## 第一章 绪论

### 1.1选题背景及意义

#### 1.1.1选题背景

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。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。据国内知名互联网安全厂商奇虎360发布的《2016年中国互联网安全报告》显示，2016年全年，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。2016年360互联网安全中心共截获Android平台新增恶意程序样本1403.3万个，其中资费消耗类程序为74.2%。同PC端相似，手机端勒索软件也开始爆发，360全年截获新增手机勒索软件17万，170万台手机遭到攻击。在截获盗取个人信息的手机恶意程序样本中，67.4%的样本会窃取短信信息，34.8%的样本会窃取手机银行信息，10.0%的样本会窃取手机联系人信息，3.7%的样本会窃取手机通话记录，2.0%的样本会窃取社交软件（例如微信、QQ等）聊天记录，1.8%的样本会窃取手机录音信息，0.1%的样本会窃取手机照片信息。根据中国互联网安全报告显示，以下是2017年上半年5大典型安全威胁事件：

1. 2017年上半年“WannaCry”、“暗云Ⅲ”、“Petya”等多种类型的病毒木马连续集中的爆发，为社会和行业敲响了网络安全的警钟。《报告》显示，仅2017年上半年，腾讯安全反病毒实验室在电脑端总计已拦截病毒已超过10亿次，平均每月拦截木马病毒近1.7亿次，相较于2016下半年病毒拦截总量增长30%。其中，受“WannaCry”刺激，勒索类病毒仅第二季度就新增了13.39%，但“WannaCry”在非感染型敲诈类病毒占比中仅排第三，而带有感染传播方式的PolyRansom勒索病毒传播力更巨大，其占所有勒索类病毒的78.84%。
2. 国务院某App的H5遭遇流量劫持。5月中旬，某国字号的App遭遇流量劫持的传闻在业界流传。有消息称，该App某H5页面被植入色情内容广告，经排查“基本确定为用户当地运营商http劫持导致H5页面被插入广告……”。
3. 12306官方网站再现安全漏洞。4月21，有媒体记者发现在12306官方网站订票时发现，当退出个人账号，网站页面竟自动转登他人账号，且与账号相关联的身份证号、联系方式等个人信息均可见，随后记者在该页面点击常用联系人选项时页面再次刷新并显示他人账号及账号涵盖的所有信息。而记者尝试在网站账户页面的个人信息栏等其他选项进行操作，点击进入后均得到不同的个人身份信息。
4. 上亿优酷信息数据在暗网售卖。4月17日，外国媒体hackread报道，100759591 条优酷账户信息数据库在暗网售卖，该数据库售卖价格定为比特币 0.2559，，人民币约 2065.56 元。
5. “土耳其犯罪家庭”的网络犯罪团伙掌握3亿苹果帐户。3月底，国外媒体报道，自称为“土耳其犯罪家庭”的网络犯罪团伙，通过电子邮件告知苹果公司他们掌握了超过3亿苹果帐户，并能远程清除所有装置的内容。他们宣表示只想苹果支付75000美元的比特币赎金，或者价值10万美元的iTunes礼品卡。若苹果公司在4月7日拒绝遵守他们的要求，他们将大量清除iCloud帐户。

#### 1.1.2选题意义

综上所述，恶意代码的危害无处不在，不仅给个人、企业带来了巨大的损失，甚至可能给国家安全带来不可预期的危害。对于个人，恶意代码的入侵会导致个人隐私的泄露，造成经济或者名誉损失；对于企业来说，企业数据一旦遭到恶意代码的入侵，会导致企业大量数据资产外泄，给企业带来无可挽回的损失，甚至因此产生一些灰色的产业链；对于国家，信心安全是国家安全的重点组成部分，网路安全已经是国策和民生的大问题，信息安全成为国家战略，随着互联网的发展，信息安全的问题将会更加突出和重要。因此研究更加有效的恶意代码检测技术是非常具有现实意义的。

<待扩展>

### 1.2国内外研究现状

### 1.3本文研究内容

本文设计并实现了一个自动化检测恶意代码的系统，并提出了一种新的静态恶意代码检测方法。主要是在程序的汇编代码中提取特征，然后根据特定的映射规则对汇编指令进行更高层次的抽象，以此更好的反映程序的行为特征。这种方法结合了基于特征码和行为的共同特征，最终使用机器学习的分类技术实现检测。这样既可以克服动态检测的不完整性，又解决了一般静态检测技术不能处理的混淆带来的困惑。首先，为了逃避病毒检测系统的检测，一般的恶意代码作者都会对恶意程序进行加壳处理，所以本文的第一步就是对恶意代码进行脱壳处理。其次，对恶意程序进行反汇编处理，得到程序的汇编代码，并提取汇编操作码序列。然后，配置特定映射规则对汇编操作吗序列进行抽象化处理。接着，使用n-gram算法提取特征并且根据配置文件选择降维方法，进行特征选择。最后，使用随机森林、支持向量机以及K邻近三种机器学习分类算法实施分类。本文的工作有以下几点：

第一：本文提出一种新的方式对汇编代码进行抽象，以此在高层次反映程序的控制流。这样可以在一定程度上避免混淆技术带给静态检测技术的困惑。

第二：采用模块化编程的技术实现检测系统，在映射规则、n-gram算法n的选取以及降维算法的选取上，采用配置文件的方式，方便了实验的进行和结果的对比。

第三：对不同的映射规则下得到的结果进行对比，然后总结出最优的映射规则。

### 1.4本文章节安排

本文总共分为四章：

第一章为绪论，主要介绍选题背景以及选题的意义，并且详细阐述了国内外对于恶意代码检测技术的研究现状和存在的问题，最后介绍了本文的研究内容和章节安排。

第二章是恶意代码检测技术概述，主要介绍恶意代码的定义、分类和现有恶意代码检测技术的原理和优缺点。

第三章是基于机器学习算法的恶意代码检测技术，按照步骤详细介绍了恶意代码的查壳与脱壳、反汇编技术、汇编码抽象方法以及中间码序列的提取与选择。然后分析中间码序列特征以及特征的提取与选择。最后简单介绍了三种机器学习分类算法。

第四章是系统架构的设计以及实验结果。介绍整个系统的架构设计以及整个系统的核心模块和各项实验结果，然后通过准确率、误报率和漏报率三个指标对实验结果对比分析，并给出结论。

## 第二章 恶意代码检测技术概述

### 2.1 恶意代码简介

#### 2.1.1 恶意代码的定义

恶意代码也成为恶意软件，是对各种敌对和入侵软件的概括性术语。包括各种形式的计算机病毒、蠕虫、特洛伊木马、勒索软件、间谍软件、广告软件以及其他的恶意软件。形式上多种多样，可以是可执行文件、脚本、插件等等。其违背使用者的意愿去执行一些操作，损害用户的利益以达到入侵者不可告人的目的。

#### 2.1.2 恶意代码的分类

根据不同的依据，恶意代码有很多种不同的分类方法，没有一种标准的分法，但是常见的种类有：计算机病毒、蠕虫、特洛伊木马、间谍软件、勒索软件等等。下面对几种恶意代码做简要介绍：

### 2.2 恶意代码的检测技术

### 2.3 本章小结

## 第三章 基于机器学习算法的恶意代码检测技术