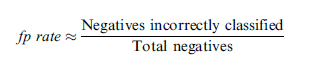
**ROC曲线**

一、基本定义

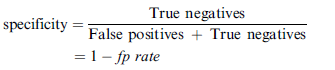
1. tp rate也叫hit rate或recall（召回率），代表所有正例被识别出来的概率



1. fp rate也叫false alarm rate（误报率），代表所有反例被错误识别的概率



1. sensitivity，灵敏度，等于召回率
2. specificity，特异度，代表反例被正确识别的概率

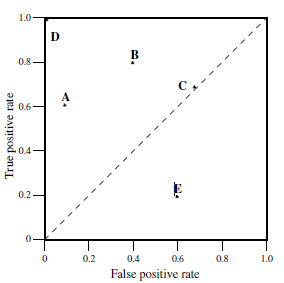


1. precision，positive predictive value，准确率

二、ROC曲线

X轴为fp，也即误报率，y轴为tp也即召回率，目的是综合衡量收益（召回率）与代价（误报率）之间的关系。

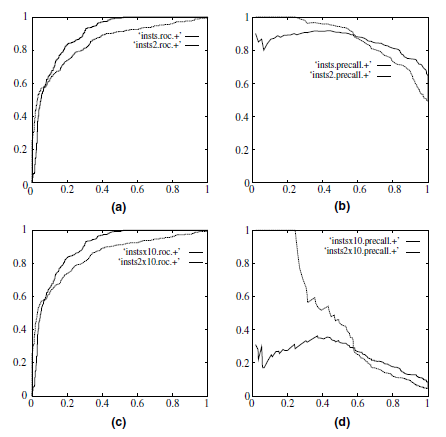
点在靠左侧说明较为保守，说明只有在足够准确率的情况下才判定为正例；点靠右侧较为自由，判断出较多的正例，但误报率也相对较高。



三、优点

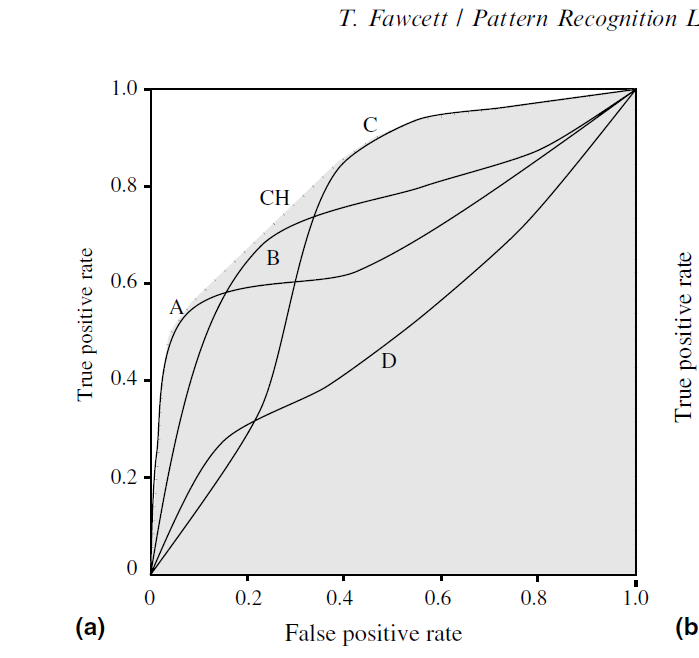
ROC曲线的特征受样本类别不均衡的影响小，如果用其他方式衡量，如准确率-召回率曲线，这样会很容易受到类别影响。

如下图ab两图是类别均匀的情况下，ROC与准确-召回曲线，cd为类别1:10的情况下曲线形式，可以很显然看出ROC曲线的稳定性。



ROC凸包

如下图所示，AC曲线明显优于BD曲线，由此可以筛去BD分类器。



四、相关算法

（1）ROC曲线画法：

先将所有样本按照F值从小到大排列

如果该样本为TP（正例且被识别出来），那么将（FP/N,TP/P）画出来，且TP自增

如果为FP，即错误被识别为正确，同样将点画出来，FP自增

期间若遇到（FP/N,TP/P）与之前值相同，则略过

（2）AUC计算：

与微积分类似，底乘以两高均值

（3）多个ROC曲线求平均：

方法一：对于每一个FP值，直接纵向求均值

方法二：基于阈值求平均

（4）多类别ROC计算：

每个类别分别作为正例，剩下作为反例，绘制ROC

（5）多类别AUC计算

方法一：按照（4）给出的曲线，对所有AUC值加权求和



方法二：原理未详细说明，但排除了概率分布的影响

