**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：

(1)计算机病毒。病毒是早期产生的最主要的恶意代码之一，病毒是能够自我繁殖并寄生在其他程序中的代码，这个被寄存的程序被称为宿主程序，但是病毒不能单独运行，必须通过激活宿主程序并满足一定条件下，病毒就能干扰电脑正常工作，扰乱或破坏己有存储的信息，甚至引起整个系统不能正常工作。一般而言计算机病毒通常由三个单元和一个标志构成：引导模块、感染模块、破坏表现 模块和感染指标。1、引导模块是指将计算机病毒感染的宿主程序设法引导安装到

计算机操作系统中，为以后的感染、破坏两个后期模块提供前期的有效准备，一般而言不同的计算机病毒有不同的引导操作，而且引导操作往往是隐蔽的，不易被用户察觉和发现的。2、感染 模块包括两个部分，一个是用来激活感染功能的判断部分。该模块提供一个感染

的标志，用来判断计算机是否被感染。另一个是执行感染功能部分。这一部分主要的功能就是监控宿主满足条件的时机，并及时的将计算机病毒存入到系统特定的位置。3、破坏表现模块与感染模块一样包括两个部分，一是具有触发破坏表现功能的判断部分。二是具有破坏表现功能的实施部分。计算机病毒一般具有寄生 性、传染性、隐藏性、破坏性、潜伏性等特征。

（2）特洛伊木马。木马分为客户端和服务端，客户端安装在攻击者的主机 上是控制端，服务端安装在受害者的机器上。木马可以使攻击者远程控制受害 者的主机，造成受害者信息丢失等问题。木马有很好的隐蔽性，通过模仿正常 的系统文件命名、与其他程序绑定、进程注入及拦截系统调用的方法伪装自己。 木马也有很好的自启动性和自恢复性。常见木马有远程访问型木马、键盘记录 型木马、密码发送型木马、FTP 型木马以及破坏型木马等。

（3）蠕虫。蠕虫是一种可以独立运行、自我复制及自动传播的恶意程序。它通过网络、共享文件、电子邮件、移动存储设备以及有漏洞的主机等自我复 制和传播。蠕虫的传播速度非常快，根据它的危害性可以简单分为无害型、消 耗型和破坏型。无害型蠕虫感染主机后会产生很多垃圾文件减少系统的可用空 间；消耗型蠕虫感染主机后，发送大量扫描数据包，消耗主机的 CPU 和内存资 源，与此同时增加了网络的负载，降低网络的性能；破坏型蠕虫感染主机后会 删除和破坏程序和文件，有时会泄露一些重要信息。

（4）后门。它是一种运行在目标系统中，能够绕过安全控制机制获得对系 统的访问权，为攻击者提供通道的恶意代码。后门可以使攻击者远程控制目标 主机，危害无穷。后门提供的通道有几种类型：本地权限提升、远程命令行访 问、单命令远程执行、远程控制等。 （5）Rootkit。它是指帮助攻击者获取主机管理权限后，实现维持拥有管 理权限的程序[42]。通常攻击者通过后门获取管理权限，并使用 Rootkit 维持管 理权限使的恶意代码能隐藏在目标系统中。Rootkit 分为用户模式和内核模式。 用户模式通过通道插入恶意代码、覆盖文件、API 钩子和 DLL 注入等方式达到 目的。而内核模式通过安装恶意的设备驱动程序、打补丁、修改内存中运行的 内核以及虚拟伪造系统的方式实现。 （6）间谍软件。它是在未授权的情况下窃取用户的信息并通过网络发送给 攻击者的一种恶意代码。这种恶意代码不仅仅能泄露目标主机的数据信息，还 可以提供恶意代码的植入接口使得被侵系统受到更加严重的破坏。 （7）广告软件。它是指在未经用户授权的情况下和别的程序捆绑在一起， 以便经常弹出一些用户不想接受的广告。这种恶意程序目的是通过这种强制的 方式做商业宣传。一些广告插件的安装会降低主机的性能。广告软件主要的危 害是弹出一些色情或者恶意的广告，这会给用户带来很大的困扰。 （8）恶意网页脚本。它是指在网页中嵌入一些用脚本语言编写的有恶意行 为的代码。当用户点击带恶意脚本的网站后，脚本通过修改目标系统的注册表、 下载病毒或者加载木马程序等方式对被侵系统实施破坏行为。

**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**

**3.1 引言**

本文研究了基于机器学习算法的恶意代码检测技术，通过对汇编操作码不同程度的抽象，以此来勾勒出程序的行为轮廓，并实现对恶意代码的检测。研究方法的主要步骤是：首先，数据预处理，对所有的恶意程序进行脱壳处理并对脱壳后的程序进行反汇编，得到其汇编文本文件；其次，根据上一步骤得到的汇编文本文件提取汇编操作码，并根据映射关系对汇编操作码进行抽象，得到三种不同抽象级别的中间码，记为Level-0、Level-1和Level-2，抽象程度由低到高；然后，分析各个类别恶意代码的中间码序列特征，分别根据2-gram、3-gram、4-gram算法提取中间码序列特征，接着使用数据挖掘技术对各个抽象级别的中间码进行关联分析，可以证明各个类别的频繁项集具有明显区别，所以本文有理由使用概率矩阵作为机器学习的输入；最后，使用随机森林、支持向量机和K近邻算法实现对恶意代码的分类检测。下面将以此对每一个步骤进行详细介绍。

**3.2 数据预处理**

**3.2.1 查壳与脱壳**

壳是指在计算机软件中一段专门保护软件不被非法修改或反编译的程序。它们一般都先于程序运行，拿到控制权，然后完成保护软件的功能。

壳通常分为两类：压缩壳和加密壳。压缩壳出现较早，可追溯到DOS时代，使用压缩壳可以帮助缩减PE文件大小，隐藏PE文件内部代码和资源，便于网络传输和隐藏。压缩壳通常有两种用途，一种是单纯用于压缩普通的PE文件，另一种则会对源文件产生较大的变形，严重破坏PE文件头，通常用于压缩恶意程序。常见的压缩壳有：Upx、ASpack、PECompat。加密壳，也称为保护壳，它主要的功能是防止逆向分析技术，保护PE文件不被逆向分析。加密壳保护的文件通常比PE源文件大得多。常见的加密壳有：ASProtector、Armadillo、 EXECryptor、Themida、VMProtect。

现阶段主要的查壳手段有两种：基于特征码和基于信息熵。

基于特征码：同样的加壳方法会使得加壳后的 PE 文件在特定的位置有相 同的字节序列。相同的序列即为该加壳方式的特征，然而这种基于特征码的查 壳方法，只能检测出已有的并且加入到特征库的加壳方法。

基于信息熵：经过加壳的PE文件，其结构会产生变化。壳一般分为加密 壳和压缩壳，加密和压缩会使 PE 文件变成随机性更大的无结构形式。然而，熵是用来衡量不确定性的，这样经过加壳的PE文件的熵一般会大于源文件PE文件的熵。

目前有很多加壳工具，常用的有ASPACK、PE PACK、UPX、PECOMPACT 等。相应的也有一些查壳软件有 PEID、FILEINFO、FILE SCANNER 等，本文 主要是用 PEID 的批量模式查壳。 脱壳与加壳是相反的过程，为了更好的分析恶意程序的真实意图，首先要 去掉外面的保护壳，找到原程序的真正入口点。脱壳分为手动和自动。手动脱 壳需要脱壳者有很深厚的技术水平，跟踪和分析加壳的方法，然后再对应的去 脱壳。而自动脱壳是运用专门的脱壳工具进行脱壳处理。随着逆向技术的发展，产生了大量的脱壳软件，有的是专门脱某一种壳，也有的能脱几种壳，然而这些软件都是只能脱掉现有的壳，对于一些新的壳还需要手工脱壳。目前常用的脱壳软件有 WASPACK、UNDBPE、UNFSG、EUNPACKER、VMUNPACKER 等。

**3.2.2 反汇编技术**

反汇编指将机器代码转换为汇编代码、低级转高级的意思，常用于逆向工程领域。目前网络上许多“免费软件”，PSP、PS、NDS游戏机的破解和苹果系统的越狱都跟反汇编息息相关。

基本的反汇编算法有两种：线性扫描反汇编和递归下降反汇编。 线性扫描反汇编算法从代码段的第一个字节开始，以线性模式逐条反汇编 每条指令。其优点在于：它能够扫描所有代码段。缺点：对数据和代码混合的 情况没有考虑。 递归下降反汇编算法重视控制流的概念，通过对可执行代码的扫描来获得 比较准确的反汇编结果，根据每条指令的引用来进行反汇编。其优点在于：它 能将代码段与数据段区分开,可以有效的跳过嵌入在程序段中的数据及其无效 的代码。缺点在于：它无法处理间接代码路径。

反汇编工具有很多，如有 IDA Pro、C32Asm、W32DASM、花指令清除器 1.2 等等。IDA Pro 是一款专业的反汇编工具，是由 DataRescue 开发的现己成为了很多 Hacker，Cracker， Reverse engineerer 的必备工具。C32Asm 的特点是集反汇编、 16 进制工具 、Hiew 修改功能于一体。静态分析中文软件的利器的优点是速度 快。花指令工具可以去除花指令，虽然目前效果不是太好，但目前也是对付静态 分析的一个重要手段。

本文使用objdump工具对恶意代码进行反汇编，主要使用objdump -d指令，环境是Windows下进行。

**3.2.3 操作码提取**

本实验使用的数据包含两部分，第一部分是所有的恶意程序反汇编后的文件，文件名为对恶意程序使用MD5哈希算法之后得到的字符串；第二部分是一个label.csv文件，包含两列，第一列为恶意程序的MD5字符串，第二列为该恶意程序所属类别。操作码提取流程如图x:



图x

首先读入label.csv文件，存储文件名和对应类别的映射关系；然后读取所有的恶意代码文件名，接着根据配置文件中的线程数，启用相应多的线程，每个线程处理一部分汇编文件，提取出对应的操作码；接着根据先前读入的文件名和类别映射关系，或者每一个文件所属类别；最后将提取出的操作码文件放置在对应类别的文件夹下。

对于操作码提取的具体过程，本文使用了两个小技巧。第一，使用了配置文件中的Level-0级别的抽象映射关系，由于Level-0级别实际并未对汇编操作码做任何程度的抽象，即就是所有汇编操作码，所以可以用于操作码的提取，并且该映射关系存储类型为unordered\_map，所以查找复杂度近似为,不会带来任何性能上的问题。第二，使用了多线程处理操作码提取，具体线程数可根据实际环境进行动态配置，进一步提高了处理速度。

**3.3 操作码抽象及特征分析**

**3.3.1 操作码抽象化**

将汇编操作码先进行抽象化处理是很多研究常用的手段之一。本文将操作码进行抽象化主要有两点考虑：

1. 汇编语言是面向机器的程序语言，是一种用文字助记符来表示机器指令的符号语言。它是目前所有编程语言中最接近机器码的一种语言。它的针对性特别强，需要对机器硬件进行精确的控制，所以它的每一条指令都是极致细化的，这也导致了汇编操作码的局限性。当使用n-gram算法时，一般来说n的取值不会太大，n条少量的汇编操作码并不能反映出程序行为的一般特征。如果对汇编操作码进行抽象，得到更高维度抽象的中间码，同样是n条中间码，其反映的行为特征将会更加明确及有意义，对于恶意代码的检测是存在积极意义的。
2. 将汇编操作码进行抽象化处理能够显著的提升实验的效率。本文使用概率矩阵作为恶意代码的特征，即机器学习的输入，而x86汇编操作码的种类就有100多种，当使用2-gram、3-gram、4-gram算法时，其特征维度分别是1002、1003、1004，即使后续对概率矩阵进行降维操作也是非常低效的。根据查阅文献以及对大量的汇编操作码进行分析后，本文分别对汇编操作码进行两种级别的抽象，抽象只有中间码的种类分别为9种和5种，其特征维度数的底数将会是9和5，这将显著提升实验效率，并且由实验分析及结果也知，这种抽象操作对恶意代码的检测是有积极意义的。

本文对汇编操作码进行了三种级别的抽象，分别是Level-0，Level-1，Level-2。对于Level-0，表示未作任何程度的抽象，即汇编操作码原本的形式；Level-1和Level-2分别对汇编操作码进行了不同程度的抽象精简。下面分别对Level-1和Level-2进行详细介绍。

Level-1将汇编操作码映射为9类中间码：

1. 数据传输，比如：MOV、PUSH、POP、IN、OUT等；
2. 算术运算，比如：ADD、SUB、CMP、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，伪指令及不常用的指令，比如：SEGMENT，ASSUME，END等。

下表表示出这9种中间码的表示方法：

Level-1抽象中间码的表示

|  |  |
| --- | --- |
| 种类 | 表示方法 |
| 数据传输 | DATA\_TRANS |
| 算术运算 | ALTH\_OPE |
| 逻辑运算 | LOGIC |
| 移位运算 | SHIFT |
| 字符串操作 | STR\_OPE |
| 过程控制 | PRO\_CTRL |
| 标志位控制 | FLAG\_CTRL |
| CPU控制 | CPU\_CTRL |
| 其他 | OTHER |

Level-2将汇编操作码映射为5类中间码：

1）数据传输，比如：MOV，PUSH，IN等等；

2）运算，其中包含算数运算指令(ADD、 SBB等)和逻辑运算指令(OR，XOR，TEST等等)；

3）程序转移，这是指影响程序正常流程的 指令，比如：JMP，CALL，RET，INT等；

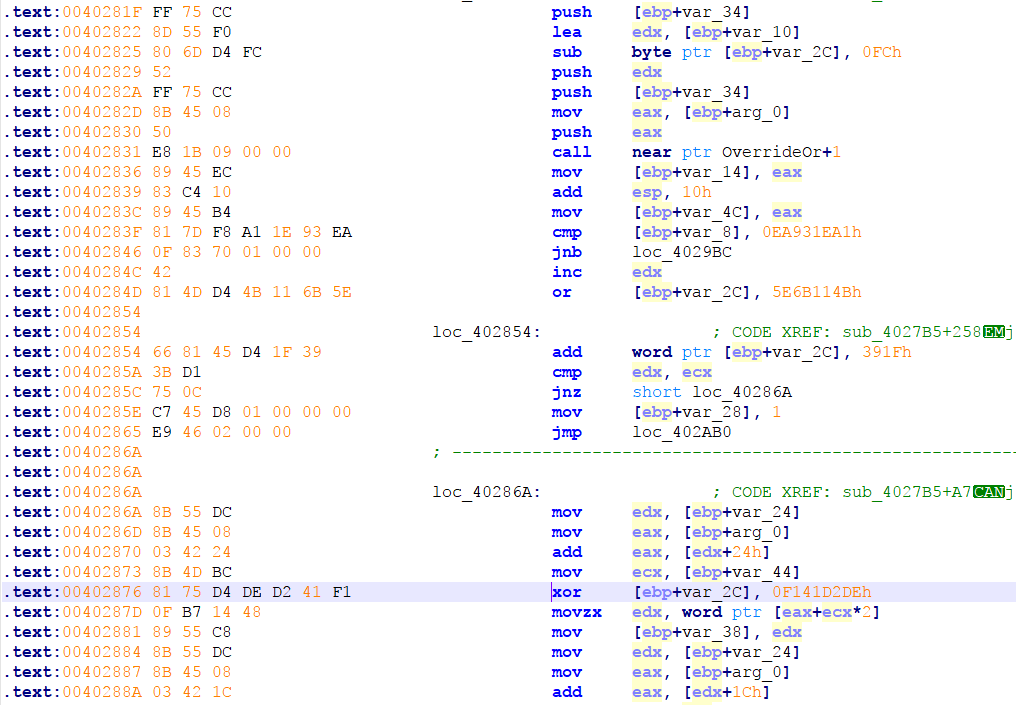
4）处理机控制，比如：NOP，HLT，WAIT 等等；

5）其他，伪指令及不常用的指令，比如：SEGMENT，ASSUME，END等。

下表表示出这5种中间码的表示方法：

|  |  |
| --- | --- |
| 种类 | 表示方法 |
| 数据传输 | ASSIGN |
| 运算 | TEST |
| 程序转移 | JMP |
| 处理机控制 | CTRL |
| 其他 | OTHER |

下图x是截取样本种的一段汇编程序，为节省篇幅，本文分别使用Level-0、Level-1、Level-2对其前15条汇编操作码进行抽象，翻译后的中间码顺序分别表示为下：



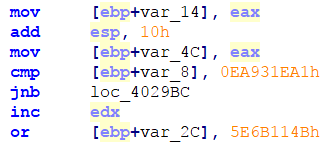
Level-0：push->lea->sub->push->push->mov->push->call->mov -> add->mov->cmp->jnb->inc->or

Level-1：DATA\_TRANS-> DATA\_TRANS-> ALTH\_OPE-> DATA\_TRAN S-> DATA\_TRANS-> DATA\_TRANS-> DATA\_TRANS-> PRO\_CTRL-> DATA\_TRANS-> ALTH\_OPE-> DATA\_TRAN-> ALTH\_OPE-> PRO\_CT RL-> ALTH\_OPE-> LOGIC

Level-2：ASSIGN-> ASSIGN-> TEST-> ASSIGN-> ASSIGN-> ASSIGN -> ASSIGN-> JMP-> ASSIGN-> TEST-> ASSIGN-> TEST-> JMP-> TEST-> TEST

**3.3.2 n-gram算法提取特征**

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。该模型基于第n个词出现只和其前n-1个词相关的假设，对于处理依靠序列关系分类的问题具有很强的优势。本文应用n-gram参考文献）算法的思想，将获得的中间码序列，设置不同的n的大小，获得一系列滑动窗口为n的中间码序列。以下图为例。



定义一个程序为p,操作码序列为o，则可表示程序，其中m表示一个程序操作码序列的长度。这样每一个特征都是它的子序列。对于图x中给出的操作码序列，应用3-gram算法可获得特征为：，，，，，然后根据不同的抽象级别得到其对应的n-gram表示。对于n-gram算法，很难确定滑动窗口n的大小，太小很难检测复杂的模块，太大又很难检测一些简单的混淆技术，本文通过实验对比，选择了特定的长度，n分别选择2、3以及4来提取相应的特征。

**3.3.3 特征分析**

应用n-gram算法分别提取2-gram、3-gram和4-gram特征，并根据不同抽象级别做映射操作，得到最终提取到的特征文本。本节将使用Eclat算法分析各类别特征文本的频繁项集，并最终证明概率矩阵可作为区分不同类别恶意代码的特征。

本文使用Eclat算法来分析特征文本的频繁项集。尽管Apriori算法是最广为人知的关联规则挖掘算法，但是算法需要多遍扫描数据库因而会产生大量的候选项集，支持度的计算也很耗时。相对于Apriori算法，Eclat算法采用了等价类、深度优先遍历、求交集等策略，支持度计算效率有很大改善。Eclat算法采用了垂直数据结构来存储每个事物的ID（tid）列表，在计算k+1-项集的支持度时，Eclat算法只需要计算两个k-项集的tid-表的交集即可。最后，Eclat算法也从频繁项集中生成关联规则。

本文使用频繁项集的平均相似度评估频繁项集对于恶意代码分类的效果，平均相似度越低，代表频繁项集分类效果越明显，反之，代表分类效果差。本文假设频繁项集平均相似度小于50%便可用于恶意代码分类。频繁项集的平均相似度定义如下：

式中分别都是集合，A表示集合的集合， 表示集合A的基数。例如，集合，其中集合频繁项集为，集合频繁项集为 ，集合频繁项集为，那么

本文选取最频繁的10个项集来计算平均相似度，结果如图x：

图x 平均相似度对比图

由图可以看出，任意抽象级别，随着n-gram中n的增大，频繁项集平均相似度越来越低；对于2-gram，随着抽象级别的增大，频繁项集平均相似度增大，3-gram中，抽象级别增大，频繁项集平均相似度减小，4-gram中，抽象级别Level-1的频繁集相似度最大，但是也未超过30%；频繁项集平均相似度最低为23.33%，抽象级别Level-2并且n-gram中n值为4；除了抽象级别Level-2中2-gram的相似度超过50%之外，其他各个类之间的频繁项集平均相似度均未超过50%。由此，本文有理由相信根据频繁项集来对恶意代码进行分类是可行的。

**3.3.4 概率矩阵**

由3.3.3小节可知，不同类别的恶意代码的频繁项集是可区分的，并且不同抽象级别的区分度也是不同的，本文针对每个特征的出现频率构建一个概率矩阵，以此作为机器学习算法的训练集和测试集。

以抽象级别Level-1为例，对图x应用2-gram算法，可得到，统计每个词组出现的次数，然后除以总的词组数6得到各个词组的频率，将该频率作为该段程序的概率分布中的值，表x给出了图x程序片段的概率分布：

表x 2-gram&Level-1概率分布（未填写表示0）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | DATA\_  TRANS | ALTH\_  OPE | LOGIC | SHIFT | STR\_O  PE | PRO\_C  TRL | FLAG\_  CTRL | CPU\_C  TRL | OTHE  R |
| DATA\_TRANS |  | 2/6 |  |  |  |  |  |  |  |
| ALTH\_OPE | 1/6 |  | 1/6 |  |  | 1/6 |  |  |  |
| LOGIC |  |  |  |  |  |  |  |  |  |
| SHIFT |  |  |  |  |  |  |  |  |  |
| STR\_OPE |  |  |  |  |  |  |  |  |  |
| PRO\_CTRL |  | 1/6 |  |  |  |  |  |  |  |
| FLAG\_CTRL |  |  |  |  |  |  |  |  |  |
| CPU\_CTRL |  |  |  |  |  |  |  |  |  |
| OTHER |  |  |  |  |  |  |  |  |  |

然后将这81种组合作为概率矩阵的列，每一个程序的概率分布作为概率矩阵的一行，由此构建本文所诉的概率矩阵。

设集合，表示抽象级别。对于本文，，其中对于，x代表所有的汇编操作码，大约140种；对于，x表示9种中间码中任意一个；对于，x表示5种中间码的任意一个。设，即集合的基数，所以。易知，对于2-gram算法，概率矩阵的列数等于的笛卡儿积的基数，即就是；对于3-gram算法，概率矩阵的列数等于三重笛卡儿积的基数，即就是；对于4-gram算法，概率矩阵的列数等于三重笛卡儿积的基数，即就是。归纳总结可知，对于n-gram算法，概率矩阵的列数就是的n重笛卡儿积的基数。

**3.4 恶意代码分类**

**3.4.1 随机森林**

随机森林是一种集成学习算法，该算法在学习过程中将产生多个决策树，每棵决策树会根据输入数据集产生相应的预测输出，算法采用投票机制选择类别众数作为预测结果。

随机森林算法的目标是通过将多个弱学习机（如单棵决策树）组合得到一个强学习机。算法处理过程与bagging方法非常相似，假设当前拥有N个特征数为M的样例，首先采用bootstrap对数据进行采样，每次随机采样N个样本作为单个决策树的训练数据集。在每个节点，算法首先随机选取m（m<<M）个变量，从它们中间找到能够提供最佳分割效果的预测属性；然后，算法在不剪枝的前提下生成单个决策树；最后从每棵决策树都得到一个分类预测结果。如果是回归分析，算法将所有预测的平均值或者加权平均值作为最后输出，若果是分类问题，则选择类别预测众数作为最终预测。

训练模型的生成流程如图x所示，检测模型流程如图x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

本文使用R语言中的randomForest包完成数据分类，并使用3.3.4小节的概率矩阵作为模型的训练集和测试集。

**3.4.2 SVM**

很多研究已经证明支持向量机（Support Vector Machine，SVM）是一种强大的分类工具，可以被广泛的应用与不同的领域。与随机森林算法不同，在SVM训练中，从输入数据到输出结果的过程并不清晰，也难以解释，因此，SVM属于黑盒算法。

SVM通过定义核函数将输出数据映射到高维特征空间上，并在此空间中构造一个最有分类超平面（或者一组超平面），使得高维特征空间内的各个类的边缘间隔最大化。定义这些超平面的向量就被称为支持向量，如图x中被正方形圈出的o和x：



图x SVM分类示意图

SVM首先构建一个超平面能够最大化间隔距离，然后将定义扩展到非线性可分问题上，最后，将数据映射到一个高维空间，使得数据能够更容易地被线性边界分离开来。简而言之，SVM可以看作高维空间的线性算法。

SVM的优势在于利用了面向工程问题的核函数，能够提供准确率非常高的分类模型，同时借助正则项可以避免模型的过度适应，用户也不必担心诸如局部最优和多重共线性难题。SVM主要弊端是对模型进行训练和测试的速度比较慢，模型处理的时间冗长，因此算法不适合，应用于规模庞大的数据集。本文采用的数据集规模不是很庞大，因此可以使用SVM算法进行实验。另外，SVM的结果很难解释，如何确定核函数也是一个难点。

本文使用R语言中的e1017包的libsvm函数训练支持向量机。通过训练函数svm，用户可以确定核函数、成本函数（即惩罚因子）和gamma函数。对于核函数，通过大量实验，本文使用效果最好的也是默认的radial（径向函数）作为核函数；对于gamma和惩罚因子，通过tune.svm来寻找最佳的gamma和惩罚因子。本文中，将gamma参数的可能范围设置为，惩罚因子选择10和100，使用tune.svm函数得到12组不同的参数组合。函数采用10遍交叉检验的方法获得每次组合的错误偏差，最后选择误差最低的最优参数组合。

**3.4.3 KNN**

K近邻（K-Nearest Neighbor，KNN）算法属于一种无参惰性学习方法学习。无参类算法不会对数据的分布做任何的假设，而惰性学习方法则不要求算法具备显性学习过程。

KNN算法采用相似性（距离）测量来训练所有样本并对新的样本进行分类标记。常用的相似性测量如下公式：  
欧氏距离：

曼哈顿距离：

KNN算法中，新到达的样本将被分配到某个在KNN中普通的类中。如果K=1，新样本将会被分配到距离其最近的类别中。算法仅要求用户输入K的值，如果K值比较小，有可能产生过度适应的问题。相反，如果给的K值过大，则有可能导致低度拟合。可以通过交叉验证的方式得到比较合适的K值。

KNN算法的优势：

1. 学习成本为0；
2. 无参意味着用户不需要对数据的分布类型做假设；
3. 只要对给定样本选取和式的距离测量方法，算法就可以处理任意类型的数据。

KNN算法的不足：

1. 算法难以理解；
2. 如果数据集比较大，算法计算代价非常高；
3. 算法性能依赖数据集维度的大小，如果要处理高维数据，应先对数据进行降维操作以提高算法的过程性能。

针对KNN算法不足2)，本文用于实验的数据，不会大到KNN处理不了的情况；对于不足3），本文事先对操作码做了抽象化处理，除了Level-0的3-gram和4-gram维度较大，其他的维度都维持在可接受的范围。

本文使用R语言class包中的knn函数来处理实验数据。

**3.5 本章小结**

本章主要介绍了基于机器学习算法的恶意代码检测技术。重点介绍了操作码抽象及特征分析的过程：首先，对数据查壳和脱壳处理，并对样本进行反汇编并得到汇编操作码序列；然后，对汇编操作码进行3种级别的抽象化，分别为Level-0、Level-1和Level-2，得到其对应的中间码；接着，分别对中间码进行n-gram算法提取特征，应用数据挖掘基数对提取到的n-gram词组进行统计分析，得到各个类别的频繁项集存在差异，因此本文使用概率矩阵作为机器学习算法的输入；最后，介绍了三种机器学习算法，以及本文如何应用这三种机器学习算法。

**第四章 系统设计与实验结果**

**4.1 系统设计**

**4.2 评价方法**

为了确保模型能对未知或新到达的对象进行正确预测，需要对模型性能进行评估，避免模型可能存在过拟合问题。

k折交叉验证技术能够解决过拟合问题，因此被广泛应用于分类器性能评估领域。k折交叉验证并不需要使用整个数据集，相反，它会将数据集划分为训练集和测试集两部分。这样，基于训练集得到的模型就可以通过测试集来完成性能测评。重复执行n次k折交叉验证后，就能够根据n次检验的平均正确率实现对模型的真实评估。

R语言的e1071包提供的tune函数，该函数提供对随机森林算法、SVM和KNN算法进行k折交叉验证。本文使用tune.randomForest、tune.svm和tune.knn分别对随机森林算法、SVM和KNN算法进行10折交叉验证。

**4.3 实验数据**

本文使用的实验数据来自Kaggle上微软发起的一个恶意代码分类的比赛，使用的数据集的大小为136GB，其中样本种类分布如表x，

表x 恶意程序数据集

|  |  |
| --- | --- |
| 恶意程序家族 | 数量（单位，个） |
| Ramnit | 1541 |
| Lollipop | 2478 |
| Kelihos\_ver3 | 2942 |
| Vundo | 475 |
| Simda | 42 |
| Tracur | 751 |
| Kelihos\_ver1 | 398 |
| Obfuscator.ACY | 1228 |
| Gatak | 1013 |
| total | 10868 |

**4.4 实验结果对比**

**4.4.1 准确率比较**

**4.4.2 误报率比较**

**4.4.3 漏报率比较**

**4.4.4 实验结果分析**

**4.5 本章小结**