基础知识

隐马尔可夫模型的介绍与数学原理

工程实践

1、数据收集(Collector)

2、数据清理(Extractor)

3、模型训练(Trainer)

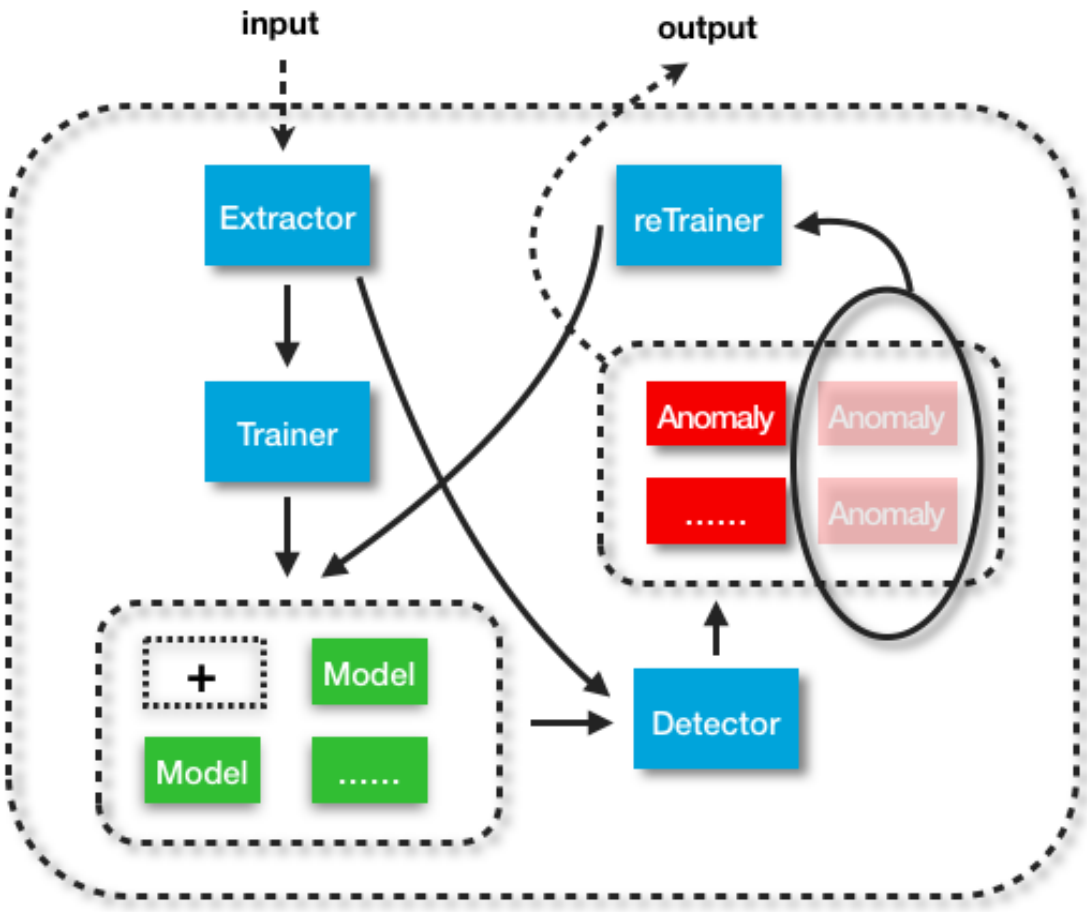
4、模型应用与评价(Detector)

5、*模型重训练器（Retrainer）

开发语言：Python

数据库：Spark or Hadoop or MYSQL

数据流向图：



项目初步实施路径

1、了解基础知识 1 天

- 阅读有关 HMM 的介绍和原理，对如何使用 HMM 识别恶意流量有一定的理解

2、部署环境 1 天

- 安装 spark，python2.7 以及相关依赖包, TcpFlow（或其他 HTTP 抓包工具）

3、数据收集与存储 2-5 天

- 从 web 应用中采集日志，使用 logstash 从日志文件中提取日志，或者从网络流量中抓包提取 http 信息

- 将得到的信息存储到 spark 数据库中，将得到的数据库封装好，写好解释文档

3、数据清理 1-2 周（详情请见任务分拆部分）

- 参数提取

- 数据泛化

4、模型训练 1 周

- 用 python 下的 hmmlearn 模块完成训练，实现对任意 HTTP 请求的分类器

5、模型初步应用，测试，项目打包整合 3-7 天

- 进行软件测试和简单 UI 制作

项目后续完善：

1、应用 Spark Streaming 实现实时监测

2、模型动态重训练模块

核心参考材料：

1. https://www.jianshu.com/p/942d1beb7fdd
2. https://www.freebuf.com/articles/web/134334.html
3. https://www.anquanke.com/post/id/107124
4. https://www.freebuf.com/column/132796.html
5. https://www.freebuf.com/column/134319.html
6. https://github.com/SparkSharly/Sharly

任务分拆

模块 1 网络流量采集

1.实现自动化监测网卡流量并记录

2.能提供较大数据量的正常流量（白样本），并以 spark 数据库的形式存储

模块 2 参数提取

对 http 请求数据进行拆解，提取如下参数：

- GET、POST、Cookie 请求参数
- GET、POST、Cookie 参数名本身
- 请求的 URL 路径
- http 请求头，如 Content_type、Content-Length(对应 strust2-045)

这部分的难点在于如何正确的识别编码方式并解码

模块 3 数据泛化

大小写英文字母泛化为”A”，对应的 unicode 数值为 65

- 数字泛化为”N”，对应的 unicode 数值为 78
- 中文或中文字符泛化为“C”，对应的 unicode 数值为 67
- 特殊字符和其他字符集的编码不作泛化，直接取 unicode 数值
- 参数值为空的取 0

更多关于数据泛化的内容请见参考材料

模块 4 模型训练

1. 熟悉 python 中的 hmmlearn 库，能做到训练小部分数据并生成模型

2. 保存模型参数到 Hdfs 中

3. 根据生成的模型实现检测模块（Detector）

模块 5 项目测试和打包

1. 将不同模块进行整合，完成单元测试和系统测试

2. 实现简单 UI，整理相关依赖包，实现一键部署