CSDN
不 止 于 代 码

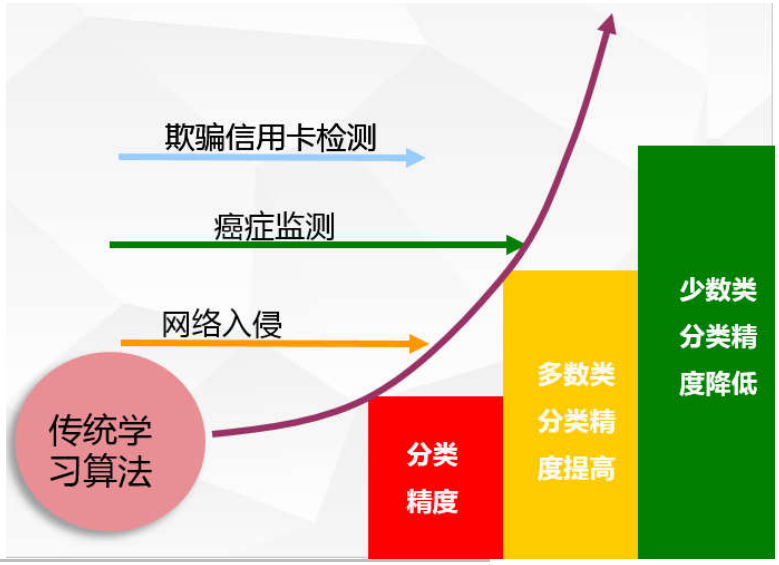
2.4 不平衡数据分类学习

CSDN学院

2017年10月

不平衡数据的出现场景

- 搜索引擎的点击预测
  - 点击的网页往往占据很小的比例
- 电子商务领域的商品推荐
  - 推荐的商品被购买的比例很低
- 信用卡欺诈检测
- 网络攻击识别
- …



解决方案

- 从数据的角度：抽样，从而使得数据相对均衡
- 从算法的角度：考虑不同误分类情况代价的差异性对算法进行优化

采样

- 随机欠采样：从多数类中随机选择少量样本再合并原有少数类样本作为新的训练数据集
  - 有放回采样
  - 无放回采样
  - 会造成一些信息缺失，选取的样本可能有偏差
- 随机过采样：随机复制少数类来样本
  - 扩大了数据集，造成模型训练复杂度加大，另一方面也容易造成模型的过拟合问题

集成学习算法

- EasyEnsemble算法：
  - 对于多数类样本，通过n次有放回抽样生成n份子集
  - 少数类样本分别和这n份样本合并训练一个模型：n个模型
  - 最终模型：n个模型预测结果的平均值
- BalanceCascade（级联）算法：
  - 从多数类中有效地选择一些样本与少数类样本合并为新的数据集进行训练
  - 训练好的模型每个多数类样本进行预测。若预测正确，则不考虑将其作为下一轮的训练样本
  - 依次迭代直到满足某一停止条件，最终的模型是多次迭代模型的组合

SMOTE: Synthetic Minority Over-sampling Technique

- 基本思想：基于“插值”来为少数类合成新的样本
- 对少数类的一个样本 i ，其特征向量为 $\mathbf{x}_i$,：
  - 1. 从少数类的全部 N 个样本中找到样本$\mathbf{x}_i$的K个近邻（如欧氏距离），记为$\mathbf{x}_{i(near)}$,$near \in \{1,...,K\}$；
  - 2. 从这 K个近邻中随机选择一个样本$\mathbf{x}_{i(nn)}$，再生成一个 0到 1之间的随机数 ζ ，从而合成一个新样本 $\mathbf{x}_{i1}$ ：
    - $\mathbf{x}_{i1} = （1-\zeta）\mathbf{x}_i + \zeta \cdot \mathbf{x}_{i(nn)}$
    - 新样本 $\mathbf{x}_{i1}$相当于是表示样本$\mathbf{x}_i$和表示样本$\mathbf{x}_{i(nn)}$ 的点之间所连线段上的一个点： 插值

SMOTE

- SMOTE算法摒弃了随机过采样复制样本的做法，可以防止随机过采样易过拟合的问题。实践证明此方法可以提高分类器的性能。
- SMOTE 对高维数据不是很有效。
- 当生成合成性实例时，SMOTE 并不会把来自其他类的相邻实例考虑进来，这导致了类重叠的增加，并会引入额外的噪音。为了解决SMOTE算法的这一缺点提出一些改进算法，如Borderline-SMOTE算法

代价敏感学习

- 在算法层面上解决不平衡数据学习的方法主要是基于代价敏感学习算法(Cost-Sensitive Learning)
- 代价敏感学习方法的核心要素是代价矩阵：不同类型的误分类情况导致的代价不一样

Predict \ True	0	1
0	C_{00}	C_{01}
1	C_{10}	C_{11}

代价敏感学习

- 基于代价矩阵分析，代价敏感学习方法主要有以下三种实现方式：
- 从贝叶斯风险理论出发，把代价敏感学习看成是分类结果的一种后处理，按照传统方法学习到一个模型，以实现损失最小为目标对结果进行调整
  - 不依赖所用具体的分类器
  - 但是缺点要求分类器输出值为概率
- 从学习模型出发，对具体学习方法的改造，使之能适应不平衡数据下的学习
  - 代价敏感的支持向量机，决策树，神经网络
  - 从预处理的角度出发，将代价用于权重的调整，使得分类器满足代价敏感的特性

Scikit learn中的不均衡样本分类处理

- 类别权重class_weight
- 样本权重sample_weight

类别权重class_weight

- class_weight参数用于标示分类模型中各类别样本的权重
- 1. 不考虑权重，即所有类别的权重相同
- 2. balanced：自动计算类别权重
  - 某类别的样本量越多，其权重越低；样本量越少，则权重越高
  - 类权重计算方法为：n_samples / (n_classes * np.bincount(*y*))
    - n_samples为样本数，n_classes为类别数量，np.bincount(*y*)输出每个类的样本数
- 3. 手动指定各个类别的权重
  - 如对于0,1二类分类问题，可以定义class_weight={0:0.9, 1:0.1}，即类别0的权重为90%，而类别1的权重为10%

样本权重sample_weight

- 模型训练：fit(*X*, *y*, *sample_weight=None*)
  - 其中参数sample_weight为样本权重参数
- 当样本高度失衡时，样本不是总体样本的无偏估计，可能导致模型预测能力下降
- 解决方案：调节样本权重
  - 一种是在class_weight使用balanced
  - 另一种是在调用fit函数时，通过sample_weight来调节每个样本权重
  - 如果两种方法都用了，那么样本的真正权重是class_weight*sample_weight

小结：Logistic回归

- 不均衡样本分类
  - 样本采样：过采样、欠采样
  - 分类器：代价敏感函数
    - 样本权重、类别权重

CSDN
不 止 于 代 码

THANK YOU

AI 100