基于关联规则的未知病毒检测方法研究

赖英旭[1], 刘增辉[2]
LAI Ying-xu[1], LIU Zeng-hui[2]

1.北京工业大学 计算机学院, 北京 100022
2.北京电子科技职业学院 工程技术系, 北京 100029
1.College of Computer Science, Beijing University of Technology, Beijing 100022, China
2.Department of Engineering Technology, Beijing Vocational College of Electronic Science, Beijing 100029, China
E-mail: laiyingxu@bjut.edu.cn

LAI Ying-xu, LIU Zeng-hui.Research of unknown virus detection based on association rules.Computer Engineering and Applications, 2008, 44(7): 133-135.

Abstract: With the development of computer science, more and more computer viruses come out, which seriously compromised the security of the computer world.Current virus scanner does not generalize well to detect unknown viruses.The paper promotes an unknown virus detection technology based on association rules method and explores the extraction of features.The paper also gives an unknown virus detection framework.The evaluations and results are given in this paper.The test shows that the data mining method has the advantage on the unknown virus detection by association rules.
Key words: detection of unknown viruses; extraction of feature; association rules

摘 要: 随着计算机技术的发展, 计算机病毒也层出不穷, 严重地危害了计算机世界的安全, 当前的病毒检测技术对未知病毒还很难做到事先检测。关联规则挖掘是数据挖掘领域中的重要技术, 经研究发现, 基于关联规则的未知病毒检测技术, 可以实现对未知病毒的分类检测。实验结果表明, 采用关联规则构建的未知病毒检测模型, 能较好地实现未知病毒检测, 具有自适应能力强、智能性好、自动化程度较高等优点, 具有一定的应用价值。
关键词: 未知病毒检测; 特征提取; 关联规则
文章编号: 1002-8331(2008)07-0133-03 **文献标识码**: A **中图分类号**: TP309.5

1 引言

计算机病毒检测技术中, 无论是基于特征检测的, 还是基于行为监测的, 其人为参与因素都非常多。防病毒专家通常关注已知的攻击行为特征和病毒特征对其进行分析研究, 造成检测工具对更多的未知病毒缺乏适应能力。因此如何建立具有较强的有效性、自适应性、可扩展性的检测工具成为未知病毒检测的重要研究课题。

数据挖掘是从大量数据中提取或'挖掘'出知识。具体说是对数据进行处理, 从而获得隐含的、事先未知的、潜在的而又非常有用的知识, 这些知识可表示为模式。数据挖掘方法有多种, 其中比较常见的有关联规则、序列模式、数据分类、聚类分析等[1,2]。利用数据挖掘在有效利用信息方面的优势, 将病毒特征视为一类数据进行分析, 能够从大量的数据中自动产生精确的适用的检测模型, 使检测系统适用于未知病毒检测。

数据挖掘技术的优点在于[3,4]:

(1) 可以处理大规模的数据量;

(2) 不需要用户提供主观的评价信息, 善于发现容易被主观忽视和隐藏的信息。

计算机病毒有着其独有的特点: 变化性强、影响因素不确定、各影响因素之间的关系复杂, 这使得关联规则发现方法相对较适合这一问题。

2 基于关联规则技术的未知病毒检测模型

关联规则(Association Rules)挖掘是数据挖掘(Data Mining)领域中的重要技术, 是从海量数据库中进行知识发现(Knowledge Discovery in Database)的有效方法, 在商业、金融、科学研究、电子商务以及电子政务等领域都有很好的应用前景。目前, 针对这方面的算法主要有: Apriori、Han and Fu、基于频繁项集的FP-树、基于粗集的PC-树以及Apriori的改进算法[5,6]。

本研究采用数据挖掘技术中的关联规则挖掘方法, 选择FP-tree算法作为频繁项目集生成算法, 得到形如(支持度, 可信度)的规则, 支持度表示同时满足规则前件和规则后件的例数占总例数的比例即概率, 可信度表示在所有满足规则前件的例数中满足规则后件所占的比例即条件概率。

未知病毒检测模型的主要组成部分如图1。

(1) 特征提取模块: 收集已知病毒文件样本和正常文件样

基金项目: 北京工业大学青年基金(No.97007011200603)。
作者简介: 赖英旭(1973-), 博士, 副教授, 主要研究方向: 网络接入控制、病毒防御。

收稿日期: 2007-06-20 修回日期: 2007-09-17

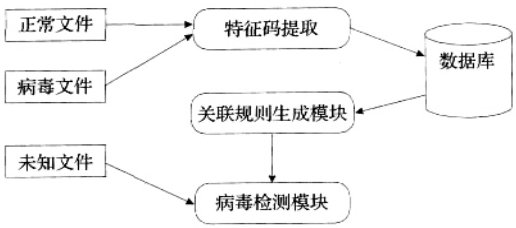

图 1 总体结构图

本的特征;

(2) 关联规则生成模块:用关联规则算法对历史信息和数据集进行特征提取,形成新的模式,更新到数据库中。这个模块是模型的核心部分,挖掘算法的选择和实现直接影响模型的检测效率和误报率;

(3) 数据库:存储挖掘好的规则;

(4) 检测模块:提取新的未知文件的特征,跟知识库中的知识特征进行比较,更新知识库,对符合规则的未知病毒进行报警。

该检测模型的工作原理是:特征提取模块负责采集正常文件和病毒文件的特征,然后进行数据预处理,形成统计数据。关联规则模块对统计数据进行挖掘,提取出特征和模式,保存到数据库中。检测模块负责从数据库中读取规则来进行未知病毒判定,同时对更新的规则进行入库管理。

这种基于数据挖掘的未知病毒检测系统有以下几个优点:

(1) 自适应能力强:由于不是基于预定义的数据库,因此自适应能力强,可以检测出一些新的变种病毒;

(2) 智能性好,自动化程度比较高:数据挖掘方法能够自动从数据中提取特征模式,减轻了人员的工作负担,提高了检测的准确性;

(3) 检测效率高:数据挖掘方法可以自动对数据进行预处理,减少了数据处理量,提高了检测效率。

3 未知病毒检测模型实现

3.1 静态文件内字符串的提取方法

由于要提取静态文件中的字符串,其中包含有大量的 16 进制的乱码和空格,就像'HT@ 內 □? ",因此用通常的字符串提取函数是不行的,如 fscanf、fread 等函数,它们遇到空格就停止或需要限定读取字符串的长度。

为解决上述问题,本文采用的方法是从文件里一个字符一个字符地读取,并判断每个字符是否是有意义的字符,如果是认为有意义的字符,像英文字母、数字或常用符号等[7],就把该字符存入一个字符数组中,该数组用来存放提取的当前字符串;否则就认为当前一个字符串已读完,截断字符数组之后,就把该字符串存入数据库中。

这种方法对于那些加了壳或者加了密的程序,脱去壳后或解密后该方法仍旧适用,因为该方法不是依靠读取文件导入节的数据来获取特征码的。但它有一个缺点,那就是会提取出大量的无意义字符串,为后面的自学习过程带来了很大的麻烦。

3.2 无意义字符串的过滤

由于上述的提取字符串方法会提取出大量的无意义字符串,如'aaabbb'","'dfu ekl'等等,这些字符串在文件中大量存在,但是对于特征码的分析是毫无用处的,而且会严重影响到后面要进行的自学习过程的结果,因此需要将这些字符串过滤掉。

过滤方法:在数据库中建立一张单词表,表中记录了所有常用的单词和词根,当运行特征码提取程序时,先把该表中的所有记录都读入内存中,之后每提取出一个字符串,就检查这些单词或词根是否被包含在该字符串中,如果有一个单词或词根在字符串中出现,那就认为该字符串是有意义的,才能把它存入数据库里,否则就认为该字符串是无意义的,将其过滤掉。

单词表设计:

(1) 加入 400 多个常见的单词词根和 5 000 多个常见单词。

(2) 统一改成小写字符串,为防止后面的处理因大小写的不同导致误判 2 个意义相同的字符串。

(3) 把其中包含以前单词的单词删掉。如,red 和 reduce,就要把 reduce 去掉。

(4) 加入".dll"字符串,以便避免动态链接库名称字符串遗漏。

过滤无意义字符串流程如图 2 所示。

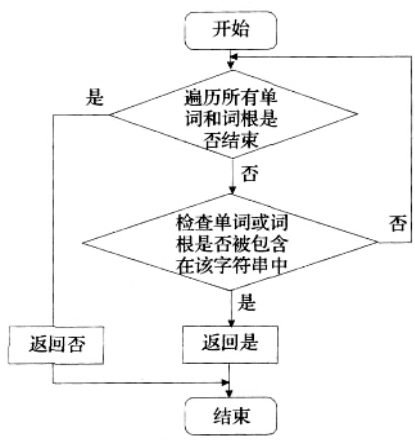

图 2 过滤无意义字符串流程

3.3 基于关联规则的未知病毒检测方法实现

如何通过已知的关联规则来判别一个文件是否是病毒,是一个非常关键的问题。现在这方面的研究总体上来讲也是没有一个公认的方法,本文通过大量的实验和反复尝试,总结了一个相对可行的办法。

该方法步骤如下:

步骤 1 通过频繁项目集生成程序,获取已知病毒样本特征码的频繁项目集和正常文件样本特征码的频繁项目集。

步骤 2 以步骤 1 获得的正常文件的项目集为参照,将病毒项目集中的与其相同集合删去。这样做可以过滤掉正常文件与病毒之间的共有的关联规则,使分类效果更加明显。

步骤 3 根据步骤 2 得到的频繁项目集生成强规则。

步骤 4 对未知文件进行检测时,先把一个文件的所有字符串放在内存中,之后从强规则的文件中读入一条强规则,看该条强规则的前项是否出现在该文件字符串中。如果不出现,则不匹配规则条数加 1,提取下一条规则继续匹配。如果都不匹配,则报该文件不是本类文件。如果前项匹配,则查找后项是否存在。如果后项存在,正常规则条数加 1,则继续提取下一条规则。如果后项不存在,异常规则条数加 1,继续提取下一条规则。最终直到所有的强规则检查完。

步骤 5 统计该文件的正常规则条数、异常规则条数和不匹配规则条数,如果正常规则条数大于 2 且异常规则条数为 0,则把该文件判为是病毒;否则如果异常规则条数大于 0,异常规则条数的百分比小于 5%而且正常规则条数多于异常规则条数 2 倍,那么也把该文件判为是病毒,否则就认为该文件

是正常文件。

未知病毒检测流程如图 3 所示。

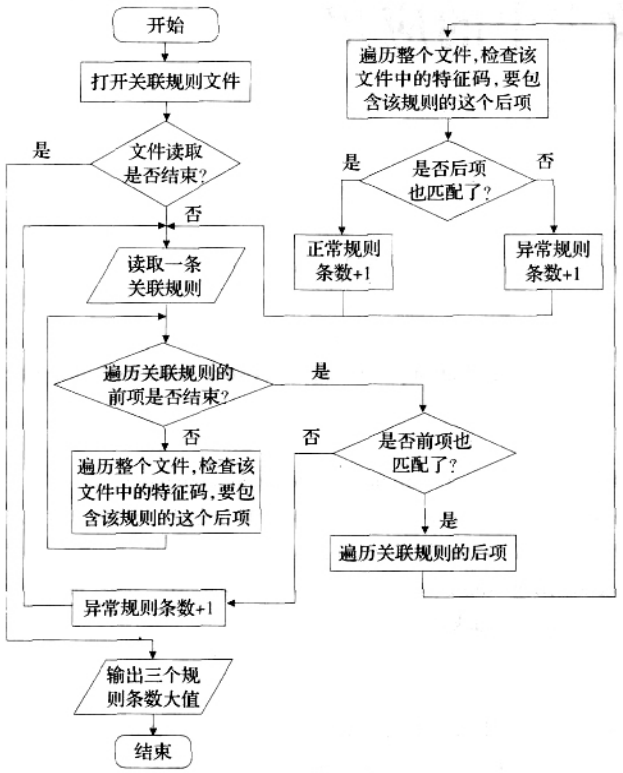

图 3 未知病毒检测程序流程

4 测试及结果

下面是对未知病毒检测程序的测试,通过反复测试与实验,可以检验该程序的执行效果。

(1) 学习样本

学习样本分为正常文件学习样本和病毒学习文件样本两类:①正常文件学习样本:江民公司服务器上 C:\WINDOWS\system32 目录里的 2 690 个 PE 格式文件中的 1 345 个;②病毒文件学习样本:江民公司提供的 334 个病毒文件。

(2) 测试样本

测试样本:江民公司服务器上 C:\WINDOWS\system 目录中其余 1 345 个 PE 文件和江民公司提供的另外 334 个病毒文件。

(3) 测试环境

①测试机:江民公司提供的服务器;②操作系统:Windows 2000;③开发环境:Visual C++ 6.0。

由于不同最小支持度和最小置信度的设定,会产生不同的测试结果,因此本文进行了多组测试,以便对它们进行对比。

在频繁项目集项数最小为 2 的条件下,不同的最小支持度与最小置信度设定对测试结果的影响见表 1。

在频繁项目集项数最小为 6 的条件下,不同的最小支持度与最小置信度设定对测试结果的影响见表 2。

上面两次实验选取的频繁项目集的项数不同,第一次项数设为了 2,因为至少 2-项集才能生成关联规则,这样是由所有的 2 项以上的频繁项目集来生成关联规则;第二次项数设为了 6,与第一次结果进行对比表明,增加关联规则的项数不能明显提高正确率。

在实验过程中,发现设定的最小支持度和最小信任度对检测效果起到了非常关键的作用,观察在最小支持度和最小信任度设定不同的条件下所得的检测效果后,不难看出,在设定较低的最小支持度的基础上,较高的最小信任度会使分类效果更加明显。分析其原因,不同种类的病毒都有其自己的特征,但由于它们都是 Windows 下的二进制可执行文件,PE 格式的文件,因此它们又都与正常文件有着许多的共同特征。较低的最小支持度正好使得每一条强关联规则不必非要在整个事务集中所占的百分比很大,这样设定是适应了病毒的独有特征,可使其更多的保留下来;较高的最小信任度使得得出的强关联规则更加具有可信性,非常有利于分类的判别。由此本文提出的低支高信的参数设定规律是适应通过关联规则检测未知病毒这个方法的。

表 1 最小项数为 2 的测试结果

S-C/%	漏报率/%	误报率/%	正确率/%
50~60	6.1	28.9	65.0
40~70	6.5	27.6	65.9
35~80	5.3	21.7	73.0
30~85	5.1	20.3	74.6
25~90	6.7	14.6	78.7
22~90	4.9	11.2	83.9
20~90	4.1	8.3	87.6

表 2 最小项数为 6 测试结果

S-C/%	漏报率/%	误报率/%	正确率/%
22~90	10.5	9.2	80.3
20~91	9.3	7.6	83.1

5 结论

对于现在的反病毒软件大多都是采用定期更新病毒库的方式来应对新产生的病毒,这种方式使得用户需要随时更新反病毒软件的病毒库以保证反病毒软件的查毒能力,一旦一段时间没有进行病毒库的更新,就将使此客户机上的反病毒软件减小甚至是失去它的防护能力。病毒库的存储方式也是反病毒技术中的一大重点,随着计算机病毒数量的不断增加,病毒库的大小有增无减,最终将成为一个庞大的数据库,并且这样一个数据库需要每一个用户都拥有一个副本。所以病毒库的存储组织形式直接影响着反病毒软件的大小。虽然现在的计算机存储介质越来越大,但是,用户还是希望有一个小巧的高性能的反病毒软件为其保驾护航。

由实验结果可以看出,本文介绍的检测未知病毒的方法是可行的,它能够通过对以往的病毒文件和正常文件的学习,提取出它们的特征码,找出其中的关联规则,进而实现对病毒的检测。

参考文献:

[1] Han Jiawei, Kamber M.数据挖掘概念与技术[M].范明,孟小峰,译.北京:北京机械工业出版社,2005.

[2] 李雄飞,范淼淼,董立岩,等.多段支持度数据挖掘算法研究[J].计算机学报,2001,24(6):661-665.

[3] 罗可,蔡碧野,卜胜贤,等.数据挖掘及其发展研究[J].计算机工程与应用,2002,38(14):182-184.

[4] 程续华,施鹏飞.快速多层次关联规则的挖掘[J].计算机学报,1998,21(11):1037-1041.

[5] Edelstein H.浅说数据挖掘[J].计算机系统应用,1998(4):56-57.

[6] 何飞,罗三定,沙莎.居于领域本体的知识关联研究[J].湖南城市学院学报,2005,14(1):69-71.

[7] Kaspersky K.黑客反汇编揭秘[M].北京:电子工业出版社,2005.