

一种训练机器学习引擎的方法、系统及相关装置

本申请要求于 2018 年 6 月 20 日提交中国专利局、申请号为 201810638258.2、发明名称为“一种训练机器学习引擎的方法、系统及相关装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及机器学习领域，更具体地说，涉及一种训练机器学习引擎的方法、系统、一种计算机可读存储介质及一种训练机器学习引擎的装置。

背景技术

10 大数据和机器学习技术是当前工业界和学术界研究的热点，机器学习和大数据技术被越来越多的用于了信息安全方面，并取得了不错的效果。其中，基于机器学习的恶意软件检测技术取得长足的进步，业界也涌现了一大批基于机器学习做恶意软件检测的公司，比如 Cylance 等，这些新兴公司颇受 VC 的青睐，也取得一些很好的成果。

15 目前的机器学习引擎一般采用软件的信息熵来作为机器学习的静态特征，通常的做法是对可执行文件的主要特征进行信息熵提取操作，将提取到的结果作为输入机器学习引擎的静态特征。但是由于这种对可执行文件的主要特征进行信息熵提取的操作是指仅对可执行文件的导入表、导出表以及节信息进行信息熵提取，这些特征比较通用，因此得到的样本特征颗粒度较粗。使用现有技术中得到的样本特征进行机器学习将会导致机器学习引擎的训练效果较差，具有较低的识别率。

20 因此，如何稳定的提取可执行文件的静态特征使得机器学习引擎生成的检测模型具有较强鲁棒性的识别率，是本领域技术人员目前需要解决的技术问题。

25

发明内容

本申请的目的是提供一种训练机器学习引擎的方法、系统、一种计算机可读存储介质及一种训练机器学习引擎的装置，能够稳定的提取可执行

-2-

文件的静态特征使得机器学习引擎生成的检测模型具有较强鲁棒性的识别率。

为解决上述技术问题，本申请提供一种训练机器学习引擎的方法，该方法包括：

- 5 获取可执行文件，并按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分；

提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征；

利用所述静态特征训练机器学习引擎。

- 10 可选的，按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分包括：

按照所述可执行文件的数据目录将所述可执行文件划分为可执行文件头部分、可执行文件节部分、可执行文件 Overlay（覆盖）数据部分和可执行文件整体部分。

- 15 可选的，提取所有所述待提取部分的信息熵特征，将所有所述信息熵特征进行数据变换得到静态特征包括：

提取所有所述待提取部分的信息熵特征；其中，所述信息熵特征的数制为二进制；

- 20 统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流；

对所述熵流进行 DCT 变换或 DWT 变换得到所述静态特征。

可选的，统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流包括：

- 25 将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块；

统计每一所述字节块中每一个字符占所属的字节块中所有字符的字节块比例 p ，并根据公式 $s = |p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量 S ；

-3-

将所有所述熵流分量累加得到所述熵流。

可选的，在获取可执行文件后，还包括：

根据所述可执行文件的通用格式信息确定所述可执行文件的文件格式，并根据所述文件格式选择对应的特征提取策略；

5 相应的，提取所有所述待提取部分的信息熵特征包括：

利用所述特征提取策略提取所有所述待提取部分的信息熵特征。

可选的，根据可执行文件的通用格式信息确定可执行文件的文件格式包括：

确定可执行文件的数据目录中.NET MetaData Directory RVA 和.NET
10 MetaData Directory Size 的值；当.NET MetaData Directory RVA 和.NET MetaData Directory Size 的值均不为空时，判定可执行文件的文件格式为.NET 格式。

可选的，按照可执行文件的数据目录将可执行文件划分为多个待提取部分包括：

15 按照可执行文件的 section 数组将可执行文件划分为多个待提取部分。

可选的，在利用静态特征训练机器学习引擎之后，还包括：

当检测到可疑软件时，提取可疑软件的目标可执行文件；

将目标可执行文件输入训练后的机器学习引擎，以便利用训练后的机器学习引擎判断目标可执行文件是否为恶意软件的可执行文件。

20 可选的，还包括：

当训练后的机器学习引擎判定目标可执行文件为恶意软件的可执行文件时，上报可疑软件的软件信息，以便对可疑软件进行人工检测；

接收人工检测结果并根据人工检测结果判断训练后的机器学习引擎的本次判定是否为误判；

25 若是，则将目标可执行文件作为新的负训练样本训练机器学习引擎。

本申请还提供了一种训练机器学习引擎的系统，该系统包括：

文件划分模块，用于获取可执行文件，并按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分；

-4-

静态特征提取模块，用于提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征；

训练模块，用于利用所述静态特征训练机器学习引擎。

可选的，所述文件划分模块包括：

- 5 文件获取单元，用于获取所述可执行文件；

划分单元，用于按照所述可执行文件的数据目录将所述可执行文件划分为可执行文件头部分、可执行文件节部分、可执行文件 Overlay 数据部分和可执行文件整体部分。

可选的，所述静态特征提取模块包括：

- 10 信息熵特征提取单元，用于提取所有所述待提取部分的信息熵特征；
其中，所述信息熵特征的数制为二进制；

熵流统计单元，用于统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流；

- 15 数据变换单元，用于对所述熵流进行 DCT 变换或 DWT 变换得到所述静态特征。

可选的，所述熵流统计单元包括：

分块子单元，用于将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块；

- 20 熵流分量计算子单元，用于统计每一所述字节块中每一个字符占所属的字节块中所有字符的字节块比例 p ，并根据公式 $S = |p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量 S ；

熵流分量累加单元，用于将所有所述熵流分量累加得到所述熵流。

可选的，还包括：

- 25 提取策略确定模块，用于根据所述可执行文件的通用格式信息确定所述可执行文件的文件格式，并根据所述文件格式选择对应的特征提取策略；

相应的，所述静态特征提取模块具体为利用所述特征提取策略提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征的模块。

—5—

可选的，所述提取策略确定模块包括：

文件格式确定单元，用于确定所述可执行文件的数据目录中 .NET MetaData Directory RVA 和 .NET MetaData Directory Size 的值；当所述 .NET MetaData Directory RVA 和所述 .NET MetaData Directory Size 的值均不为空
5 时，判定所述可执行文件的文件格式为 .NET 格式；

策略选择单元，用于根据所述文件格式选择对应的特征提取策略。

可选的，所述文件划分模块具体为用于获取可执行文件并按照所述可执行文件的 section 数组将所述可执行文件划分为多个待提取部分的模块。

可选的，还包括：

10 提取模块，用于当检测到可疑软件时，提取所述可疑软件的目标可执行文件；

判断模块，用于将所述目标可执行文件输入训练后的机器学习引擎，以便利用所述训练后的机器学习引擎判断所述目标可执行文件是否为恶意软件的可执行文件。

15 可选的，还包括：

上报模块，用于当所述训练后的机器学习引擎判定所述目标可执行文件为恶意软件的可执行文件时，上报所述可疑软件的软件信息，以便对所述可疑软件进行人工检测；

20 二次训练模块，用于接收人工检测结果并根据所述人工检测结果判断所述训练后的机器学习引擎的本次判定是否为误判；若是，则将所述目标可执行文件作为新的负训练样本训练所述机器学习引擎。

本申请还提供了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序执行时实现上述训练机器学习引擎的方法执行的步骤。

25 本申请还提供了一种训练机器学习引擎的装置，包括存储器和处理器，所述存储器中存储有计算机程序，所述处理器调用所述存储器中的计算机程序时实现上述训练机器学习引擎的方法执行的步骤。

本申请提供了一种训练机器学习引擎的方法，包括获取可执行文件，并按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部

-6-

分；提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征；利用所述静态特征训练机器学习引擎。

本申请通过将可执行文件划分为多个待提取部分，针对每一个待提取部分都进行信息熵特征提取操作，再将所有的信息熵特征变换得到机器学习引擎训练时所需的静态特征以便进行机器学习引擎的训练。相对于现有技术中只针对可执行文件的主要特征进行信息熵提取的办法，本申请对可执行文件的进行按照数据目录的划分，通过化整为零的方式，针对每个待提取部分进行信息熵特征提取。由于现有技术是针对主要特征进行统一的信息熵特征提取，而这些主要特征比较通用，因此得到的静态特征的粒度较粗。本申请是相当于对可执行文件整体进行信息熵特征提取，进一步由通过将可执行文件划分为多个待提取部分进行提取，能够保证得到的静态特征的粒度较细。本申请能够稳定的提取可执行文件的静态特征使得机器学习引擎生成的检测模型具有较强鲁棒性的识别率。本申请同时还提供了一种训练机器学习引擎的系统、一种计算机可读存储介质和一种训练机器学习引擎的装置，具有上述有益效果，在此不再赘述。

附图说明

为了更清楚地说明本申请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申请的一些实施例，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。

图 1 为本申请实施例所提供的一种训练机器学习引擎的方法的流程图；

图 2 为本申请实施例所提供的一种训练机器学习引擎的方法的提取信息熵特征并变换得到静态特征的流程图；

图 3 为本申请实施例所提供的另一种训练机器学习引擎的方法的流程图；

图 4 为本申请实施例所提供的一种优选的训练机器学习引擎的方法的

流程图;

图5为本申请实施例所提供的一种训练机器学习引擎的系统的结构示意图。

5 具体实施方式

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

下面请参见图 1，图 1 为本申请实施例所提供的一种训练机器学习引擎的方法的流程图。

具体步骤可以包括：

15 S101：获取可执行文件，并按照可执行文件的数据目录将可执行文件划分为多个待提取部分；

其中，本实施例是针对训练机器学习引擎提出的方案，训练机器学习引擎所需要的样本为可执行文件的静态特征。可执行文件是指是可以由操作系统进行加载执行的文件，在不同的操作系统环境下，可执行文件的呈现方式不一样，例如在 Windows 系统下可执行文件的格式为 PE（Portable Executable，可移植的可执行）文件，在 Linux、Mac、Android 等系统中均有其对应的可执行文件格式。此处不对可执行文件所在的系统进行具体的限定，本方案可以对多种系统下的可执行文件进行操作。

25 本步骤是按照可执行文件的数据目录将可执行文件划分为多个待提取部分，具体来说，可以是按照可执行文件结构中的 section 数组来划分，还可以按照可执行文件结构中其他数据目录进行划分，此处不进行具体的限定。例如，可以按照可执行文件的数据目录将可执行文件划分为可执行文件头部分、可执行文件节部分、可执行文件 Overlay 数据部分和可执行文件整体部分。作为一种优选的实施方案，可以在确定可执行文件的通用格

式信息后，根据其相对应的通用格式的某些部分或者某些结构进行待提取部分的划分。

通过本步骤中将可执行文件划分为多个待提取部分的操作，能够将可执行文件按照数据目录细分为多个待提取部分，在进行信息熵特征提取的过程中，可以针对每一待提取部分的具体情况选择特定的特征提取方法，使本方案的信息熵特征的粒度更细，进而增加了静态特征的稳定性。

S102: 提取所有待提取部分的信息熵特征，并将所有信息熵特征进行数据变换得到静态特征；

其中，本步骤是建立在 S101 已经将可执行文件划分为多个待提取部分的基础上，提取每一个待提取部分的信息熵特征。信息熵是指某种特定信息的出现概率，可以将信息熵和热力学熵进行类比。在本实施例中提到的信息熵特征就是待提取部分中各个信息出现的概率，通常来讲在计算机中信息是以字符的形式存在的，本步骤中提到的信息熵特征可以认为是可执行文件中关于各个字符出现概率的特征值。

可以理解的是，机器学习引擎在进行训练时所采用的样本为可执行文件的静态特征，本步骤通过数据变换将信息熵特征转化为静态特征，以便下一步进行机器学习引擎的训练。

S103: 利用静态特征训练机器学习引擎；

本实施例通过将可执行文件划分为多个待提取部分，针对每一个待提取部分都进行信息熵特征提取操作，再将所有的信息熵特征变换得到机器学习引擎训练时所需的静态特征以便进行机器学习引擎的训练。相对于现有技术中只针对可执行文件的主要特征进行信息熵提取的办法，本实施例对可执行文件的进行按照数据目录的划分，通过化整为零的方式，针对每个待提取部分进行信息熵特征提取。由于现有技术是针对主要特征进行统一的信息熵特征提取，而这些主要特征比较通用，因此得到的静态特征的粒度较粗。本实施例是相当于对可执行文件整体进行信息熵特征提取，进一步由通过将可执行文件划分为多个待提取部分进行提取，能够保证得到的静态特征的粒度较细。本实施例能够稳定的提取可执行文件的静态特征使得机器学习引擎生成的检测模型具有较强鲁棒性的识别率。

下面请参见图 2，图 2 为本申请实施例所提供的一种训练机器学习引擎的方法的提取信息熵特征并变换得到静态特征的流程图。

提取信息熵特征并变换得到静态特征的方式包括但不限于下述方式：

5 S201：提取所有待提取部分的信息熵特征；其中，所述信息熵特征的数制为二进制；

S202：统计信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到信息熵特征的熵流；

10 其中，在本方案中可以将信息熵和热力学熵进行类比，故在本方案中将借助信息流的净熵传递称为熵流。

在本步骤中仅仅是对二进制数据构成的信息熵特征进行了整体的熵流计算，作为一种优选的实施方案，可以将该二进制数据进行一定程度的分块操作，更加便于熵流的计算。例如：首先将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块；再统计每一所述字节块
15 中每一个字符占所属的字节块中所有字符的字节块比例 p ，并根据公式 $S = |p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量 S ；最后将所有所述熵流分量累加得到所述熵流。

S203：对熵流进行 DCT 变换或 DWT 变换得到静态特征。

20 其中，DCT (Discrete Cosine Transform，离散余弦变换) 是与傅里叶变换相关的一种变换，DWT (Discrete Walsh Transform，离散沃尔什变换) 是一种以沃尔什函数为基本函数的一种非正弦正交变换，DCT 变换或 DWT 变换可以将熵流变换为静态特征。

25 下面请参见图 3，图 3 为本申请实施例所提供的另一种训练机器学习引擎的方法的流程图。

具体步骤可以包括：

S301：根据可执行文件的通用格式信息确定可执行文件的文件格式，并根据文件格式选择对应的特征提取策略；

本步骤的目的在于根据可执行文件的通用格式信息，确定可执行文件

的具体小类（如为.NET 格式、Py2exe 格式、auotit 格式、Nsis 格式等），针对可执行文件所述的文件格式确定特征提取策略（如确定提取哪些特征）。

下面针对确定文件格式进行举例：在确定可执行文件的文件格式时，
5 可以采用检测可执行文件数据目录中名称为 “.NET MetaData Directory RVA” 和 “.NET MetaData Directory Size” 的值是否为空，如果不为空则为.NET 程序；还可以采用查看可执行文件中是否包含 python 运行时库的引用的方式，比如检查可执行文件导入表是否包含 python*.dll 的方式或者其它方式（资源中是否有 python 编译后的代码或者 python 代码等）判断是
10 否为 Py2exe 格式的可执行文件。当然，不同文件格式的可执行文件有不同的确定方法，此处不对如何确定文件格式的具体方法进行限定，本领域的技术人员可以根据实际应用情况进行设置。

S302：根据特征提取策略提取可执行文件的静态特征；

其中，尽管不同特征提取策略是不同的，但是提取静态特征的目的可
15 以是：把文档里面规定的结构、元数据都提取出来，按照定义的结构逐个将属性读取和解析出来，然后将数值型的数据转换到某个范围中，字符串型数据的截取一定长度、替换一些特殊字符等操作。

S303：利用静态特征训练机器学习引擎；

通过本实施中的方案能够获取粒度较细的静态特征，可以较好地训练
20 机器学习引擎，提高及其学习引擎的识别率。当然，本实施例中的操作可以与第一个实施例或第二个实施例相结合，使用本实施例中根据所述文件格式选择对应的特征提取策略对前面两个实施例中信息熵特征进行提取操作。即，利用特征提取策略提取所述待提取部分的信息熵特征。

25 下面请参见图 4，图 4 为本申请实施例所提供的一种优选的训练机器学习引擎的方法的流程图。

S401：获取可执行文件，并按照可执行文件的数据目录将可执行文件划分为多个待提取部分；

S402：根据可执行文件的通用格式信息确定可执行文件的文件格式，

并根据文件格式选择对应的特征提取策略;

S403: 利用特征提取策略提取所有所述待提取部分的信息熵特征; 其中, 所述信息熵特征的数制为二进制;

5 S404: 将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块;

S405: 统计每一字节块中每一个字符占所属的字节块中所有字符的字节块比例 p , 并根据公式 $S = |p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量 S ;

S406: 将所有熵流分量累加得到所述熵流;

10 S407: 对熵流进行 DCT 变换或 DWT 变换得到所述静态特征;

S408: 利用静态特征训练机器学习引擎。

在利用所述静态特征训练机器学习引擎之后, 还可以包括以下操作:

步骤 1: 当检测到可疑软件时, 提取所述可疑软件的目标可执行文件;

15 步骤 2: 将所述目标可执行文件输入训练后的机器学习引擎, 以便利用所述训练后的机器学习引擎判断所述目标可执行文件是否为恶意软件的可执行文件。

步骤 3: 当所述训练后的机器学习引擎判定所述目标可执行文件为恶意软件的可执行文件时, 上报所述可疑软件的软件信息, 以便对所述可疑软件进行人工检测;

20 步骤 4: 接收人工检测结果并根据所述人工检测结果判断所述训练后的机器学习引擎的本次判定是否为误判; 若是, 则将所述目标可执行文件作为新的负训练样本训练所述机器学习引擎。

上述实施例是在机器学习引擎训练完毕后使用机器学习引擎监测恶意软件的操作, 由于可能存在训练后机器学习引擎的检测准确率不高的情况, 25 本实施例在机器学习引擎检测到恶意软件后, 上报恶意软件进行人工检测 (即复检)。若根据人工检测判定该可疑软件不为恶意软件, 则说明机器学习引擎出现了误检, 因此可以将可疑软件的目标可执行文件作为新的负训练样本再次训练机器学习引擎, 以提升检测准确率。训练样本可以分为正样本和负样本, 正样本指恶意软件的可执行文件对应的样本, 负样本指非

恶意软件的可执行文件对应的样本。

请参见图 5，图 5 为本申请实施例所提供的一种训练机器学习引擎的系统的结构示意图；

5 该系统可以包括：

文件划分模块 100，用于获取可执行文件，并按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分；

静态特征提取模块 200，用于提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征；

10 训练模块 300，用于利用所述静态特征训练机器学习引擎。

可选的，所述文件划分模块 100 包括：

文件获取单元，用于获取所述可执行文件；

划分单元，用于按照所述可执行文件的数据目录将所述可执行文件划分为可执行文件头部分、可执行文件节部分、可执行文件 Overlay 数据部分和可执行文件整体部分。

15 可选的，所述静态特征提取模块 200 包括：

信息熵特征提取单元，用于提取所有所述待提取部分的信息熵特征；

其中，所述信息熵特征的数制为二进制；

熵流统计单元，用于统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流；

数据变换单元，用于对所述熵流进行 DCT 变换或 DWT 变换得到所述静态特征。

可选的，所述熵流统计单元包括：

25 分块子单元，用于将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块；

熵流分量计算子单元，用于统计每一所述字节块中每一个字符占所属的字节块中所有字符的字节块比例 p ，并根据公式 $|p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量；

熵流分量累加单元，用于将所有所述熵流分量累加得到所述熵流。

可选的，还包括：

提取策略确定模块，用于根据所述可执行文件的通用格式信息确定所述可执行文件的文件格式，并根据所述文件格式选择对应的特征提取策略；

- 5 相应的，所述静态特征提取模块具体为利用所述特征提取策略提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征的模块。

可选的，所述提取策略确定模块包括：

- 10 文件格式确定单元，用于确定所述可执行文件的数据目录中 .NET MetaData Directory RVA 和 .NET MetaData Directory Size 的值；当所述 .NET MetaData Directory RVA 和所述 .NET MetaData Directory Size 的值均不为空时，判定所述可执行文件的文件格式为 .NET 格式；

策略选择单元，用于根据所述文件格式选择对应的特征提取策略。

- 15 可选的，所述文件划分模块具体为用于获取可执行文件并按照所述可执行文件的 section 数组将所述可执行文件划分为多个待提取部分的模块。

可选的，还包括：

提取模块，用于当检测到可疑软件时，提取所述可疑软件的目标可执行文件；

- 20 判断模块，用于将所述目标可执行文件输入训练后的机器学习引擎，以便利用所述训练后的机器学习引擎判断所述目标可执行文件是否为恶意软件的可执行文件。

可选的，还包括：

- 25 上报模块，用于当所述训练后的机器学习引擎判定所述目标可执行文件为恶意软件的可执行文件时，上报所述可疑软件的软件信息，以便对所述可疑软件进行人工检测；

二次训练模块，用于接收人工检测结果并根据所述人工检测结果判断所述训练后的机器学习引擎的本次判定是否为误判；若是，则将所述目标可执行文件作为新的负训练样本训练所述机器学习引擎。

由于系统部分的实施例与方法部分的实施例相互对应，因此系统部分

的实施例请参见方法部分的实施例的描述，这里暂不赘述。

本申请还提供了一种计算机可读存储介质，其上存有计算机程序，该计算机程序被执行时可以实现上述实施例所提供的步骤。该存储介质可以包括：U 盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

本申请还提供了一种训练机器学习引擎的装置，可以包括存储器和处理器，所述存储器中存有计算机程序，所述处理器调用所述存储器中的计算机程序时，可以实现上述实施例所提供的步骤。当然所述训练机器学习引擎的装置还可以包括各种网络接口，电源等组件。

说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的系统而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

还需要说明的是，在本说明书中，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。在没有更多限制的状况下，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。

权 利 要 求

1、一种训练机器学习引擎的方法，其特征在于，包括：

获取可执行文件，并按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分；

5 提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征；

利用所述静态特征训练机器学习引擎。

2、根据权利要求 1 所述方法，其特征在于，按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分包括：

10 按照所述可执行文件的数据目录将所述可执行文件划分为可执行文件头部分、可执行文件节部分、可执行文件 Overlay 数据部分和可执行文件整体部分。

3、根据权利要求 1 所述方法，其特征在于，提取所有所述待提取部分的信息熵特征，将所有所述信息熵特征进行数据变换得到静态特征包括：

15 提取所有所述待提取部分的信息熵特征；其中，所述信息熵特征的数制为二进制；

统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流；

对所述熵流进行 DCT 变换或 DWT 变换得到所述静态特征。

20 4、根据权利要求 3 所述方法，其特征在于，统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流包括：

将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块；

25 统计每一所述字节块中每一个字符占所属的字节块中所有字符的字节块比例 p ，并根据公式 $S = |p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量 S ；

将所有所述熵流分量累加得到所述熵流。

5、根据权利要求 1 所述方法，其特征在于，在获取可执行文件后，还

包括:

根据所述可执行文件的通用格式信息确定所述可执行文件的文件格式, 并根据所述文件格式选择对应的特征提取策略;

相应的, 提取所有所述待提取部分的信息熵特征包括:

5 利用所述特征提取策略提取所有所述待提取部分的信息熵特征。

6、根据权利要求 5 所述方法, 其特征在于, 根据所述可执行文件的通用格式信息确定所述可执行文件的文件格式包括:

确定所述可执行文件的数据目录中 .NET MetaData Directory RVA 和 .NET MetaData Directory Size 的值; 当所述 .NET MetaData Directory RVA
10 和所述 .NET MetaData Directory Size 的值均不为空时, 判定所述可执行文件的文件格式为 .NET 格式。

7、根据权利要求 1 所述方法, 其特征在于, 按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分包括:

按照所述可执行文件的 section 数组将所述可执行文件划分为多个待
15 提取部分。

8、根据权利要求 1 至 7 任一项所述方法, 其特征在于, 在利用所述静态特征训练机器学习引擎之后, 还包括:

当检测到可疑软件时, 提取所述可疑软件的目标可执行文件;

将所述目标可执行文件输入训练后的机器学习引擎, 以便利用所述训练后的机器学习引擎判断所述目标可执行文件是否为恶意软件的可执行文
20 件。

9、根据权利要求 8 所述方法, 其特征在于, 还包括:

当所述训练后的机器学习引擎判定所述目标可执行文件为恶意软件的可执行文件时, 上报所述可疑软件的软件信息, 以便对所述可疑软件进行
25 人工检测;

接收人工检测结果并根据所述人工检测结果判断所述训练后的机器学习引擎的本次判定是否为误判;

若是, 则将所述目标可执行文件作为新的负训练样本训练所述机器学习引擎。

10、一种训练机器学习引擎的系统，其特征在于，包括：

文件划分模块，用于获取可执行文件，并按照所述可执行文件的数据目录将所述可执行文件划分为多个待提取部分；

静态特征提取模块，用于提取所有所述待提取部分的信息熵特征，并将所有所述信息熵特征进行数据变换得到静态特征；

训练模块，用于利用所述静态特征训练机器学习引擎。

11、根据权利要求 10 所述系统，其特征在于，所述文件划分模块包括：文件获取单元，用于获取所述可执行文件；

划分单元，用于按照所述可执行文件的数据目录将所述可执行文件划分为可执行文件头部分、可执行文件节部分、可执行文件 Overlay 数据部分和可执行文件整体部分。

12、根据权利要求 10 所述系统，其特征在于，所述静态特征提取模块包括：

信息熵特征提取单元，用于提取所有所述待提取部分的信息熵特征；

15 其中，所述信息熵特征的数制为二进制；

熵流统计单元，用于统计所述信息熵特征对应的二进制数据中每一个字符占所述二进制数据中所有字符的比例，根据所述比例得到所述信息熵特征的熵流；

数据变换单元，用于对所述熵流进行 DCT 变换或 DWT 变换得到所述静态特征。

13、根据权利要求 12 所述系统，其特征在于，所述熵流统计单元包括：

分块子单元，用于将所述信息熵特征对应的二进制数据按照 256 个字节进行分块得到多个字节块；

熵流分量计算子单元，用于统计每一所述字节块中每一个字符占所属的字节块中所有字符的字节块比例 p ，并根据公式 $S = |p \cdot \log(p)|$ 计算每一字节块中每一个字节的熵流分量 S ；

熵流分量累加单元，用于将所有所述熵流分量累加得到所述熵流。

14、根据权利要求 10 所述系统，其特征在于，还包括：

提取策略确定模块，用于根据所述可执行文件的通用格式信息确定所

述可执行文件的文件格式,并根据所述文件格式选择对应的特征提取策略;

相应的,所述静态特征提取模块具体为利用所述特征提取策略提取所有所述待提取部分的信息熵特征,并将所有所述信息熵特征进行数据变换得到静态特征的模块。

- 5 15、根据权利要求 14 所述系统,其特征在于,所述提取策略确定模块包括:

文件格式确定单元,用于确定所述可执行文件的数据目录中 .NET MetaData Directory RVA 和 .NET MetaData Directory Size 的值;当所述 .NET MetaData Directory RVA 和所述 .NET MetaData Directory Size 的值均不为空

- 10 时,判定所述可执行文件的文件格式为 .NET 格式;

策略选择单元,用于根据所述文件格式选择对应的特征提取策略。

16、根据权利要求 10 所述系统,其特征在于,所述文件划分模块具体为用于获取可执行文件并按照所述可执行文件的 section 数组将所述可执行文件划分为多个待提取部分的模块。

- 15 17、根据权利要求 10 至 16 任一项所述系统,其特征在于,还包括:

提取模块,用于当检测到可疑软件时,提取所述可疑软件的目标可执行文件;

判断模块,用于将所述目标可执行文件输入训练后的机器学习引擎,以便利用所述训练后的机器学习引擎判断所述目标可执行文件是否为恶意软件的可执行文件。

20

- 18、根据权利要求 17 所述系统,其特征在于,还包括:

上报模块,用于当所述训练后的机器学习引擎判定所述目标可执行文件为恶意软件的可执行文件时,上报所述可疑软件的软件信息,以便对所述可疑软件进行人工检测;

- 25 二次训练模块,用于接收人工检测结果并根据所述人工检测结果判断所述训练后的机器学习引擎的本次判定是否为误判;若是,则将所述目标可执行文件作为新的负训练样本训练所述机器学习引擎。

- 19、一种训练机器学习引擎的装置,其特征在于,包括:

存储器,用于存储计算机程序;

处理器，用于执行所述计算机程序时执行如权利要求 1 至 9 任一项所述的训练机器学习引擎的方法的步骤。

- 20、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求 1 至 9 任一项所述的训练机器学习引擎的方法的步骤。