带权重的CART模型

改善样本非平衡性的方法之一是简单过采样法，即对少类样本进行重复抽样。该方法在样本量较小的情况下可以使用。但是当样本量较大时，会增加样本构建的时间开销，并且易形成内存溢出等错误。为了避免上述缺陷，我们可以将抽样次数转换成权重带入部分模型中去，例如CART模型。该模型在构建的时候，Gini指数是非常重要的统计量，我们可以将权重带入Gini指数的计算。该计算方法与简单重复抽样法在期望上是一致的。

考虑以下数据集：

|  |  |  |
| --- | --- | --- |
| **var** | **y** | **Sampling size(w)** |
| A | 1 | k |
| A | 0 | 1 |
| A | 1 | k |
| B | 1 | k |
| B | 0 | 1 |

k是坏样本的重复抽样率，k>1。每个样本的抽样倍率为。

考虑var=A的子集，总共有个。其y=1的个数为：

占比为

则Gini指数为

其中

类似的，

其中

同时，var=A的子集的占比为

var=B的子集的占比为

所以最终的Gini为