

(74) 专利代理机构 中国专利代理(香港)有限公司
72001

专利代理师 姜冰 张金金

(51) Int.Cl.

(56) 对比文件

CN 103984953 A,2014.08.13
WO 2007092518 A2,2007.08.16
EP 2182458 A1,2010.05.05
US 2016127402 A1,2016.05.05
US 2009287620 A1,2009.11.19

Uri Shaham et al..Understanding Adversarial Training: Increasing Local Stability of Neural Nets through Robust Optimization.《arxiv》.2016,第1-6页.

Marco Barreno.The security of machine learning.《Springer》.2010,第121-146段.

审查员 李召卿

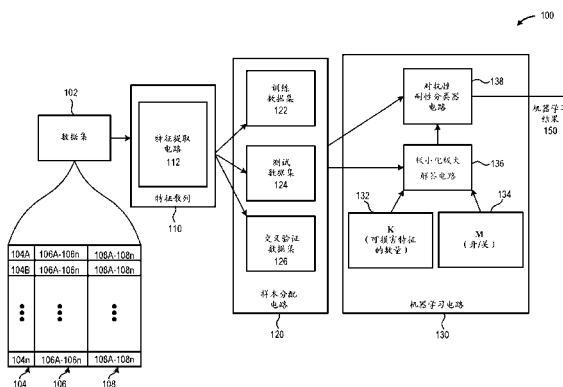
权利要求书3页 说明书12页 附图3页

权利要求书3页 说明书12页 附图3页

权利要求书3页 说明书12页 附图3页

对抗性环境中的机器学习

一种对抗性环境分类器训练系统包括特征提取电路,用于识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征。该系统还包括样本分配电路,用于将包括在初始数据集中的至少一部分样本分配给至少训练数据集;机器学习电路可通信地耦合到样本分配电路,机器学习电路用于:识别至少一部分初始数据集的可损害特征的至少一个集合;定义分类器损失函数 $l(x_i, y_i, w)$,其包括:包括在初始数据集中的每个样本的特征向量 (x_i) ;包括在初始数据集中的每个样本的标签 (y_i) ;以及与分类器相关联的权重向量 (w) ;以及确定分类器损失函数的极小化极大 $(\min_{\max} l(x_i, y_i, w))$ 。



1. 一种用于检测恶意代码的对抗性机器学习系统,包括:

特征提取电路,用于识别与包括在包括多个元素的原始数据集中的每个元素相关联的多个特征;

元素分配电路,用于将包括在所述原始数据集中的至少一部分所述元素分配给至少训练数据集;

机器学习电路,可通信地耦合到所述元素分配电路,所述机器学习电路用于:

识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合;

定义分类器损失函数 $l(x_i, y_i, w)$,所述分类器损失函数包括:

包括在所述训练数据集中的每个元素的特征向量 (x_i) ;

包括在所述训练数据集中的每个元素的标签 (y_i) ;以及

与所述分类器相关联的权重向量 (w) ;以及

确定所述分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

2. 如权利要求1所述的系统,其中所述元素分配电路还包括用于将包括在所述原始数据集中的至少一部分所述元素分配给以下数据集中的至少一个数据集的电路:训练数据集;测试数据集;或交叉验证数据集。

3. 如权利要求1所述的系统,所述机器学习电路用于自主地识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合。

4. 如权利要求1所述的系统,所述机器学习电路用于接收至少一个输入来手动识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合。

5. 如权利要求1所述的系统,所述机器学习电路用于:

如果相应元素表示非恶意元素,则定义包括与元素相关联的标签的第一逻辑值的损失函数;以及

如果相应元素表示恶意元素,则定义包括与元素相关联的标签的第二逻辑值的损失函数。

6. 如权利要求1所述的系统,所述机器学习电路还用于:

识别由包括在所述训练数据集中的至少一部分所述元素的固定数量的可损害特征组成的集合。

7. 如权利要求1所述的系统,所述机器学习电路还用于:

识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的多个集合,所述多个集合中的每个集合包括所述训练数据集中包括的至少一部分所述元素的不同数量的可损害特征。

8. 一种用于检测恶意代码的分类器训练方法,包括:

识别与包括在包括多个元素的训练数据集中的每个元素相关联的多个特征;

识别包括所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合;

定义分类器损失函数 $l(x_i, y_i, w)$,所述分类器损失函数包括:

包括在所述训练数据集中的每个元素的特征向量 (x_i) ;

包括在所述训练数据集中的每个元素的标签 (y_i) ;以及

与所述分类器相关联的权重向量 (w) ;以及

确定所述分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

9. 如权利要求8所述的方法,其中识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合包括:

自主地识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合。

10. 如权利要求8所述的方法,其中识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合包括:手动识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合。

11. 如权利要求8所述的方法,其中定义包括在所述训练数据集中包括的每个元素的标签 (y_i) 的分类器损失函数 [$l(x_i, y_i, w)$] 包括:

如果相应元素表示非恶意元素,则定义包括与元素相关联的标签的第一逻辑值的损失函数;以及

如果相应元素表示恶意元素,则定义包括与元素相关联的标签的第二逻辑值的损失函数。

12. 如权利要求8所述的方法,其中识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合包括:

识别由包括在所述训练数据集中的至少一部分所述元素的固定数量的可损害特征组成的集合。

13. 如权利要求8所述的方法,其中识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合包括:

识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的多个集合,所述多个集合中的每个集合包括所述训练数据集中包括的至少一部分所述元素的不同数量的可损害特征。

14. 一种用于检测恶意代码的分类器训练系统,包括:

用于识别与包括在包括多个元素的训练数据集中的每个元素相关联的多个特征的部件;用于识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合的部件;

用于定义分类器损失函数 [$l(x_i, y_i, w)$] 的部件,所述分类器损失函数包括:

包括在所述训练数据集中的每个元素的特征向量 (x_i);

包括在所述训练数据集中的每个元素的标签 (y_i); 以及

与所述分类器相关联的权重向量 (w); 以及

用于确定所述分类器损失函数的极小化极大的部件 ($\min_w \max_i l(x_i, y_i, w)$)。

15. 如权利要求14所述的系统,其中用于识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合的部件包括:

用于自主识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合的部件。

16. 如权利要求14所述的系统,其中用于识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合的部件包括:

用于手动识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合的部件。

17. 如权利要求14所述的系统,其中用于定义包括在所述训练数据集中包括的每个元素的标签(y_i)的分类器损失函数 $[l(x_i, y_i, w)]$ 的部件包括:

用于如果相应元素表示非恶意元素则定义包括与元素相关联的标签的第一逻辑值的损失函数的部件;以及

用于如果相应元素表示恶意元素则定义包括与元素相关联的标签的第二逻辑值的损失函数的部件。

18. 如权利要求14所述的系统,其中用于识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合的部件包括:

用于识别由包括在所述训练数据集中的至少一部分所述元素的固定数量的可损害特征组成的集合的部件。

19. 如权利要求14所述的系统,其中识别包括在所述训练数据集中的至少一部分所述元素的可损害特征的至少一个集合包括:

用于识别与包括在训练数据集中的至少一部分所述元素的可损害特征的多个集合的部件,所述多个集合中的每个集合包括所述训练数据集中包括的至少一部分所述元素的不同数量的可损害特征。

20. 一种用于训练对抗性机器学习分类器的系统,所述系统布置成执行如权利要求8至13中的任一项所述的方法。

21. 一种布置成执行如权利要求8至13中的任一项所述的方法的芯片集。

22. 至少一种机器可读介质,包括多个指令,所述多个指令响应于被执行在计算装置上而使所述计算装置执行根据权利要求8至13中的任一项所述的方法。

23. 一种配置用于训练对抗性机器学习分类器的装置,所述装置布置成执行如权利要求8至13中的任一项所述的方法。

对抗性环境中的机器学习

技术领域

[0001] 本公开涉及机器学习,并且更特定地涉及对抗性环境中的机器学习。

背景技术

[0002] 网络攻击对于大多数计算机系统表现出越来越大的威胁。机器学习系统提供了用于防范此类网络攻击的相对复杂和强大的工具。通用机器学习工具在自动恶意软件分析中已经显示出成功。然而,这种机器学习工具可能易受针对被采用以训练机器检测恶意软件的学习过程的攻击。例如,对手可以用包含少量的恶意软件的训练示例来“加盐(salt)”或否则“污染”训练数据集,使得训练过的在线系统基于被加盐的或被污染的训练数据集错误地将未来的恶意软件攻击识别为合法活动。攻击者利用学习算法的特定漏洞并小心的操纵训练数据集的能力损害整个机器学习系统。使用被加盐的或被污染的训练数据可能导致训练数据集和任何后续测试数据集之间的差异。

附图说明

[0003] 所要求保护的的主题的各种实施例的特征和优点将随着以下具体实施方式进行并且参考附图而变得显而易见,其中相似的数字表示相似的部分,并且其中:

[0004] 图1描绘了根据本公开的至少一个实施例的说明性机器学习系统,其包括特征散列电路、样本分配电路和机器学习电路;

[0005] 图2提供了根据本公开的至少一个实施例的在其上可以实现机器学习电路和对抗性耐性分类器的说明性系统的框图;以及

[0006] 图3提供了根据本公开的至少一个实施例的用于在潜在的对抗性环境中使用机器学习训练对抗性环境分类器的说明性方法的高级逻辑流程图。

[0007] 尽管以下具体实施方式将用对说明性实施例做出参考来进行,但是其许多替换、修改和变化对本领域技术人员将清楚将是显然的。

具体实施方式

[0008] 本文描述的系统和方法通过最小化用于经由机器学习来训练对抗性环境分类器系统的被加盐或被污染的训练数据的影响来提供更好的系统安全性和性能。从博弈论的角度来看,对抗性环境分类器系统应该成功地预测恶意软件生成器在将恶意代码插入系统时所做的努力。这种对抗性环境分类器系统可以围绕极小化极大算法来设计。极小化极大是在决策理论和博弈论中使用的决策规则,用于最小化最坏情况的可能损失。在博弈论中,极小化极大值是其他玩家在不知道玩家行为的情况下可以强制该玩家接收的最小值。同样地,它是当玩家知道其他玩家的行为时该玩家可以肯定地接收的最大值。

[0009] 在本文公开的系统和方法中,当特征被敌方损害时,分类器的准确性被损害。换句话说,向对抗性环境分类器机器学习系统供应改变的、“被加盐的”或被污染的训练数据损害对抗性环境分类器的后续准确性。本文描述的系统和方法使所有可能受损特征上的最坏

情况损失最小化。这样的系统和方法通过提供降低的总体错误率和对对手更健壮的防御来改善对抗性环境分类器的性能。

[0010] 本文描述的系统和方法考虑到当特征的集合被对手损害时的分类准确度的最坏情况损失。当已知的解决方案可能在其他情况下失败时,本文公开的系统和方法使对抗性逃避的影响最小化。本文描述的系统和方法导致在特征的集合被攻击者损害时优化对抗环境中的最坏情况场景的分类器。元数据可用于定义可能被损害的特征的最大数量。可以使用单个位来允许系统设计者将对抗模式开关切换为开和关。当对抗性模式为开时,对抗性耐性分类器的性能对最差情况准确度损失被优化。当对抗性模式为关时,对抗性耐性分类器用作传统分类器。

[0011] 提供了一种对抗性环境分类器训练系统。对抗性环境分类器训练系统可以包括:特征提取电路,用于识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征;样本分配电路,用于将包括在初始数据集中的至少一部分样本分配给至少一个训练数据集;机器学习电路,可通信地耦合到样本分配电路,机器学习电路用于:识别对于包括在初始数据集中的至少一部分的样本的可损害特征的至少一个集合;定义分类器损失函数 $[l(x_i, y_i, w)]$, 其包括:包含在初始数据集中的每个样本的特征向量 (x_i) ; 包含在初始数据集中的每个样本的标签 (y_i) ; 和与分类器相关联的权重向量 (w) ; 以及确定分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

[0012] 提供了一种对抗性环境分类器训练方法。该方法可以包括:将包括在初始数据集中的多个样本的至少一部分分配给至少一个训练数据集;识别与包括在训练数据集中的每个样本相关联的多个特征;识别对于包括在初始数据集中的至少一部分样本的可损害特征的至少一个集合;定义分类器损失函数 $[l(x_i, y_i, w)]$, 其包括:包括在初始数据集中的每个元素的特征向量 (x_i) ; 包括在初始数据集中的每个元素的标签 (y_i) ; 和与分类器相关联的权重向量 (w) ; 以及确定分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

[0013] 一种存储装置,包括机器可读指令,机器可读指令在被执行时将可配置电路物理地变换为对抗性环境分类器训练电路,所述对抗性环境分类器训练电路用于识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征;将包括在初始数据集中的至少一部分样本分配给至少一个训练数据集;识别对于包括在初始数据集中的至少一部分样本的可损害特征的至少一个集合;定义分类器损失函数 $[l(x_i, y_i, w)]$, 其包括:包括在初始数据集中的每个样本的特征向量 (x_i) ; 包括在初始数据集中的每个样本的标签 (y_i) ; 和与分类器相关联的权重向量 (w) ; 以及确定分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

[0014] 提供了一种对抗性环境分类器训练系统。对抗性环境分类器训练系统可以包括:用于识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征的部件;用于将包括在初始数据集中的至少一部分样本分配给至少一个训练数据集的部件;用于识别对于训练数据集中包括的至少一部分样本的可损害特征的至少一个集合的部件;用于定义分类器损失函数 $[l(x_i, y_i, w)]$ 的部件,其包括:对于包括在初始数据集中的每个样本的特征向量 (x_i) ; 对于包括在初始数据集中的每个样本的标签 (y_i) ; 和与分类器相关联的权重向量 (w) ; 以及确定分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

[0015] 在本文中使用时,术语“顶部”、“底部”、“上”、“下”、“向上”、“向下”、“向上地”、“向

下地”以及类似的方向术语应该以其相对意义而不是绝对意义被理解。因此,如果携带组件的装置被旋转90度,则被描述为“向上地移位”的组件可以被认为是“横向移位”,并且如果携带组件的装置被倒置,则可以认为是“向下地移位”。此类实现应该认为被包括在本公开的范围之内。

[0016] 在本说明书和所附权利要求中使用,单数形式“一个(a/an)”和“该”包括复数对象,除非内容另有明确规定。还应注意的是,术语“或”通常以包括“和/或”的其含义被采用,除非内容另有明确规定。

[0017] 图1描绘了根据本公开的至少一个实施例的说明性对抗性环境分类器机器学习系统100,其包括特征散列电路110、样本分配电路120、和机器学习电路130。特征提取电路112可以形成或包括特征散列电路110的一部分,特征散列电路110接收包括多个样本104A-104n(统称为“样本104”)的初始数据集102。特征散列电路110可以包括特征提取电路112。样本分配电路120可以将从特征提取电路112接收的数据集分配到一个或多个训练数据集122、一个或多个测试数据集124、以及一个或多个交叉验证数据集126中。机器学习电路130包括极小化极大解答电路136和对抗性耐性分类器电路138。机器学习电路至少提供机器学习结果150作为输出。在实施例中,初始数据集102可以源自任何数量的位置。例如,训练数据可以来自McAfee高级威胁检测(ATD)盒的API日志文件或来自McAfee Chimera(Intel Corp.,圣克拉拉,CA)的CF日志的形式被提供。

[0018] 包括在初始数据集102中的每个样本104可以具有与其逻辑上相关联的相应数量的特征106A-106n(统称为“特征106”)。多个这些特征106可以手动或自主地被识别为潜在可损害特征108A-108n(统称为“可损害特征108”)。这种可损害特征108代表有可能被对抗性代理损害的那些特征。变量“K”可被用于表示初始数据集102中存在的这种可损害特征108的最大数量。

[0019] 特征提取电路112生成数据集,其包括在逻辑上与包括在初始数据集102中的每个样本相关联的相应特征向量(x_i ,对于样本 $i = 1 \dots n$,其中 $X_i \in \mathbb{R}^d$, $i = 1, 2, \dots, n$)和标签(y_i ,对于样本 $i = 1 \dots n$)。特征提取电路112可以包括能够形成或以其他方式提供能够识别和/或提取与每个样本104逻辑上相关联的一个或多个特征106A-106n的一个或多个逻辑装置、状态机、处理器、和/或控制器的任何数量和/或组合的电子组件和/或半导体装置。任何当前或未来开发的特征提取技术可以在特征提取电路中被实现。

[0020] 在一些实现中,特征提取电路112可采用一个或多个筛选特征选择方法(诸如,卡方检验、信息增益和相关性得分)来识别初始数据集102中存在的至少一些特征106。这种筛选特征选择方法可以采用一个或多个统计测量来为每个特征应用或以其他方式分配评分。然后通过分数对特征进行排序,并且特征至少部分地基于分数被选择包括在数据集中或被拒绝包括在数据集中。

[0021] 在一些实现中,特征提取电路112可采用一个或多个包装器方法(诸如递归特征消除选择方法)来识别初始数据集102中存在的至少一些特征106。此类包装器特征选择方法将特征的集合的选择视为搜索问题,其中准备、评估和与其他组合比较不同的特征组合。

[0022] 在一些实现中,特征提取电路112可以采用一个或多个嵌入式特征选择方法(诸如LASSO法、弹性网络法或岭回归法)来识别初始数据集102中存在的至少一些特征106。这样的嵌入特征选择方法可以在创建分类器时了解哪些特征106最有助于分类器的准确性。

[0023] 样本分配电路120从特征提取电路112接收数据集,并将接收的数据集的内容分配给一个或多个中:训练数据集122、测试数据集124、和交叉验证数据集126。样本分配电路120可以包括能够形成或以其他方式提供能够将包括在初始数据集102中的样本104分配到多个子集中的一个或多个逻辑装置、状态机、处理器、和/或控制器的任何数量和/或组合的电子组件和/或半导体装置。例如,如图1中所描绘,样本104可以被分配到一个或多个训练数据子集122、一个或多个测试数据子集124、和/或一个或多个交叉验证子集126中。在一些实现中,样本分配电路120可以在一个或多个训练数据子集122、一个或多个测试数据子集124、和/或一个或多个交叉验证子集126中的一些或全部之中随机分配样本。在一些实现中,样本分配电路120可以在一个或多个训练数据子集122、一个或多个测试数据子集124、和/或一个或多个交叉验证子集126中的一些或全部之中均匀地或不均匀地分配样本。

[0024] 训练数据集122可以被供应给极小化极大解答电路136,其形成机器学习电路130的全部或一部分。极小化极大解答电路136可以包括能够形成或以其他方式提供能够解决极小化极大问题的一个或多个逻辑装置、状态机、处理器、和/或控制器的任何数量和/或组合的电子组件和/或半导体装置。极小化极大问题可以包括其中问题的一部分在整个样本群体上被最小化(例如,可归因于诸如恶意软件等恶意代码的接受的损失风险)的任何问题。当最大数量的特征被加盐、污染或以其他方式损坏时(即,当“K”最大化时)存在最坏情况。例如,如图1所描绘,极小化极大解答电路136可以为以下极小化极大问题提供解答:

[0025] $\min_w \max_i l(x_i, y_i, w) \quad (1)$

[0026] 其中: w =分类器的权重向量 $w \in \mathbb{R}^d$

[0027] $l(x_i, y_i, w)$ =由分类器定义的损失函数

[0028] 在这样的实现中,当对抗性耐性分类器比特134(“M”)被设置为定义的二进制逻辑状态时,损失函数 $l(x_i, y_i, w)$ 被最小化以产生、生成或以其他方式提供对抗性耐性分类器电路138。例如,当分类器比特134处于逻辑关状态时,分类器电路138以传统的、非对抗模式运行。当分类器比特134处于逻辑开状态时,分类器电路138以对抗性耐性模式运行。

[0029] 支持向量机(SVM)可被用于提供本文所描述的对抗耐性机器系统和方法的实现的说明性示例。支持向量机是具有相关学习算法的监督学习模型,其分析用于分类和回归分析的数据。给定包含多个样本104的训练数据集122,其中每个样本被标记为有效样本子集或恶意软件样本子集的成员,SVM训练算法组装或以其他方式构建确定后续样本104是有效还是恶意软件的模型。可以认为这样的SVM提供非概率二元线性分类器。SVM模型是作为空间中的点样本104的表示,样本104被映射使得单独类别的示例被尽可能宽的明确间隙划分。然后将新样本104映射到相同的空间并取决于样本落于的间隙的侧来预测为有效或恶意软件。SVM的目标是最小化由以下等式给出的铰链损失:

$$[0030] \quad \min \left[\frac{1}{2} \|w\|^2 + c \sum_i [1 - y_i w \cdot x_i] \right] \quad (2)$$

[0031] 当K个特征被损害时(即,“最坏情况”情景),样本“i”的铰链损失由以下等式给出:

$$[0032] \quad \max \left[1 - y_i w \cdot (x_i * (1 - a_i)) \right] \quad (3)$$

[0033] 其中 $a_i \in \{0,1\}^d$
 $\sum_j a_{ij} = K$

[0034] 等式 (3) 中提供的最大化问题演示了对于样本 i 的最坏情况损失。为了获得稳健的分类器,应最小化所有样本上的最坏情况损失,从而形成极小化极大问题。因此,通过解决以下问题可以获得对抗性耐性分类器:

$$[0035] \quad \min_w \max_{a_1, a_2, \dots, a_n} \left[\frac{1}{2} \|w\|^2 + C \sum_i \left[1 - y_i w \cdot (x_i * (1 - a_i)) \right] \right] \quad (4)$$

[0036] 等式 (4) 中呈现的极小化极大问题可以通过改变所有可能的场景来直接解决。更有效的备选解决方案是将极小化极大问题转换为具有凸对偶变换的二次程序使得可以应用一个或多个快速算法。在一些实现中,“ K ”的值可以是预定的并且解决固定“ K ”值的极小化极大问题产生了比传统分类器对恶性数据具有更大弹性的分类模型。在另一个实施例中,“ K ”的值可以针对至少一些场景而变化,并且针对每个场景解决极小化极大问题来确定最佳“ K ”值。

[0037] 有利地,解决极小化极大问题通过简化和改进恶意代码的检测的准确度来改进主机系统的操作和性能,特别是当训练数据可能包含被加盐的、污染的或其他不正确修改的代码时。通过简化和改进恶意代码检测的准确性,可能降低分配给系统安全性的计算资源,为增值处理能力释放资源。在实施例中,可以将一个或多个测试数据集 124 和/或一个或多个交叉验证数据集 126 可以被提供给对抗性耐性分类器 138,以测试和/或验证系统性能。

[0038] 图 2 提供了根据本公开的至少一个实施例的说明性对抗性环境分类器系统 200 的框图,其中可以实现机器学习电路 130 和对抗性耐性分类器 138。尽管在个人计算机 202 的上下文中进行了描述,但是图 2 中描绘的各种组件和功能块可以全部或部分地在各种平台上并且使用诸如在服务器、工作站、膝上型计算装置、便携式计算装置、可穿戴计算装置、平板计算装置、手持计算装置、以及类似的单核或多核处理器或基于微处理器的装置中找到那些组件的组件来实现。

[0039] 可配置电路 204 可包括任何数量和/或组合的电子组件和/或半导体器件。可配置电路 204 可以部分或全部硬连线。可配置电路 204 可包括能够读取和/或执行机器可读指令集的任何数量和/或组合的逻辑处理单元,诸如一个或多个中央处理单元 (CPU)、微处理器、数字信号处理器 (DSP)、图形处理器/处理单元 (GPU)、专用集成电路 (ASIC)、现场可编程门阵列 (FPGA) 等。除非另有描述,否则图 2 中所示的各种块的结构和操作属于常规设计。因此,这些块不需要在本文进一步详细描述,因为相关领域的技术人员将很容易理解它们。

[0040] 图 2 中描绘的说明性系统的一般架构包括可通信地耦合到南桥 220 的北桥 210。北桥 210 可以可通信地耦合到可配置电路、一个或多个非暂态存储器、以及一个或多个视频输出装置、处理单元、和/或控制器。南桥 220 可以可通信地耦合到非暂时性存储器、一个或多个串行和/或并行输入/输出总线结构、以及任何数量的输入/输出装置。

[0041] 一个或多个可配置电路 204 可以包括能够形成能够执行一个或多个机器可读指令集的电路或多个电路的任何数量和/或组合的系统和/或装置。在实施例中,一个或多个可配置电路 204 可以包括但不限于以下中的一个或多个:硬连线电子组件和/或半导体器件;可编程门阵列 (PGA);精简指令集计算机 (RISC);数字信号处理器;单核或多核处理器;单核或多核微处理器;专用集成电路 (ASIC);片上系统 (SoC);数字信号处理器 (DSP);图形处理

单元 (GPU) ;或其任何组合。一个或多个可配置电路204可以包括能够多线程操作的处理器或微处理器。

[0042] 在一些实现中,一个或多个可配置电路204可以执行一个或多个机器可读指令集,其使得一个或多个可配置电路204的全部或部分提供特征提取电路110、样本分配电路120、和机器学习电路130。一个或多个可配置电路204可以经由一个或多个总线与北桥(以及其他系统组件)通信。

[0043] 系统存储器可以可通信地耦合到北桥210。这样的系统存储器可以包括任何数量和/或组合的任何当前和/或未来开发的存储器和/或存储装置。系统存储器的全部或部分可以以可移除存储器或存储装置的形式提供,其可以从系统200分离或以其他方式解耦。在一些实现中,系统存储器可以包括随机存取存储器 (RAM) 230。RAM 230可以包括以任何数量和/或组合的静电、电磁、光学、分子和/或量子存储器。当系统200在操作中时,可以将信息存储或以其他方式保留在RAM 230中。这样的信息可以包括但不限于一个或多个应用232、操作系统234和/或程序数据236。系统存储器可以经由北桥210与可配置电路204交换信息和/或数据。

[0044] 在一些实现中,一个或多个应用232可以包括一个或多个特征提取应用,其可用于从包括在初始数据集102中的样本104中的每一个识别和/或提取一个或多个特征106。一个或多个特征提取应用可以由可配置电路204和/或特征提取电路112执行。在一些实现中,一个或多个应用232可以包括一个或多个样本分配应用,其可用于将包括在初始数据集102中的样本102分配到一个或多个训练数据集122、一个或多个测试数据集124、和/或一个或多个交叉验证数据集126中。一个或多个样本分配应用可以由可配置电路204和/或样本分配电路120执行。在一些实现中,一个或多个应用232可以包括一个或多个极小化极大问题解答应用。一个或多个极小化极大问题解答应用可以由可配置电路204、机器学习电路130和/或极小化极大解答电路136执行。

[0045] 操作系统234可以包括任何当前或未来开发的操作系统。这种操作系统234的示例可以包括但不限于Windows® (Microsoft Corp, Redmond, WA); OSx (Apple Inc., Cupertino, CA); iOS (Apple Inc., Cupertino, CA); Android® (Google, Inc., Mountain View, CA); 和类似的。操作系统234可以包括一个或多个开源操作系统,包括但不限于Linux、GNU和类似的。

[0046] 数据236可以包括由特征散列电路110、样本分配电路120、和/或机器学习电路130使用、生成或以其他方式耗用的任何信息和/或数据。数据236可以包括初始数据集102、一个或多个训练数据集122、一个或多个测试数据集124、和/或一个或多个交叉验证数据集126。

[0047] 一个或多个视频控制电路240可以经由一个或多个导电构件(诸如一个或多个总线215)可通信地耦合到北桥210。在一些实现中,一个或多个视频控制电路240可以可通信地耦合到插座(例如,AGP插座或类似的),其进而经由一个或多个总线215可通信地耦合到北桥。一个或多个视频控制电路240可包括一个或多个独立装置,例如一个或多个图形处理单元 (GPU)。一个或多个视频控制电路240可以包括一个或多个嵌入式装置,诸如设置在片上系统 (SOC) 上的一个或多个图形处理电路。一个或多个视频控制电路240可以经由北桥210从可配置电路204接收数据。

[0048] 一个或多个视频输出装置242可以可通信地耦合到一个或多个视频控制电路。这样的视频输出装置242可以无线地可通信地耦合到一个或多个视频控制电路240或者栓(即,有线的)到一个或多个视频控制电路242。一个或多个视频输出装置242可以包括但不限于一个或多个液晶(LCD)显示器;一个或多个发光二极管(LED)显示器;一个或多个聚合物发光二极管(PLED)显示器;一个或多个有机发光二极管(OLED)显示器;一个或多个阴极射线管(CRT)显示器;或其任何组合。

[0049] 北桥210和南桥220可以经由诸如一个或多个总线212(例如,链路通道或类似结构)的一个或多个导电构件可通信地耦合。

[0050] 一个或多个只读存储器(ROM)可以经由一个或多个导电构件(诸如一个或多个总线222)可通信地耦合到南桥220。在一些实现中,基本输入/输出系统(BIOS) 252可以整个或部分被存储在ROM 250内。BIOS 252可以向系统200提供基本功能性,并且可以在系统200启动时被执行。

[0051] 通用串行总线(USB)控制器260可以经由一个或多个导电构件(诸如一个或多个总线224)可通信地耦合到南桥。USB控制器260提供用于附接多个USB兼容I/O装置262的网关。示例输入装置可以包括但不限于键盘、鼠标、轨迹球、触摸板、触摸屏、和类似的。示例输出装置可以包括打印机、扬声器、三维打印机、音频输出装置,视频输出装置、触觉输出装置、或其组合。

[0052] 多个通信接口可以经由一个或多个导电构件(诸如一个或多个总线226)可通信地耦合到南桥220。通信和/或外围接口可以包括但不限于一个或多个外围组件互连(PCI)接口270;一个或多个PCI高速接口272;一个或多个IEEE 1384(Firewire)接口274;一个或多个THUNDERBOLT®接口276;一个或多个小型计算机串行(SCSI)接口278;或其组合。

[0053] 多个输入/输出(I/O)装置280可以经由一个或多个导电构件(诸如一个或多个总线223)可通信地耦合到南桥220。I/O装置可以包括任何数量和/或组合的手动或自主I/O装置。例如,I/O装置中的至少一个可以可通信地耦合到提供初始数据集102的全部或一部分的一个或多个外部装置。这样的外部装置可以包括例如服务器或其他可通信地耦合的系统,所述系统收集、保留、或以其他方式存储样本104。

[0054] I/O装置280可以包括一个或多个文本输入或键入(例如,键盘)装置282;一个或多个指向(例如,鼠标、轨迹球)装置284;一个或多个音频输入(例如,麦克风)和/或输出(例如,扬声器)装置286;一个或多个触觉输出装置288;一个或多个触摸屏装置290;一个或多个网络接口292;或其组合。一个或多个网络接口292可以包括一个或多个无线(例如,IEEE 802.11、NFC、蓝牙®、Zigbee)网络接口、一个或多个有线(例如,IEEE 802.3、以太网)接口;或其组合。

[0055] 多个存储装置294可以经由一个或多个导电构件(诸如一个或多个总线225)可通信地耦合到南桥220。一个或多个存储装置294可以使用任何当前或未来开发的接口技术通信地耦合到南桥220。例如,存储装置294可以经由一个或多个集成驱动电子装置(IDE)接口或一个或多个增强IDE接口296通信地耦合。在一些实现中,多个存储装置可以包括但不限于存储装置的阵列。存储装置阵列的一个这样的示例包括但不限于廉价磁盘冗余阵列(RAID)存储装置阵列298。

[0056] 图3提供了根据本公开的至少一个实施例的用于在潜在对抗性环境中使用机器学习

习训练分类器的说明性方法300的高级逻辑流程图。如上所讨论,机器学习为训练系统提供了有价值的机制,以将输入的代码、数据和/或信息识别和分类为SAFE或MALICIOUS。机器学习系统依赖于准确且可信赖的训练数据集,来正确地将输入的代码、数据和/或信息识别为SAFE,并且还不将输入的代码、数据和/或信息误识别为MALICIOUS。可能“加盐”或以其他方式污染被用于训练分类器的训练数据样本,使得分类器将包含恶意代码的在线或实时样本误识别为SAFE。识别包括在初始训练数据102中的可损害特征并最小化这种样本的误分类为SAFE的可能性,因此改进了分类器的准确性、效率 and 安全性。方法300开始于302。

[0057] 在304处,特征提取电路112识别与包括在初始数据集102中的样本逻辑相关联的特征106。在一些实现中,初始数据集102可以全部或部分地由一个或多个可通信地耦合的系统或装置提供。可以由特征提取电路112采用或以其他方式执行任何当前或未来开发的特征提取方法,来识别与包括在初始数据集102中的样本104逻辑相关联的特征106。

[0058] 在306处,识别与包括在初始数据集102中的至少一部分样本104逻辑相关联的潜在可损害特征108。潜在可损害特征是被识别为是以损害系统安全性的方式改变和/或污染的高风险的那些特征106。例如,这样的特征106可以表示其中难以检测的特征或小变化可以由训练的对抗性耐性分类器电路138简化随后将包含暗中插入的恶意软件或其他软件、木马、病毒或其组合的实时样本分类为SAFE而不MALICIOUS。

[0059] 在一些实现中,特征提取电路112可自主地识别至少一部分潜在可损害特征108。在一些实现中,可手动识别至少一部分潜在可损害特征108。在一些实现中,样本分配电路120可以在将样本104分配到一个或多个训练数据集122、一个或多个测试数据集124和/或一个或多个交叉验证数据集126之前,自主地识别至少一部分潜在可损害特征108。

[0060] 在308,定义分类器损失函数。分类器损失函数是计算上可能的函数,其表示在分类问题中为预测的不准确性所支付的代价(例如,贝叶斯误差或误分类的概率)。在至少一些实现中,分类器损失函数可以由以下表达式定义:

[0061] $l(x_i, y_i, w)$ (5)

[0062] 其中 w =分类器的权重向量, $w \in \mathbb{R}^d$

[0063] x_i =样本 $i = 1 \dots n$ 的特征向量,其中 $w \in \mathbb{R}^d, i = 1, 2, \dots, n$

[0064] y_i =样本 $i = 1 \dots n$ 的标签(例如,SAFE/MALICIOUS)

[0065] 通常,损失函数将由用于对样本104执行分类操作的分类器的类型来给出。

[0066] 在310处,分类器损失函数的极小化极大由极小化极大解答电路136确定。极小化极大解答电路136可以使用一种或多种分析技术自主地解决与对抗性耐性分类器电路138逻辑地相关联的极小化极大,以在最大数量的潜在可损害特征108实际上已经受到了损害时评估相应对抗性耐性分类器电路138的最坏情况损失情形。在一些实现中,可以将该最坏情况损失值与一个或多个定义的阈值进行比较,来确定是否已经选择了适当的对抗性耐性分类器电路138。方法300在312处结束。

[0067] 另外,已经参考以上附图和所附示例进一步描述了实施例的操作。一些图可能包括逻辑流程。尽管本文呈现的这些附图可以包括具体的逻辑流程,但是可以理解,逻辑流程仅提供了如何可以实现本文描述的一般功能的示例。此外,除非另有说明,否则给定的逻辑流程不一定必须以所呈现的顺序执行。另外,给定逻辑流程可以由硬件元件、由处理器执行的软件元件、或其任何组合来实现。实施例不限于此上下文。

[0068] 本文已经描述了各种特征、方面和实施例。如由本领域技术人员将理解的,所述特征、方面和实施例易于彼此组合以及变化和修改。因此,本公开应被视为包含这样的组合、变化和修改。因此,本发明的广度和范围不应被任何上述示例性实施例的限制,而应仅根据所附权利要求及其等同物来限定。

[0069] 本文已经采用的术语和表达用作描述而非限制的术语,并且在使用这些术语和表达时,无意排除所示和所描述的特征的任何等同物(或其部分),并且认识到在权利要求的范围内各种修改是可能的。因此,权利要求旨在涵盖所有这些等同物。本文已经描述了各种特征、方面和实施例。如由本领域技术人员将理解的,所述特征、方面和实施例易于彼此组合以及变化和修改。因此,本公开应被视为包含这样的组合、变化和修改。

[0070] 如本文所描述,可以使用硬件元件、软件元件或其任何组合来实现各种实施例。硬件元件的示例可以包括处理器、微处理器、电路、电路元件(例如,晶体管、电阻器、电容器、电感器、线圈、传输线、慢波传输线、变压器等)、集成电路、专用集成电路(ASIC)、无线接收器、发送器、收发器、用于波束成形的智能天线阵列和用于无线宽带通信的电子束控制器或用于自动驾驶的雷达传感器或用作替代键盘装置以用于触觉互联网体验的手势传感器、用于安全应用的屏幕传感器、医疗传感器(癌症筛查)、可编程逻辑装置(PLD)、数字信号处理器(DSP)、现场可编程门阵列(FPGA)、逻辑门、寄存器、半导体装置、芯片、微芯片、芯片集等。

[0071] 贯穿本说明书对“一个实施例”或“实施例”的引用意味着结合该实施例描述的具体特征、结构或特性包括在至少一个实施例中。因此,贯穿本说明书在各个地方出现的短语“在一个实施例中”或“在实施例中”不一定都指的是相同的实施例。此外,具体特征、结构或特性可以在一个或多个实施例中以任何合适的方式组合。

[0072] 以下示例属于其他实施例。本公开的以下示例可以包括诸如用于提供适合在训练数据可能被损害的不利环境中使用的机器学习系统的装置、系统、方法和部件的主要内容。

[0073] 根据示例1,提供了一种对抗性机器学习系统。所述对抗性机器学习系统可以包括:特征提取电路,用于识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征;样本分配电路,用于将包括在初始数据集中的至少一部分样本分配给至少训练数据集;机器学习电路,可通信地耦合到样本分配电路,机器学习电路用于:识别包括在初始数据集中的至少一部分样本的可损害特征的至少一个集合;定义分类器损失函数 $l(x_i, y_i, w)$,分类器损失函数包括:包括在初始数据集中的每个样本的特征向量(x_i);包括在初始数据集中的每个样本的标签(y_i);以及与分类器相关联的权重向量(w);以及确定分类器损失函数的极小化极大($\min_w \max_i l(x_i, y_i, w)$)。

[0074] 示例2可以包括示例1的元件,其中样本分配电路还包括用于将包括在初始数据集中的至少一部分样本分配给以下数据集中的至少一个数据集的电路:训练数据集;测试数据集;或交叉验证数据集。

[0075] 示例3可以包括示例1的元件,其中机器学习电路可以自主地识别包括在至少训练数据集中的至少一部分样本的可损害特征的至少一个集合。

[0076] 示例4可以包括示例1的元件,其中机器学习电路可以接收至少一个输入来手动识别包括在至少训练数据集中的至少一部分样本的可损害特征的至少一个集合。

[0077] 示例5可以包括示例1的元件,其中机器学习电路可以如果相应样本表示非恶意样本,则定义包括与样本相关联的标签的第一逻辑值的损失函数以及可以如果相应样本表示

恶意样本,则定义包括与样本相关联的标签的第二逻辑值的损失函数。

[0078] 示例6可以包括示例1的元件,其中机器学习电路还可以识别由包括在至少训练数据集中的至少一部分样本的固定数量的可损害特征组成的集合。

[0079] 示例7可以包括示例1的元件,其中机器学习电路还可以识别包括在至少训练数据集中的至少一部分样本的可损害特征的多个集合,多个集合中的每个集合包括训练数据集中包括的至少一部分样本的不同数量的可损害特征。

[0080] 根据示例8,提供了一种分类器训练方法。所述方法可以包括:将包括在初始数据集中的多个样本的至少一部分分配到至少训练数据集;识别与包括在训练数据集中的每个样本相关联的多个特征;识别包括训练数据集中的至少一部分样本的可损害特征的至少一个集合;定义分类器损失函数 $l(x_i, y_i, w)$,分类器损失函数包括:包括在训练数据集中的每个元素的特征向量 (x_i) ;包括在训练数据集中的每个元素的标签 (y_i) ;以及与分类器相关联的权重向量 (w) ;以及确定分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

[0081] 示例9可以包括示例8的元素,其中识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合包括:自主地识别包括在训练数据集中的至少一部分元素的可损害特征的至少一个集合。

[0082] 示例10可以包括示例8的元素,其中识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合可以包括手动识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合。

[0083] 示例11可以包括示例8的元素,其中定义包括在训练数据集中包括的每个样本的标签 (y_i) 的分类器损失函数 $l(x_i, y_i, w)$ 可以包括如果相应样本表示非恶意样本,则定义包括与样本相关联的标签的第一逻辑值的损失函数;以及如果相应样本表示恶意样本,则定义包括与样本相关联的标签的第二逻辑值的损失函数。

[0084] 示例12可以包括示例8的元素,其中识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合可以包括识别由包括在训练数据集中的至少一部分样本的固定数量的可损害特征组成的集合。

[0085] 示例13可以包括示例8的元素,其中识别包括在训练数据集中的至少一部分元素的可损害特征的至少一个集合可以包括识别包括在训练数据集中的至少一部分样本的可损害特征的多个集合,多个集合中的每个集合包括训练数据集中包括的至少一部分样本的不同数量的可损害特征。

[0086] 根据示例14,提供了一种存储装置,包括机器可读指令,机器可读指令在被执行时将可配置电路物理地变换为对抗性机器学习训练电路,对抗性机器学习训练电路用于:

[0087] 识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征;将包括在初始数据集中的至少一部分样本分配到至少训练数据集;识别包括训练数据集中的至少一部分样本的可损害特征的至少一个集合;定义分类器损失函数 $l(x_i, y_i, w)$,分类器损失函数包括:包括在训练数据集中的每个样本的特征向量 (x_i) ;包括在训练数据集中的每个样本的标签 (y_i) ;以及和与分类器相关联的权重向量 (w) ;以及确定分类器损失函数的极小化极大 $(\min_w \max_i l(x_i, y_i, w))$ 。

[0088] 示例15可以包括示例14的元件,其中使对抗性机器学习训练电路识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的机器可读指令使对抗性机器

学习训练电路:自主地识别包括训练数据集中的至少一部分样本的可损害特征的至少一个集合。

[0089] 示例16可以包括示例14的元件,其中,使对抗性机器学习训练电路识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的机器可读指令使对抗性机器学习训练电路:接收包括手动识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的数据的输入。

[0090] 示例17可以包括示例14的元件,其中使对抗机器学习训练电路定义包括在训练数据集中包括的每个样本的标签(y_i)的分类器损失函数 $[l(x_i, y_i, w)]$ 的机器可读指令,进一步使对抗机器学习训练电路:如果相应样本表示非恶意样本,则定义包括与样本相关联的标签的第一逻辑值的损失函数;以及如果相应样本表示恶意样本,则定义包括与样本相关联的标签的第二逻辑值的损失函数。

[0091] 示例18可以包括示例14的元件,其中使对抗性机器学习训练电路识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的机器可读指令,进一步使对抗性机器学习训练电路:识别由包括在训练数据集中的至少一部分样本的固定数量的可损害特征组成的集合。

[0092] 示例19可以包括示例14的元件,其中使对抗性机器学习训练电路识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的机器可读指令,进一步使对抗性机器学习训练电路:识别包括在训练数据集中的至少一部分样本的可损害特征的多个集合,多个集合中的每个集合包括训练数据集中包括的至少一部分样本的不同数量的可损害特征。

[0093] 根据示例20,提供了一种分类器训练系统。所述分类器训练系统可以包括:用于识别与包括在包括多个样本的初始数据集中的每个样本相关联的多个特征的部件;用于将初始数据集中的至少一部分样本分配到至少训练数据集的部件;用于识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的部件;用于定义分类器损失函数 $[l(x_i, y_i, w)]$ 的部件,分类器损失函数包括:包括在训练数据集中的每个样本的特征向量(x_i);包括在训练数据集中的每个样本的标签(y_i);以及与分类器相关联的权重向量(w);以及用于确定分类器损失函数的极小化极大的部件($\min_w \max_i l(x_i, y_i, w)$)。

[0094] 示例21可以包括示例20的元件,其中用于识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的部件包括:用于自主识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的部件。

[0095] 示例22可以包括示例20的元件,其中用于识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的部件包括:用于手动识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的部件。

[0096] 示例23可以包括示例20的元件,其中用于定义包括在训练数据集中包括的每个样本的标签(y_i)的分类器损失函数 $[l(x_i, y_i, w)]$ 的部件包括:用于如果相应样本表示非恶意样本则定义包括与样本相关联的标签的第一逻辑值的损失函数的部件;以及用于如果相应样本表示恶意样本则定义包括与样本相关联的标签的第二逻辑值的损失函数的部件。

[0097] 示例24可以包括示例20的元件,其中用于识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合的部件包括:用于识别由包括在训练数据集中的至少一

部分样本的固定数量的可损害特征组成的集合的部件。

[0098] 示例25可以包括示例20的元件,其中识别包括在训练数据集中的至少一部分样本的可损害特征的至少一个集合包括:用于识别与包括在训练数据集中的至少一部分样本的可损害特征的多个集合的部件,多个集合中的每个集合包括训练数据集中包括的至少一部分样本的不同数量的可损害特征。

[0099] 根据示例26,提供了一种用于训练对抗性环境分类器的系统,所述系统布置成执行根据示例8至13中的任一个中的方法。

[0100] 根据示例27,提供了一种布置成执行根据示例8至13中的任一个中的方法的芯片集。

[0101] 根据示例28,提供了一种包括多个指令的至少一个机器可读介质,所述指令响应于被执行在计算装置上,使该计算装置执行根据示例8至13中的任一个中的方法。

[0102] 根据示例29,提供了一种配置成训练对抗性环境分配器的装置,所述装置布置成执行根据示例8至13中的任一个中的方法。

[0103] 本文已经采用的术语和表达用作描述而非限制的术语,并且在使用这些术语和表达时,无意排除所示和所描述的特征的任何等同物(或其部分),并且认识到在权利要求的范围内各种修改是可能的。因此,权利要求旨在涵盖所有这些等同物。

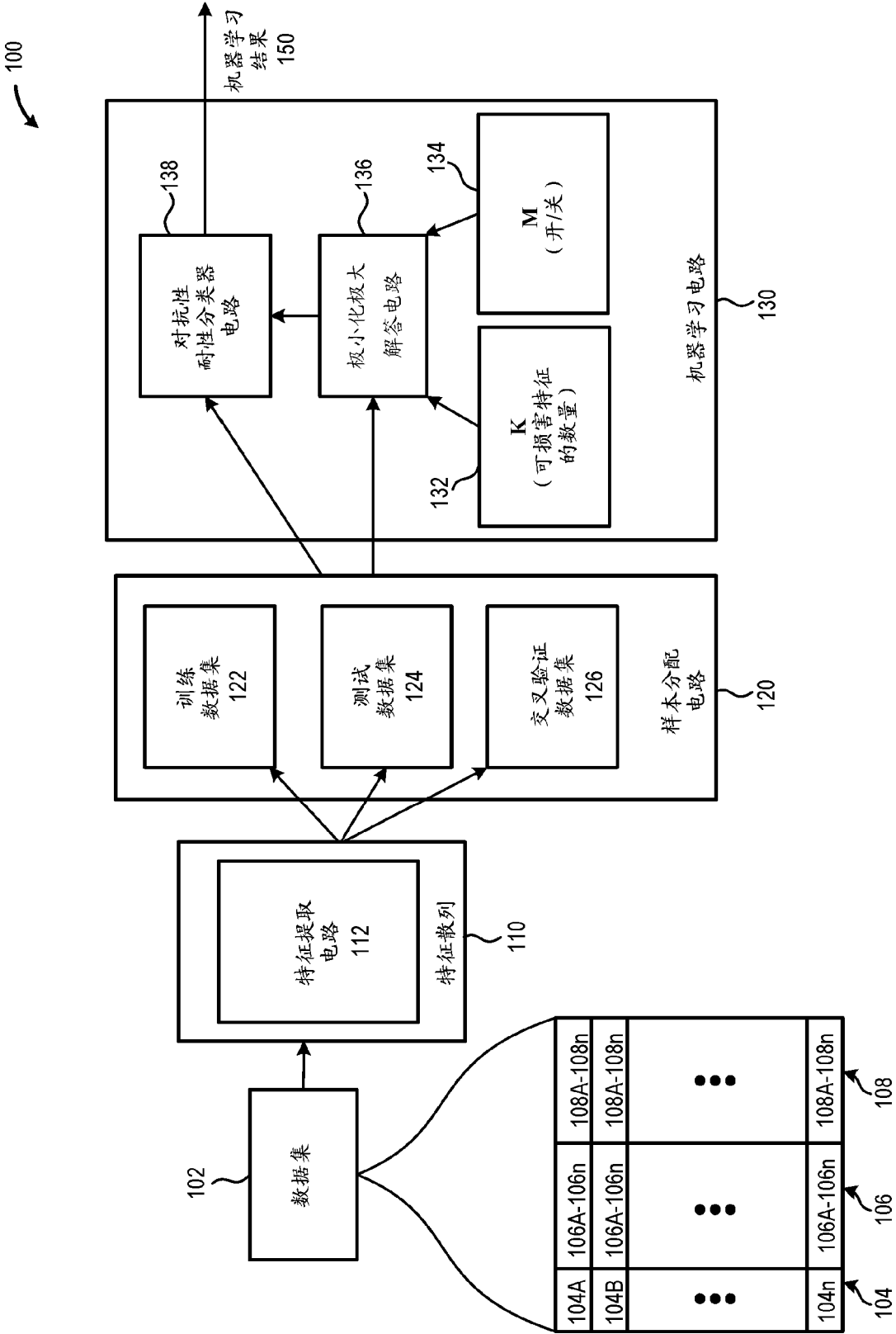


图 1

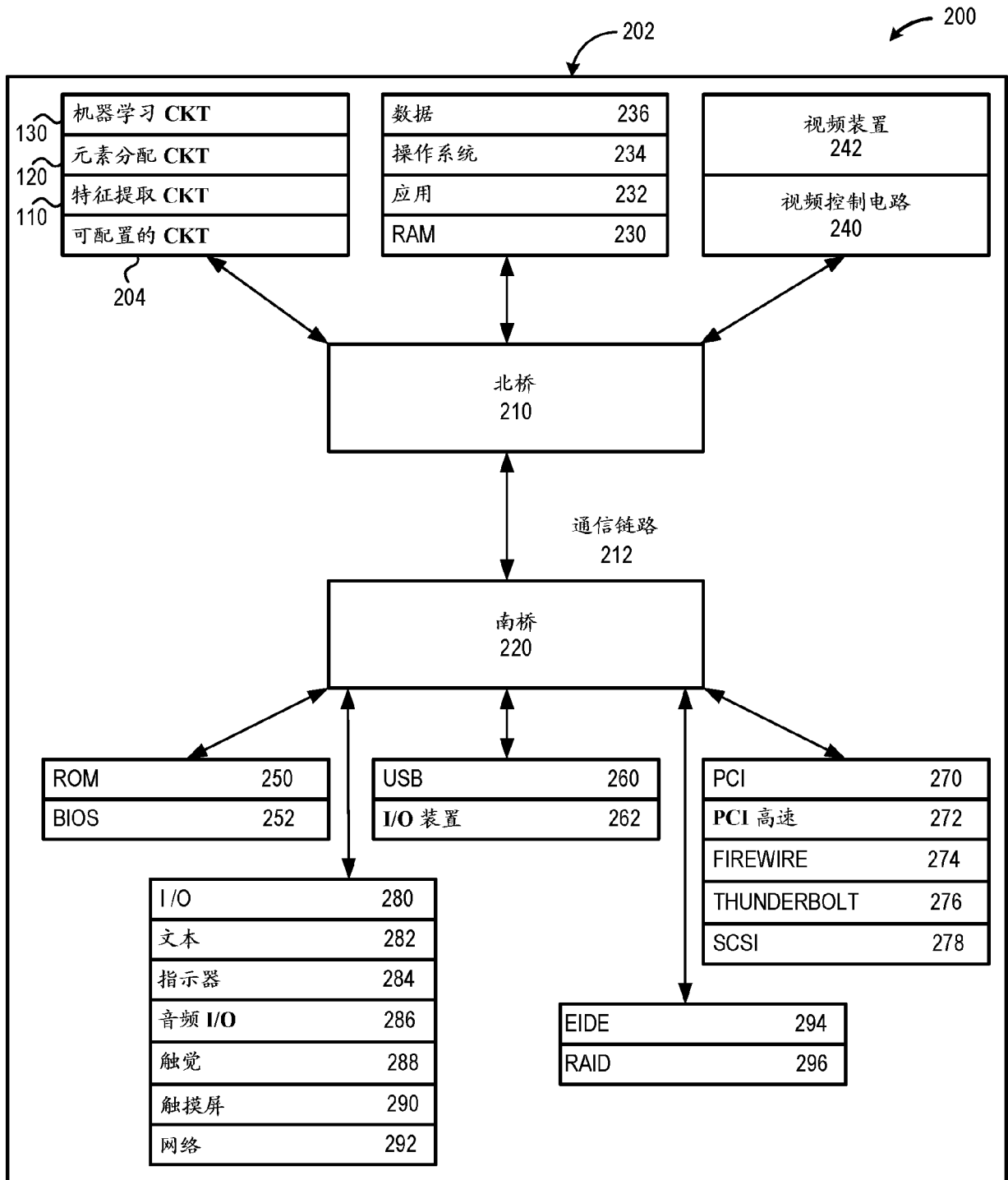


图 2

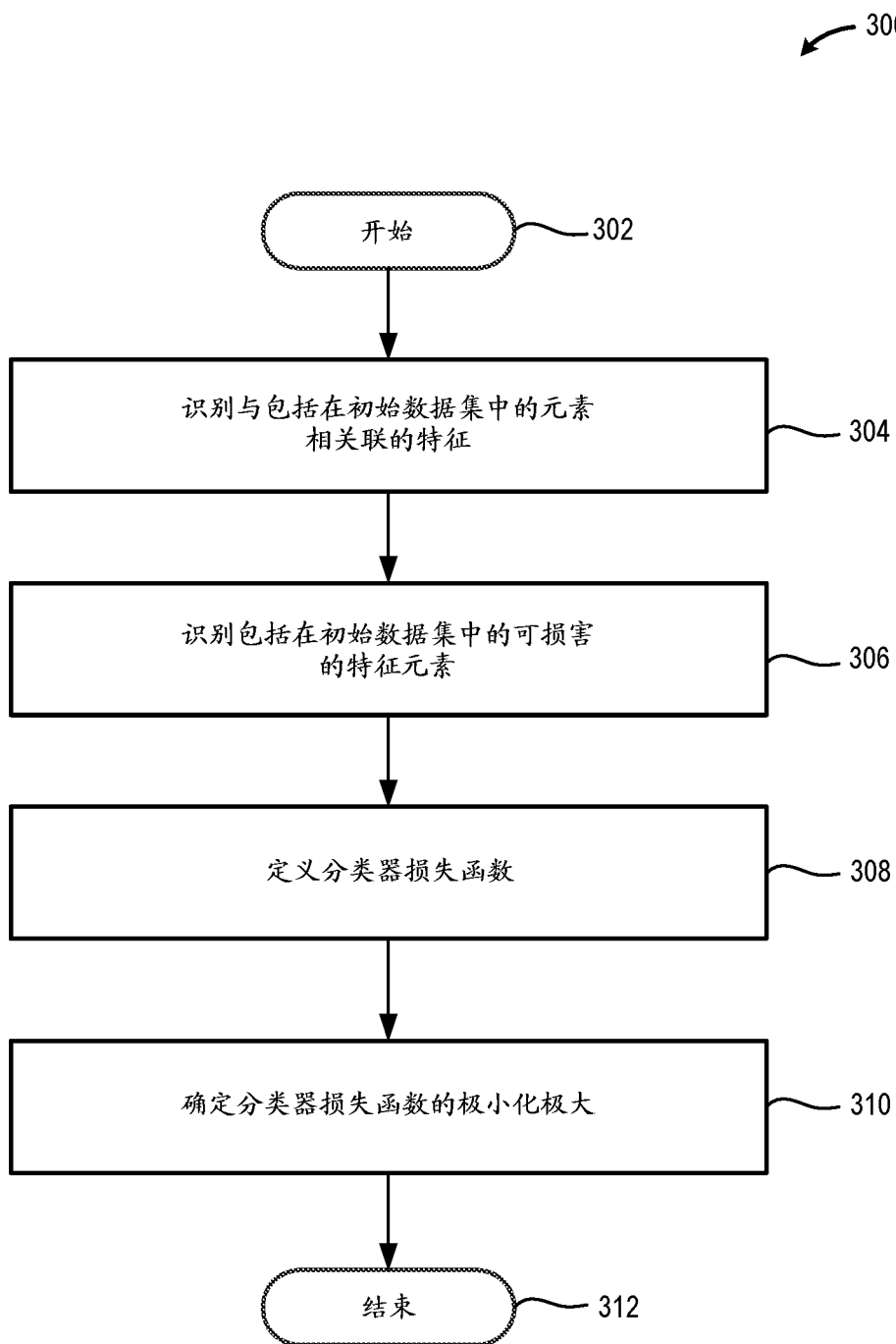


图 3