测试计划

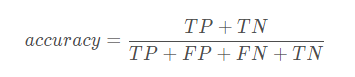
1. 说明：本小组项目为机器学习的应用，测试方法与常规的功能性软件区别较大，主要集中在训练导出的模型的评价上。另外根据设计，也会进行针对提供给用户的使用接口的功能和界面测试。
2. 测试内容
   1. 模型评价

本组课题使用随机森林和神经网络的方法，对恶意代码所属的APT组织进行检测，属于分类模型。常采用的指标有：

* 四个基础比例：
* 正确地预测为正例：TP
* 正确地预测为反例：TN
* 错误地预测为正例：FP
* 错误地预测为反例：FN
* **混淆矩阵Confusion Matrix：**以上四个比例构成的矩阵：



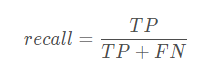
* **准确率Accuracy**：被预测正确的比例：



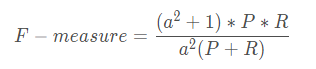
* **精确率(查准率) Precision**："正确被预测为正(TP)"占所有"实际被预测为正的(TP+FP)"的比例：

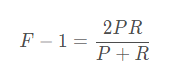


* **召回率(查全率) Recall**："正确被预测为正(TP)"占所有"应该被预测为正(TP+FN)"的比例：



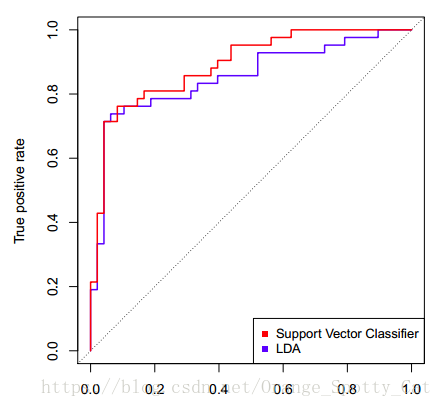
* **F-measure**：查准率和查全率通常是一对矛盾的度量。F-measure就是一个二者调和的指标。如果A取1，那么F-measure实际上就是precison和recall的调和平均值，此时的F-measure就称为F1值：





* **ROC曲线：**以模型的“False positive rate”与“True positive rate”值分别为横纵轴的曲线。曲线越往左上角贴近说明模型效果越好：

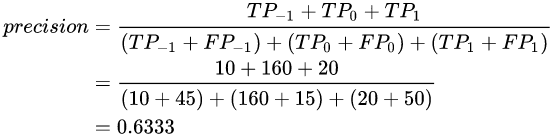




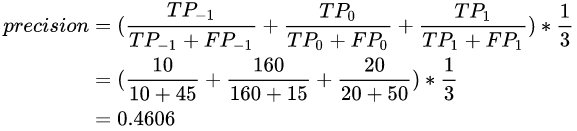
* **AUC（Area Under Curve）：**指ROC曲线下的面积是多大，取值在0与1之间。AUC = 1，代表ROC曲线在纵轴上，预测完全准确。

以上指标最初用于二分类模型的评价。而APT组织判定是多分类的问题，所以相应的指标含义会发生一些变化，并引入适用于多分类的新指标：

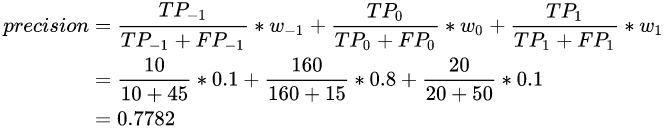
* 属于类C的样本被正确分类到类C，记这一类样本数为 TP
* 不属于类C的样本（可能是A也可能是B…）被错误分类到类C，记这一类样本数为 FP
* 属于类别C的样本被错误分类到类C的其他类，记这一类样本数为 FN
* 不属于类别C的样本被正确分类到了类别C的其他类，记这一类样本数为 TN
* 假设有N个类，混淆矩阵相应变为NxN的矩阵
* **micro方法**：指把所有的类放在一起算。例如对于precision，就是把所有类的TP加和，再除以所有类的TP和FN的加和。因此micro方法下的precision和recall都等于accuracy。



* **macro方法**：先分别求出每个类的precision，再算术平均。



* **weighted方法**：是macro算法的改良版，不再取算术平均、乘以固定weight，而是乘以该类在总样本数中的占比。



计划使用以上所有指标对我们的模型进行测试评价。

* 1. 功能及界面测试