基于机器学习APT检测

难点一：

|  |  |
| --- | --- |
| 名称 | 从原始病毒文件到可分析数据的预处理 |
| 解决方法 | 先将病毒文件放到cuckoo沙盒中运行，得到病毒文件行为的原始报告，然后用文本处理方法，提取病毒行为的关键词，形成词频统计，以矩阵形式作为模型的输入。 |
| 提出者 | 张以宁 |

难点二：

|  |  |
| --- | --- |
| 名称 | 训练模型的导出和读取 |
| 解决方法 | 使用scikit-learn的joblib模块的dump方法将模型保存到pkl文件，load方法读取文件，还原出分类器对象 |
| 提出者 | 张以宁 |

难点三：

|  |  |
| --- | --- |
| 名称 | 对用户输入的测试样例与训练出的模型进行匹配（模型种类多，不同的模型只能测某种样例） |
| 解决方法 | 暂未解决 |
| 提出者 | 杜梁楷 |

难点四：

|  |  |
| --- | --- |
| 名称 | 软件管理期间，时间较短，代码工作较多，测试时间较少 |
| 解决方法 | 由于此软件开发模型为敏捷开发，因此测试模型采用H模型，将软件测试独立开来，一旦完成某一部分的要求即开始测试。 |
| 提出者 | 罗潘亚欣 |

难点五：

|  |  |
| --- | --- |
| 名称 | 模型参数优化部分出现问题，部分参数遍历时即使全遍历效果也差强人意，甚至有逆向优化的现象 |
| 解决方法 | 重选参数讨论，并在基于训练集数量的情况下选择n\_features为auto模式，选择其余参数进行遍历，并重新界定范围来寻找更优参数。 |
| 提出者 | 林湛 |

难点六：

|  |  |
| --- | --- |
| 名称 | 在完成模型的训练和导出读取功能，准备做单样本的预测时，存在问题 |
| 解决方法 | 其一是将样本做文本预处理后作为模型输入时，出现了特征数量不一致的报错。这是因为测试样本与训练集的文本特征确实不一定匹配，对此我们提出了两种解决方案：一种是在构造词频统计器时先读取原训练集进行fit\_transform，再对测试样本进行transform，最后导出词频矩阵；另一种是提前提取一个训练集所有关键词的文件，并作为一项补在测试样本中，这样直接对测试集fit\_transform时也能输出完整特征的词频矩阵。经测试两种方法是等价的，为了预测效率更高我们选用了第二种方法。  其二是即使补齐了特征，输入到模型中预测值也非常不准确。经检查发现是因为原参考代码在训练时对数据根据最小最大值做了标准化处理，而测试集没有训练集的最小最大信息，自然也无法做相同的缩放处理。查询资料发现我们使用的随机森林分类器是非线性的，本来就不需要做标准化，修改代码重新训练模型后该问题解决。 |
| 提出者 | 张以宁 |