成这一目标的方法很多：



不平衡分类数据

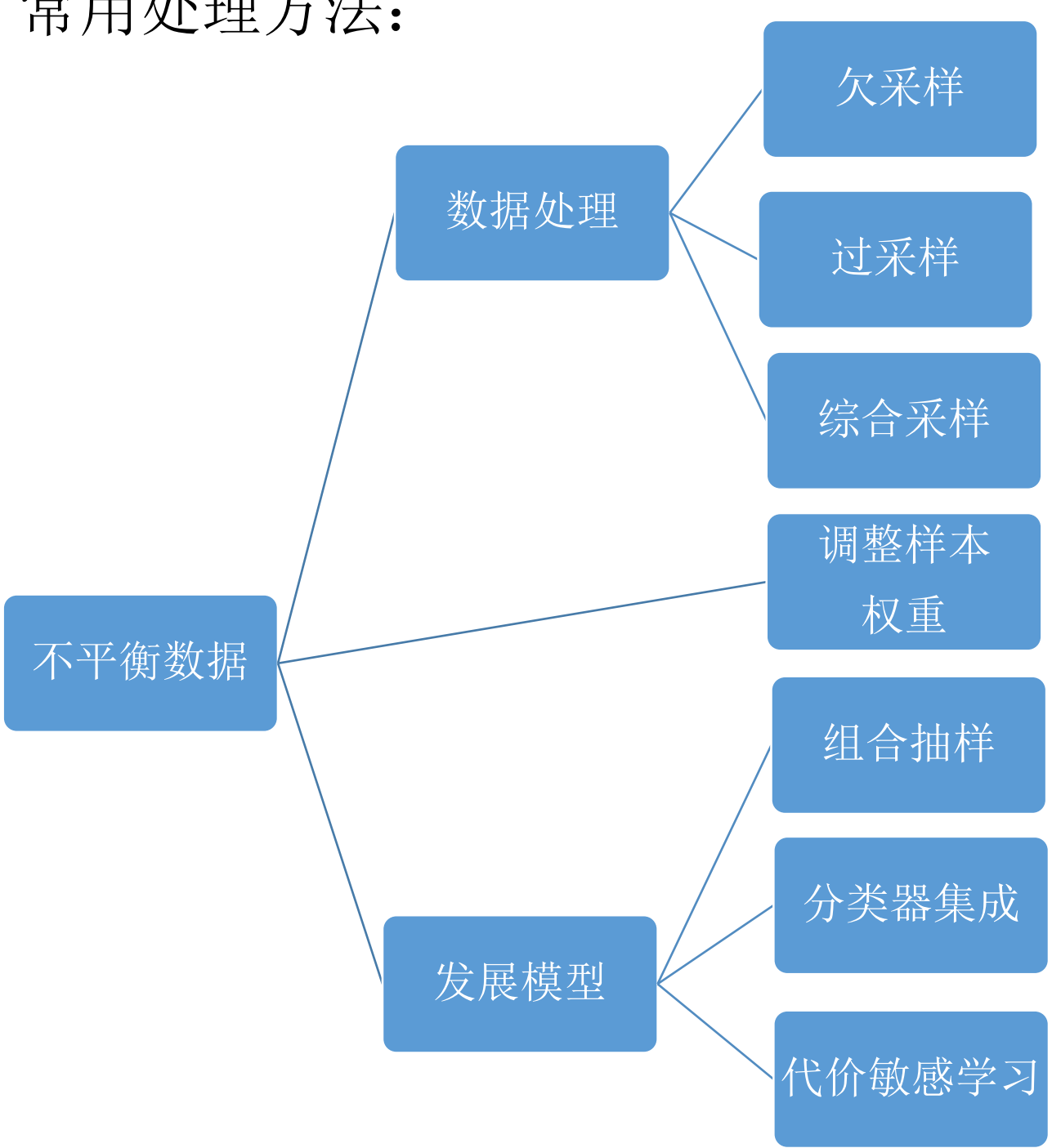
然而，当你开始面对真实的、未加工过的数据时，你会马上注意到，这些数据要嘈杂且不平衡得多。真实数据的散点图看起来更像是这样的：



对于不平衡类的研究通常认为「不平衡」意味着少数类只占 10% 到 20% 。而在现实中，数据库甚至能够比上面的例子更加不平衡。以下是一些例子：

- 每年，约 2% 的信用卡账户是伪造的（多数的欺诈检测领域是极其不平衡的）；
- 针对某一病征的医学筛查通常涵盖了许多没有此病征的人，以检查出少数患者（例：美国的 HIV 感染率约为 0.4%）；
- 每年，硬盘驱动器故障的发生率约为 1%；
- 在线广告的转化率在 10^-3 到 10^-6 的范围区间内；
- 工厂的产品缺陷率一般在 0.1% 左右。

常用处理方法：



随机欠采样法
Tomek Link方法
NearMiss方法
One-Sided Selection方法

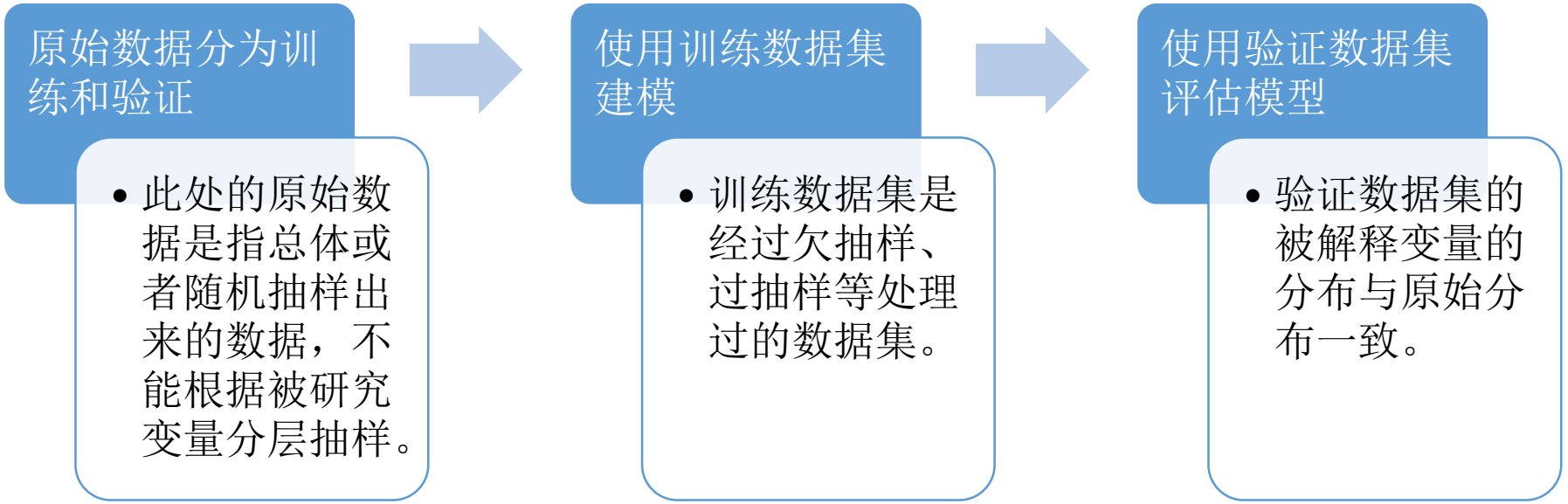
随机过采样
SMOTE方法

SMOTE+Tomek Links方法
SMOTE+KNN方法

Easy Ensemble 方法
(训练集划分法)

GBDT/Xgboost等

数据处理方法的全流程：



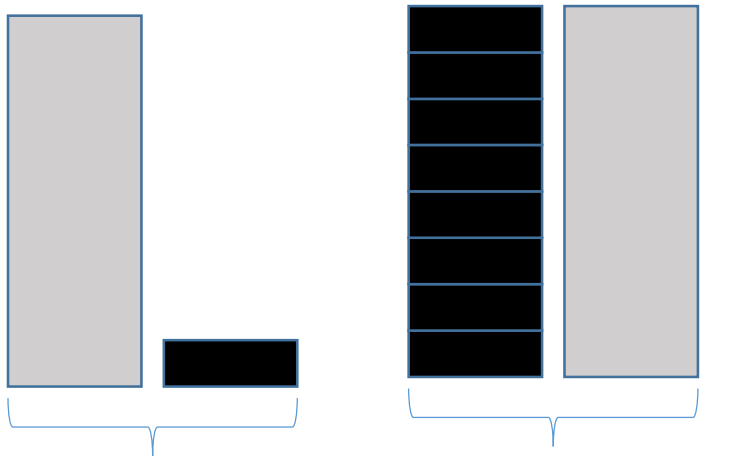
注意事项：

1、评估指标：使用精确度（Precise Rate）、召回率（Recall Rate ）、Fmeasure或ROC曲线、准确度召回曲线（precision-recall curve）；不要使用准确度（Accurate Rate ）

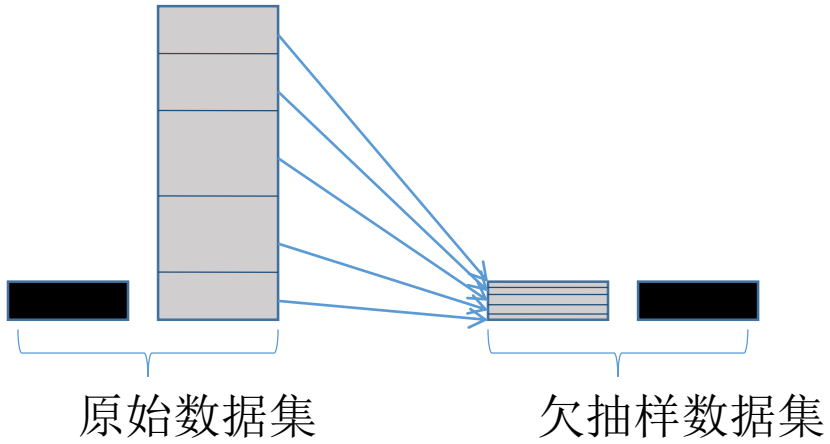
2、不要使用模型给出的标签，而是要概率估计；得到概率估计之后，不要盲目地使用0.50 的决策阈值来区分类别，应该再检查表现曲线之后再自己决定使用哪个阈值。

随机过抽样与欠抽样

过采样

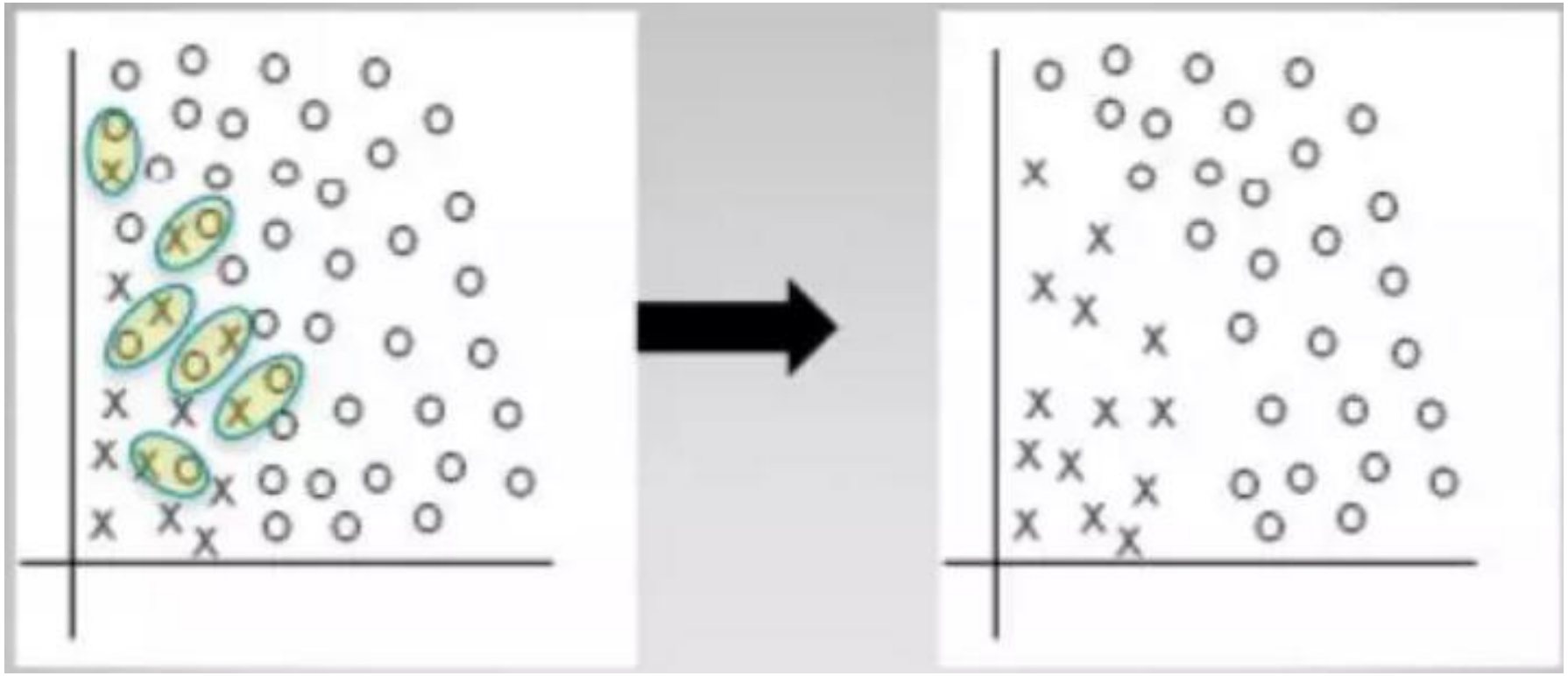


欠采样

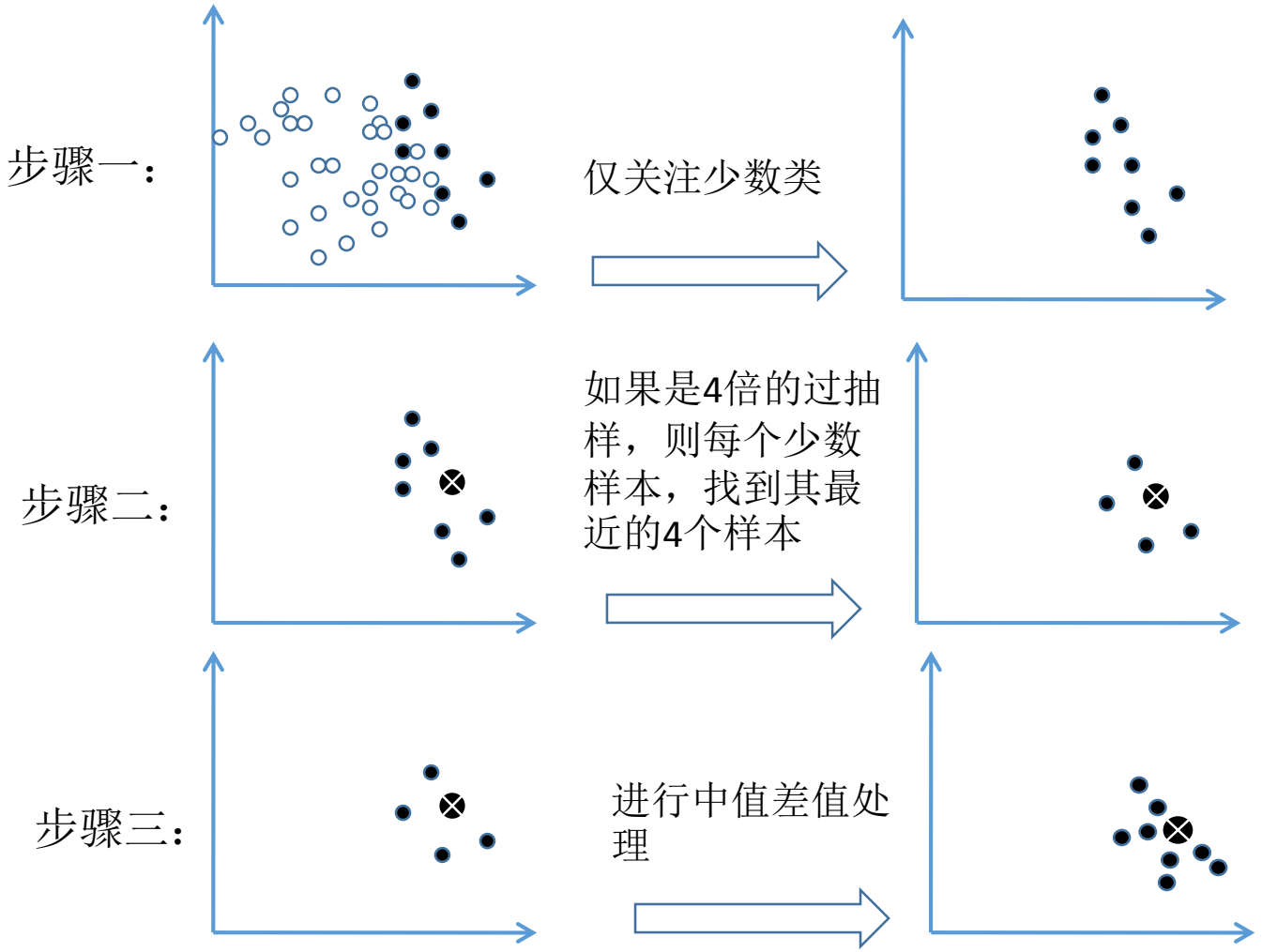


过采样会随机复制少数样例以增大它们的规模。欠采样则随机地少采样主要的类。一些数据科学家（天真地）认为过采样更好，因为其会得到更多的数据，而欠采样会将数据丢掉。但请记住复制数据不是没有后果的——因为其会得到复制出来的数据，它就会使变量的方差表面上比实际上更小。而过采样的好处是它也会复制误差的数量：如果一个分类器在原始的少数类数据集上做出了一个错误的负面错误，那么将该数据集复制五次之后，该分类器就会在新的数据集上出现六个错误。相对地，欠采样会让独立变量（independent variable）的方差看起来比其实际的方差更高。

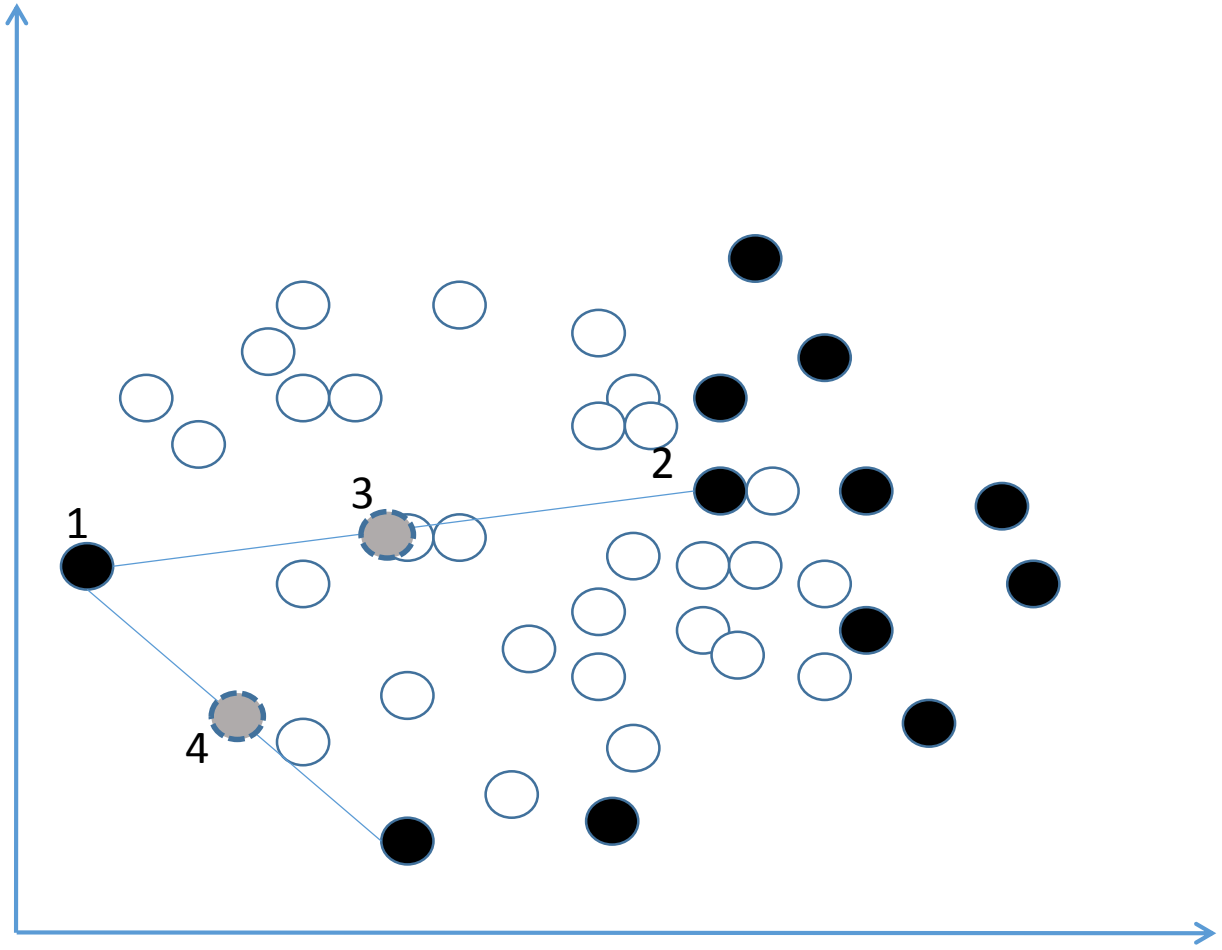
欠采样：Tomek Link方法



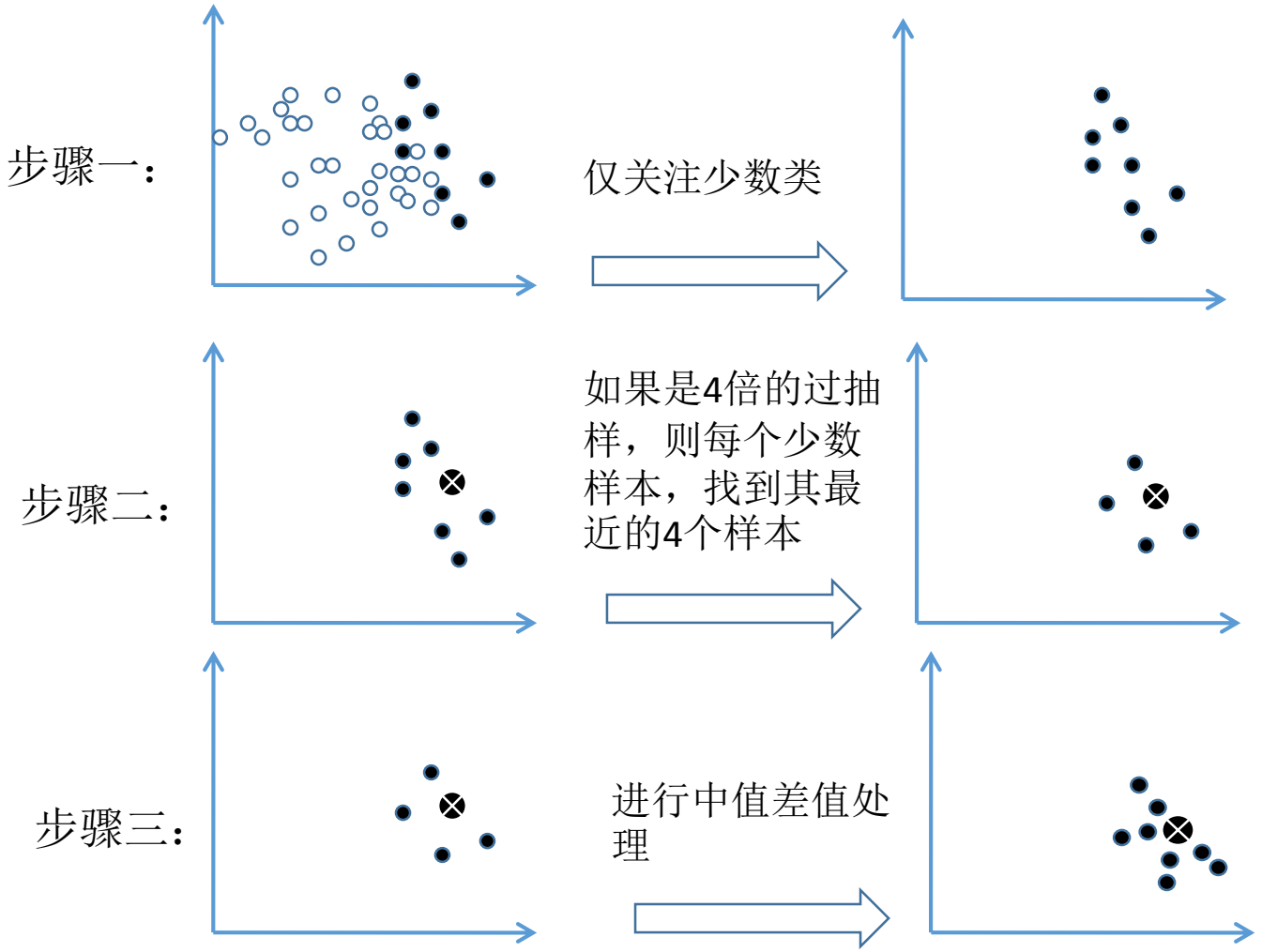
过采样：SMOTE方法的实现



过采样：SMOTE方法的问题

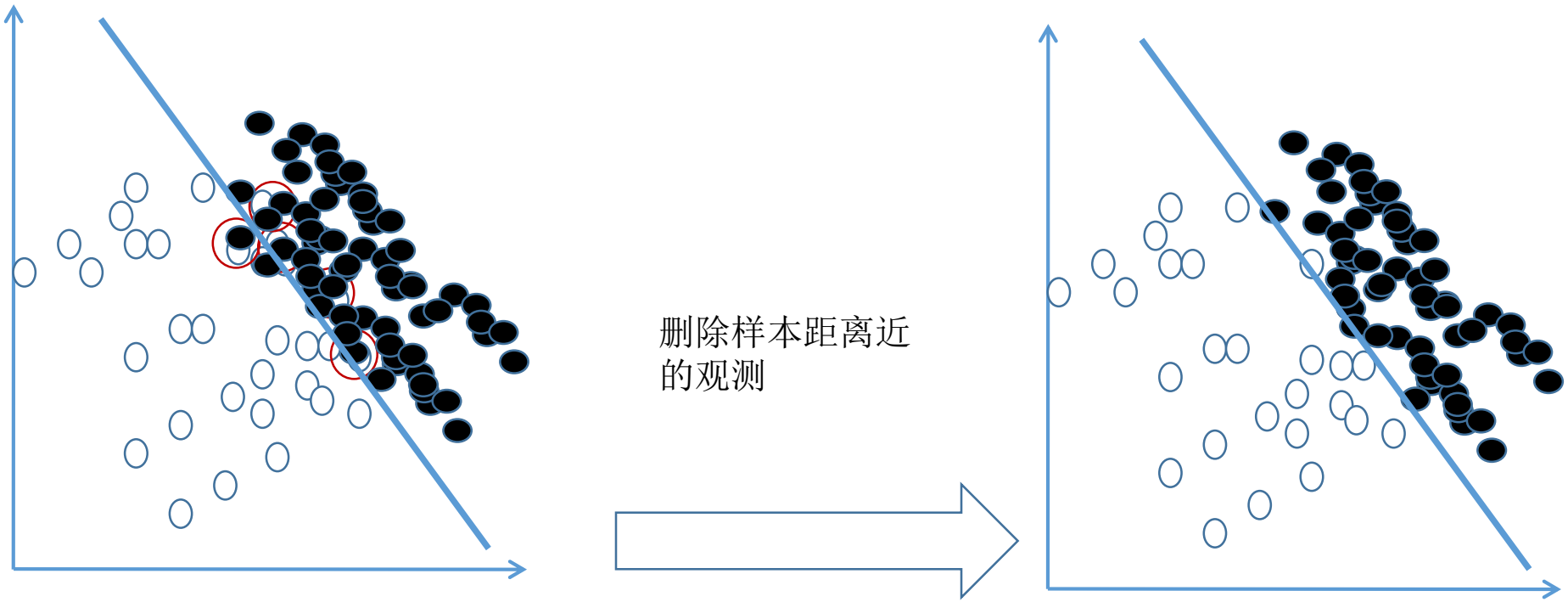


综合采样

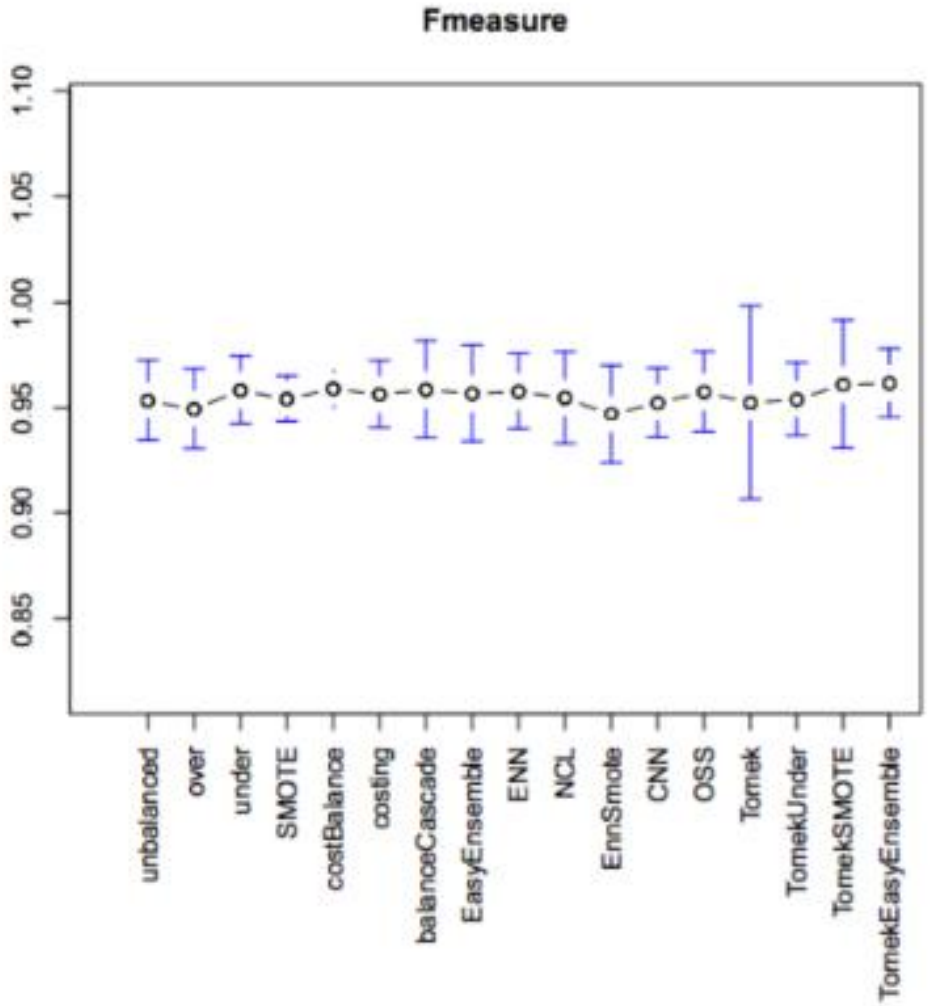


综合采样

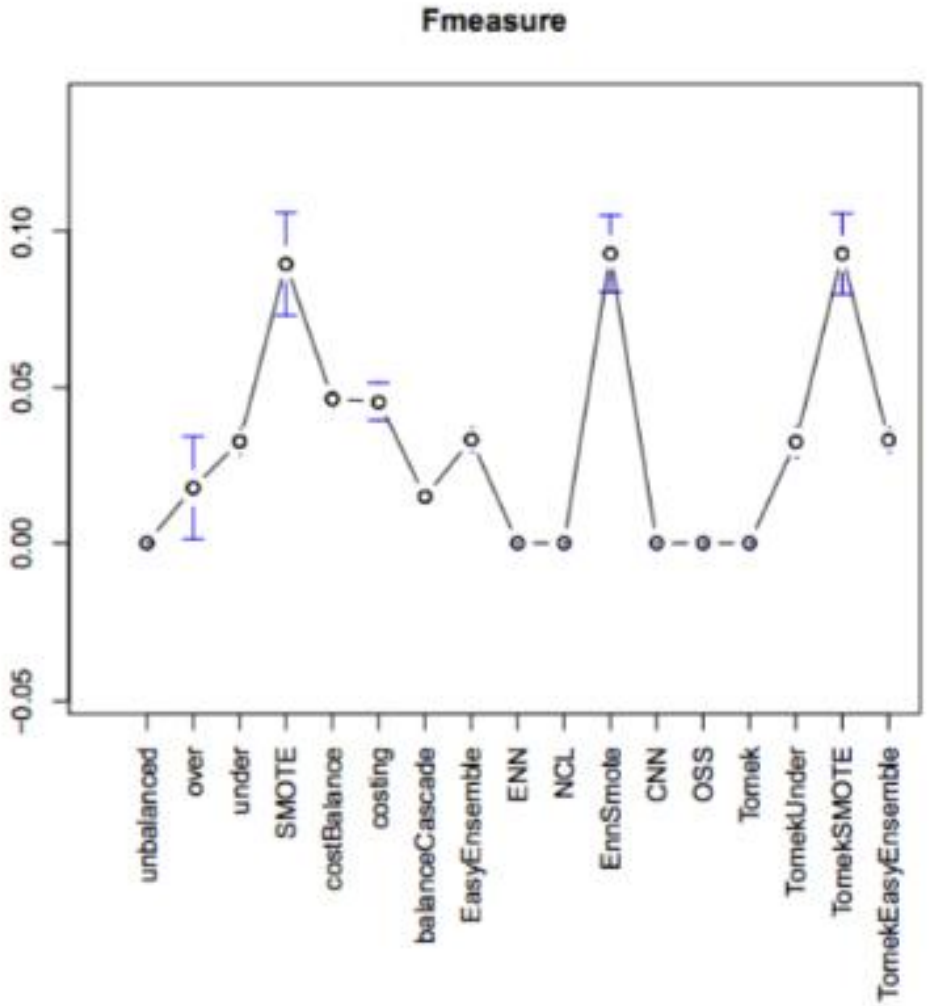
步骤四：



只有检验才能知道



UCI breast cancer dataset



Atos fraud dataset

更多商业智能BI和大数据精品视频尽在 www.hellobi.com

BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI

天善智能
TIANSHAN SOFT