1. 引言

随着互联网的普及，恶意代码的危害也变得越来越来越难以控制。201基于随机森林的恶意代码检测7年5 月12 日晚上20 时左右，全球爆发大规模蠕虫勒索软件感染事件，仅仅几个小时内，该勒索软件已经攻击了99个国家近万台电脑。英国、美国、俄罗斯、德国、土耳其、意大利、中国、菲律宾等国家都已中招，且攻击仍在蔓延。据报道，勒索攻击导致16家英国医院业务瘫痪，西班牙某电信公司有85%的电脑感染该恶意程序。至少1600家美国组织，11200家俄罗斯组织和6500家中国组织和企业都受到了攻击。国内也有大量教学系统瘫痪，包括校园一卡通系统。根据国内知名互联网安全厂商奇虎360在2017年2月发布的安全报告，2016年360互联网安全中心共截获PC端新增恶意程序样本1.9亿个。敲诈者病毒在国内发生两次大规模传播，全国至少有497多万台用户电脑遭到了敲诈者病毒攻击。通过对受害者调研，42.6%的受害者不知道感染病毒的原因。预计在2017年敲诈者会增长10倍，且利用挂马攻击也将再次爆发。各种病毒、木马、蠕虫等恶意代码在网络间广泛传播，已经给个人、企业甚至政府带来了难以估量的损失， 因此研究更有效的恶意代码检测技术是十分有意义的。

恶意代码的各种混淆和变形手段是其难以检测的主要原因。传统的基于特征匹配的检测方法对于各种变形的恶意代码显得无能为力。随着机器学习和数据挖掘技术的快速发展，已经有许多研究将这些技术用于恶意代码的检测，并且取的了不错的效果。其中，大部分研究都是首先对操作码进行特征的提取，此时特征维度会特别高，然后再进行降维处理以选择特征，最后对选择出的特征应用机器学习算法进行训练和检测。这种方法的检测结果依赖于特征选择算法的有效性，而且由于汇编操作码的局部性，对于程序总体流程的反映有着严重的局限性。针对以上存在的问题，本文通过对大量恶意代码在汇编操作码层次的统计，提出对汇编操作码进行粗粒度划分的方法，对汇编操作码进行集合映射，以便能很好的反映了程序的总体流程，又不会维度太高而需要进行复杂的降维过程；最后，基于随机森林算法对提取出的特征进行快速精准的分类。主要成果：1）提出一种对汇编操作码进行集合映射的方法，建立一种改进的n-gram模型；2）基于hadoop分布式环境，实现了分布式操作码集合特征模型训练；3）有助于研究人员了解恶意代码对汇编操作码的使用情况，可用于进一步恶意代码家族分类。

1. 恶意代码检测方法整体设计

本文所述的恶意代码检测技术是一种静态检测技术。检测过程主要分为四步，分别是反汇编、汇编操作码集合映射、n-gram特征提取、使用机器学习算法对样本进行训练和测试。

本文所采用的恶意程序样本来自于卡饭论坛以及360截获的恶意软件，非恶意程序样本在360软件库中随机下载。接下来对收集的软件进行反汇编，以便从中提取所需的静态汇编操作码。反汇编的工具有很多，本文采用objdump工具对所有PE文件进行反汇编。首先，将反汇编的结果以文件形式保存；接下来对所有文件进行汇编操作码到集合的映射操作，根据原有汇编操作码的顺序获得集合的顺序；最后根据保存集合顺序信息的文件提取n-gram信息，并且使用n-gram信息获得每种特征所占比例作为特征。至此，特征提取工作结束。

在Malware Detection based on Random Forests提取了样本特征后，将这些特征作为机器学习算法的输入，从而获得训练检测模型。其总体设计如图1所示：

图恶意代码检测模型

1. 恶意代码检测方法详细设计与实现
   1. 恶意代码检测技术概述

现有的恶意代码检测技术，根据恶意代码的运行状态，分为静态检测技术和动态检测技术。动态的方法根据行为来判断程序是否恶意，首先动态方法只能判断程序一次的执行路径，而程序的分支很多无法预知在某一次执行会是哪一条路径，这样的判断方法不够全面；其次，很多恶意代码会检测其所处环境是否是虚拟机环境，可以根据实际环境隐藏或者改变自己的行为来逃避检测动态检测技术由于其实现的复杂性及效果的不明显性。由于动态检测技术的不完整性，现有的检测技术大多以静态检测为主。当然静态检测技术也有缺陷，静态检测技术对于恶意代码的混淆技术束手无策，将机器学习算法应用于恶意代码检测，对于对抗混淆技术具有很显著的效果。

不难发现，机器学习算法的有效性依赖于特征提取和特征选择的有效性，特征选择的结果一般是特征提取后通过降维过程获得，降维过程是一个非常复杂的过程，耗时且具有非常大不确定性，不能保证降维后的结果一定是对于分类最有效的特征。

基于以上存在的问题，本文提出一种新的静态检测技术，将汇编操作码进行集合映射，每种集合代表了不同的动作，按照汇编指令的顺序得到集合的顺序，通过集合的顺序可以大致反映应用的总体流程，同时会过滤掉一些无用的可能会影响分类的因素；然后对该集合顺序应用n-gram算法并获得词组的频度信息，考虑到程序之间汇编操作码数量的差异，本文使用每词组的概率作为特征；最后将此概率矩阵计作为机器学习的输入，生成训练检测模型。其具体流程如图2所示：

图机器学习检测一般过程

* 1. 特征提取
     1. 汇编映射及特征提取

将汇编操作码进行集合映射后再进行后续分析，看起来好像是一种牺牲精确性，以换取效率提升的办法，但是如果映射方法得当，提升效率的同时又不损失精确性也不是不可能。本文通过分析大量恶意代码和非恶意代码的汇编操作码，将汇编指令分为9个集合，每种汇编操作码都唯一的映射到一个集合中。

1. 数据传输，比如：MOV、PUSH、POP、IN、OUT等；
2. 算术运算，比如：ADD、SUB、COM、DIV、MUL等；
3. 逻辑运算，比如：NOT、AND、OR、TEST等；
4. 移位运算，比如：SAL、SHL、SAR、SHR等；
5. 字符串操作，比如：MOVS、STOS、CMPS等；
6. 过程控制，比如：JMP、JZ、JP等；
7. 标志位控制，比如：CLC、STD、STC等；
8. CPU控制，比如：HLT、WAIT、NOP等；
9. 其他，不常用的指令。

使用objdump获取到程序的汇编源码之后，将其转换为以上9种集合指令，按照原汇编指令的顺序对集合指令进行排序，以下代码列出了这9种集合指令的表示方法：

|  |
| --- |
| #define DATA\_TRANS 1  #define ALTH\_OPE 2  #define LOGIC 3  #define SHIFT 4  #define STR\_OPE 5  #define PRO\_CTRL 6  #define FLAG\_CTRL 7  #define CPU\_CTRL 8  #define OTHER 0 |

图3是截取样本种的一段代码，将其进行集合映射后结果如图4所示：

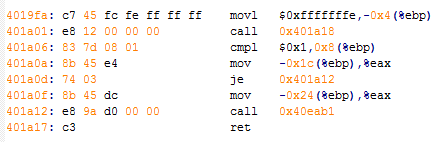
图3 指令样本

图4 集合映射后的指令顺序

* + 1. 改进型n-gram特征提取

n-gram是计算机语言学和概率论范畴内的概念，目前n-gram被广泛应用于自然语言的自动分类功能。一般的恶意代码检测算法直接对操作码使用n-gram算法，导致获得的特征维度特别大，从而不得不进行降维处理，降维的办法有很多，大部分都是使用信息增益或者概率的算法进行淘汰，达到降维的目的。

本文中对汇编操作码进行集合映射，除非n取值特别大，一般不会出现维度特别大的情况，所以降维过程是不必要的。对于图4应用2-gram算法，提取其特征集m，其结果为；然后根据特征集m统计每种组合的出现次数；最后由统计结果求出每种组合的概率分布，表1给出了图4的概率分布结果。

表图概率分布（未填写表示）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0 |  |  |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  | 3/7 |  |  |
| 2 |  | 1/7 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |  |
| 6 |  | 1/7 | 1/7 |  |  |  | 1/7 |  |  |
| 7 |  |  |  |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |  |  |  |

本文中将汇编操作码映射为9类，对于1-gram特征共有9维，相应的2-gram、3-gram以及4-gram特征维度分别是81、729和6561。本文将分别使用2-gram、3-gram和4-gram对实验数据进行特征提取。

* 1. Random Forests算法

本文使用了Random Forests算法来构建训练测试型。下面就对这种算法加以简单介绍。

Random Forests算法，是一个包含多个决策树的分类器，其输出的类别是由个别树输出的类别的众数决定的。决策树由一个决策图和可能的结果组成，用来达到创建目标的规划。决策树是一种特殊的树结构，在这棵树上非叶子节点都是属性节点，叶子结点时类别节点。用户根据现有的条件（输入数据）首先针对决策树的根节点对现有条件进行分析决定现有条件的走向。在建立决策树的过程中，使用熵和增益两个参数评估属性与类别之间的关系。

Random Forests算法工作原理：针对输入数据进行如下算法，输入数据一般包括训练集个数N、训练集特征属性数目M、随机选取的特征属性数目m（m<M）、随机森林包含的决策树总数T。

1. 从N个训练集中以有放回抽样的方式、取样N次，形成一组N个训练集（即bootstrap取样）；
2. 循环步骤1），形成T组训练集；
3. 对于每一组训练集，随机选取m个训练集特征属性。根据这m个特征值计算其最佳的分裂方式；
4. 每棵树都要在完整成长而不减枝的前提下完成；
5. 当所有的T棵树建立完成后，算法结束。
6. 实验数据与结果

实验中恶意样本来自于卡饭论坛和本机360拦截的少量程序，非恶意代码从360软件库中下载。其中恶意程序个数为420个，非恶意程序个数为389个，并且所有程序都反汇编成功。我们选择320个恶意程序和320个非恶意程序作为训练样本，剩下的100个恶意样本和69个非恶意样本作为测试集。实验平台为伪分布式Hadoop平台，操作系统是Windows 64位专业版。

本文主要通过正确率、准确率和召回率三项来评估实验结果。假设恶意样本和非恶意样本的总数量设为N，非恶意样本被正确分类的样本个数位TN，恶意样本被正确分类的样本个数为TP，非恶意样本被分类为恶意样本的个数为FP，恶意样本被分类为非恶意样本的个数为FT。正确率反映了判别系统对所有样本的总的判定能力，计算公式如下：

准确率表示被检测出的恶意程序占所有被判定为恶意程序的比率，计算公式如下：

召回率反映了分类成功的恶意程序占所有恶意程序的比率，计算公式如下：

分别对2-gram样本、3-gram样本、4-gram样本利用随机森林算法进行训练，再利用各自的测试集进行检测，最终获得的效果如下表2、表3和表4：

表2 依据2-gram实验结果

|  |  |
| --- | --- |
| 指标 | 结果 |
| 正确率（Accuracy） | 90.53% |
| 准确率（Precision） | 92.86% |
| 召回率（Recall） | 91.00% |

表3 依据3-gram实验结果

|  |  |
| --- | --- |
| 指标 | 结果 |
| 正确率（Accuracy） | 92.90% |
| 准确率（Precision） | 94.00% |
| 召回率（Recall） | 94.00% |

表4 依据4-gram实验结果

|  |  |
| --- | --- |
| 指标 | 结果 |
| 正确率（Accuracy） | 87.57% |
| 准确率（Precision） | 95.40% |
| 召回率（Recall） | 83.00% |

对比三个实验结果，发现除了依据4-gram进行训练检测的结果稍差，其他两个实验结果，对于恶意代码的分类的三种指标都在90%以上。2-gram对于非恶意程序的分类误报率相对较高，4-gram对恶意程度的误报率相对较高，但是对于3-gram，其对于非恶意程序和恶意程序的误报率都计较低。

实验结果表明，将汇编操作码首先进行集合映射，以代替或者补充降维算法，可以达到良好的分类效果。实验种所有程序是彼此独立的，结果表明，对于未知样本该模型也可以达到良好的检测效果。

1. 总结

本文提出的检测方法是一种静态检测方法，首先对汇编操作码的集合映射，大大简化了对于程序的分析和建模过程，避免或者简化了复杂的降维过程；然后对于映射结果直接根据n-gram算法提取特征，本文分别使用的2-gram、3-gram和4-gram进行特征提取；最后对于获得的特征直接使用随机森林算法进行训练检测。实验结果表明，三种特征对于样本都有比较好的分类效果，尤其是3-gram特征，有较高的检测精准度和较低的误报率。

在以后的学习工作中，继续研究恶意程序的汇编操作码规律，研究更加精确的集合映射方法，来简化或者优化特征提取和特征选择算法，并且继续提高检测精度。