(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116192448 A
(43)申请公布日 2023.05.30

(21)申请号 202211649321.5

(22)申请日 2022.12.21

(71)申请人 安天科技集团股份有限公司
地址 150028 黑龙江省哈尔滨市高新技术产业开发区科技创新城创新创业广场7号楼(世坤路838号)

(72)发明人 毕光耀 康学斌 肖新光

(74)专利代理机构 北京科衡知识产权代理有限公司 11928
专利代理师 王淑静

(51)Int.Cl.
H04L 9/40(2022.01)

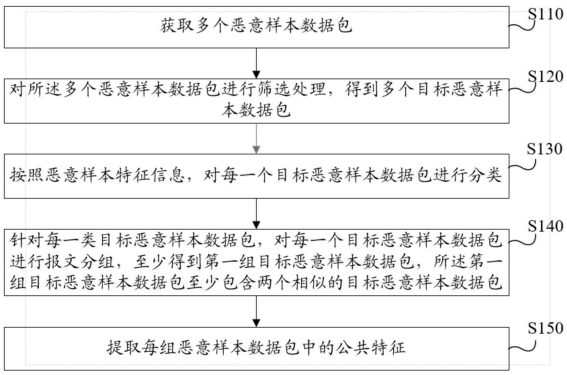
权利要求书3页 说明书10页 附图4页

(54)发明名称

一种恶意样本数据包分析方法、装置及电子设备

(57)摘要

本发明实施例公开一种恶意样本数据包分析方法、装置及电子设备，涉及网络安全技术领域。所述方法包括：获取多个恶意样本数据包；对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样本数据包；按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类；针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包；提取每组恶意样本数据包中的公共特征。本发明通过上述方法步骤，便于提取出未知协议的恶意样本中的公共特征，从而可以提升识别发现未知协议的恶意代码威胁的能力。



CN 116192448 A

1.一种恶意样本数据包分析方法,其特征在于,包括步骤:

获取多个恶意样本数据包;

对所述多个恶意样本数据包进行筛选处理,得到多个目标恶意样本数据包;

按照恶意样本特征信息,对每一个目标恶意样本数据包进行分类;

针对每一类目标恶意样本数据包,对每一个目标恶意样本数据包进行报文分组,至少得到第一组目标恶意样本数据包,所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包;

提取每组恶意样本数据包中的公共特征。

2.根据权利要求1所述的恶意样本数据包分析方法,其特征在于,所述公共特征包括:不变字段域、可变字段域和/或常量域。

3.根据权利要求1所述的恶意样本数据包分析方法,其特征在于,所述恶意样本特征信息包括:恶意样本所属的病毒家族;

所述按照恶意样本特征信息,对每一个目标恶意样本数据包进行分类,包括:对每一个目标恶意样本数据包,按照恶意样本所属的病毒家族分成相应的多个类别,每个类别中包含多个目标恶意样本数据;

所述在对每一个目标恶意样本数据包进行报文分组之前,所述方法还包括:针对每一类目标恶意样本数据包,对每一个目标恶意样本数据包进行报文分类,确定所述目标恶意样本数据包的报文类型;

所述对每一个目标恶意样本数据包进行报文分类,确定所述目标恶意样本数据包的报文类型,包括:针对每一个病毒家族类别,获取该类别包含的多个恶意样本上线数据包;所述目标恶意样本数据包包含:恶意样本上线数据包;

对于每一个恶意样本上线数据包,对所述恶意样本上线数据包的字节类型进行提取;所述字节类型包括:字符型和/或二进制型;

判断所述字符型的字节数量是否超过预定阈值;

若是,则将该恶意样本上线数据包确定为文本类报文;

若否,则将该恶意样本上线数据包确定为二进制类报文。

4.根据权利要求3所述的恶意样本数据包分析方法,其特征在于,所述对每一个目标恶意样本数据包进行报文分类,确定所述目标恶意样本数据包的报文类型,还包括:针对每一个病毒家族类别,获取该类别包含的多个恶意样本指令数据包;所述目标恶意样本数据包包含:恶意样本指令数据包;

对于每一个恶意样本指令数据包,对所述恶意样本指令数据包的字节类型进行提取;所述字节类型包括:字符型和/或二进制型;

判断所述字符型的字节数量是否超过预定阈值;

若是,则将该恶意样本指令数据包确定为文本类报文;

若否,则将该恶意样本指令数据包确定为二进制类报文。

5.根据权利要求3或4所述的恶意样本数据包分析方法,其特征在于,所述恶意样本特征信息还包括:通信数据包长度及通信数据包传输方向,所述对每一个目标恶意样本数据包进行报文分组,至少得到第一组目标恶意样本数据包包括:在确定出所述目标恶意样本数据包的报文类型之后,得到所述目标恶意样本数据包的初步分组;

针对每一个目标恶意样本数据包，将该目标恶意样本数据包确定为第一目标恶意样本数据包；

基于通信数据包长度及通信数据包传输方向，通过局部序列比对算法，将第一目标恶意样本数据包与不同报文类型的目标恶意样本数据包进行比较，确定是否存在与所述第一目标恶意样本数据包匹配的报文分类；

如果存在，则将所述第一目标恶意样本数据包加到匹配的报文分类对应的初步分组里，构成第一报文分组；

如果不存在，则创建新的报文分组，将所述第一目标恶意样本数据包加到该新的报文分组，构成第二报文分组。

6.根据权利要求5所述的恶意样本数据包分析方法，其特征在于，对于二进制类报文，在确定是否存在与所述第一目标恶意样本数据包匹配的报文分类时，对第一目标恶意样本数据是否能加到该类型报文分类中的限制条件包括：

长度限制：两个报文之间的长度距离不大于规定的阈值；

内容限制：两个报文之间的编辑距离不大于规定的阈值；

格式限制：相同类型的两个报文具有相同数量和次序的文本域、二进制域和统一编码域。

7.根据权利要求5所述的恶意样本数据包分析方法，其特征在于，对于文本类型的报文，在确定是否存在与所述第一目标恶意样本数据包匹配的报文分类时，对第一目标恶意样本数据是否能加到该类型报文分类中的限制条件包括：

片段限制：以换行符为分隔符将报文划分为不同的片段构成的组合，报文的序列的数量差异不大于规定阈值；

长度限制：报文之间对应的两个片段的长度差距不超过规定阈值；

内容限制：两个相同类型报文间对应的两个片段的编辑距离不超过规定阈值；

字符串限制：对于预定片段以特殊符号为界限划分出字符集合，两个相同类型报文间相对应的片段之间存在的共同字符串的数量满足预定阈值条件。

8.根据权利要求1所述的恶意样本数据包分析方法，其特征在于，所述对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包还包括：当存在两个目标恶意样本数据满足二进制类报文分类的限制条件或文本类型报文的限制条件时，则判断该两个目标恶意样本数据具有相似片段；

将该两个目标恶意样本数据分至同一报文分组；

所述提取每组恶意样本数据包中的公共特征包括：

提取该组两个目标恶意样本数据的相似片段，将其确定为公共特征；或者，

当存在两个以上的目标恶意样本数据满足二进制类报文分类的限制条件或文本类型报文的限制条件时，则判断该两个以上的目标恶意样本数据具有相似片段；

提取所述两个以上的目标恶意样本数据中的相似片段，将相似片段之间编辑距离较小的片段确定为公共特征。

9.一种恶意样本数据包分析装置，其特征在于，包括：

获取程序模块，用于获取多个恶意样本数据包；

筛选程序模块，用于对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样

本数据包；

初分类程序模块，用于按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类；

分组程序模块，用于针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包；

特征提取程序模块，用于提取每组恶意样本数据包中的公共特征。

10.一种电子设备，其特征在于，所述电子设备包括：壳体、处理器、存储器、电路板和电源电路，其中，电路板安置在壳体围成的空间内部，处理器和存储器设置在电路板上；电源电路，用于为上述电子设备的各个电路或器件供电；存储器用于存储可执行程序代码；处理器通过读取存储器中存储的可执行程序代码来运行与可执行程序代码对应的程序，用于执行前述权利要求1至8任一所述的恶意样本数据包分析方法。

一种恶意样本数据包分析方法、装置及电子设备

技术领域

[0001] 本发明涉及网络安全技术领域，尤其涉及一种恶意样本数据包分析方法、装置及电子设备。

背景技术

[0002] 随着网络信息技术的发展，围绕网络和数据的服务与应用呈现爆发式增长，丰富的应用场景下暴露出越来越多的网络安全风险和问题，并产生广泛而深远的影响。网络环境的复杂性、多变性，以及信息系统的脆弱性，决定了网络安全威胁持续客观存在。

[0003] 一些单位面临着通过社会工程学、公网业务漏洞利用等攻击手段入侵的风险。由于单位一般具有一定的规模，恶意代码在内部进行漫游传播后，将造成大面积的感染，致使内部核心数据、机密数据将可能面临着被窃取泄露的风险。如果不能及时发现潜在的未知协议的恶意代码(也称为恶意样本)威胁，识别其特征，将造成重大的损失。

发明内容

[0004] 有鉴于此，本发明实施例提供一种恶意样本数据包分析方法、装置及电子设备，便于提取出未知协议的恶意样本中的公共特征，从而可以提升识别发现未知协议的恶意代码威胁的能力。

[0005] 第一方面，本发明实施例提供的恶意样本数据包分析方法，包括步骤：获取多个恶意样本数据包；对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样本数据包；按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类；针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包；提取每组恶意样本数据包中的公共特征。

[0006] 可选地，所述公共特征包括：不变字段域、可变字段域和/或常量域。

[0007] 可选地，所述恶意样本特征信息包括：恶意样本所属的病毒家族；

[0008] 所述按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类，包括：对每一个目标恶意样本数据包，按照恶意样本所属的病毒家族分成相应的多个类别，每个类别中包含多个目标恶意样本数据；

[0009] 所述在对每一个目标恶意样本数据包进行报文分组之前，所述方法还包括：针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分类，确定所述目标恶意样本数据包的报文类型；

[0010] 所述对每一个目标恶意样本数据包进行报文分类，确定所述目标恶意样本数据包的报文类型，包括：针对每一个病毒家族类别，获取该类别包含的多个恶意样本上线数据包；所述目标恶意样本数据包包含：恶意样本上线数据包；对于每一个恶意样本上线数据包，对所述恶意样本上线数据包的字节类型进行提取；所述字节类型包括：字符型和/或二进制型；判断所述字符型的字节数量是否超过预定阈值；若是，则将该恶意样本上线数据包

确定为文本类报文;若否,则将该恶意样本上线数据包确定为二进制类报文。

[0011] 可选地,述对每一个目标恶意样本数据包进行报文分类,确定所述目标恶意样本数据包的报文类型,还包括:针对每一个病毒家族类别,获取该类别包含的多个恶意样本指令数据包;所述目标恶意样本数据包包含:恶意样本指令数据包;对于每一个恶意样本指令数据包,对所述恶意样本指令数据包的字节类型进行提取;所述字节类型包括:字符型和/或二进制型;判断所述字符型的字节数量是否超过预定阈值;若是,则将该恶意样本指令数据包确定为文本类报文;若否,则将该恶意样本指令数据包确定为二进制类报文。

[0012] 可选地,所述恶意样本特征信息还包括:通信数据包长度及通信数据包传输方向,所述对每一个目标恶意样本数据包进行报文分组,至少得到第一组目标恶意样本数据包包括:在确定出所述目标恶意样本数据包的报文类型之后,得到所述目标恶意样本数据包的初步分组;

[0013] 针对每一个目标恶意样本数据包,将该目标恶意样本数据包确定为第一目标恶意样本数据包;

[0014] 基于通信数据包长度及通信数据包传输方向,通过局部序列比对算法,将第一目标恶意样本数据包与不同报文类型的目标恶意样本数据包进行比较,确定是否存在与所述第一目标恶意样本数据包匹配的报文分类;如果存在,则将所述第一目标恶意样本数据包加到匹配的报文分类对应的初步分组里,构成第一报文分组;如果不存在,则创建新的报文分组,将所述第一目标恶意样本数据包加到该新的报文分组,构成第二报文分组。

[0015] 可选地,对于二进制类报文,在确定是否存在与所述第一目标恶意样本数据包匹配的报文分类时,对第一目标恶意样本数据是否能加到该类型报文分类中的限制条件包括:长度限制:两个报文之间的长度距离不大于规定的阈值;内容限制:两个报文之间的编辑距离不大于规定的阈值;格式限制:相同类型的两个报文具有相同数量和次序的文本域、二进制域和统一编码域。

[0016] 可选地,对于文本类型的报文,在确定是否存在与所述第一目标恶意样本数据包匹配的报文分类时,对第一目标恶意样本数据是否能加到该类型报文分类中的限制条件包括:片段限制:以换行符为分隔符将报文划分为不同的片段构成的组合,报文的序列的数量差异不大于规定阈值;长度限制:报文之间对应的两个片段的长度差距不超过规定阈值;内容限制:两个相同类型报文间对应的两个片段的编辑距离不超过规定阈值;字符串限制:对于预定片段以特殊符号为界限划分出字符集合,两个相同类型报文间相对应的片段之间存在的共同字符串的数量满足预定阈值条件。

[0017] 可选地,所述对每一个目标恶意样本数据包进行报文分组,至少得到第一组目标恶意样本数据包还包括:当存在两个目标恶意样本数据满足二进制类报文分类的限制条件或文本类型报文的限制条件时,则判断该两个目标恶意样本数据具有相似片段;将该两个目标恶意样本数据分至同一报文分组;

[0018] 所述提取每组恶意样本数据包中的公共特征包括:提取该组两个目标恶意样本数据的相似片段,将其确定为公共特征;或者,

[0019] 当存在两个以上的目标恶意样本数据满足二进制类报文分类的限制条件或文本类型报文的限制条件时,则判断该两个以上的目标恶意样本数据具有相似片段;提取所述两个以上的目标恶意样本数据中的相似片段,将相似片段之间编辑距离较小的片段确定为

公共特征。

[0020] 第二方面，本发明实施例提供的恶意样本数据包分析装置，包括：获取程序模块，用于获取多个恶意样本数据包；筛选程序模块，用于对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样本数据包；初分类程序模块，用于按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类；分组程序模块，用于针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包；特征提取程序模块，用于提取每组恶意样本数据包中的公共特征。

[0021] 第三方面，本发明实施例提供的电子设备，包括：壳体、处理器、存储器、电路板和电源电路，其中，电路板安置在壳体围成的空间内部，处理器和存储器设置在电路板上；电源电路，用于为上述电子设备的各个电路或器件供电；存储器用于存储可执行程序代码；处理器通过读取存储器中存储的可执行程序代码来运行与可执行程序代码对应的程序，用于执行第一方面任一所述的恶意样本数据包分析方法。

[0022] 本发明实施例提供的恶意样本数据包分析方法、装置及电子设备，通过对采集的大量的未知网络协议的恶意样本进行逆向报文解析操作，对各种病毒家族的恶意样本进行报文分组，并从中提取每组恶意样本数据包中的公共特征，便于提取出未知协议的恶意样本中的公共特征，进而识别出某一病毒家族的恶意样本的恶意属性，通过提取出的公共特征，可以丰富各种病毒家族的恶意样本的特征库，从而可以提升识别发现未知协议的恶意代码威胁的能力。

附图说明

[0023] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

[0024] 图1为本发明恶意样本数据包分析方法一实施例的流程示意图；

[0025] 图2为本发明恶意样本数据包分析方法又一实施例的流程示意图；

[0026] 图3为本发明中基于Smith-Waterman算法得到的一个示意性的偏序比对图；

[0027] 图4为本发明网络资产指纹特征的识别装置一实施例架构图；

[0028] 图5为本发明电子设备一个实施例的结构示意图。

具体实施方式

[0029] 下面结合附图对本发明实施例进行详细描述。

[0030] 应当明确，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其它实施例，都属于本发明保护的范围。

[0031] 本实施例提供的恶意样本数据包分析方法，通过采集大量的威胁样本(也称为恶意样本，本文中大多数时候都称为“恶意样本”)通信数据包，对其进行病毒家族初步分类之后，采用报文分组方式对其进一步分组，通过未知网络协议逆向分析技术，从中提取出威胁

样本通信数据包的报文结构与公共特征，从而可以丰富威胁样本特征库，进而可有效提升对网络数据流量中各类潜伏的未知协议恶意代码程序的识别能力。

[0032] 实施例一

[0033] 图1为本发明恶意样本数据包分析方法一实施例的流程示意图，请参看图1所示，本发明实施例提供的恶意样本数据包分析方法，可应用于恶意样本特征库补全及网络威胁检测发现场景中，用于对未知网络协议的恶意样本数据包进行分析，从而提取出相应的表征恶意属性的公共特征。

[0034] 需要说明的是，该方法可以以软件的形式固化于某一制造的实体产品中，当用户需要对恶意样本数据包进行分析时，可以触发并再现本申请的方法流程。

[0035] 参看图1所示，所述恶意样本数据包分析方法，可包括步骤：

[0036] S110、获取多个恶意样本数据包。

[0037] 本实施例例中，可以通过Pcap数据包抓取工具，通过大量的威胁活性样本站点或恶意样本库，例如，VirusShare、Wireshark等，抓取收集Pcab数据包。

[0038] 其中，Pcab也是一个数据包抓取库，很多软件都是用于它来作为数据包抓取工具的，例如前述的WireShark。Pcab数据包一般是一种不同于原始数据流格式的新数据格式。

[0039] S120、对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样本数据包。

[0040] 在获取到大量Pcap数据包之后，对所述Pcap数据包进行处理，例如，数据清洗筛选等，从Pcap数据包中提取大量的上线、心跳、指令等数据包，根据需要可以将上线数据包、心跳数据包和指令数据包作为所述目标恶意样本数据包。

[0041] 应当理解，恶意样本数据包很多时候，由于在恶意样本库中一般是静态存在的，大部分没有心跳数据包。

[0042] S130、按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类。

[0043] 其中，恶意样本特征信息一般包括：所属病毒家族、通讯数据包长度、通讯数据包传输方向等，在得到目标恶意样本数据包之后，可以按照病毒家族对所有目标恶意样本数据包进行一个初步简单分类。至于通讯数据包长度、通讯数据包传输方向等可以作为后续报文分组的依据。

[0044] S140、针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包。

[0045] S150、提取每组恶意样本数据包中的公共特征。

[0046] 应当理解，该公共特征可以用于表征某一具体类型的恶意样本的共同恶意属性特征，当遇到未知网络协议的恶意样本时，通过提取其携带的恶意属性特征，就可以实现对该未知网络协议的恶意样本的快速检测识别，从而提升对潜伏的恶意样本的检测能力。

[0047] 为了实现对未知网络协议的恶意样本进行识别，必须得到类似样本或相同样本的特征或恶意属性，因此，本实施例中，通过对每一个目标恶意样本数据包进行报文分组，将具有至少两个相似的目标恶意样本数据包分到统一报文分组中，并进而进行逆向报文解析操作，从中提取出其公共特征，作为某一具体类型的恶意样本的恶意属性特征，并进而构建出丰富的恶意属性特征库，可以实现对某一未知网络协议的恶意样本的检测，从而提升网络流量中各类潜伏的未知协议恶意代码程序(即恶意样本)的检测识别能力。

[0048] 应当理解，实现对恶意样本检测识别的前提是针对大量恶意样本，进行网络流量数据逆向分析，挖掘出某一类恶意样本的共同恶意属性，因此，在逆向分析过程中，对网络流量数据按照病毒家族进行报文分类及报文分组，就是一个可以分析出未知网络协议的恶意样本的恶意属性特征的有效手段。

[0049] 图2为本申请另一实施例提供的恶意样本数据包分析方法流程示意图。在一些实施例中，所述恶意样本特征信息包括：恶意样本所属的病毒家族；其中，每一个病毒家族的病毒都具有特定的传播特点，当然不同病毒家族的病毒也可以具有相同的传播特点。例如，在一项研究中发现：某僵尸网络病毒家族的大部分样本，都具有启动新进程的行为特点，而且该新进程存在异常行为，包括设置延迟启动线程，且会尝试与另一个控制端进行远程连接以通信。

[0050] 所述按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类(步骤S130)，包括：对每一个目标恶意样本数据包，按照恶意样本所属的病毒家族分成相应的多个类别，每个类别中包含多个目标恶意样本数据；

[0051] 步骤S140中，所述在对每一个目标恶意样本数据包进行报文分组之前，所述方法还包括：S135、针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包(由于数据包在通信领域一般称为报文，因此，本文中也以报文代指)进行报文分类，确定所述目标恶意样本数据包的报文类型。

[0052] 其中，报文分类，作为报文格式解析和协议状态机推断的基础，报文分类(也可以认为是初步分组)是非常重要的一个步骤。其具体做法可以为：对于获取到的报文，首先进行一个初步的字节类型提取的处理过程，并通过这个处理结果来进行报文的初步分组。

[0053] 具体的，请参看图2所示，所述对每一个目标恶意样本数据包进行报文分类，确定所述目标恶意样本数据包的报文类型，包括：针对每一个病毒家族类别，获取该类别包含的多个恶意样本上线数据包；所述目标恶意样本数据包包含：恶意样本上线数据包；对于每一个恶意样本上线数据包，对所述恶意样本上线数据包的字节类型进行提取；所述字节类型包括：字符型和/或二进制型；判断所述字符型的字节数量是否超过预定阈值；若是，则将该恶意样本上线数据包确定为文本类报文；若否，则将该恶意样本上线数据包确定为二进制类报文。

[0054] 由于目标恶意样本数据包还包含：指令数据包和/或心跳包，基于与上述上线数据包同样的报文分类方式执行对指令数据包和/或心跳包的的报文分类。具体的，所述对每一个目标恶意样本数据包进行报文分类，确定所述目标恶意样本数据包的报文类型(步骤S130)，还包括：针对每一个病毒家族类别，获取该类别包含的多个恶意样本指令数据包；所述目标恶意样本数据包包含：恶意样本指令数据包；

[0055] 步骤S140中，对于每一个恶意样本指令数据包，对所述恶意样本指令数据包的字节类型进行提取；所述字节类型包括：字符型和/或二进制型；判断所述字符型的字节数量是否超过预定阈值；若是，则将该恶意样本指令数据包确定为文本类报文；若否，则将该恶意样本指令数据包确定为二进制类报文。

[0056] 根据本实施例上述上线数据包和指令数据包的报文分类方案可知，在报文分类过程中，需要首先根据报文中字节的类型来对报文类型进行判断，主要基于如下两个原则：

[0057] a、对每一条报文中的每一个不同的字节进行分析，假如是可打印字符，一般表示

为换行符0X0D0A,则判定该字节是字符类型,否则判定为二进制字节。

[0058] b、判断了每条报文的每个字节类型之后,再分析整个报文,当发现报文中出现的字符型字节超过预定阈值时,则判定这条报文是文本类报文,否则判定为二进制类报文。

[0059] 对于原则b,判定为文本类报文的规则例如可以将预定阈值设置为报文全文中出现的字节全部为字符型字节,当然也可以设置为超过预定数量阈值个字节为字符型字节,则可以判断为文本类报文,否则判定为二进制类报文。

[0060] 本实施例中,通过对不同类型的目标恶意样本数据包按照病毒家族,并基于报文特征进行报文分类,初步确定了报文分组,将一个或多个恶意样本初步贴上了较为具体的类别标签,便于后续某一类恶意样本的公共特征的准确提取。

[0061] 在初步处理了各个报文,并进行了报文分类之后,接下来的工作是对报文进行分组(相对于报文分类的概念,更为具体的划分),请继续参看图2所示,在一些实施例中,所述恶意样本特征信息还包括:通信数据包长度及通信数据包传输方向。步骤S140中,所述对每一个目标恶意样本数据包进行报文分组,至少得到第一组目标恶意样本数据包,包括:在确定出所述目标恶意样本数据包的报文类型之后,得到所述目标恶意样本数据包的初步分组;针对每一个目标恶意样本数据包,将该目标恶意样本数据包确定为第一目标恶意样本数据包;基于通信数据包长度及通信数据包传输方向,通过局部序列比对算法,将第一目标恶意样本数据包与不同报文类型的目标恶意样本数据包进行比较,确定是否存在与所述第一目标恶意样本数据包匹配的报文分类;如果存在,则将所述第一目标恶意样本数据包加到匹配的报文分类对应的初步分组里,构成第一报文分组;如果不存在,则创建新的报文分组,将所述第一目标恶意样本数据包加到该新的报文分组,构成第二报文分组。

[0062] 其中,所述局部序列比对算法,用于找出两个报文序列中具有相似高度的片段,以确定是否存在与所述第一目标恶意样本数据包匹配的报文分类,从而判断目标恶意样本数据包的具体分组。

[0063] 在一些实施例中,所述局部序列比对算法采用Smith-Waterman局部比对算法,根据其算法从一个点斜向左上角进行运算,也可以形象地称为偏序比对算法。

[0064] 本实施例中,具体分组的策略可以为:获得一条报文A的字节类型之后,通过Smith-Waterman偏序比对算法将这条报文A与已经划分出来的报文分组(最开始是前面的报文分类,即初步分组)进行比较,看是否存在与该报文匹配的报文分类。如果存在则直接将报文加到相应的报文分组里,否则创建新的报文分组。在这个过程中可以设定循环执行比较的迭代次数,以实现更为精准的分组。

[0065] 具体的,通过Smith-Waterman偏序比对算法,得到多条报文对应的偏序比对图,根据偏序比对图比对可以判断两条报文是否具有相同的片段,若有则可以分至同一个报文分组中。

[0066] 其中,二进制类报文只能和二进制类报文在同一个分组,文本类报文只能和文本类报文同一分组。同理,相同传输方向的报文才可能存放在同一个报文分组之中。不同传输方向的报文也不可能存放在同一报文分组中。

[0067] 具体的,对于二进制类报文,在确定是否存在与所述第一目标恶意样本数据包匹配的报文分类时,对第一目标恶意样本数据是否能加到该类型报文分类中的限制条件包括:1、长度限制:两个报文之间的长度距离不大于规定的阈值;2、内容限制:两个报文之间

的编辑距离不大于规定的阈值;3、格式限制:相同类型的两个报文具有相同数量和次序的文本域、二进制域和统一编码(Unicode)域。

[0068] 而对于文本类型的报文,在确定是否存在与所述第一目标恶意样本数据包匹配的报文分类时,对第一目标恶意样本数据是否能加到该类型报文分类中的限制条件包括:1、片段限制:以换行符0x0D0A为分隔符将报文划分为不同的片段构成的组合,报文的序列的数量差异不大于规定阈值;2、长度限制:报文之间对应的两个片段的长度差距不超过规定阈值;3、内容限制:两个相同类型报文间对应的两个片段的编辑距离不超过规定阈值;4、字符串限制:对于预定片段以特殊符号为界限划分出字符集合,两个相同类型报文间相对应的片段之间存在的共同字符串的数量满足预定阈值条件。

[0069] 所述对每一个目标恶意样本数据包进行报文分组,至少得到第一组目标恶意样本数据包还包括:当存在两个目标恶意样本数据满足二进制类报文分类的限制条件或文本类型报文的限制条件时,则判断该两个目标恶意样本数据具有相似片段;将该两个目标恶意样本数据分至同一报文分组。

[0070] 所述提取每组恶意样本数据包中的公共特征包括:提取该组两个目标恶意样本数据的相似片段,将其确定为公共特征。

[0071] 或者,在另一些实施例中,当存在两个以上的目标恶意样本数据满足二进制类报文分类的限制条件或文本类型报文的限制条件时,则判断该两个以上的目标恶意样本数据具有相似片段;提取所述两个以上的目标恶意样本数据中的相似片段,将相似片段之间编辑距离较小的片段确定为公共特征。

[0072] 其中,编辑距离,也叫莱文斯坦距离(Levenshtein),是针对两个字符串(例如英文字符)的差异程度的量化量测,量测方式是看至少需要多少次的处理才能将一个字符串变成另一个字符串,编辑距离越近,则表明二者差异化程度越小。

[0073] 本实施例中,针对某一类报文,在同时满足以上相应限制条件之后才能判断两个报文的相应片段是相似的,进而判断两条报文可能是同一组报文:如果有多个报文判断为相似报文,则选择编辑距离最小的片段作为该类报文的最佳匹配特征,并将该类报文加到相应的报文类型分组中,并将该最佳匹配特征作为该类型分组中的不变字段域。

[0074] 请继续参看图2所示,在一些实施例中,在基于偏序比对图,获得多条通信数据包的可变字段域和不变字段域之后,还包括:从所述可变字段域和不变字段域中,对每一条通信数据包中的常见字符进行过滤,得到最终的常量域、可变字段域和不变字段域。

[0075] 其中,常见字符为一般的正常样本中也常见的字符,例如1.ASCII字符集&编码、GB2312字符集&编码、Unicode字符集&编码、UTF-8编码等。不变字符域即为通讯数据包中携带的不可变的对象,可变字符域即为通讯数据包中携带的可变的对象。

[0076] 本实施例中,在通过Smith-Waterman偏序比对算法,对每一条(个)通讯数据包执行局部序列比对运算之后,根据得到的偏序比对图可以确定出报文分组中的多条通讯数据包的可变、不变字符域,从而从该字符域中可以得到某一类恶意样本的报文结构和公共特征,从而可以用于丰富未知网络协议的恶意样本特征库。

[0077] 为了帮助理解本发明实施例提供的技术方案及其技术效果,请参看图3,现结合一具体示例对其中一个实施例进行详细说明如下:

[0078] 某网络安全部门需要对未知网络协议的网络攻击事件进行检测识别,为了提高对

未知网络协议的网络攻击事件的检出率，需要收集大量的恶意样本(威胁样本)，对其进行逆向分析，从而提取出某一类型的恶意样本的恶意属性特征，以构建检测特征库。具体的逆向分析过程如下：

[0079] 步骤S2：数据采集，通过从具有大量恶意样本数据库中，获取大量的威胁活性样本，收集pcap数据包；

[0080] 步骤S2中：对pcap数据包处理，具体为从pcap数据包中提取大量的上线、心跳、指令等指定类别的数据包；

[0081] 步骤S2之前，可以包括步骤S1、提供大量恶意样本Pcab数据包库。

[0082] 步骤S4：对数据包初步分类，按照病毒家族等对所有的通讯包进行一个初步简单的分类；当然，也可以进一步基于通讯包长度、传输方向等信息进行部分再分类。

[0083] 在步骤S4之前，可以包括步骤S3：对每条通讯报的基本信息进行计算，例如，数据包长度和传输方向等样本特征信息，以用于后续分组的依据。当然，该步骤也可以在后续分组时再进行执行。

[0084] 步骤S5、报文分类，包括步骤S5a和S5b，根据获取的目标恶意样本数据的类型，也可以只执行其中一个步骤。报文初步分类作为报文格式解析和协议状态机推断的基础，是非常重要的一个步骤。其具体做法可以为：对于获取到的报文，首先进行一个初步的字节类型提取的处理过程，并通过这个处理结果来进行报文的初步分组。在字节类型提取过程中。第一步是根据报文中字符的类型来对报文类型进行判断，主要由以下两个原则：

[0085] a、对每一条报文中的每一个不同的字节进行分析，假如是可打印字符的话则判定该字符是字符类型，否则判定为二进制字节；

[0086] b、判断了每条报文的每个字节之后，分析整个报文，当发现报文中出现的全都是字符字节，包括0X0D0A时，则判定这条报文是文本报文，否则判定为二进制类报文。

[0087] 步骤S6至S8：报文分组及每条通讯数据包的特征提取的过程，在初步处理了各个报文之后，接下来的工作是对报文进行分组，请参看图2所示，具体分组的策略为：获得一条报文的字节类型之后，通过Smith-Waterman偏序比对算法将这条报文与前面已经划分出来的报文分组进行比较，看是否存在与该报文匹配的报文分类，如果存在则直接将报文加到相应的报文分组里。其中，一个比对示意性地偏序比对图如图3所示，根据偏序比对图可以得到两条通信数据包中的相同片段为：GTT和AC，作为不可变字段域。可变字段域可以为图中不同的片段部分。常量域可以为过滤掉的ASCII码。

[0088] 步骤9：最终公共特征提取，通过上述步骤可以获取每一个家族名下指令特征的不变域，可变域和常量域，从不便域、可变域和常量域能够得到该具体恶意样本报文分组的报文结构及公共特征。

[0089] 在得到所述报文结构和公共特征之后，可以丰富恶意属性特征库，用于进行某一具体类别的恶意样本的检测识别。

[0090] 由此，本发明实施例提供的恶意样本数据包分析方法，通过对采集的大量的未知网络协议的恶意样本进行逆向报文解析操作，对各种病毒家族的恶意样本进行报文分组，并从中提取每组恶意样本数据包中的公共特征，便于提取出未知协议的恶意样本中的公共特征，进而识别出某一病毒家族的恶意样本的恶意属性，通过提取出的公共特征，可以丰富各种病毒家族的恶意样本的特征库，从而可以提升识别发现未知协议的恶意代码威胁的能

力。

[0091] 实施例二

[0092] 图4为本发明恶意样本数据包分析装置一实施例的架构图。参看图4所示，本实施例的网络资产指纹特征的识别装置，包括：获取程序模块210，用于获取多个恶意样本数据包；筛选程序模块220，用于对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样本数据包；初分类程序模块230，用于按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类；分组程序模块240，用于针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包；特征提取程序模块260，用于提取每组恶意样本数据包中的公共特征。

[0093] 本实施例的装置，可以用于执行图1所示方法实施例的技术方案，其实现原理和技术效果类似，此处不再赘述。

[0094] 此外，本实施例的装置，其还可以用于执行前述实施例一中相应恶意样本数据包分析方法的其它实施例，未详细述及之处，可以相互参看，此处不再赘述。

[0095] 实施例三

[0096] 图5为本发明电子设备一个实施例的结构示意图，基于前述实施例一提供的方法、实施例二提供的装置，本发明实施例还提供一种电子设备，如图5所示，可以实现本发明实施例一中任一所述的实施例的步骤流程，上述电子设备可以包括：壳体41、处理器42、存储器43、电路板44和电源电路45，其中，电路板44安置在壳体41围成的空间内部，处理器42和存储器43设置在电路板44上；电源电路45，用于为上述电子设备的各个电路或器件供电；存储器43用于存储可执行程序代码；处理器42通过读取存储器43中存储的可执行程序代码来运行与可执行程序代码对应的程序，用于执行前述任一实施例所述的恶意样本数据包分析方法。

[0097] 处理器42对上述步骤的具体执行过程以及处理器42通过运行可执行程序代码来进一步执行的步骤，可以参见本发明实施例一的描述，在此不再赘述。

[0098] 本发明还实施例提供一种计算机可读存储介质，所述计算机可读存储介质存储有实施例一中任一所述的加密数据，所述加密数据包含的可执行解密程序可被一个或者多个处理器执行，以实现前述权利要求实施例一中任一所述的恶意样本数据包分析方法。

[0099] 综上，相比于现有的基于特征匹配的资产数据识别方案，本发明实施例提供的恶意样本数据包分析方法及装置，通过对采集的大量的未知网络协议的恶意样本进行逆向报文解析操作，对各种病毒家族的恶意样本进行报文分组，并从中提取每组恶意样本数据包中的公共特征，便于提取出未知协议的恶意样本中的公共特征，进而识别出某一病毒家族的恶意样本的恶意属性，通过提取出的公共特征，可以丰富各种病毒家族的恶意样本的特征库，从而可以提升识别发现未知协议的恶意代码威胁的能力。

[0100] 进一步地，在进行逆向分析过程中，通过输入一系列通信数据报文，在报文分类的基础上。进一步进行报文分组，从而推断出每个报文类的格式结构信息，得到协议报文的格式模型，并从中提取出公共特征，可以用于精准识别某一具体类型的恶意样本。

[0101] 上述电子设备以多种形式存在，包括但不限于：

[0102] (1)移动通信设备：这类设备的特点是具备移动通信功能，并且以提供话音、数据

通信为主要目标。这类终端包括:智能手机(例如iPhone)、多媒体手机、功能性手机,以及低端手机等。

[0103] (2)超移动个人计算机设备:这类设备属于个人计算机的范畴,有计算和处理功能,一般也具备移动上网特性。这类终端包括:PDA、MID和UMPC设备等,例如iPad。

[0104] (3)便携式娱乐设备:这类设备可以显示和播放多媒体内容。该类设备包括:音频、视频播放器(例如iPod),掌上游戏机,电子书,以及智能玩具和便携式车载导航设备。

[0105] (4)服务器:提供计算服务的设备,服务器的构成包括处理器、硬盘、内存、系统总线等,服务器和通用的计算机架构类似,但是由于需要提供高可靠的服务,因此在处理能力、稳定性、可靠性、安全性、可扩展性、可管理性等方面要求较高。

[0106] (5)其他具有数据交互功能的电子设备。

[0107] 需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0108] 本说明书中的各个实施例均采用相关的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。

[0109] 尤其,对于装置实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0110] 为了描述的方便,描述以上装置是以功能分为各种单元/模块分别描述。当然,在实施本发明时可以把各单元/模块的功能在同一个或多个软件和/或硬件中实现。

[0111] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access Memory,RAM)等。

[0112] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以权利要求的保护范围为准。

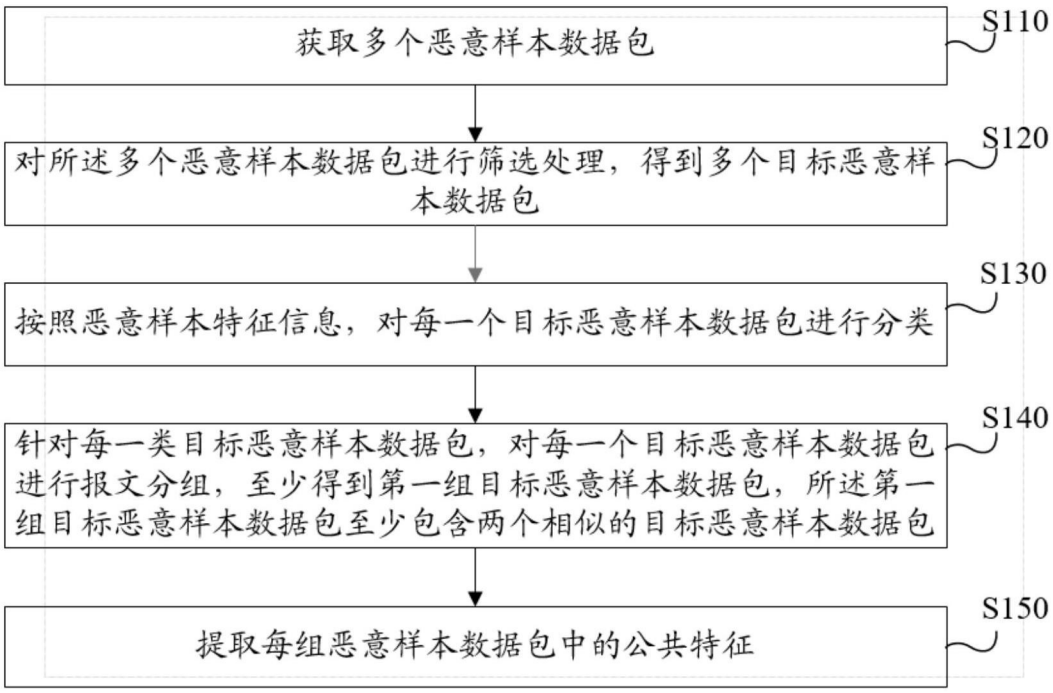

获取多个恶意样本数据包
S110
对所述多个恶意样本数据包进行筛选处理，得到多个目标恶意样本数据包
S120
按照恶意样本特征信息，对每一个目标恶意样本数据包进行分类
S130
针对每一类目标恶意样本数据包，对每一个目标恶意样本数据包进行报文分组，至少得到第一组目标恶意样本数据包，所述第一组目标恶意样本数据包至少包含两个相似的目标恶意样本数据包
S140
提取每组恶意样本数据包中的公共特征
S150

图1

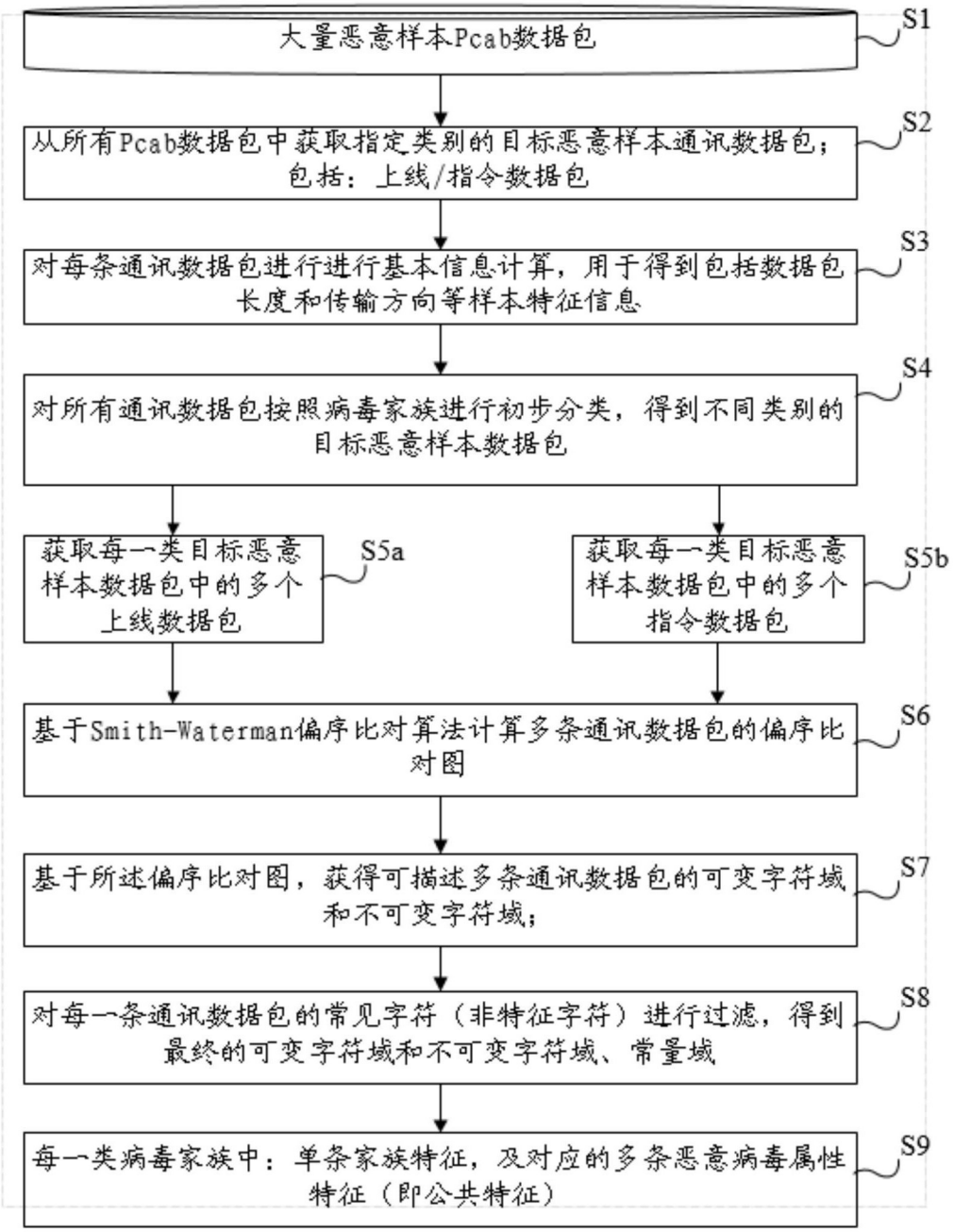

大量恶意样本Pcab数据包
S1
从所有Pcab数据包中获取指定类别的目标恶意样本通讯数据包；包括：上线/指令数据包
S2
对每条通讯数据包进行进行基本信息计算，用于得到包括数据包长度和传输方向等样本特征信息
S3
对所有通讯数据包按照病毒家族进行初步分类，得到不同类别的目标恶意样本数据包
S4
获取每一类目标恶意样本数据包中的多个上线数据包
S5a
获取每一类目标恶意样本数据包中的多个指令数据包
S5b
基于Smith-Waterman偏序比对算法计算多条通讯数据包的偏序比对图
S6
基于所述偏序比对图，获得可描述多条通讯数据包的可变字符域和不可变字符域；
S7
对每一条通讯数据包的常见字符（非特征字符）进行过滤，得到最终的可变字符域和不可变字符域、常量域
S8
每一类病毒家族中：单条家族特征，及对应的多条恶意病毒属性特征（即公共特征）
S9

图2

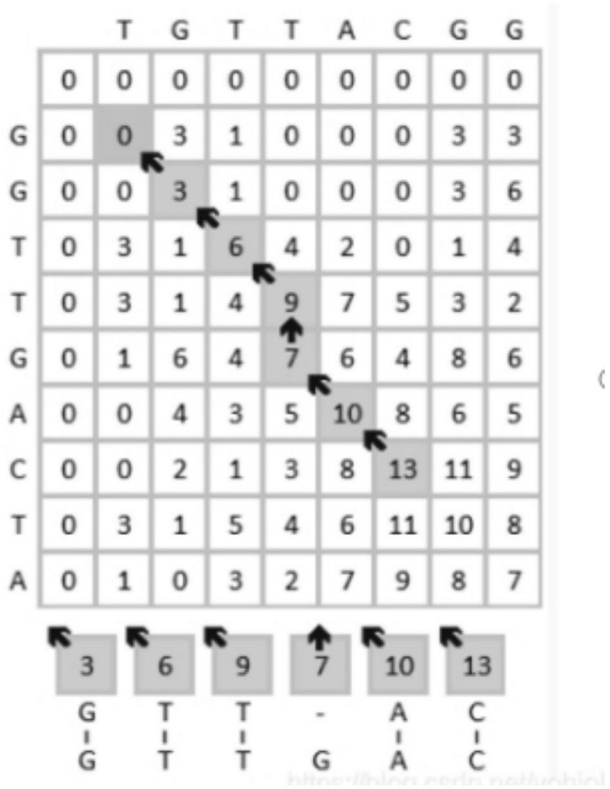

T G T T A C G G
0 0 0 0 0 0 0 0 0
G 0 0 3 1 0 0 0 3 3
G 0 0 3 1 0 0 0 3 6
T 0 3 1 6 4 2 0 1 4
T 0 3 1 4 9 7 5 3 2
G 0 1 6 4 7 6 4 8 6
A 0 0 4 3 5 10 8 6 5
C 0 0 2 1 3 8 13 11 9
T 0 3 1 5 4 6 11 10 8
A 0 1 0 3 2 7 9 8 7
3 6 9 7 10 13
G T T - A C
I I I I I I
G T T G A C

图3

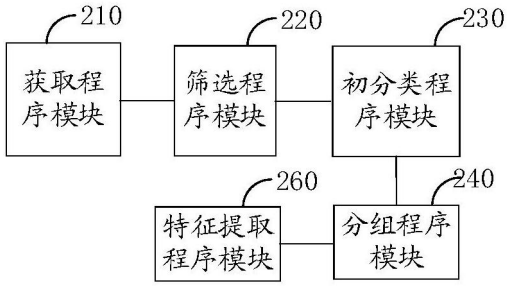

210
获取程序模块
220
筛选程序模块
230
初分类程序模块
260
特征提取程序模块
240
分组程序模块

图4

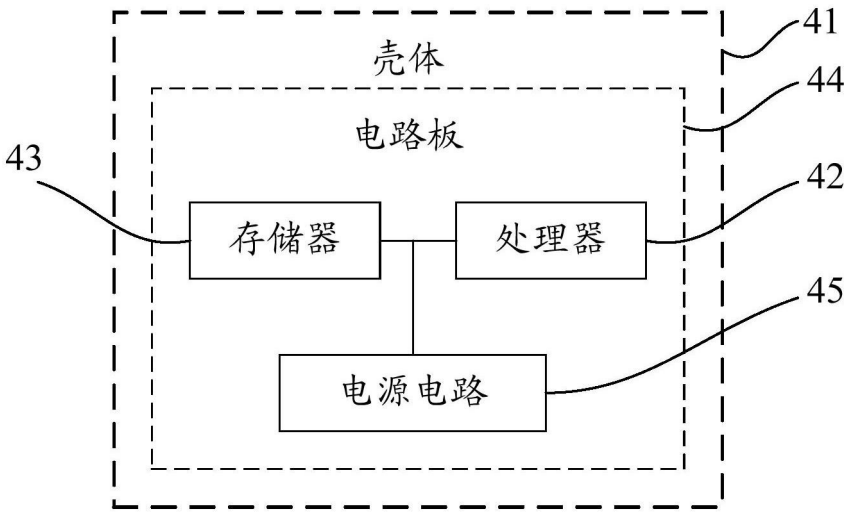


图5