# 解决类别不平衡问题的方法综述

## 一、数据不平衡

### 1.1 什么是数据不平衡

在学术研究与教学中，很多算法都有一个基本假设，那就是数据分布是均匀的。当我们把这些算法直接应用于实际数据时，大多数情况下都无法取得
果。因为实际数据往往分布得很不均匀，都会存在“长尾现象”，也就是所谓的“二八原理”。

以二分类问题为例，假设正类的样本数量远大于负类的样本数量，通常情况下把样本类别比例超过4:1（也有说3:1）的数据就可以称为不平衡数据。

不平衡程度相同（即正负样本比例类似）的两个问题，解决的难易程度也可能不同，因为问题难易程度还取决于数据量。可以把问题根据难度从小至
**大数据+分布均衡<大数据+分布不均衡<小数据+数据均衡<小数据+数据不均衡。**

### 1.2 数据不平衡会产生什么问题

样本不平衡会使得我们的分类模型存在很严重的偏向性，但是从一些常用的指标上又无法看出来。举一个极端一点的例子，如果正负样本比例为100
把全部样本都判定为正样本就有99%+的分类准确率了。从测试结果上来看，就表现为有太多的False Positive。

## 二、解决方法

在机器学习中，处理样本不平衡问题，主要有3种策略：从数据角度、从算法层面和模型评价层面。
从数据角度出发，通常的方法包括了：

### 2.1 采样

采样方法是通过对训练集进行处理使其从不平衡的数据集变成平衡的数据集，在大部分情况下会对最终的结果带来提升。采样分为过采样和欠采样。
大的优点是简单。

#### 过采样

过采样是把小众类复制多份。

过采样后的数据集中会反复出现一些样本，训练出来的模型会有一定的过拟合。

过采样会把小众样本复制多份，一个点会在高维空间中反复出现，这会导致一个问题，那就是运气好就能分对很多点，否则分错很多点。为了解决这
**以在每次生成新数据点时加入轻微的随机扰动**，经验表明这种做法非常有效。

#### 欠采样

欠采样是从大众类中剔除一些样本，或者说只从大众类中选取部分样本。

欠采样的缺点是最终的训练集丢失了数据，模型只学到了总体模式的一部分。丢弃大量数据，和过采样一样会存在过拟合的问题。

因为欠采样会丢失信息，如何减少信息的损失呢?

第一种方法叫做**EasyEnsemble**，利用模型融合的方法（Ensemble）：多次欠采样（放回采样，这样产生的训练集才相互独立）产生多个不同的训
训练多个不同的分类器，通过组合多个分类器的结果得到最终的结果。

第二种方法叫做**BalanceCascade**，利用增量训练的思想（Boosting）：先通过一次欠采样产生训练集，训练一个分类器，对于那些分类正确的大众
回，然后对这个更小的大众样本欠采样产生训练集，训练第二个分类器，以此类推，最终组合所有分类器的结果得到最终结果。

第三种方法是利用KNN试图挑选那些最具代表性的大众样本，叫做**NearMiss**，这类方法本质上是一种原型选择(prototype selection)方法，即从多数
取最具代表性的样本用于训练，计算量很大。

还可以采用**聚类**的方法，假设少数类样本数量为N，那就将多数类样本分为N个簇，取每个簇的中心点作为多数类的新样本，再加上少数类的所有样
练。这样就可以保证了多数类样本在特征空间的分布特性。

还可以通过某种规则来清洗重叠的数据，从而达到欠采样的目的，而这些规则往往也是启发性的。如**Tomek Link**和**Edited Nearest Neighbours(EI
**数据清洗方法**最大的缺点是无法控制欠采样的数量。

### 2.2 数据合成

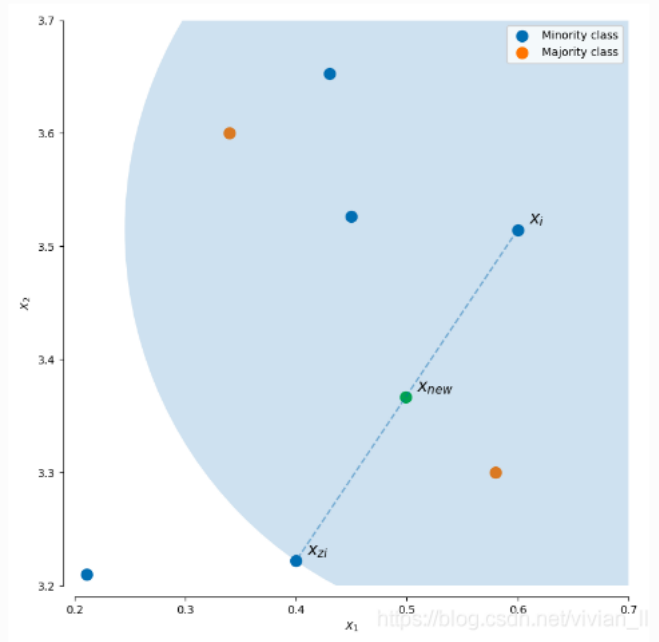
数据合成方法是利用已有样本生成更多样本，其实**也可以归类于过采样**方法，这类方法在小数据场景下有很多成功案例，比如医学图像分析等。

现在的主流过采样方法就是通过某种方式人工合成一些少数类样本，从而达到类别平衡的目的，而这其中的鼻祖就是SMOTE。

**SMOTE** (synthetic minority oversampling technique) 的思想概括起来就是在少数类样本之间进行插值来产生额外的样本。具体地，对于一个少数类样
K近邻法(k值需要提前指定)，求出离$x_i$距离最近的k个少数类样本，其中距离定义为样本之间n维特征空间的欧氏距离。然后从k个近邻点中随机选取
下列公式生成新样本：

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta$$

其中$\hat{x}_i$为选出的k近邻点，$\delta \in [0, 1]$是一个随机数。下图就是一个SMOTE生成样本的例子，使用的是3-近邻，可以看出SMOTE生成的样本一般就在
连的直线上：




SMOTE为每个小众样本合成相同数量的新样本，这带来一些潜在的问题：一方面是增加了类之间重叠的可能性，另一方面是生成一些没有提供有益
本。为了解决这个问题，出现两种方法：Borderline-SMOTE与ADASYN。

**Borderline-SMOTE**的解决思路是寻找那些应该为之合成新样本的小众样本。即为每个小众样本计算K近邻，只为那些K近邻中有一半以上大众样本的
生成新样本。直观地讲，只为那些周围大部分是大众样本的小众样本生成新样本，因为这些样本往往是边界样本。确定了为哪些小众样本生成新样本
SMOTE生成新样本。

**ADASYN**名为自适应合成抽样(adaptive synthetic sampling)，其最大的特点是根据数据分布情况自动决定不同少数类样本需要产生多少不同数量的合
而不是像SMOTE那样对每个少数类样本合成同数量的样本。
其流程为：首先计算需要合成的样本总量G，然后对于每个少类别样本xi，找出其K近邻个点，并计算分布比例$\Gamma$，最后对每个少类别样本xi计算需要
数量gi，再用SMOTE算法合成新样本。

### 2.2.1 文本数据的合成

后文中的focal loss 方法，针对像深度神经网络这些复杂的模型，具有很好的使用价值，但是针对传统分类器，小样本集情况下，实施有一定的难度
采样和欠采样对与文本分类问题效果几乎为0。

对于文本数据，可以采用文本生成的方式，解决文本样本不均衡的问题。首先分析样本数少的类别，通过文本句法依赖分析，文本词性标记分析词的
然后采用**同义词替换**的方式生成新的文本。实验结果证明，方法简单有效。

还可以进行一些**句子顺序打乱**以及**句内词序打乱**的操作进行小类的数据增强。 将数据增强结合过采样是比较直观有效的做法。

### 2.2.2 图像数据的合成

属于**图像数据增强**范畴，包括图像翻转、平移、旋转、缩放，分离单个r、g、b三个颜色通道以及添加噪声等等。

## 2.3 加权

我们还可以通过加权的方式来解决数据不平衡的问题，即**对不同类别分错的代价不同**，在训练分类器时，为少数类样本赋予更大的权值，为多数类样
的权值。

这种方法的难点在于设置合理的权重，实际应用中一般让各个分类间的加权损失值近似相等。当然这并不是通用法则，还是需要具体问题具体分析。

## 2.4 改变模型评价

### 2.4.1 选择合适的评估方式

准确度这个评价指标在类别不均衡的分类任务中并不能work，甚至进行误导（分类器不work，但是从这个指标来看，该分类器有着很好的评价指标

对于极端的类别不平衡的评估问题，我们一般用的指标有（前面是全局评估，最后一个是点评估）：

- 混淆矩阵
- Precision和Recall
- F1得分
- Kappa（Cohen kappa）
- ROC曲线和AUC
- mean Average Precesion（**mAP**），指的是在不同召回下的最大精确度的平均值
- **Precision@Rank k**。假设共有n个点，假设其中k个点是少数样本时的Precision。这个评估方法在推荐系统中也常常会用。

选择哪个评估标准需要取决于具体问题。

### 2.4.2 调整阈值

大部分模型的默认阈值为输出值的中位数，如逻辑回归的输出范围为[0,1]，当某个样本的输出大于0.5就会被划分为正例，反之为反例。当类别不平
默认的分类阈值可能会导致输出全部为反例，产生虚高的准确度，导致分类失败。因此，可以选择**调整阈值**，使得模型对于较少的类别更为敏感。

### 2.4.3 改变损失函数（OHEM和Focal loss）

以下两个方法最开始适用于图像上，但是NLP领域也可以借鉴。

#### OHEM

OHEM（online hard example miniing）算法的核心思想是根据输入样本的损失进行筛选，筛选出hard example，表示对分类和检测影响较大的样本
选得到的这些样本应用在随机梯度下降中训练。在实际操作中是将原来的一个ROI Network扩充为两个ROI Network，这两个ROI Network共享参数。
一个ROI Network只有前向操作，主要用于计算损失；后面一个ROI Network包括前向和后向操作，以hard example作为输入，计算损失并回传梯度

算法优点：1、对于数据的类别不平衡问题不需要采用设置正负样本比例的方式来解决，这种在线选择方式针对性更强。2、随着数据集的增大，算
加明显（作者是通过在COCO数据集上做实验和VOC数据集做对比，因为前者的数据集更大，而且提升更明显，所以有这个结论）。

#### Focal loss

Focal loss主要是为了解决one-stage目标检测中正负样本比例严重失衡的问题。主旨是：ssd按照ohem选出了loss较大的，但忽略了那些loss较小的
本，虽然这些easy负样本loss很小，但数量多，加起来的loss较大，对最终loss有一定贡献。作者想把这些loss较小的也融入到loss计算中。但如果直
有的loss，loss会被那些easy的负样本主导，因为数量太多，加起来的loss就大了。也就是说，作者是想融入一些easy example，希望他们能有助于
不希望他们主导loss。这个时候就用了公式进行衰减那些easy example，让他们对loss做贡献，但又不至于主导loss，并且通过balanced crossentro
别。

OHEM是只取3:1的负样本去计算loss，之外的负样本权重置零，而focal loss取了所有负样本，根据难度给了不同的权重。

focal loss相比OHEM的提升点在于，3:1的比例比较粗暴，那些有些难度的负样本可能游离于3:1之外。之前实验中曾经调整过OHEM这个比例，发现
的，现在可以试试focal loss了。

可以参考本人的另一篇博文：论文笔记：Focal Loss for Dense Object Detection

## 2.5 一分类/无监督/半监督

对于正负样本极不平衡的场景，我们可以换一个完全不同的角度来看待问题：把它看做**一分类**（One Class Learning）或**异常检测**（Novelty Detecti
**变化趋势检测**问题。

一分类方法的重点不在于捕捉类间的差别，而是为其中一类进行建模，经典的工作包括One-class SVM等。

**异常检测**指的是从数据中找到那些异常值，比如你案例中的“广告”。无监督的异常检测一般依赖于对于数据的假设，比如广告和正常的文章内容很不
一种假设是广告和正常文章间的欧式距离很大。无监督异常检测最大优势就是在于**不需要数据标签**，如果在对数据假设正确时效果甚至可以比监督学
其是当获取标签成本很高时。

**变化趋势检测**类似于异常点检测，不同在于其通过检测不寻常的变化趋势来识别。如通过观察用户模式或银行交易来检测用户行为的不寻常改变。

此外还可以尝试**半监督异常集成学习**，简单而言，可以现在原始数据集上使用多个无监督异常方法来抽取数据的表示，并和原始的数据结合作为新的
在新的特征空间上使用集成树模型，比如xgboost，来进行监督学习。无监督异常检测的目的是提高原始数据的表达，监督集成树的目的是降低数据

最终预测结果的影响。这个方法还可以和主动学习结合起来，进一步提升系统的性能。当然，这个方法最大的问题是运算开销比较大，需要进行深度
做法可以参考：

> Zhao, Y.; Hryniewicki, M.K. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Rio, Brazil, 8–13 July 2018.

### 2.5.1 不平衡文本数据的半监督

**高维数据上的半监督异常检测**：考虑到文本文件在转化后往往维度很高，可以尝试一下最近的一篇KDD文章，主要是找到高维数据在低维空间上的表
助基于距离的异常检测方法。文章如下：

> Pang, G., Cao, L., Chen, L. and Liu, H., 2018. Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier De arXiv preprint arXiv:1806.04808.

## 三、如何选择

如何根据实际问题选择合适的方法呢？接下来谈谈一些我的经验。

1、在正负样本都非常之少的情况下，应该采用数据合成的方式；

2、在负样本足够多，正样本非常之少且比例及其悬殊的情况下，应该考虑一分类方法；

3、在正负样本都足够多且比例不是特别悬殊的情况下，应该考虑采样或者加权的方法。

4、采样和加权在数学上是等价的，但实际应用中效果却有差别。尤其是采样了诸如Random Forest等分类方法，训练过程会对训练集进行随机采样
况下，如果计算资源允许过采样往往要比加权好一些。

5、另外，虽然过采样和欠采样都可以使数据集变得平衡，并且在数据足够多的情况下等价，但两者也是有区别的。实际应用中，我的经验是如果计
且小众类样本足够多的情况下使用过采样，否则使用欠采样，因为过采样会增加训练集的大小进而增加训练时间，同时小的训练集非常容易产生过拟

6、对于欠采样，如果计算资源相对较多且有良好的并行环境，应该选择Ensemble方法。

## 四、总结

1、怎样解决样本不平衡问题：

主要三个方面，**数据，模型和评估方法。**

从数据的角度出发，通常的方法包括：

- 扩充数据集
- 过采样
- 欠采样
- 数据合成
- 基于异常检测的方式

从算法的角度出发，通常的方法包括：

- 尝试不同的分类算法

  决策树往往在类别不均衡数据上表现不错。它使用基于类变量的划分规则去创建分类树，因此可以强制地将不同类别的样本分开。

- 对小类错分进行加权惩罚

  如penalized-SVM和penalized-LDA算法。

- 从重构分类器的角度出发

  仔细对你的问题进行分析与挖掘，是否可以将你的问题划分成多个更小的问题，而这些小问题更容易解决。你可以从这篇文章In classification, h handle an unbalanced training set?中得到灵感。例如：

  - 将你的大类压缩成小类
  - 使用One Class分类器（将小类作为异常点）

- 使用集成方式，训练多个分类器，然后联合这些分类器进行分类
- 将二分类问题改成多分类问题

从评估的角度出发，通常的方法包括：

- 选择合适的评估指标
- 选择合适的损失函数
- 选择合适的阈值
- 设置不同类别的权重

2、经验：

1. 采样方法一般比直接调整阈值的效果要好。
2. 使用采样方法（过采样和欠采样）一般可以提升模型的泛化能力，但有一定的过拟合的风险，应搭配使用正则化模型
3. 过采样的结果较为稳定，作为一种升级版的过采样，SMOTE也是不错的处理方式，大部分时候和过采样的效果相似
4. 过采样大部分时候比欠采样的效果好，但很难一概而论哪种方法最好，还是需要根据数据的特性（如分布）具体讨论
5. 实验结果在（L2正则的逻辑回归、随机森林、xgboost）一致，因此和采样法搭配使用的模型最好可以很好的处理过拟合



参考网址：
【小夕精选】如何优雅而时髦的解决不均衡分类问题 - 夕小瑶的卖萌屋
怎样解决样本不平衡问题？（较全，有经验总结）
聊一聊深度学习中的样本不平衡问题
机器学习之类别不平衡问题 (3) —— 采样方法（详细讲了多种过采样和欠采样的方法，并对比了效果）
欠采样（undersampling）和过采样（oversampling）会对模型带来怎样的影响？ - 微调的回答 - 知乎（含部分方法在不同数据集上的实验结果）

**相关推荐**