论文阅读

11.A state-of-the-art survey of malware

简介：恶意软件分析主要分为动态分析和静态分析。新的文献综述可以影响研究的发展以及探索数据挖掘在恶意软件检测上的一些细节。但是以往的调查研究有一些缺陷：例如有些文章是过时发表的，没有与新文章做过比较与分析。一些调查没有对检测系统进行分类以及文章选择的研究。而本文使用了数据挖掘方法最新的恶意软件检测技术进行系统的文献综述，并且将恶意软件检测方法分为两类：基于签名和基于行为的。本文贡献分为：总结了数据挖掘中恶意检测方法当前的挑战；对方法中机器学习的机制进行系统和分类得概述；探索恶意软件检测方法中重要方法的结构；讨论分类恶意软件的重要因素。

恶意检测方法：

机器学习策略分为有监督类和无监督类。恶意软件检测方法分为基于行为的方法和基于签名的方法。此外查找恶意程序时常用静态和动态分析两种方法。API调用特征，汇编特征以及二进制特征是是现存恶意检测常用

基于签名的恶意软件检测：

能够根据预先定义的预知攻击列表来识别恶意软件。其优点是快速识别，易于运用以及可广泛访问。但是需要依赖数据库进行检索，并且这种方法是建立在已知恶意软件上的，无法识别恶意软件的变体，无法提供零日保险。

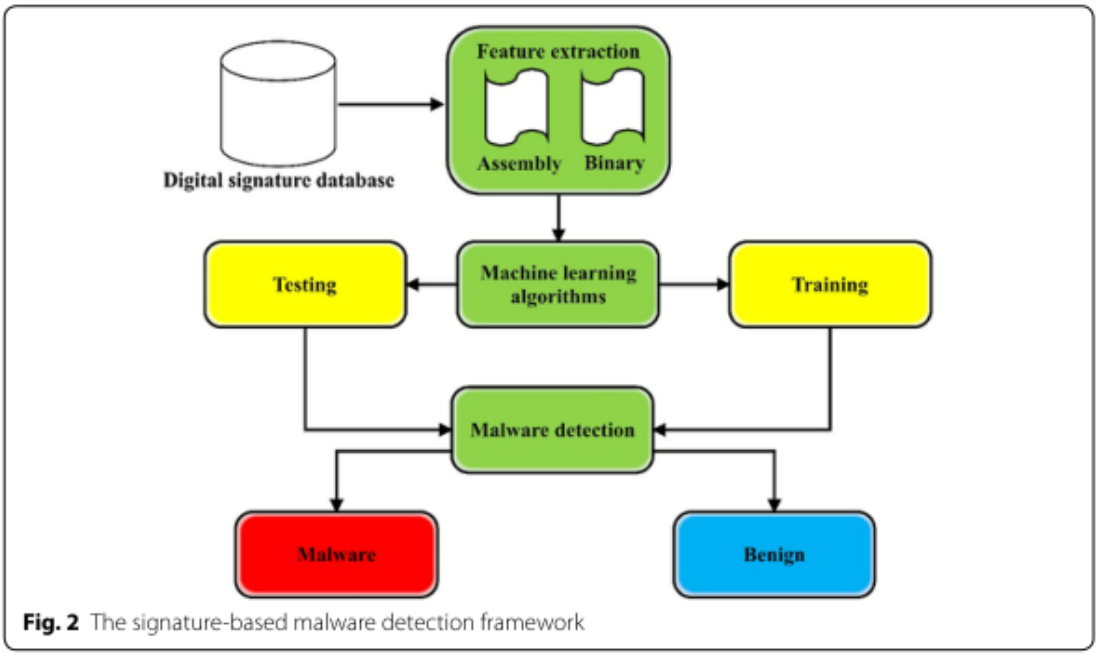

Fig. 2 The signature-based malware detection framework

对于基于特征的恶意软件检测有两种主要的方法用于在机器学习方法中应用恶意软件检测方法，包括汇编特征和二进制特征。

基于行为的恶意软件检测：

基于行为的方法是在沙箱中执行样本，并检查和记录运行时行为。在该方法中，基于可疑对象在系统中无法执行的活动来评估样本。使用动态分析可以识别恶意行为。在基于行为的检测中，应用编程接口调用和汇编特性是应用机器学习算法的两种主要方法

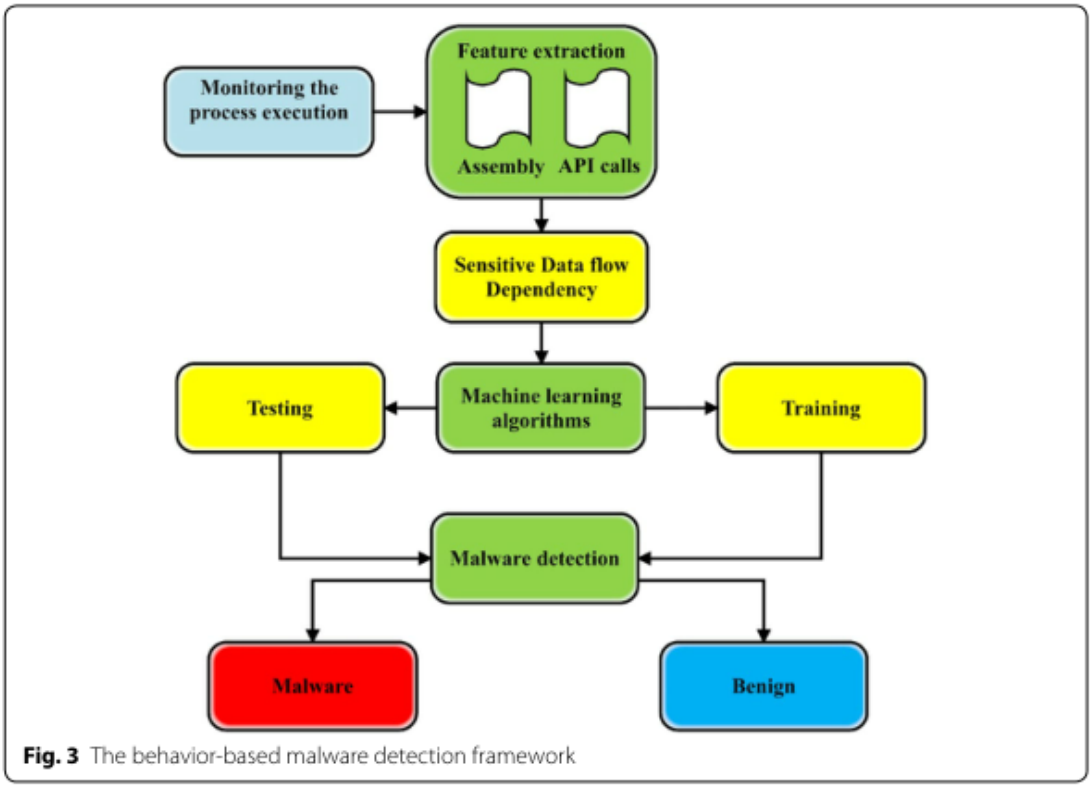

Fig. 3 The behavior-based malware detection framework

优点：可以检测到没有发现过的恶意软件攻击；基于数据流的检测器；可以检测多态恶意软件

缺点：行为模式的存储复杂性；

回顾恶意软件检测技术（思想，优缺点，算法类型，评估类型）

基于签名：Wu等人创建了基于人体免疫系统的手机恶意软件检测模型

Bat-Erdene等人提出了一种策略来描述给定的未知打包可执行文件的打包算法

Cui等人在云环境和包检查的基础上提出了一种新的识别框架。该框架利用信息挖掘策略，通过恶意移动恶意软件的捆绑来识别恶意移动恶意软件行为

Fan等人[32]提出了一种强制排列挖掘计算方法来寻找有报复性的五次样本，然后针对已建立样本中的恶意位置构造了All-近邻(ANN)分类器

Hellal和Ben Romdhane [33]展示了另一种图表挖掘技术，利用静态检查来识别恶意软件的变体，同时覆盖当前的缺陷

Martín等人[34]说明了外部调用来回避这些伪装方法的影响，因为它们不能被混淆。我们结合聚类和多目标推进来产生一个以外部调用聚类为特征的分类器

Santos等人提出了另一种策略来识别模糊的恶意软件家族。这个模型依赖于操作码分组的重复出现。此外，他们描述了一个系统来挖掘每个操作码的重要性和评估每个操作码分组的递归性。此外，他们提供了实验证明，这种新策略适合于识别模糊的恶意软件

Wang和Wang[24]提出了一个恶意软件识别框架，利用支持向量模型(svm)的推测能力，通过机器学习来确保少量的顺序错误。本文通过准备一个基于行为标记的SVM分类器，构建了一个程序化的恶意软件定位框架。通过审批，计划利用与60组真正的恶意软件连接的支持向量机来解决分组的准确性问题。

总结：基于签名的检测方法的主要优点是使用模式检测，减少了恶意软件预测的系统开销和执行时间。基于签名的检测方法的主要缺点是忽略了特征选择。在基于签名的检测中，大多数研究使用基于windows的环境来表示所提出的恶意软件检测方法。

回顾基于行为的恶意软件技术

Altaher [38]提出了一种进化的混合神经模糊分类器(EHNFC)，用于利用基于同意的组件对基于安卓的恶意软件进行分组。所提出的EHNFC不仅具有利用模糊原则来区分模糊恶意软件的能力，而且同样可以通过采用新的恶意软件识别模糊原则来改进其结构，以在被用作更多恶意软件应用的位置的一部分时增强其发现的准确性。

Mohaisen等人提出了一种计算机化的、基于恶意软件检测和标记的框架AMAL

Yuan等人[40]提出了一种深度学习方法，将来自静态调查的组件与来自安卓应用程序动态调查的元素联系起来。此外，他们还实现了一个基于深度学习方法的安卓恶意软件检测引擎，从而可以区分文件是否有恶意行为

Boukhtouta等人[41]提出了以识别和安排为最终目标的活动的指纹危害性问题。本研究首先利用两种方法指出指纹的危害性:深度包检测和IP包头排列。为此，我们认为由元素恶意软件检测产生的恶性活动是运动危害性的基础事实。根据这一假设，他们展示了如何利用这两种方法来识别更有甚者，将恶意归因于各种威胁。

Ding等人提出了一种基于API调用的关联挖掘策略来识别恶意软件。

Eskandari等人提出了一种新的混合方法，HDM-Analyzer，该方法采用了动态和静态调查技术的兴趣点，以提高速度，同时在合理的水平上保护精度

苗等[44]根据动态API序列的语义检查，提出了双层传导反射策略。

Ming等人[45]提出了一种替代攻击，通过损害基于行为的规范来覆盖类似的实践。攻击的关键策略是将系统调用依赖图替换为其语义相同的变体，从而使可比较的恶意软件测试机密的唯一系列最终成为特征

Nikolopoulos和Polenakis[46]提出了一个基于图的模型，该模型利用系统调用集合之间的关系，区分一个未知的软件样本是恶意的还是良性的，并将一个恶意软件归类到一组已知恶意软件家族中的一个。

Sheen等人已经考虑使用基于android的恶意软件进行检查，并计划使用一个适应性识别组件来利用多特征协同决策融合(MCDF)。

Norouzi等人提出了独特的分类技术，根据每个恶意软件的元素和行为，有一个特定的最终目标来识别恶意软件。为了识别恶意软件的特征，提出了一种动态调查技术。引入了一个推荐的程序，用于将恶意软件行为执行历史XML文档更改为适当的WEKA工具输入。为了体现执行能力和准备信息与测试，作者应用所提出的方法，利用WEKA工具处理真实的上下文调查信息集

Galal等人提出了一种基于行为的特征模型，该模型定义了恶意软件实例所展示的恶意行为。为了消除提出的模型，作者首先在一个受控的虚拟环境中对一个通常迟来的恶意软件数据集进行动态检查，并捕获恶意软件示例所召唤的API调用的踪迹。然后，跟踪被概括为高级特性，称为操作。利用随机森林、决策树和支持向量机等著名的分类方法对该方法进行了评价。实验结果表明，该分类器在恶意软件变体检测方面具有较高的准确率和较好的检测效果。

总结：基于行为的检测方法的主要优点是通过调用行为检测所有可疑文件，从而提高恶意软件预测的准确性。基于签名的检测方法的主要缺点是运行时开销。目标环境分为三个主要平台，包括嵌入式系统、基于windows和智能手机。大多数基于行为检测的研究都使用智能手机环境来代表所提出的恶意软件检测方法。

未来的挑战

- 对于信息隐藏恶意软件的检测
- 元启发式检测:使用元启发式算法的恶意软件检测分析可以影响数据挖掘过程的执行速度和总精度因子。
- 实时恶意软件检测:是基于混合分析，安全的多目标进化恶意软件检测，安全的电子银行环境和安全的医疗系统是非常具有挑战性的识别恶意文件和隐藏的攻击使用数据挖掘方法。

技术建议

- 在电子钱包应用中，一些改进的方法可以提高恶意软件检测能力，预测多态攻击。例如，在电子移动支付中，元启发式算法为多态恶意软件攻击找到了最优的签名检测方法
- 上下文感知检测是一种基于语义签名的物联网应用动态恶意软件检测方法的新思路，该方法根据最终用户和物联网应用层之间的交互对API调用进行分类。当智能设备无法在用户设备和数据中心之间进行交互时，智能服务的可靠性和可用性会降低
- 针对恶意软件的攻击，为大数据等海量数据的采集提供安全的环境是恶意软件检测导航大数据安全的关键挑战。因此，为了选择恶意软件破坏的最小样本空间，可以使用数据挖掘和合成的方法对数据采集和存储大数据进行导航。

结论

大多数关于数据挖掘的文章都是基于行为的技术。在恶意软件分析阶段，大多数案例研究是针对android智能手机提出的。此外，在恶意软件检测分析中使用元启发式算法可以加快和提高数据挖掘过程的执行时间和整体准确性。通过数据挖掘发现，SVM方法在基于签名的恶意软件检测方法中具有最好的准确率。此外，DPIM方法的最大精度百分比为99.2%，DMDAM方法的最小精度百分比为86%。此外，我们观察到，最近的研究认为android智能手机分析恶意软件检测方法有40%。基于windows平台的符号代码聚合案例研究占23%，模式挖掘占11%，系统调用占8%。最后，我们看到30%的基于签名的方法使用了动态数据分析。65%的基于行为的恶意软件检测方法使用了动态数据分析方法。作为一个重要的开放性问题，安全多目标恶意软件、电子银行环境和医疗系统恶意软件攻击等都是恶意文件识别和隐藏攻击的挑战领域