Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

机器学习中的数据不平衡问题

胡雅娴　硕士生

2018年11月25日

内容提要

- 背景简介
- 基本概念
- 主要方法
- 方法选择
- 未来研究发展
- 参考文献

背景简介

- 在生活中，我们解决分类问题时，总会遇到类似以下情况的问题
  - 每年移动硬盘驱动器故障的发生率约为1%
  - 工厂产品的缺陷率一般在0.1%左右
  - 美国的HIV感染率约为0.4%
  - 在信用卡欺诈数据中，每年约2%的信用卡账户是伪造的
  - 研究哈佛大学申请人情况时，只有2%的申请人能够被录取
- 以上问题被称为“大海捞针”问题

背景简介

- 以上问题的共同点
  - 一个类别的数据个数要远远少于另一类别的数据个数，通常认为少数类小于20%时为“不平衡”
  - 数据不平衡问题
  - 在这种情况下，机器学习分类器要从庞大的负面（不相关）样本中，寻找少量正面（相关）样本所蕴含的信息

基本概念

- 数据不平衡分类问题
  - 指训练样本数量在类间分布不平衡的分类问题，即某些类别样本数量远远少于其他类别时的分类问题
  - 信用卡欺诈问题检测，医疗诊断，文本分类等
  - 少数类的识别率更为重要
  - 在二分类中更为常见

基本概念

- 数据不平衡分类问题
  - 在疾病诊断二分类问题中，共有样本100个，其中1个样本属于class1（患病），其余99个样本属于class2（未患病）
  - class1：class2=患病：未患病=1:99

	Diagnosed Sick	Diagnosed Healthy
Sick	0 True Positive	1 False Negative
Healthy	0 False Positive	99 True Negative

基本概念

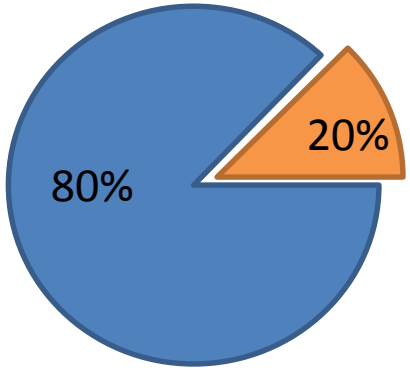
- 数据不平衡在构建模型时会造成怎样的问题呢?
  - 导致模型的预测效果变差，特别是对于样本个数较少的类别，难以达到预期的分类效果

	Diagnosed Sick	Diagnosed Healthy
Sick	0 True Positive	1 False Negative
Healthy	0 False Positive	99 True Negative

  - 灵敏度=TP/(TP+FN)=0/(0+1)=0
  - 特异度=TN/(FP+TN)=99/(0+99)=1
  - 低灵敏度=高漏诊率　高特异度=低误诊率

基本概念

- 为什么大部分机器学习算法在不平衡数据集上表现不佳?
  - 算法本身假设数据集的类分布均衡，同时它们假定不同类别的误差带来相同的损失
  - 算法本身是精度驱动的，即该模型的目标是最小化总体误差，而小类对于总体误差的贡献很低，使模型更偏向于预测结果为比例更大的类别



主要方法

- 扩大数据集
- 尝试其他评价指标
- 尝试不同的分类算法
- 对数据集进行重采样
- 合成人工数据
- 代价敏感学习法

扩大数据集

- 当遇到类别不均衡问题时，首先应该想到，是否可以再增加数据（一定要有小类样本数据）
- 增加小类样本数据时可能又增加了大类数据，对大类数据欠采样

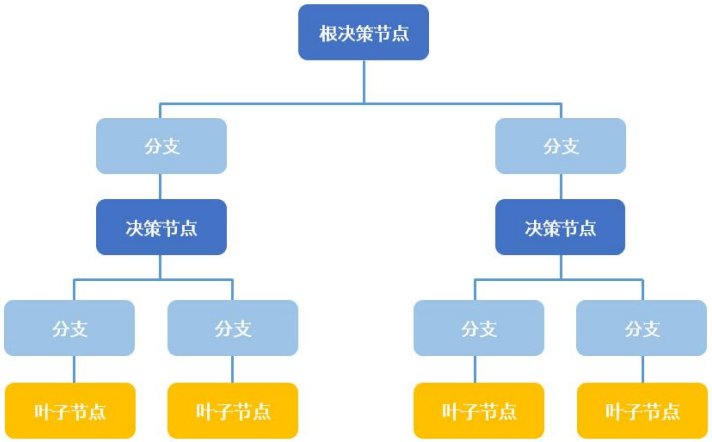
尝试其他评价指标

- 不能简单地使用分类准确率
- 混淆矩阵、精确度、召回率
- F1得分
- ROC曲线
- ……

	Diagnosed Sick	Diagnosed Healthy
Sick	True Positive	False Negative
Healthy	False Positive	True Negative

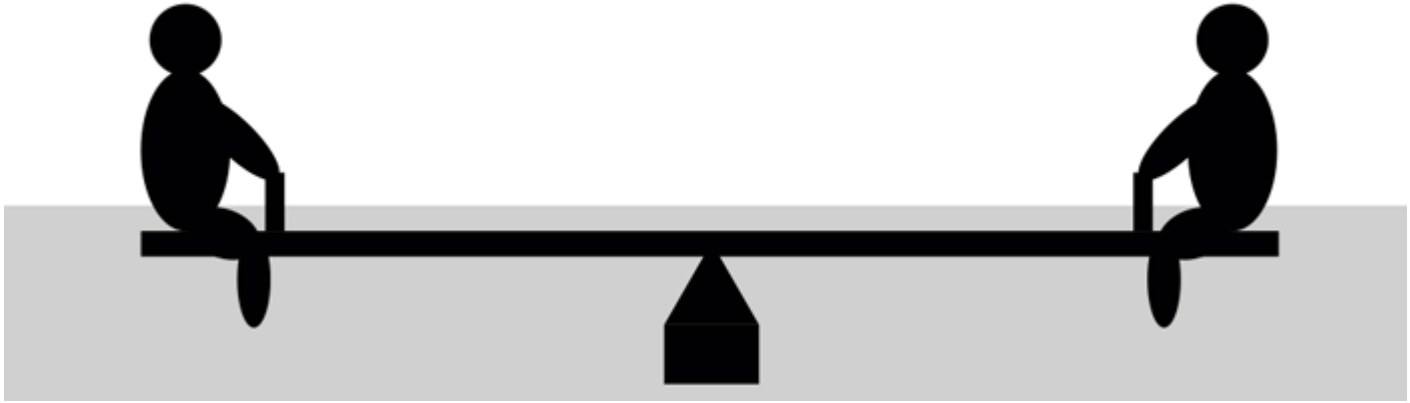
尝试不同的分类算法

- 使用在类别不均衡数据上表现较好的算法
- 决策树，如C4.5、C 5.0、随机森林等



对数据集进行重采样

- 调整原数据集的样本量，把不平衡数据修正为平衡数据，使得不同类的数据比例一致
  - 欠采样法(Undersampling)
  - 过采样法(Oversampling)



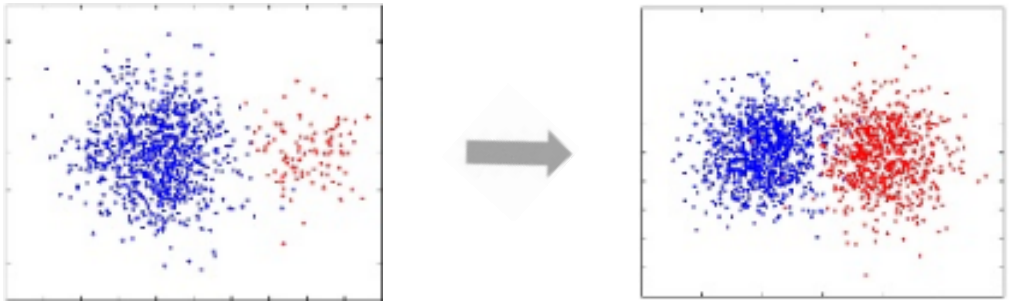
欠采样(Undersampling)

- 对大类的数据样本进行采样减少该类样本的个数
  - 随机欠采样：随机删除大类的训练样本直至数据集平衡
  - 简易集成算法：多次有放回的欠采样，产生多个不同的训练集，进而训练多个不同的分类器，组合多个分类器的结果

- 缺陷：大类损失很多重要信息，模型只学到总体模式的一部分

过采样(Oversampling)

- 对小类的数据样本进行采样来增加小类的数据样本个数
  - 随机过采样：将小类数据随机重复



  - 优势：没有任何信息损失
  - 缺陷：计算时间和存储开销相应增大，并且很有可能导致过拟合，泛化能力较差

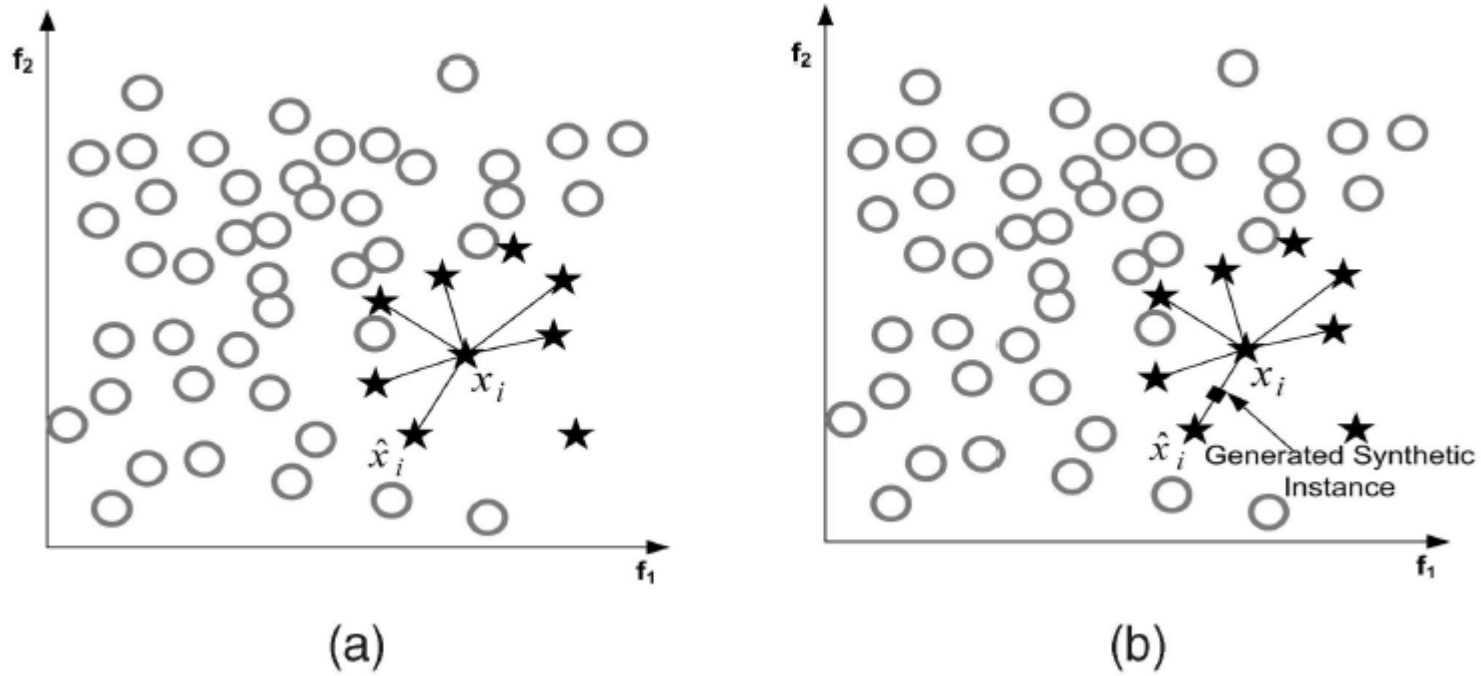
合成人工数据

- SMOTE算法：对少数类样本进行分析，根据少数类样本人工合成新样本添加到数据集中，流程如下：
  - 对于少数类中的每一个样本x，以欧式距离为标准计算它到少数类样本集S_{min}中所有样本的距离，得到其k近邻
  - 根据样本不平衡比例设置一个采样比例，对于每一个少数类样本x，从其k近邻中随机选择若干个样本，假设选择的近邻为x_n
  - 对于每一个随机选出的近邻x_n，分别与原样本按照如下公式构建新的样本：

$$x_{new} = x + rand(0,1) * |x - x_n|$$

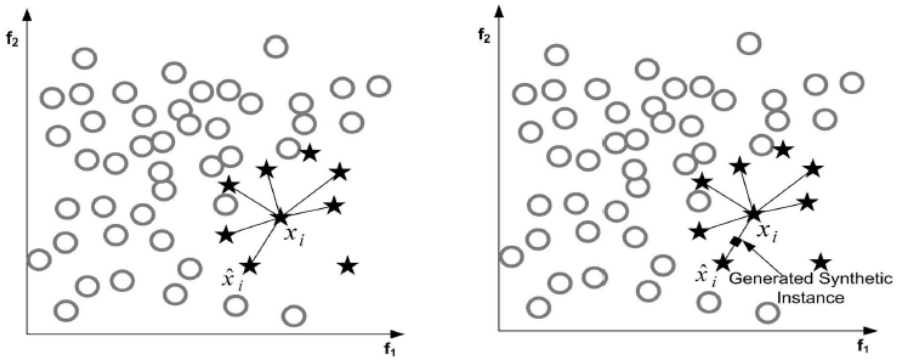
合成人工数据

- SMOTE算法



合成人工数据

- SMOTE算法的优缺点
  - 优点：不是采取简单的复制样本策略增加少数类样本，解决了模型容易过拟合的问题
  - 缺点：容易产生分布边缘化问题，如果一个样本在少数类样本集的分布边缘，由此产生的“人造”样本也会处在这个边缘，模糊了正负类样本的边界，加大了分类难度



代价敏感学习法

- 代价矩阵

Actual	Predicted		
		Positive	Negative
	Positive	0	C(FN)
	Negative	C(FP)	0

- 代价敏感学习
  - 让损失函数考虑到误分类的情况不同，产生的代价不同，最后达到损失代价最小化的目标即可，比如直接修正损失函数，使其对于不同的误分类情况拥有不同的代价

方法选择

- 计算资源足够，且小类样本足够多——过采样
- 计算资源不够，或小类样本不够多——欠采样
- 正负样本都非常少——数据合成
- 在正负样本都足够多，且比例不是特别悬殊——代价敏感学习

未来研究发展

- 如何提高算法的效率并降低时间开销
- 如何自适应地确定最好的抽样比例
- 对两类数据的数目相当，但类分布差异较大的情况研究
- 如何将属性选择方法融入到不平衡分类算法中

参考文献

- https://www.jianshu.com/p/3e8b9f2764c8
- https://blog.csdn.net/zhili8866/article/details/69704727
- https://blog.csdn.net/heyongluoyao8/article/details/49408131
- 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012:25-35

知人者智，自知者明。
胜人者有力，自胜者
强。知足者富。强行
者有志。不失其所者
久。死而不亡者，寿。

谢谢！