



(12)发明专利申请

(10)申请公布号 CN 106709345 A

(43)申请公布日 2017.05.24

(21)申请号 201611024547.0

(22)申请日 2016.11.16

(66)本国优先权数据

201510787438.3 2015.11.17 CN

(71)申请人 武汉安天信息技术有限责任公司

地址 430000 湖北省武汉市东湖新技术开发区关山大道光谷软件园F4栋12楼

(72)发明人 潘宣辰 孙岩 马志远

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201

代理人 张大威

(51)Int.Cl.

G06F 21/56(2013.01)

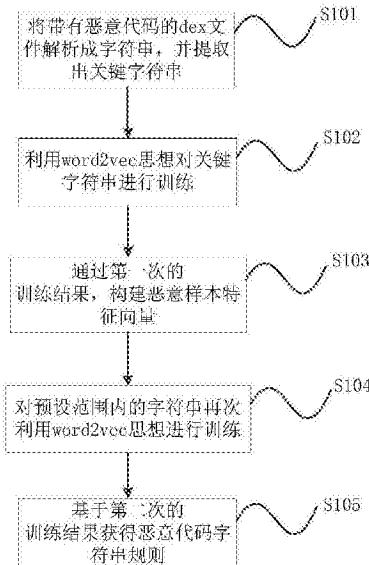
权利要求书2页 说明书9页 附图4页

(54)发明名称

基于深度学习方法推断恶意代码规则的方法、系统及设备

(57)摘要

本发明公开了基于深度学习方法推断恶意代码规则的方法、系统及终端设备，创造性的将word2vec思想应用于恶意代码分析领域，对已知的恶意代码的字符串进行训练，得到恶意代码字符串关联性最大的字符串，进而得到该恶意代码的字符串规则。本发明能充分利用恶意样本的特征推断出误报率低、覆盖率高的恶意代码规则，以优化现有病毒检测引擎，提升恶意代码检测效率。本发明能广泛应用于恶意代码检测及恶意代码分析领域。



1. 一种基于深度学习方法推断恶意代码规则的方法,其特征在于,包括:

将带有恶意代码的dex文件解析成字符串,并根据预设规则从所述字符串中提取出关键字符串;

利用word2vec思想对所述关键字符串进行训练,得到第一训练结果;

通过所述第一训练结果,构建恶意样本特征向量;

根据所述恶意样本特征向量从所述关键字符串中提取预设范围内的字符串,并对所述预设范围内的字符串再次利用word2vec思想进行训练,得到第二训练结果;

基于所述第二训练结果获得恶意代码字符串规则。

2. 如权利要求1所述的方法,其特征在于,所述利用word2vec思想对所述关键字符串进行训练,具体为:对所述关键字符串进行特征提取,并根据所述特征推断出关联性最大的字符串,其中,所述关联性最大的字符串用于指示针对所述关键字符串的解释性的字符串。

3. 如权利要求1所述的方法,其特征在于,所述根据预设规则从所述字符串中提取出关键字符串,包括:

选取对照样本,并计算所述对照样本中的字符串与所述带有恶意代码的dex文件的字符串之间的距离;

根据所述距离以及预设的距离阈值,从所述带有恶意代码的dex文件的字符串中提取出所述关键字符串。

4. 如权利要求1所述的方法,其特征在于,所述关键字符串包括函数调用关系和代码结构上下文中的用于描述所述dex文件内容字符串关系的内容。

5. 如权利要求1所述的方法,其特征在于,所述构建恶意样本特征向量为:基于样本属性数据,归纳n个样本属性数据,并基于所述n个样本属性数据,计算属性数据的向量,并计算所述属性数据的向量两两相似度矩阵,保留主要向量,累加所有所述主要向量的各维度分量。

6. 如权利要求5所述的方法,其特征在于,所述保留主要向量,累加所有所述主要向量的各维度分量,包括:

将两两相似度矩阵之间的差异性大于预设阈值的属性数据的向量进行保留;

将所述保留的属性数据的向量作为所述恶意样本特征向量,其中,针对所述保留的属性数据的向量,对相同的属性数据的向量的各维度分量进行累加。

7. 如权利要求1所述的方法,其特征在于,所述基于训练结果获得恶意代码字符串规则,具体为:通过获得关联性最大的字符串,找出关联性字符串之间的字符串特征,最终根据所述关联性字符串之间的字符串特征获得恶意代码字符串规则。

8. 如权利要求7所述的方法,其特征在于,通过文档主题生成模型LDA将所述关联性字符串之间的字符串特征进行特征提取以获得所述恶意代码字符串规则。

9. 一种基于深度学习方法推断恶意代码规则的系统,其特征在于,包括:

解析模块,用于将带有恶意代码的dex文件解析成字符串,并根据预设规则从所述字符串中提取出关键字符串;

训练模块,用于利用word2vec思想对所述关键字符串进行训练,得到第一训练结果;

构建模块,用于通过所述第一训练结果,构建恶意样本特征向量;

再训练模块,用于根据所述恶意样本特征向量从所述关键字符串中提取预设范围内的

字符串，并对所述预设范围内的字符串再次利用word2vec思想进行训练，得到第二训练结果；

获取模块，用于基于所述第二训练结果获得恶意代码字符串规则。

10. 如权利要求9所述的系统，其特征在于，所述训练模块具体用于：对所述关键字符串进行特征提取，并根据所述特征推断出关联性最大的字符串，其中，所述关联性最大的字符串用于指示针对所述关键字符串的解释性的字符串。

11. 如权利要求9所述的系统，其特征在于，所述解析模块具体用于：

利用蒙特卡洛方法选取对照样本，并计算所述对照样本中的字符串与所述带有恶意代码的dex文件的字符串之间的欧式距离；

根据所述欧式距离以及预设的欧式距离阈值，从所述带有恶意代码的dex文件的字符串中提取出所述关键字符串。

12. 如权利要求9所述的系统，其特征在于，所述构建模块中的构建恶意样本特征向量为基于样本属性数据，归纳n个样本属性数据，并基于所述n个样本属性数据，计算属性数据的向量，并计算所述属性数据的向量两两相似度矩阵，保留主要向量，累加所有所述主要向量的各维度分量。

13. 如权利要求12所述的系统，其特征在于，所述构建模块中的保留主要向量，累加所有所述主要向量的各维度分量具体为：将两两相似度矩阵之间的差异性大于预设阈值的属性数据的向量进行保留，并将所述保留的属性数据的向量作为所述恶意样本特征向量，其中，针对所述保留的属性数据的向量，对相同的属性数据的向量的各维度分量进行累加。

14. 如权利要求9所述的系统，其特征在于，所述获取模块具体用于：通过获得关联性最大的字符串，找出关联性字符串之间的字符串特征，最终根据所述关联性字符串之间的字符串特征获得恶意代码字符串规则。

15. 如权利要求14所述的系统，其特征在于，所述获取模块通过文档主题生成模型LDA将所述关联性字符串之间的字符串特征进行特征提取以获得所述恶意代码字符串规则。

16. 一种终端设备，其特征在于，包括：

一个或者多个处理器；

存储器；

一个或多个程序，所述一个或者多个程序存储在所述存储器中，当被所述一个或者多个处理器执行时进行如下操作：

将带有恶意代码的dex文件解析成字符串，并根据预设规则从所述字符串中提取出关键字符串；

利用word2vec思想对所述关键字符串进行训练，得到第一训练结果；

通过所述第一训练结果，构建恶意样本特征向量；

根据所述恶意样本特征向量从所述关键字符串中提取预设范围内的字符串，并对所述预设范围内的字符串再次利用word2vec思想进行训练，得到第二训练结果；

基于所述第二训练结果获得恶意代码字符串规则。

基于深度学习方法推断恶意代码规则的方法、系统及设备

[0001] 相关申请的交叉引用

[0002] 本申请要求武汉安天信息技术有限责任公司于2015年11月17日提交的、发明名称为“基于深度学习方法推断恶意代码规则的方法及系统”的、中国专利申请号“201510787438.3”的优先权。

技术领域

[0003] 本发明涉及移动网络安全技术领域，尤其涉及基于深度学习方法推断恶意代码规则的方法及系统。

背景技术

[0004] 近年来，随着Android系统的流行，针对Android平台的攻击也日益增加。2016年第三季度，以安天AVLSDK服务为基础的中国绝大多数ROM安全扫描截获的新增恶意软件包每天超过10万个，其中Android平台的恶意软件包占据了总数的92%，呈连续递增趋势。Android系统面临着严重的安全威胁。与此同时，针对智能手机的安全研究也成为了全球安全研究的重点。M.Miettinen和P.Halonen对移动智能设备面临的安全威胁以及智能设备安全检测中的主要挑战和不足做了详尽的分析。Abhijit Bose等人在论文中提出了一个全新的智能手机异常检测模型，与M.Miettinen和P.Halonen最大的不同是，该模型采用的异常检测对象是智能手机上正在运行的应用，他们采用基于因果关系的时间逻辑描述智能手机应用的行为模式，并使用SVM机器学习方法进行异常检测分析。可以看到现有的智能手机安全研究基本集中在通用的基于行为模式的异常检测。对于特定智能手机平台，例如Android，并没有研究人员对该平台的恶意软件行为模式（或称为恶意代码规则）进行研究和总结。

[0005] 可以理解的，基于海量的新增样本，充分解决筛选问题成了最重要的事情，从工程经验来看，一般先利用可靠的本地病毒检测引擎（后文简称“引擎”）解决筛选问题，即将已经发现的规则和样本进行过滤，检出已知恶意样本；而剩下的引擎难以判断的样本，一般还需要通过机器学习或人工分析再检测一次。从长远看来，这种检测方式效率低下，对于难以判断的恶意样本，应该通过科学合理的技术手段对其进行分析，以推断出能够优化现有引擎的恶意代码规则。可以理解的，引擎内包含的恶意代码检测规则越多则检测效率越高。另外，用于优化引擎的好的规则，一般还希望能同时满足两个指标：(1) 误报率不能高，即提取的规则不能太宽泛；(2) 覆盖率要高，即提取的规则要充分覆盖疑似样本集。

发明内容

[0006] 针对上述技术问题，本发明提供了基于深度学习方法推断恶意代码规则的方法及系统，能充分利用恶意样本的特征推断出误报率低、覆盖率高的恶意代码规则，以优化现有病毒检测引擎，提升恶意代码检测效率。

[0007] 本发明提出了一种基于深度学习方法推断恶意代码规则的方法，包括：

- [0008] 将带有恶意代码的dex文件解析成字符串，并根据预设规则从所述字符串中提取出关键字符串；
- [0009] 利用word2vec思想对所述关键字符串进行训练，得到第一训练结果；
- [0010] 通过所述第一训练结果，构建恶意样本特征向量；
- [0011] 根据所述恶意样本特征向量从所述关键字符串中提取预设范围内的字符串，并对所述预设范围内的字符串再次利用word2vec思想进行训练，得到第二训练结果；
- [0012] 基于所述第二训练结果获得恶意代码字符串规则。
- [0013] 进一步的，所述利用word2vec思想对所述关键字符串进行训练，具体为：对所述关键字符串进行特征提取，并根据所述特征推断出关联性最大的字符串，其中，所述关联性最大的字符串用于指示针对所述关键字符串的解释性的字符串。
- [0014] 进一步的，所述根据预设规则从所述字符串中提取出关键字符串，包括：选取对照样本，并计算所述对照样本中的字符串与所述带有恶意代码的dex文件的字符串之间的距离；根据所述距离以及预设的距离阈值，从所述带有恶意代码的dex文件的字符串中提取出所述关键字符串。
- [0015] 进一步的，所述关键字符串包括函数调用关系和代码结构上下文中的用于描述所述dex文件内容字符串关系的内容。
- [0016] 进一步的，所述构建恶意样本特征向量为：基于样本属性数据，归纳n个样本属性数据，并基于所述n个样本属性数据，计算属性数据的向量，并计算所述属性数据的向量两两相似度矩阵，保留主要向量，累加所有所述主要向量的各维度分量。
- [0017] 进一步的，所述保留主要向量，累加所有所述主要向量的各维度分量，包括：将两两相似度矩阵之间的差异性大于预设阈值的属性数据的向量进行保留；将所述保留的属性数据的向量作为所述恶意样本特征向量，其中，针对所述保留的属性数据的向量，对相同的属性数据的向量的各维度分量进行累加。
- [0018] 进一步的，所述基于训练结果获得恶意代码字符串规则，具体为：通过获得关联性最大的字符串，找出关联性字符串之间的字符串特征，最终根据所述关联性字符串之间的字符串特征获得恶意代码字符串规则。
- [0019] 进一步的，通过文档主题生成模型LDA将所述关联性字符串之间的字符串特征进行特征提取以获得所述恶意代码字符串规则。
- [0020] 基于深度学习方法推断恶意代码规则的系统，包括：
- [0021] 解析模块，用于将带有恶意代码的dex文件解析成字符串，并根据预设规则从所述字符串中提取出关键字符串；
- [0022] 训练模块，用于利用word2vec思想对所述关键字符串进行训练，得到第一训练结果；
- [0023] 构建模块，用于通过所述第一训练结果，构建恶意样本特征向量；
- [0024] 再训练模块，用于根据所述恶意样本特征向量从所述关键字符串中提取预设范围内的字符串，并对所述预设范围内的字符串再次利用word2vec思想进行训练，得到第二训练结果；
- [0025] 获取模块，用于基于所述第二训练结果获得恶意代码字符串规则。
- [0026] 进一步的，所述训练模块具体用于：对所述关键字符串进行特征提取，并根据所述

特征推断出关联性最大的字符串,其中,所述关联性最大的字符串用于指示针对所述关键字符串的解释性的字符串。

[0027] 进一步的,所述解析模块具体用于:利用蒙特卡洛方法选取对照样本,并计算所述对照样本中的字符串与所述带有恶意代码的dex文件的字符串之间的欧式距离;根据所述欧式距离以及预设的欧式距离阈值,从所述带有恶意代码的dex文件的字符串中提取出所述关键字符串。

[0028] 进一步的,所述构建模块中的构建恶意样本特征向量为基于样本属性数据,归纳n个样本属性数据,并基于所述n个样本属性数据,计算属性数据的向量,并计算所述属性数据的向量两两相似度矩阵,保留主要向量,累加所有所述主要向量的各维度分量。

[0029] 进一步的,所述构建模块中的保留主要向量,累加所有所述主要向量的各维度分量具体为:将两两相似度矩阵之间的差异性大于预设阈值的属性数据的向量进行保留,并将所述保留的属性数据的向量作为所述恶意样本特征向量,其中,针对所述保留的属性数据的向量,对相同的属性数据的向量的各维度分量进行累加。

[0030] 进一步的,所述获取模块具体用于:通过获得关联性最大的字符串,找出关联性字符串之间的字符串特征,最终根据所述关联性字符串之间的字符串特征获得恶意代码字符串规则。

[0031] 进一步的,所述获取模块通过文档主题生成模型LDA将所述关联性字符串之间的字符串特征进行特征提取以获得所述恶意代码字符串规则。

[0032] 本发明提出了一种终端设备,包括:一个或者多个处理器;存储器;一个或多个程序,所述一个或者多个程序存储在所述存储器中,当被所述一个或者多个处理器执行时进行如下操作:

[0033] 将带有恶意代码的dex文件解析成字符串,并根据预设规则从所述字符串中提取出关键字符串;

[0034] 利用word2vec思想对所述关键字符串进行训练,得到第一训练结果;

[0035] 通过所述第一训练结果,构建恶意样本特征向量;

[0036] 根据所述恶意样本特征向量从所述关键字符串中提取预设范围内的字符串,并对所述预设范围内的字符串再次利用word2vec思想进行训练,得到第二训练结果;

[0037] 基于所述第二训练结果获得恶意代码字符串规则。

[0038] 本发明还提出了一种存储介质,用于存储应用程序,所述应用程序用于执行本发明所述的基于深度学习方法推断恶意代码规则的方法。

[0039] 本发明利用创造性的将word2vec思想应用于恶意代码分析领域,对已知的恶意代码的字符串进行训练,得到恶意代码字符串关联性最大的字符串,进而得到该恶意代码的字符串规则。本发明能充分利用恶意样本的特征推断出误报率低、覆盖率高的恶意代码规则,以优化现有病毒检测引擎,提升恶意代码检测效率。本发明还能应用于恶意代码检测及恶意代码分析的领域。

附图说明

[0040] 为了更清楚地说明本发明的技术方案,下面将对实施例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中记载的一些实施例,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他的附图。

[0041] 图1为本发明提供的基于深度学习方法推断恶意代码规则的方法实施例流程图;

[0042] 图2为本发明提供的一个带有恶意代码的dex文件的示例图;

[0043] 图3为根据本发明一个具体实施例的基于深度学习方法推断恶意代码规则的方法的流程图;

[0044] 图4为本发明提供的基于深度学习方法推断恶意代码规则的系统实施例结构图。

具体实施方式

[0045] 本发明给出了基于深度学习方法推断恶意代码规则的方法及系统。一般的,深度学习方法起源应用于图像识别,基于文本上下文特征提取最近几年才有了部分成果。传统的文本特征提取算法模型为N-gram,但是N-gram存在一个问题,若训练语料里面有些n元组没有出现过,其对应的条件概率就是0,这会导致计算一整句话的概率为0。语料的不足使得无法训练更高阶的语言模型。另外,这种模型无法建模出词之间的相似度。

[0046] 神经网络模型也广受关注,基于神经网络的语言模型在效果上表现的很不错,但是其训练和预测的时间较长,影响实际的应用。

[0047] 综上,为了克服以上缺点,可以采用以word2vec为代表的深度学习文本方法,例如word2vec文本特征提取模型。但是,传统的word2vec都是采用文章分词然后进行上下文关系训练,是不能直接应用于恶意代码分析技术领域的,因此本发明做了创造性的研究。

[0048] 为了使本技术领域的人员更好地理解本发明实施例中的技术方案,并使本发明的上述目的、特征和优点能够更加明显易懂,下面结合附图对本发明中技术方案作进一步详细的说明:

[0049] 本发明首先提供了基于深度学习方法推断恶意代码规则的方法,如图1所示,包括:

[0050] S101,将带有恶意代码的dex文件解析成字符串,并根据预设规则从字符串中提取出关键字符串;其中,该预设规则可以是根据实际需求来设置的,例如,该预设规则可包括:URL地址、具有跳转或调用等功能的代码、包名、类名等。

[0051] 具体地,可以提取带有恶意代码的dex文件中的函数调用关系、代码结构上下文等能够描述dex文件内容字符串关系的内容作为关键字符串。并且还需选择相应数量的对照样本(或称白样本,即官方发布没有恶意代码的正常样本),其中,本发明采用了蒙特卡洛抽样方法(Monte Carlo method,又称统计模拟法)的思想,对带有恶意代码样本(要提取规则的样本)的对照样本进行选择,计算出多组等数量的对照样本以便再接下来的步骤来使用。

[0052] 例如,以预设规则包括URL地址、包名和类名为例,假设如图2所示,为一个带有恶意代码的dex文件,可将该dex文件解析成字符串,并根据该预设规则从该字符串中分别提取出包含有URL地址、包名或类名为的字符串,并将这些包含有URL地址、包名或类名为的字符串作为该关键字符串,即经过提取后得到的关键字符串可如下:

[0053] http://127.0.0.1:8787(内部访问ip,端口)

[0054] http://52.71.240.169/api/pxkj(外部访问ip)

[0055] http://schemas.android.com/apk/res/android(外部访问url)

[0056] http://xmlpull.org/v1/doc/features.html#indent-output(外部访问url)

[0057] com.sams.listviewdemo(包名)

[0058] ListViewDemo(类名)。

[0059] S102,利用word2vec思想对关键字符串进行训练,得到第一训练结果。

[0060] 举例而言,可利用以word2vec为代表的深度学习文本思想对上述从带有恶意代码的dex文件中提取出的关键字符串进行训练,得到带有恶意代码的dex文件的训练结果;同样,将S101中的对照样本进行word2vec训练,得到对照样本的多组训练结果。

[0061] 下面简单解释一下word2vec的产出结果,可以理解,该训练结果可包含了针对关键字符串的解释性的字符串(即可用于解释该关键字符串的字符串),该解释性的字符串即为与该关键字符串关联性最大的字符串。例如,以关键字字符串“孟非”为例,利用word2vec思想可对该“孟非”进行训练,可以得到该“孟非”的解释性的字符串,即与该关键字符串“孟非”关联性最大的字符串可为:非诚勿扰、乐嘉、黄磊、…、黄菡等,并根据权重的降序方式对该解释性的字符串进行排序,以得到该与“孟非”的关联性最大的字符串:{乐嘉,非诚勿扰,黄磊,黄菡,…}。

[0062] 也就是说,基于利用word2vec思想,对关键字符串“孟非”进行特征提取,如提取出该“孟非”的特征为“主持人”、“所主持节目”、“所主持节目中的嘉宾”等,通过这些特征推断出与该关键字符串“孟非”的关联性最大的字符串,并按照关联性从大到小的顺序,将这些字符串进行排序,并将排序后的字符串以一个向量的形式进行表示,从而实现了通过一组与该关键字符串“孟非”的关联性最大的字符串来表示该“孟非”,即基于word2vec思想将关键字符串用向量的形式进行表示。

[0063] 需要说明的是,上述以关键字符串“孟非”为例,仅是为了方便本领域的技术人员能够了解本发明的一种示例,不能作为对本发明的具体限定。

[0064] 同理,举个例子,以如图2所示的带有恶意代码的dex文件为例,假设提取出的关键字符串为“<http://52.71.240.169/api/pxki>(外部访问ip)”,则对该关键字符串进行word2vec训练,生成该关键字符串的向量关系举例如下:

[0065] {52.71.240.169,

[0066] <http://52.71.240.169/api>,

[0067] <http://52.71.240.169/api/iexs>,

[0068] <http://52.71.240.169/api/mksw>,

[0069] <http://52.71.240.168/api/omes>,

[0070]

[0071] }

[0072] S103通过第一训练结果,构建恶意样本特征向量。

[0073] 可以理解,通过上述步骤S101和S102可以将带有恶意代码的dex文件解析成字符串(其中,保留调用关系,代码结构上下文),并利用蒙特卡洛方法选取对照样本(或称白样本)与恶意样本(即上述带有恶意代码的dex文件)做对照,根据对照的欧式距离设定阈值,选取并提取出关键字符串,成为第一次训练和特征提取;

[0074] 基于样本属性数据,归纳n个样本属性数据,并基于该n个样本属性数据,计算属性数据的向量,并计算属性数据的向量两两相似度矩阵,保留主要向量,累加所有主要向量的各维度分量;其中,该样本属性数据包括对照样本属性数据和恶意样本属性数据,该带有恶

意代码的样本属性数据即为从上述带有恶意代码的dex文件中提取出的关键字符串；而该对照样本属性数据为从合法文件中提取出的关键字符串。

[0075] 作为一种示例，该对照样本属性数据可通过以下步骤获得：获取上述dex文件所对应的官方发布的dex文件，该官方发布的dex文件可认为是合法文件，并将该合法文件解析成字符串，并根据预设规则从该字符串中提取出关键字符串，将该从合法文件中提取出的关键字符串作为对照样本的属性数据。

[0076] 举例而言，假设上述n为10，针对上述对照样本属性数据和恶意样本属性数据，分别归纳出10个对照样本属性数据和10个恶意样本属性数据，并基于10个对照样本属性数据和10个恶意样本属性数据，分别计算10个对照样本中属性数据的向量两两相似度矩阵，并分别计算10个恶意样本中属性数据的向量两两相似度矩阵，针对同一个属性数据，比较对照样本中该属性数据的向量两两相似度矩阵与恶意样本中该属性数据的向量两两相似度矩阵之间的差异性，并将该差异性比较大（例如大于预设阈值）的属性数据的向量进行保留，之后，可针对这些保留的属性数据的向量，对相同的属性数据进行向量的各维度分量进行累加，最后，将这些差异性比较大的属性数据的向量作为恶意样本特征向量。

[0077] 作为一种示例，对恶意样本属性数据和对照样本属性数据进行差异性对比，并根据该差异性对比结果从恶意样本属性数据中提取最重要的n个词，做特征提取以获取下述预设范围内的字符串（即敏感字符串），具体实现过程可如下：如果对照样本集合和恶意样本集合同时出现某个字符串时，计算该字符串的两个向量的欧氏距离，设置一个阈值n（暂定n=0.7），如果该欧氏距离大于阈值n，那么可认为该字符串对于两个集合差异性大，那么该词为特征词；如果一个字符串存在于对照样本，而不存在于恶意样本，那么舍弃该字符串；如果一个字符串存在于恶意样本，而不存在于对照样本，那么该字符为特征词，距离定为1。

[0078] S104根据恶意样本特征向量从关键字符串中提取预设范围内的字符串，并对预设范围内的字符串再次利用word2vec思想进行训练，得到第二训练结果。

[0079] 根据所述恶意样本特征向量从多个含有恶意代码的样本中分别在所述关键字符串中提取预设范围内（第一次训练结果）的字符串，并对所述预设范围内的字符串再次利用word2vec思想进行训练，得到第二训练结果，对所述关键字符串进行特征提取。

[0080] 其中，该预设规则即上述设定的欧式距离阈值。

[0081] S105基于第二训练结果获得恶意代码字符串规则；

[0082] 通过获得关联性最大的字符串，找出关联性最大的字符串之间的字符串特征，最终根据关联性最大的字符串之间的字符串特征获得恶意代码字符串规则。其中，所述关联性最大的字符串是指该向量在跟该样本向量的欧式距离上小于其他的n-1个样本向量。通过获得与某个恶意代码样本关联性最大的字符串（多个），找出关联性字符串之间的字符串特征，最终根据所述关联性字符串之间的字符串特征获得恶意代码字符串规则。

[0083] 作为一种示例，根据关联性最大的字符串之间的字符串特征，通过人工规则的提取方式从该dex文件中提取出恶意代码字符串规则。例如，以如图2所示的关键字符串为例，在利用word2vec思想对这些关键字符串进行训练，并通过第一次训练结果，构建恶意样本特征向量，并根据恶意样本特征向量从关键字符串中提取预设范围内的字符串，并对预设范围内的字符串再次利用word2vec思想进行训练，最后基于该第二次训练结果，通过人工

规则提取方式,获得恶意代码字符串规则为:`*52.71.240.169*OR*52.71.240.168*OR*52.71.240.169*OR*abc_btn_radio*`。其中,“*”表示模糊匹配;该规则“`*52.71.240.169*`”可理解为包含“`52.71.240.169`”的字符串均视为恶意代码字符串。

[0084] 综上,本发明先将带有恶意代码的dex文件解析成字符串,并根据预设规则从字符串中提取出关键字符串,之后,可利用word2vec思想对该这些关键字符串进行训练,以将这些关键字符串用向量进行表示,然后将这些关键字符串的向量作为恶意样本属性数据,基于恶意样本属性数据和白样本属性数据,归纳n个恶意样本属性数据和对照样本属性数据,并基于n个恶意样本属性数据和对照样本属性数据,计算属性数据的向量,并计算属性数据的向量两两相似度矩阵,保留主要向量,累加所有向量的各维度分量,从而构建出恶意样本特征向量,即该恶意样本特征向量即为差异性比较大的属性数据的向量。然后,根据恶意样本特征向量从关键字符串中提取预设范围内的字符串,并对该预设范围内的字符串再次利用word2vec思想进行训练,得到第二训练结果,即得到敏感字符串,最后基于这些敏感字符串,通过规则提取方式,从该dex文件中提取出恶意代码字符串规则。可以理解的,从该dex文件中提取出恶意代码字符串规则的方法可以应用于恶意代码检测及恶意代码分析领域。

[0085] 图3为根据本发明一个具体实施例的基于深度学习方法推断恶意代码规则的方法的流程图。其中,假设有N个带有恶意代码的dex文件,并将该带有恶意代码的dex文件称为恶意样本,可以理解,每个恶意样本应具有对应的对照样本,该对照样本即为官方发布没有恶意代码的正常样本。

[0086] 如图3所示,该基于深度学习方法推断恶意代码规则的方法可以包括:

[0087] S301,根据N个恶意样本选择N*M个对照样本,其中,N和M分别为大于0的整数。

[0088] 作为一种示例,可以采用蒙特卡洛抽样的思想来选择该对照样本。

[0089] S302,分别将各恶意样本和对照样本的dex文件解析成字符串,以N个一组为单位,即M+1个样本簇,提取出各样本(包括恶意样本和对照样本)的函数调用关系、代码结构上下文等能够描述样本内容字符串关系的内容(即恶意样本中的关键字符串和对照样本中的关键字符串)。

[0090] S303,分别对各恶意样本簇的函数调用关系、代码结构上下文(即恶意样本中的关键字符串)进行word2vec训练,按照第一预设规则提取各样本簇的特征向量,生成恶意样本簇的T1个(通常会有10000+个)字符串的特征向量(特征向量的维度数量可以限制在20)集合。即,对整个恶意样本簇做训练得到T1个字符串特制向量。

[0091] S304,分别对各对照样本簇的函数调用关系、代码结构上下文(即对照样本中的关键字符串)进行word2vec训练,按照第一预设规则提取各样本的特征向量,生成对照样本簇(一共有M份)的T2个字符串的特征向量(不一定跟T1数量相同,因为样本结构不一样,M个簇的T2都不一样)。即对整个对照样本簇做训练,M簇即是做了M次训练,然后将训练结合合起来,得到T2个字符串特制向量。

[0092] 优选地,可以将M个对照样本簇的相同的字符串的特征向量计算平均特征向量,作为对照样本的字符串的特征向量的代表。

[0093] S305,将恶意样本簇的T1个特征向量分别与对照样本簇T2个的特征向量进行相似度计算,按照第二预设规则找到恶意样本簇中与对照样本簇内特征向量相似度低的字符串,如果该字符串与对照样本中的差别大,则可以认为该相似度低的字符串为敏感字符串,

恶意代码的规则就从该敏感字符串中提取的。

[0094] 其中,上述计算相似度的方法可包括但不限于欧式距离、曼哈顿距离等。

[0095] 另外,由于对照样本簇中具有M组数据,所以可以对这M组分别设置权重计算相似度,累加所有相似度的值得到恶意样本簇中各特征向量的相似度。

[0096] 作为一种示例,上述确定相似度低可以采用设置阈值的方法,例如,可预先设定一个阈值,通过计算欧式距离,如果恶意样本簇中与对照样本簇内特征向量相似度小于该阈值,则可确定该特征向量相似度低。

[0097] 至此,即步骤S301-S305主要完成的是从N个恶意样本中找到敏感字符串的工作。

[0098] 下面就开始从该敏感字符串内提取恶意代码字符串的规则:

[0099] S306,将属于恶意样本簇的N个样本分别单独做word2vec训练(即对每个样本做训练),并且将训练出来的字符串结果属于上述步骤S305中的字符串进行过滤,即找出训练结果中属于步骤S305中敏感字符串的特征向量,并对找出的特征向量进行特征提取,最终得到恶意代码规则。

[0100] 作为一种示例,上述对找出的特征向量进行特征提取的具体实现过程可如下:可以将上述找出的特征向量看做一个正常的文档,采用LDA(Latent Dirichlet Allocation是一种文档主题生成模型,也称为一个三层贝叶斯概率模型)方法进行特征提取。

[0101] 与上述几种实施例提供的基于深度学习方法推断恶意代码规则的方法相对应,本发明的一种实施例还提供一种基于深度学习方法推断恶意代码规则的系统,由于本发明实施例提供的基于深度学习方法推断恶意代码规则的系统与上述几种实施例提供的基于深度学习方法推断恶意代码规则的方法相对应,因此在前述基于深度学习方法推断恶意代码规则的方法的实施方式也适用于本实施例提供的基于深度学习方法推断恶意代码规则的系统,在本实施例中不再详细描述。图4为本发明提供的基于深度学习方法推断恶意代码规则的系统实施例结构图。如图4所示,该基于深度学习方法推断恶意代码规则的系统包括:

[0102] 解析模块401,用于将带有恶意代码的dex文件解析成字符串,并提取出关键字符串;其中,关键字符串包括函数调用关系和代码结构上下文中的用于描述所述dex文件内容字符串关系的内容。

[0103] 作为一种示例,解析模块401提取出关键字符串的具体实现过程可如下:利用蒙特卡洛方法选取对照样本,并计算对照样本中的字符串与带有恶意代码的dex文件的字符串之间的欧式距离;根据欧式距离以及预设的欧式距离阈值,从带有恶意代码的dex文件的字符串中提取出关键字符串。

[0104] 训练模块402,用于利用word2vec思想对关键字符串进行训练,得到第一训练结果;作为一种示例,训练模块402可对关键字符串进行特征提取,并根据特征推断出关联性最大的字符串,其中,关联性最大的字符串用于指示针对关键字符串的解释性的字符串。

[0105] 构建模块403,用于通过第一训练结果,构建恶意样本特征向量;作为一种示例,构建模块403可基于样本属性数据,归纳n个样本属性数据,并基于n个样本属性数据,计算属性数据的向量,并计算属性数据的向量两两相似度矩阵,保留主要向量,累加所有主要向量的各维度分量。

[0106] 作为一种示例,构建模块403中的保留主要向量,累加所有主要向量的各维度分量具体为:将两两相似度矩阵之间的差异性大于预设阈值的属性数据的向量进行保留,并将

保留的属性数据的向量作为恶意样本特征向量,其中,针对保留的属性数据的向量,对相同的属性数据的向量的各维度分量进行累加。

[0107] 再训练模块404,用于根据恶意样本特征向量从关键字符串中提取预设范围内的字符串,并对该预设范围内的字符串再次利用word2vec思想进行训练,得到第二训练结果;

[0108] 获取模块405,用于基于第二训练结果获得恶意代码字符串规则。作为一种示例,获取模块405可通过获得关联性最大的字符串,找出关联性字符串之间的字符串特征,最终根据关联性字符串之间的字符串特征获得恶意代码字符串规则。

[0109] 具体而言,在本发明的一个实施例中,获取模块405可通过文档主题生成模型LDA将关联性字符串之间的字符串特征进行特征提取以获得恶意代码字符串规则。

[0110] 为了实现上述实施例,本发明还提出了一种终端设备,包括:一个或者多个处理器;存储器;一个或多个程序,一个或者多个程序存储在存储器中,当被一个或者多个处理器执行时进行如下操作:

[0111] S101',将带有恶意代码的dex文件解析成字符串,并根据预设规则从字符串中提取出关键字符串。

[0112] S102',利用word2vec思想对关键字符串进行训练,得到第一训练结果。

[0113] S103',通过第一训练结果,构建恶意样本特征向量。

[0114] S104',根据恶意样本特征向量从关键字符串中提取预设范围内的字符串,并对预设范围内的字符串再次利用word2vec思想进行训练,得到第二训练结果。

[0115] S105',基于第二训练结果获得恶意代码字符串规则。

[0116] 为了实现上述实施例,本发明还提出了一种存储介质,用于存储应用程序,该应用程序用于执行本发明上述实施例所述的基于深度学习方法推断恶意代码规则的方法。

[0117] 综上所述,本发明涉及基于深度学习方法推断恶意代码规则的方法,本方法的核心是基于word2vec思想,利用深度学习方法对已知的恶意代码的字符串进行两次训练,得到恶意代码字符串关联性最大的字符串,进而得到该恶意代码的字符串规则,最终获得恶意样本的关联性。本发明能充分利用恶意样本的特征推断出误报率低、覆盖率高的恶意代码规则,以优化现有病毒检测引擎,提升恶意代码检测效率,还能应用于恶意代码分析。

[0118] 以上实施例用以说明而非限制本发明的技术方案。不脱离本发明精神和范围的任何修改或局部替换,均应涵盖在本发明的权利要求范围当中。

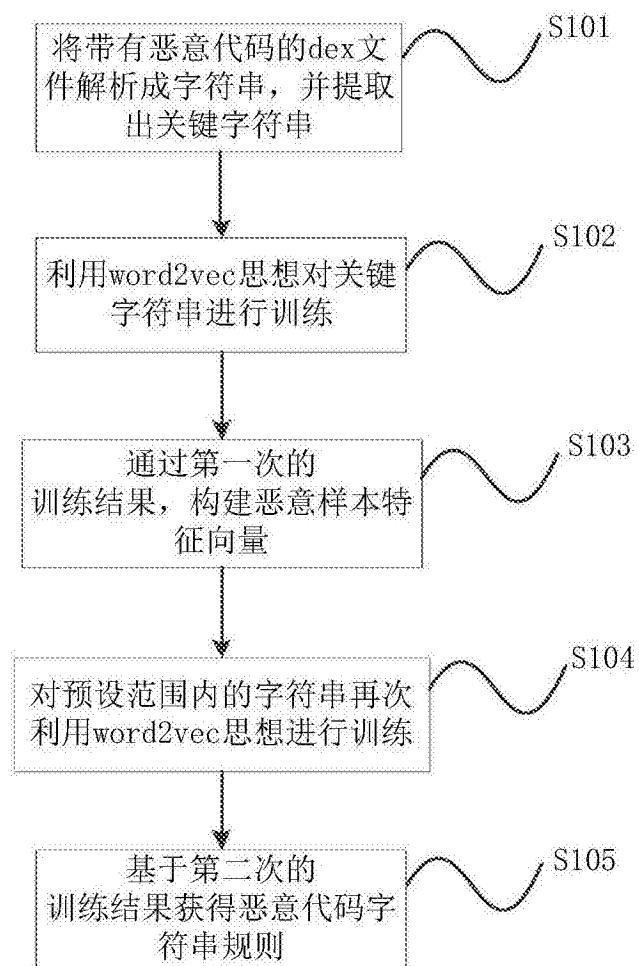


图1

http://127.0.0.1:8787
http://52.71.240.169/api/pxkj
http://52.71.240.169/api/iexs
http://52.71.240.169/api/mksw
http://52.71.240.168/api/omes
http://schemas.android.com/apk/res/android
http://xmlpull.org/v1/doc/features.html#indent-output
httpCode
https://myhosts.sinaapp.com/
blacklist.txtvirtual.android.intent.action.PACKAGE_ADDED
virtual.android.intent.action.PACKAGE_CHANGED
virtual.android.intent.action.PACKAGE_REMOVED
abc_btn_default_mtrl_shape
abc_btn_radio_material
abc_btn_radio_to_on_mtrl_000
com.sams.listviewdemo
ListViewDemo

图2

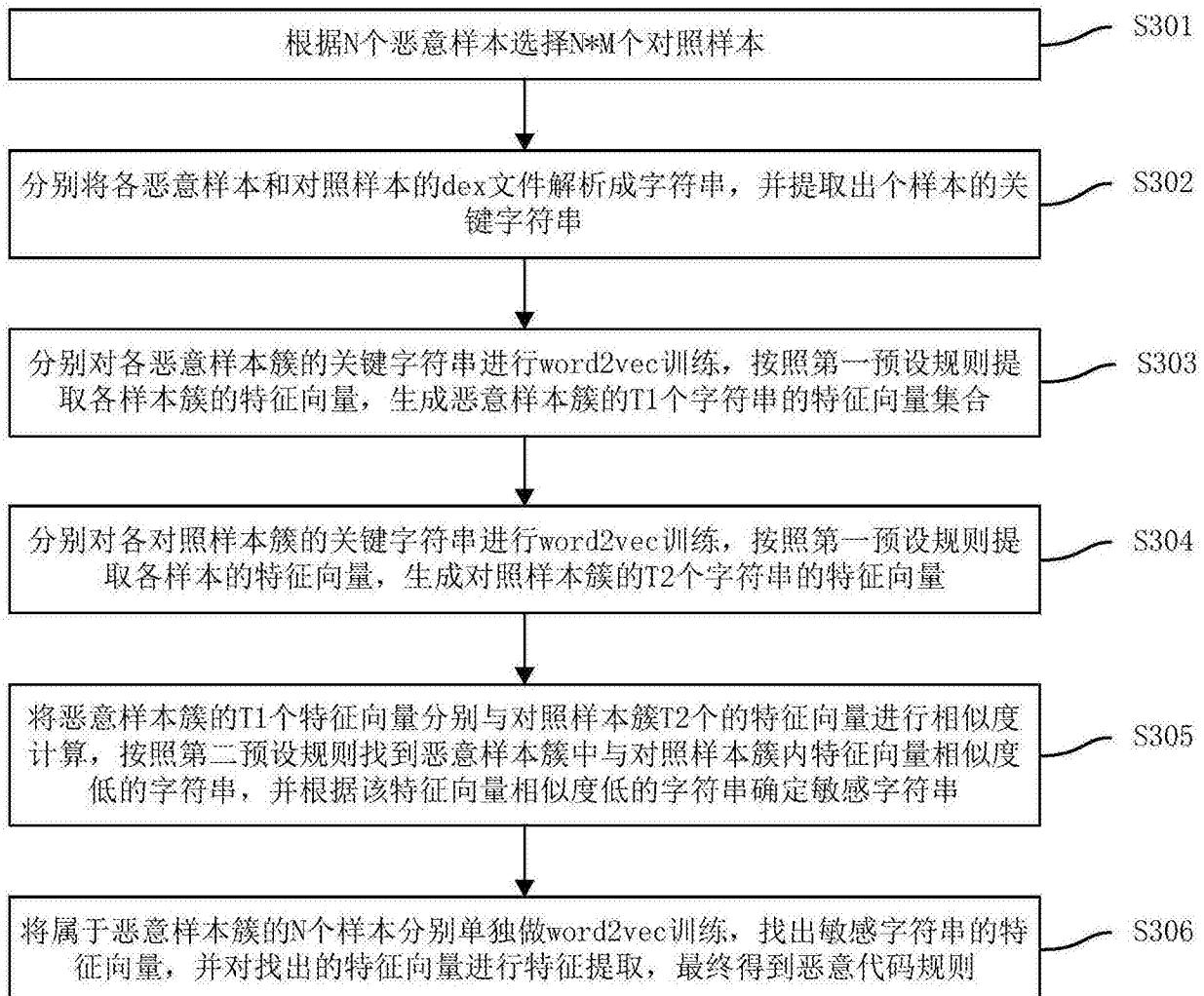


图3

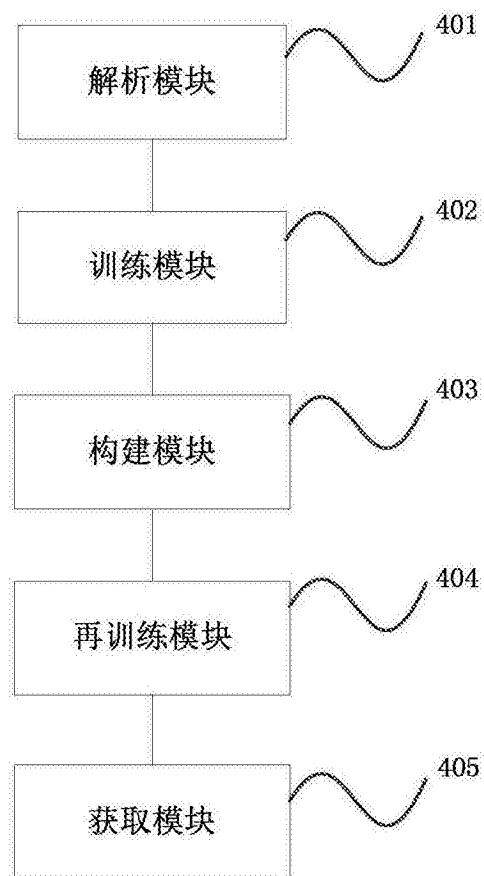


图4