Web进程链恶意入侵及SSh异常操作检测技术合作问题

1. 时间周期：截至2020年7月31日
2. 经费及工作量：待苏教授预估
3. 样本数据：
   1. 由GTS提供从业务环境采集的白样本数据，和少量的人工构造黑样本数据。
   2. 如需更多维度的数据信息，可沟通
4. 编程环境依赖：
   1. 算法应满足平台运行的基本要求：Spark(v1.5) PySpark(v1.0)
   2. 现有开发平台可支持两种语言：Java JDK(v1.8) Python(v2.7.14)
   3. 算法可使用已集成的机器学习库：SKlearn Tensorflow(v1.10.0) Keras(2.2.4)
   4. 如需依赖第三方库，需提供源码安装方式。
5. 合作交付件及验收标准：
   1. 软件交付：Python或Java的算法组件，源码交付。
   2. 交付功能：
      1. Web进程链异常检测功能
      2. SSH命令操作序列异常检测功能
      3. 落地现网后可运营：关联历史行为进行告警抖动控制，保证告警可跟进，检测策略不失效。
   3. 性能指标：对已知攻击召回率>95%，准确率>95%，误报率<8%。

附件：机器学习算法指标进行的调研

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Accuracy/Auc | Recall | FAR(FPR) | Ref |
| 数据集 | 正确率/Auc值 | 召回率 | 误报率 | 参考文献 |
| Sendmail | -/- | - | 1.73% | Software Abnormal Behavior Detection Based on Hidden Markov Model |
| Malware zoo | 89.4%/- | 89.4% | - | Deep Learning for Classification of Malware  System Call Sequences |
| KDD98  UNM | -/99.4%  -/96.9% | 100%  99.8% | 12.3%  5.5% | LSTM-Based System-call Language Modeling Method for Designing Host-Based Intrusion Detection Systems. |
| Google Play | -/97% | 97% | 3% | Detecting Android Malware using Sequences of System Calls |
| Drebin project’s dataset | 93.7%/- | 96.6% | 9.3% | Android malware detection based on system call sequences and LSTM |

1. 监督模型，k-fold交叉验证后（训练集-测试集-验证集）根据上述指标衡量k折平均值。
2. 无监督模型，随机抽取一部分聚类结果人工添加标签后，以相同的指标衡量。

综上参考文献，最终交付验收指标：

**Accuracy/Auc: > 95%**

**Recall: > 95%**

**FAR(FPR): < 8%**