关于样本选择的方法

1. 基于最近邻的方法
2. 基于密度的方法
3. 基于随机采样的方法

随机采样是指从原始训练集中选出某个子集，不断交换子集和剩余集合中的样本，即在这个条件下得到局部最优解。

基于覆盖程度的相关性计算

1. 数据集的特征覆盖程度计算

对于数据集的每个特征：

计算其与类标的相关程度weight

计算正类样本和负类样本的交叉区间，并计算其在总区间长度中的比例

保存该加权比例值

计算该数据集的加权平均值

该方法的优势：

直接根据数据集计算该指标

缺点：

单个样本无法计算该指标

1. 样本的特征覆盖程度计算

针对以上方案进行改进

对于数据集中的每个特征：

计算该特征与类标的相关程度weight

保存该特征的交叉区间

对于数据集中的每个样本：

对于每维特征：

如果该特征落入这个维度的交叉区间中，记为1，否则记为0

计算其加权平均值

返回平均值

结果：从结果中可以看到，由于两种overlapping的计算方式并没有较大差别，因而在相关性上并没有明显优劣，因为第二种方式中，可以计算每个样本的overlapping指标，因此采用第二种计算结果。



有个问题，更新幅度不大。。

根据该指标进行数据集的筛选

1. 对每个样本，计算其指标数值
2. 根据该指标，计算分布，将其归一化即可，并根据概率值采样去除某个容易分类的样本，计算更新后的指标
3. 直到指标降低，产生N个训练子集
4. 对于这N个训练子集，各自训练K个分类器，并将其总结为N\*K的形式，即两层结构的。