**机器学习中正负样本分布不均衡的问题及解决**

1. **正负样本分布不均问题：**

在很多算法都有一个基本假设，即数据分布是均匀的。但当这些算法直接应用于实际数据时，大多数情况下无法取得理想结果。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”。严格地讲，任何数据集上都有数据不平衡现象，这往往由问题本身决定的，但只需要关注那些分布差别比较悬殊。

经验表明，训练数据中每个类别有5000个以上样本，数据量是足够的，正负样本差一个数量级以内是可以接受的，不太需要考虑数据不平衡问题。

但在医学场景中，本身阳性样本数就可能很少，数据不但不均衡，而且样本数还比较少，即小数据+数据不均衡，使用机器学习的方法是比较棘手的。

1. **解决方法示例：**

1. **抽样方法**

**上采样方法**：增加分类中少数类样本的数量来实现样本均衡。

最直接的方法是简单复制少数类样本形成多条记录，这种方法的缺点是如果样本特征少而可能导致过拟合的问题；

经过改进的过抽样方法通过在少数类中加入随机噪声、干扰数据或通过一定规则产生新的合成样本，例如SMOTE算法。

**下采样方法**：通过减少分类中多数类样本的样本数量来实现样本均衡。

最直接的方法是随机地去掉一些多数类样本来减小多数类的规模，缺点是会丢失多数类样本中的一些重要信息。

**2）正负样本惩罚权重**

通过正负样本的惩罚权重解决样本不均衡的。

在算法实现过程中，对于分类中不同样本数量的类别分别赋予不同的权重（一般思路分类中的小样本量类别权重高，大样本量类别权重低），然后进行计算和建模。

使用这种方法时需要在算法模型的参数中进行相应设置，很多模型和算法中都有基于类别参数的调整设置，以scikit-learn中的SVM为例，通过在class\_weight: {dict, 'balanced'}中针对不同类别针对不同的权重，来手动指定不同类别的权重。如果使用其默认的方法balanced，那么SVM会将权重设置为与不同类别样本数量呈反比的权重来做自动均衡处理，计算公式为：n\_samples / (n\_classes \* np.bincount(y))。

如果算法本身支持，这种思路是更加简单且高效的方法。

1. **组合或集合方法**

组合/集成方法指的是在每次生成训练集时使用所有分类中的小样本量，同时从分类中的大样本量中随机抽取数据来与小样本量合并构成训练集，这样反复多次会得到很多训练集和训练模型。最后在应用时，使用组合方法（例如投票、加权投票等）产生分类预测结果。

例如，在数据集中的正、负例的样本分别为100和10000条，比例为1:100。此时可以将负例样本（类别中的大量样本集）随机分为100份（当然也可以分更多），每份100条数据；然后每次形成训练集时使用所有的正样本（100条）和随机抽取的负样本（100条）形成新的数据集。如此反复可以得到100个训练集和对应的训练模型。这种解决问题的思路类似于随机森林。在随机森林中，虽然每个小决策树的分类能力很弱，但是通过大量的“小树”组合形成的“森林”具有良好的模型预测能力。

如果计算资源充足，并且对于模型的时效性要求不高的话，这种方法比较合适。

1. **特征选择方法**

考虑使用或辅助于基于列的特征选择方法。一般情况下，样本不均衡也会导致特征分布不均衡，但如果小类别样本量具有一定的规模，那么意味着其特征值的分布较为均匀，可通过选择具有显著型的特征配合参与解决样本不均衡问题，也能在一定程度上提高模型效果。