PaperFree检测报告简明打印版

相似度：10.88%
编　号：QTIYLARMVQJESVEI
标　题：基于随机森林得恶意代码检测技术
作　者：-
长　度：4301字符
时　间：2017-06-20 14:29:53
比对库：中国学位论文全文数据库；中国学术期刊数据库；中国重要会议论文全文数据库；英文论文全文数据库；互联网资源；自建比对库

相似资源列表(学术期刊，学位论文，会议论文，英文论文等本地数据库资源)
1. 相似度：1.15%　篇名：《随机森林理论浅析》
来源：《集成技术》　年份：2013　作者：董师师
2. 相似度：0.56%　篇名：《基于机器学习的android恶意软件检测模型研究》
来源：《青岛理工大学硕士论文》　年份：2013　作者：蔡泽廷
3. 相似度：0.44%　篇名：《基于C5．0决策树的税务稽查研究》
来源：《连云港职业技术学院学报》　年份：2011　作者：陈仕鸿
4. 相似度：0.38%　篇名：《一种针对Android平台恶意代码的检测方法及系统实现》
来源：《西安交通大学学报》　年份：2013　作者：胡文君
5. 相似度：0.35%　篇名：《移动智能终端应用软件安全检测》
来源：《信息安全与通信保密》　年份：2012　作者：袁春阳
6. 相似度：0.35%　篇名：《基于iv属性选择的随机森林模型研究》
来源：《合肥工业大学硕士论文》　年份：2015　作者：杨秋洁
7. 相似度：0.32%　篇名：《数据挖掘技术在网络安全中的应用》
来源：《计算机仿真》　年份：2011　作者：郑艳君
8. 相似度：0.29%　篇名：《求解古典概型常见错误剖析》
来源：《理科考试研究：高中版》　年份：2013　作者：谭斌
9. 相似度：0.26%　篇名：《Hadoop大数据平台的搭建与测试》
来源：《山东农业大学学报：自然科学版》　年份：2013　作者：崔文斌

相似资源列表(百度文库，豆丁文库，博客，新闻网站等互联网资源)
1. 相似度：2.85%　标题：《基于操作码N¯Gram的Windows恶意软件检测_中国论文网》
来源：http://www.xzbu.com/4/view-6759873.htmiphone
2. 相似度：1.23%　标题：《N-gram的原理、用途和研究 - cdsj - 博客园》
来源：http://www.cnblogs.com/cdsj/p/5720391.html
3. 相似度：1.15%　标题：《决策树模型组合之随机森林(Random Forest) - niujin1212的专栏 - 博...》
来源：http://blog.csdn.net/niujin1212/article/details/50804110
4. 相似度：1.15%　标题：《随机森林 - 搜狗百科》
来源：
http://baike.sogou.com/v70245256.htm?fromTitle=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%
5. 相似度：0.79%　标题：《过拟合、均值漂移、决策树概念学习笔记_suppose_rWySp_新浪博客》
来源：http://blog.sina.com.cn/s/blog_14d68bd870102wy8z.html
6. 相似度：0.79%　标题：《决策树理论汇总_20区》
来源：http://www.20qu.com/jingxuan/2016/316/1314547.html
7. 相似度：0.44%　标题：《面向邮件系统的意代码检测技术研究.pdf文档全文免费阅读、在线看 》
来源：http://m.book118.com/html/2016/0105/32745781.shtm
8. 相似度：0.29%　标题：《基于数据挖掘和机器学习的恶意代码检测技术研究.pdf -max上传文档...》
来源：http://m.book118.com/html/2015/0922/25933808.shtm

全文简明报告

随着互联网的普及,恶意软件的危害也变得越来越来越难以控制。{ 63% : 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。由于恶意代码的各种混淆和变形手段,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。{ 55% : 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。比如Woo[1]等人通过分析操作码频率检测,Imran[2] 等人基于HMM的特征提取方法进行检测,J.Zico Kolter[3] 等人提出的基于字节码的检测,Christodorescu[4]等人通过反映恶意代码行为的指令模板检测恶意代码,Kolbitsch[5]等人使用了基于行为的检测技术等等。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,{ 55% : 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软件在汇编操作码层次的统计,对汇编操作码进行粗粒度划分,将汇编操作码映射到9类集合。首先根据汇编操作码序列获取集合序列,这样既能反映程序的大体流程,又不会因为维度太高而需要进行复杂的降维过程;{ 64% : 然后对集合序列应用n-gram算法提取特征, }根据提取到的特征得到频率矩阵作为随机森林算法的输入进行建模;最后根据该模型对样本进行快速精准的分类。

本文的主要成果:1)提出一种对汇编操作码进行集合映射的方法,{ 68% : 建立一种改进的n-gram模型; }{ 62% : 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软件对汇编操作码的使用情况,可用于进一步恶意软件家族分类。

恶意代码检测模型整体设计

本文所述检测模型的总体思路是根据样本的集合频率矩阵生成随机森林检测模型,然后根据将检测样本的集合频率矩阵作为随机森林检测模型的输入,获得分类结果。其总体流程如图1:

图1 恶意代码检测模型

如图1所示,首先通过对样本数据进行反汇编,获得汇编操作码,根据映射矩阵对汇编操作码进行集合映射,得到集合序列,该阶段称为预处理模块;然后对集合序列应用n-gram算法,获得所有可能的集合词组出现的次数,由于不同样本存在大小差异,所以汇编操作码的数量也会有很大差异,导致统计的集合词组频度受到样本大小的影响,所以本文采用频率矩阵来解决这种问题,根据集合词组频度获得其频率矩阵;最后将集合频率矩阵作为随机森林的输入,获得随机森林的训练检测模型。同样对检测样本进行预处理,获得集合频率矩阵并作为检测模型的输入,得到分类结果。

恶意代码检测模型设计与实现

现有的恶意代码检测技术,本质上都是通过各种方法对样本进行特征提取,{ 56% : 然后将其作为机器学习算法的输入, }得到检测模型。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,本文通过改进现有的静态检测技术,从而避免降维处理等复杂过程。本文提出的检测模型主要包括两大模块:

特征提取。特征提取包括反汇编、汇编操作码集合映射以及n-gram频率矩阵的获取;

随机森林检测模型。将步骤1)得到的频率矩阵作为随机森林模型的输入,得到检测模型。

特征提取

汇编映射

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。

数据传输,比如:MOV、PUSH、POP、IN、OUT等;

算术运算,比如:ADD、SUB、COM、DIV、MUL等;

逻辑运算,比如:NOT、AND、OR、TEST等;

移位运算,比如:SAL、SHL、SAR、SHR等;

字符串操作,比如:MOVS、STOS、CMPS等;

过程控制,比如:JMP、JZ、JP等;

标志位控制,比如:CLC、STD、STC等;

CPU控制,比如:HLT、WAIT、NOP等;

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,按照原汇编操作码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2是截取样本中的一段代码,将其进行集合映射后结果如图3所示:

图2 指令样本

图3 集合映射后的指令顺序

改进型n-gram特征提取

{98% : n-gram是计算机语言学和概率论范畴内的概念, }{ 76% : 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汇编操作码或者字节码使用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,大部分都是基于概率分布[6]或者信息熵[7~8]的算法进行特征选择,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S,其结果为{^1,6^,{ 61% : ^6,2^,^2,1^,^1,6^, }^6,1^,^1,6^,^6,6^};然后根据特征集S统计每种词组的出现次数;最后由统计结果求出每种词组的频率分布,表1给出了图4的频率分布结果。

表1 词组频率分布表(未填写表示i€ºº

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,{ 56% : 相应的2-gram、3-gram以及4-gram特征维度分别是81、729和6561。 }{ 60% : 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。 }

随机森林算法

本文使用了随机森林算法来构建训练测试模型。{88% : 随机森林算法,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。 }{91% : 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。

随机森林算法工作原理: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。{ 55% : 训练模型的生成流程如图4所示,检测模型流程如图5所示。 }

图4 随机森林训练模型流程图 图5 随机森林检测模型流程图

实验数据与结果

本文给出实验中恶意样本来自于卡饭论坛和本机360拦截的程序,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,{ 69% : 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:

Accuracy=((TP+TN))/N

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:

Precision=TP/(TP+FP)

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:

Recall=TP/(TP+FN)

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%

对比实验结果,如图6所示,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较低。

图6 结果对比图

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。实验中所有样本是彼此独立的,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。

结论

{ 58% : 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;{ 57% : 然后对于映射结果直接根据n-gram算法提取特征, } { 65% : 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,{ 76% : 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