**入侵检测与数字取证大作业**

57119108吴桐

57119126傅寒青

**T1. 恶意域名家族之全员通缉**

**题目描述**

对给出域名的流量数据进行分析，按照请求特征对域名进行分类。

有标签的域名只占所有恶意域名的50%，请选手找出另外50%的恶意域名并正确分类。

域名数量：总量约20000个，其中恶意域名约1000个。

**题目分析**

本题的数据信息量极大，需要进行全面的数据分析和特征提取。题目只给出了0-8共9种恶意家族标签，却没有给出良性样本标签，无法直接应用监督学习模型训练；存在样本数据缺失的情况，完整的样本数据只有fqdn.csv文件，其余文件都存在不同程度的数据缺失情况。

针对第一个问题，我们可以从无标签样本中随机选择一部分作为正常样本进行监督训练。题目要求找到另外500个恶意域名，因此可以推测20000个样本中正常样本出现的概率要远大于恶意样本。我们对所有无标签样本进行随机采样，选取约1500个样本作为正常样本。

针对样本数据缺失的情况，我们简单地使用零填充方法解决。

**解题过程**

**特征提取**

1、access.csv

（1）域名的总访问次数

（2）访问域名的IP集合

（3）按小时统计域名的访问次数

（4）按日期统计域名的访问次数

（5）访问域名的IP所在的国家

（6）访问域名的IP所在的城市

（7）访问域名的IP所属的ISP

2、flint.csv

（1）域名解析出的IP集合

（2）域名解析出的域名集合

（3）按日期统计域名的解析次数

（4）域名解析出的IP所在的国家

（5）域名解析出的IP所在的城市

（6）域名解析出的IP所属的ISP

3、fqdn.csv

（1）域名长度

（2）域名的字符占比

（3）域名的数字占比

（4）域名中包含的单词数量

（5）域名中包含的单词长度占比

（6）域名深度

4、whois.json

（1）域名创建时间

（2）域名过期时间

（3）域名更新时间

（4）域名的管理员邮箱

（5）域名的注册国家

（6）域名的注册省份

（7）域名的注册邮箱

（8）域名的技术邮箱

（9）域名的DNS服务器

（10）域名的注册商

**模型训练**

最初我们使用随机森林模型进行训练，然而效果并不理想；之后我们选择了XGBoost模型进行训练。

首先，将随机选取的无标签样本作为正样本，将所有家族的恶意样本作为负样本，将全部域名作为测试集，训练二分类模型，预测得到恶意域名集合。

将第一步预测得到的恶意样本数据作为测试集，题目中给出有标签恶意域名样本作为训练集，训练多分类模型，对测试集中的恶意样本进行家族种类预测，最终得到家族类别标签。

**T2. 猜猜我是谁之黑白域名现形记**

**题目描述**

对给出域名的流量数据进行分析，按照请求特征对域名进行二分类。

**题目分析**

本题主要从access.csv，flint.csv和fqdn.csv三个文件中提取特征。题目给出了0和1两种有标签样本，需要对全部样本进行二分类，因此可以直接使用监督分类。针对数据缺失的问题，我们简单地使用零填充方法解决。

**解题过程**

**特征提取**

1、access.csv

（1）按小时统计域名的访问IP个数

（2）按日期统计域名的访问IP个数

（3）按小时统计域名的请求次数

（4）按日期统计域名的请求次数

2、flint.csv

（1）域名解析出的IP集合

（2）域名解析出的域名集合

（3）按日期统计域名的解析次数

（4）域名解析出的IP所在的国家

（5）域名解析出的IP所在的城市

（6）域名解析出的IP所属的ISP

3、fqdn.csv

（1）域名长度

（2）域名的字符占比

（3）域名的数字占比

（4）域名中包含的单词数量

（5）域名中包含的单词长度占比

（6）域名深度

**模型训练**

使用随机森林模型进行训练。

首先，将有标签样本作为训练集，将全部域名作为测试机，训练二分类模型，预测得到恶意域名集合。

将第一步预测得到的恶意样本数据作为测试集，题目中给出的恶意域名样本作为训练集，训练二分类模型，对测试集中的恶意样本进行家族种类预测，最终得到家族类别标签。

**T3. 猜猜我是谁之域名大分类**

**题目描述**

对给出域名的流量数据进行分析，按照请求特征对域名进行聚类，要求簇数不能小于3。

**题目分析**

本题给出了聚类的簇数为5，分别是botnet，apt，trojan，gamble和white，因此我们直接选择KMeans算法实现分类。需要注意的是，使用KMeans算法时不宜选择过多特征，我们可以从类别名称中猜测部分有效特征，如域名的总访问次数，境内境外访问次数，域名解析类型等等。

**解题过程**

**特征提取**

1、access.csv

（1）域名的总访问次数

（2）访问域名的IP集合

（3）访问域名的IP所在的国家

（4）访问域名的IP所在的城市

（5）访问域名的IP所属的ISP

（6）访问域名的境内IP集合

（7）访问域名的境外IP集合

2、flint.csv

（1）域名解析的类型

3、fqdn.csv

（1）域名长度

（2）域名的字符占比

（3）域名的数字占比

（4）域名中包含的单词数量

（5）域名中包含的单词长度占比

（6）域名深度

**模型训练**

使用KMeans算法对样本进行聚类，类簇数为5。

**总结**

本次实验极大地锻炼了我们的团队协作能力，进一步加深了我们对于机器学习域名分类方法的理解，提升了我们的代码能力和分析解决问题的能力。实验过程中，我们尝试了各种特征子集作为训练根据，并采用了多种训练模型，最终在我们的能力范围之内选择了最好的结果。