1 Confusion Matrix

Confusion Matrix 矩阵如下表所示：

预测值-实际值	True	False
True	True Positive(真阳性)	False Positive(假阳性)
False	False Negative(假阴性)	True Negative(真阴性)

Table 1: Confusion Matrix

2 各种率的定义

正确率（Precision）：

$$Precision = \frac{TP}{TP + FP}$$

真阳性率 (True Positive Rate，TPR)，灵敏度 (Sensitivity)，召回率 (Recall):

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

真阴性率 (True Negative Rate，TNR)，特异度 (Specificity)：

$$Specificity = Recall = \frac{TN}{FP + TN}$$

假阴性率 (False Negatice Rate，FNR)，漏诊率 (= 1 - 灵敏度)：

$$FNR = \frac{FN}{TP + FN}$$

假阳性率 (False Positice Rate，FPR)，误诊率 (= 1 - 特异度)：

$$FPR = \frac{FP}{FP + TN}$$

3 ROC 和 AUC [1]

3.1 ROC

对于分类器，或者说分类算法，评价指标主要有 precision，recall，F-score 等，以及这里要讨论的 ROC 和 AUC。

ROC 曲线：接收者操作特征曲线（receiver operating characteristic curve），是反映敏感性和特异性连续变量的综合指标，ROC 曲线上每个点反映着对同一信号刺激的感受性。下图是一个 ROC 曲线的

示例：

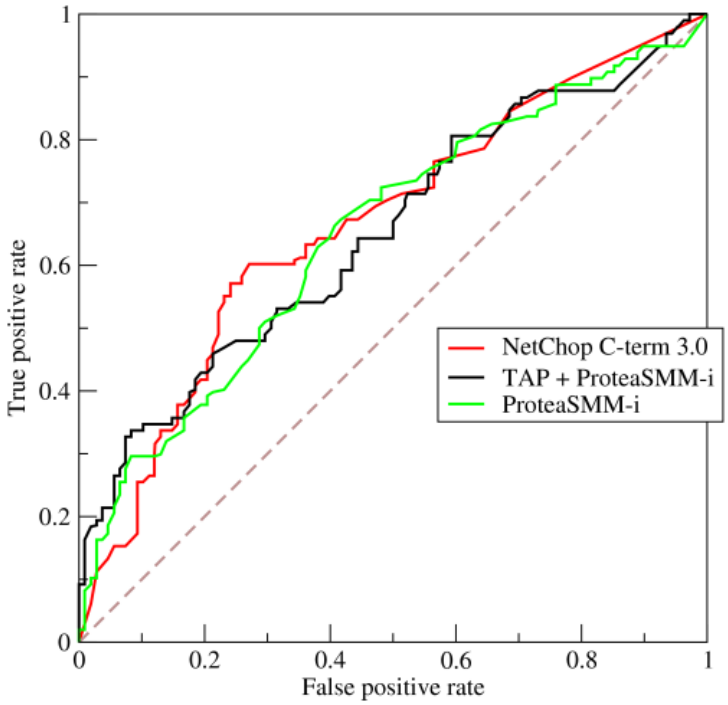

Figure 1: ROC 曲线示意

ROC 曲线的横纵坐标分别为：

横坐标：1-Specificity，伪正类率 (False positive rate，FPR)，预测为正但实际为负的样本占所有负例样本的比例 (负例中预测错了的比例)；

纵坐标：Sensitivity，真正类率 (True positive rate，TPR)，预测为正且实际为正的样本占所有正例样本的比例 (正例中预测对了的比例)。

在一个二分类模型中，假设采用逻辑回归分类器，其给出针对每个实例为正类的概率，那么通过设定一个阈值如 0.6，概率大于等于 0.6 的为正类，小于 0.6 的为负类。对应的就可以算出一组 (FPR,TPR)，在平面中得到对应坐标点。随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样也掺杂着真正的负实例，即 TPR 和 FPR 会同时增大。阈值最大时，对应坐标点为 (0,0)，阈值最小时，对应坐标点 (1,1)。

3.2 AUC(Area Under Curve)

AUC (Area Under Curve) 被定义为 ROC 曲线下的面积，显然这个面积的数值不会大于 1。又由于 ROC 曲线一般都处于 y=x 这条直线的上方（如果不是，那么可以交换阈值上下对应的分类，即可得到更好的分类结果），所以 AUC 的取值范围一般在 0.5 和 1 之间。使用 AUC 值作为评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好。

3.3 为什么使用 ROC 曲线

既然已经这么多评价标准 (如 precision-recall 等)，为什么还要使用 ROC 和 AUC 呢? 因为 ROC 曲线有个很好的特性：当测试集中的正负样本的分布变化的时候，ROC 曲线能够保持不变。在实际的数据集中经常会出现类不平衡（class imbalance）现象，即负样本比正样本多很多（或者相反），而且测试数据中的正负样本的分布也可能随着时间变化。下图是 ROC 曲线和 Precision-Recall 曲线的对比：

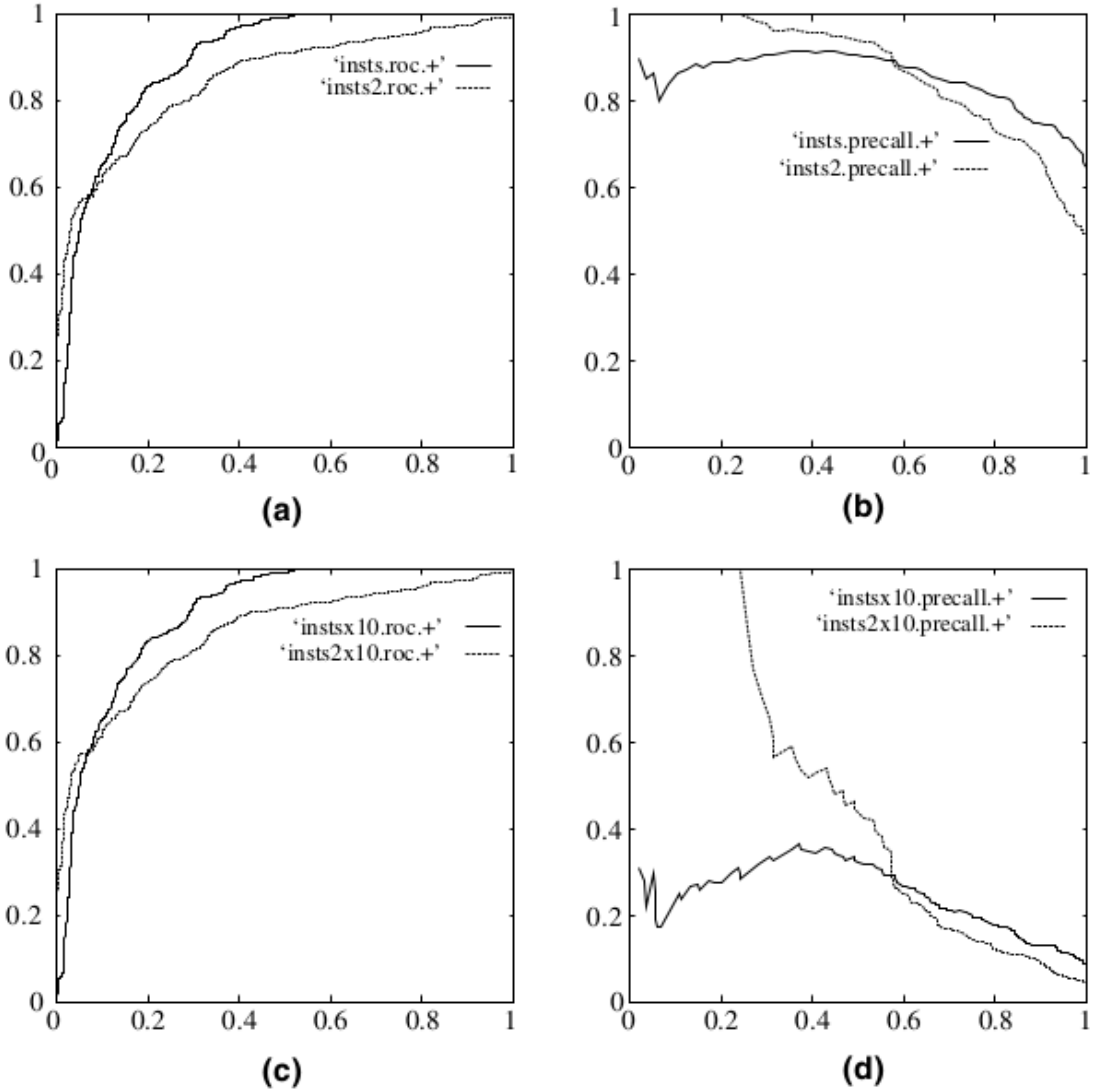

Figure 2: ROC vs Precision-Recall

在上图中，(a) 和 (c) 为 ROC 曲线，(b) 和 (d) 为 Precision-Recall 曲线。(a) 和 (b) 展示的是分类其在原始测试集（正负样本分布平衡）的结果，(c) 和 (d) 是将测试集中负样本的数量增加到原来的 10 倍后，分类器的结果。可以明显的看出，ROC 曲线基本保持原貌，而 Precision-Recall 曲线则变化较大。

3.4 精确率、准召率、F1 值各自的优缺点 [2]

3.4.1 精确率 Accuracy

Accuracy 是最常见也是最基本的 evaluation metric。但在 binary classification 且正反例不平衡的情况下，尤其是我们对 minority class 更感兴趣的时候，accuracy 评价基本没有参考价值。什么 fraud detection（欺诈检测），癌症检测，都符合这种情况。举个栗子：在测试集里，有 100 个 sample，99 个反例，只有 1 个正例。如果我的模型不分青红皂白对任意一个 sample 都预测是反例，那么我的模型的 accuracy 是正确的个数 / 总个数 = 99/100 = 99%。你拿着这个 accuracy 高达 99% 的模型屁颠儿屁颠儿的去预测新 sample 了，而它一个正例都分不出来，有意思么……也有人管这叫 accuracy paradox。

3.4.2 precision 和 recall

准招率是比 Accuracy 更有用的 metric。

recall 是相对真实的答案而言：true positive / golden set 。假设测试集里面有 100 个正例，你的模型能预测覆盖到多少，如果你的模型预测到了 40 个正例，那你的 recall 就是 40%。

precision 是相对你自己的模型预测而言：true positive / retrieved set。假设你的模型一共预测了 100 个正例，而其中 80 个是对的正例，那么你的 precision 就是 80%。我们可以把 precision 也理解为，当你的模型作出一个新的预测时，它的 confidence score 是多少，或者它做的这个预测是对的的可能性是多少。

一般来说，鱼与熊掌不可兼得。如果你的模型很贪婪，想要覆盖更多的 sample，那么它就更有可能犯错。在这种情况下，你会有很高的 recall，但是较低的 precision。如果你的模型很保守，只对它很 sure 的 sample 作出预测，那么你的 precision 会很高，但是 recall 会相对低。

这样一来，我们可以选择只看我们感兴趣的 class，就是 minority class 的 precision，recall 来评价模型的好坏。

3.4.3 F1-score

F1-score 就是一个综合考虑 precision 和 recall 的 metric：

$$\text{F1-score} = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$$

如果两个模型，一个 precision 特别高，recall 特别低，另一个 recall 特别高，precision 特别低的时候，F1-score 可能是差不多的，也不能基于此来作出选择。

References

[1] 机器学习之分类性能度量指标: Roc 曲线、auc 值、正确率、召回率.

[2] 精确率、召回率、f1 值、roc、auc 各自的优缺点是什么.