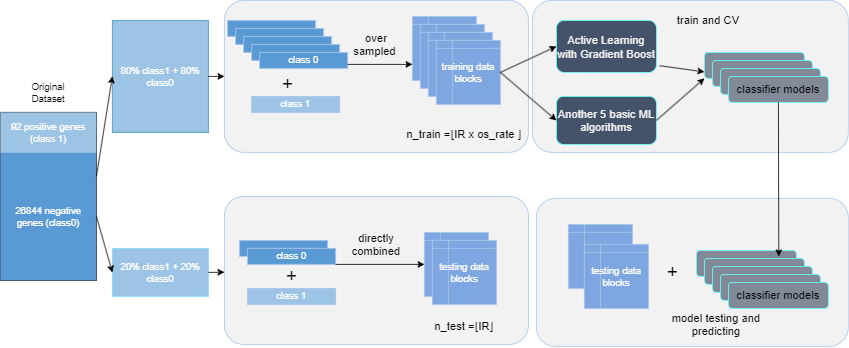
在预测与RNA甲基化通路相关基因这个工作时，我们认为原始数据集中样本类别不平衡问题非常突出，其不平衡因子（IR）达到了200以上，这一数据比大部分公开数据集的不平衡因子都要高。在查阅大量文献后，我们认为这一问题是可以解决的。我们尝试了Active learning，数据过采样方法SMOTE，结合基础的机器学习分类器(SVM,GB,GNB,RF,LR)进行实验。

数据收集和处理:

我们的数据来源于Harmonizome，共有26936个基因数据，每个基因数据有1517个特征。每个特征的取值是离散的0,1,-1。正样本（已知与RNA甲基化通路相关的基因）共有92个。其他26844个基因假定均为负样本。如此做的原因是因为人体内的基因大部分都有其他功能，也就是说与RNA甲基化通路相关的基因较少，因此可以假定训练数据中的假阴性数量应该非常少。

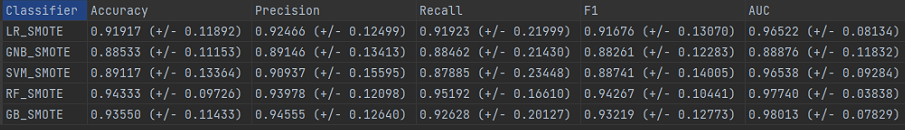
我们分别取正样本和负样本的80%用于训练和交叉验证，其余20%用于测试。计算出不平衡因子IR,它等于训练样本中多数类与少数类数量的比值。由于我们使用的的数据的不平衡因子超过绝大多数公开数据集的不平衡因子，达到了惊人的290，因此有必要对训练数据进行过采样，我们使用基础SMOTE方法进行过采样。假定参数os\_rate，这是一个介于0~1之中的小数，它用于控制训练数据的过采样程度，在我们的实验中这一参数设定为0.6。接着我们将训练样本中多数类（负样本）均匀的分为n\_train = IR\*os\_rate（向下取整）份，将每份负样本与相同的正样本结合，得到n\_train个数据块，接着再用SMOTE对每个数据块进行类别平衡操作，其中生成的多数类特征取值范围依然是离散的0,-1,1。最终得到训练集。

对于测试数据集，同样计算出他的IR，这一值应该和训练数据集相等，直接将测试数据集均匀分成IR份(向下取整)，每份与相同的测试正样本结合。得到n\_test = IR个数据块并且以此作为测试集。如图1所示

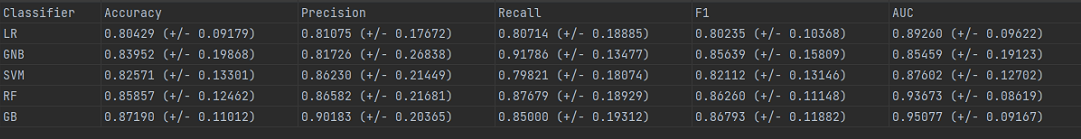


【FIG1】

模型训练和交叉验证:

我们使用Active Learning(GradientBoost based)，基础ML分类器SVM,GB,GNB,RF,LR在n\_train = 173份数据上进行训练。其中Active Learning方法中使用默认GradientBoost参数，而基础ML分类器中的SVM,GB,RF模型使用GridSearch在给定超参数组合中选择最佳(Accuracy)参数。一共训练出1038个模型。在交叉验证中，Active Learning(GB based)在未使用SMOTE进行数据平衡操作下，达到了平均89.74%的accuracy，进行类别平衡后这一值达到了91%。基本ML分类其中GradientBoost可以达到87%的accuracy(FIG3)，进行类别平衡后这一数值最高可以到93%(FIG2)。要注意的是，AL和ML的分类器在过采样后得到的最终数据集并不一定相同，这是因为AL和ML的训练脚本并不是一个。

【FIG2】



【FIG3】