(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111401475 A
(43)申请公布日 2020.07.10

(21)申请号 202010296739.7

(22)申请日 2020.04.15

(71)申请人 支付宝(杭州)信息技术有限公司
地址 310000 浙江省杭州市西湖区西溪路556号8层B段801-11

(72)发明人 林建滨

(74)专利代理机构 成都七星天知识产权代理有限公司 51253
代理人 杨永梅

(51)Int.Cl.
G06K 9/62(2006.01)
G06F 21/57(2013.01)
G06N 20/00(2019.01)

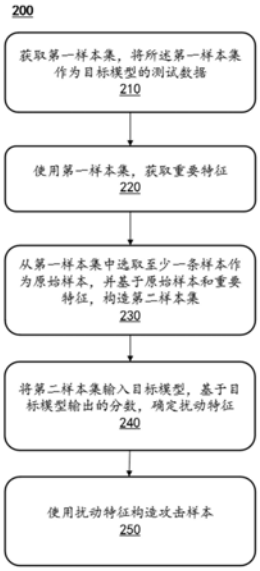
权利要求书3页 说明书9页 附图6页

(54)发明名称

一种生成攻击样本的方法和系统

(57)摘要

本说明书实施例公开了一种生成攻击样本的方法及系统，包括：获取第一样本集，将第一样本集作为目标模型的测试数据；使用第一样本集，获取重要特征；从第一样本集中选取至少一条样本作为原始样本，并基于原始样本和重要特征，构造第二样本集；其中，第二样本集包括正第二样本和负第二样本；将第二样本集输入目标模型，基于目标模型输出的结果，确定扰动特征；其中，扰动特征包括正扰动特征和负扰动特征，扰动特征由重要特征中的至少一个组成；使用扰动特征构造攻击样本；攻击样本用于攻击目标模型，从而根据攻击结果确定目标模型的鲁棒性，判断目标模型是否可以抵御数据中毒，保护个人数据。



CN 111401475 A

1.一种生成攻击样本的方法,所述方法包括:

获取第一样本集,将所述第一样本集作为目标模型的测试数据;

使用所述第一样本集,获取重要特征;其中,所述重要特征为对目标模型测试结果的影响超过预设影响阈值的特征,所述重要特征包括正重要特征和负重要特征;

从所述第一样本集中选取至少一条样本作为原始样本,并基于所述原始样本和所述重要特征,构造第二样本集;其中,所述第二样本集包括正第二样本和负第二样本;

将所述第二样本集输入所述目标模型,基于所述目标模型输出的结果,确定扰动特征;其中,所述扰动特征包括正扰动特征和负扰动特征,所述扰动特征由所述重要特征中的至少一个组成;

使用所述扰动特征构造攻击样本;其中,所述攻击样本用于攻击所述目标模型。

2.根据权利要求1所述的方法,其中,所述获取第一样本集包括:

将特征空间随机划分为多个特征序列,每个所述特征序列构成一条样本;

分别将多条所述样本输入所述目标模型,将所述目标模型输出的结果作为所述样本的标签,将带标签的多条所述样本作为第一样本集。

3.根据权利要求2所述的方法,其中,所述使用所述第一样本集,获取重要特征包括:

使用所述第一样本集训练线性模型,获取多个权重,所述多个权重对应于组成所述第一样本集的所述多个特征;

将所述多个权重中符号为正的权重按照值由大到小的顺序进行排序,选取前N个权重对应的特征作为所述正重要特征;

将所述多个权重中符号为负的权重按照值由大到小的顺序进行排序,选取前N个权重对应的特征作为所述负重要特征。

4.根据权利要求3所述的方法,其中,所述基于所述原始样本和所述重要特征,构造第二样本集包括:

从所述重要特征中的正重要特征中任意选取至少一个按照不同的组合添加到所述原始样本中,构造至少一条所述正第二样本;

从所述重要特征中的负重要特征中任意选取至少一个按照不同的组合添加到所述原始样本中,构造至少一条所述负第二样本。

5.根据权利要求4所述的方法,其中,所述将所述第二样本集输入所述目标模型,基于所述目标模型输出的结果,确定扰动特征包括:

将所述第二样本集中的所述至少一条正第二样本输入所述目标模型,获取所述目标模型输出的至少一个分数;

将所述至少一个分数中最高分数对应的正第二样本中包含的所述正重要特征作为所述正扰动特征;

将所述第二样本集中的至少一条负第二样本输入所述目标模型,获取所述目标模型输出的至少一个分数;

将所述至少一个分数中最低分数对应的负第二样本中包含的所述负重要特征作为所述负扰动特征。

6.根据权利要求5所述的方法,其中,所述使用所述扰动特征构造攻击样本包括:

从所述目标模型的测试数据中选取至少一条样本作为测试样本,将所述正扰动特征添

加到所述测试样本中，获取正攻击样本；

将所述负扰动特征添加到所述测试样本中，获取负攻击样本。

7.根据权利要求1所述的方法，其中，所述目标模型基于实体对象的数据进行预测，并根据预测结果确定后续的操作，使用所述方法构造所述目标模型的攻击样本。

8.一种生成攻击样本的系统，所述系统包括：

第一获取模块，用于获取第一样本集，将所述第一样本集作为目标模型的测试数据；

第二获取模块，用于使用所述第一样本集，获取重要特征；其中，所述重要特征为对目标模型测试结果的影响超过预设影响阈值的特征，所述重要特征包括正重要特征和负重要特征；

第一构造模块，用于从所述第一样本集中选取至少一条样本作为原始样本，并基于所述原始样本和所述重要特征，构造第二样本集；其中，所述第二样本集包括正第二样本和负第二样本；

扰动特征确定模块，用于将所述第二样本集输入所述目标模型，基于所述目标模型输出的结果，确定扰动特征；其中，所述扰动特征包括正扰动特征和负扰动特征，所述扰动特征由所述重要特征中的至少一个组成；

第二构造模块，用于使用所述扰动特征构造攻击样本；其中，所述攻击样本用于攻击所述目标模型。

9.根据权利要求8所述的系统，其中，所述获取第一样本集包括：

将特征空间随机划分为多个特征序列，每个所述特征序列构成一条样本；

分别将多条所述样本输入所述目标模型，将所述目标模型输出的结果作为所述样本的标签，将带标签的多条所述样本作为第一样本集。

10.根据权利要求9所述的系统，其中，所述使用所述第一样本集，获取重要特征包括：

使用所述第一样本集训练线性模型，获取多个权重，所述多个权重对应于组成所述第一样本集的所述多个特征；

将所述多个权重中符号为正的权重按照值由大到小的顺序进行排序，选取前N个权重对应的特征作为正重要特征；

将所述多个权重中符号为负的权重按照值由大到小的顺序进行排序，选取前N个权重对应的特征作为负重要特征。

11.根据权利要求10所述的系统，其中，所述基于所述原始样本和所述重要特征，构造第二样本集包括：

从所述重要特征中的正重要特征中任意选取至少一个按照不同的组合添加到所述原始样本中，构造至少一条所述正第二样本；

从所述重要特征中的负重要特征中任意选取至少一个按照不同的组合添加到所述原始样本中，构造至少一条所述负第二样本。

12.根据权利要求11所述的系统，其中，所述将所述第二样本集输入所述目标模型，基于所述目标模型输出的结果，确定扰动特征包括：

将所述第二样本集中的所述至少一条正第二样本输入所述目标模型，获取所述目标模型输出的至少一个分数；

将所述至少一个分数中最高分数对应的正第二样本中包含的所述正重要特征作为所

述正扰动特征；

将所述第二样本集中的至少一条负第二样本输入所述目标模型，获取所述目标模型输出的至少一个分数；

将所述至少一个分数中最低分数对应的负第二样本中包含的所述负重要特征作为所述负扰动特征。

13.根据权利要求12所述的系统，其中，所述使用所述扰动特征构造攻击样本包括：

从所述目标模型的测试数据中选取至少一条样本作为测试样本，将所述正扰动特征添加到所述测试样本中，获取正攻击样本；

将所述负扰动特征添加到所述测试样本中，获取负攻击样本。

14.一种生成攻击样本的装置，其中，所述装置包括至少一个处理器以及至少一个存储器；

所述至少一个存储器用于存储计算机指令；

所述至少一个处理器用于执行所述计算机指令中的至少部分指令以实现如权利要求1～6中任一项所述的方法。

一种生成攻击样本的方法和系统

技术领域

[0001] 本说明书涉及对抗机器学习领域，特别涉及一种生成针对输入为离散特征的模型的攻击样本的方法和系统。

背景技术

[0002] 目前，机器学习已经广泛应用于各行各业，承载着大量国计民生的重要应用，例如信息检索、金融支付、智能驾驶和智能安防等。然而，这些机器学习模型天然携带着容易被“攻击”的漏洞。如果模型不能抵御攻击样本的攻击，甚至会威胁资金安全和公共安全。通常把机器学习模型抵御攻击样本的能力称为机器学习模型的鲁棒性。如何测试和提高机器学习模型的鲁棒性尤为重要。目前此项研究在学术界和工业界还处于比较初期的探索阶段，大部分算法是针对输入为连续特征的模型。

[0003] 因此需要一种基于输入为离散、稀疏特征的模型的黑盒攻击算法，用于生成攻击样本，评估模型的鲁棒性。

发明内容

[0004] 本说明书实施例之一提供一种生成攻击样本的方法。所述方法包括：

[0005] 获取第一样本集，将所述第一样本集作为目标模型的测试数据；使用所述第一样本集，获取重要特征；其中，所述重要特征为对目标模型测试结果的影响超过预设影响阈值的特征，所述重要特征包括正重要特征和负重要特征；从所述第一样本集中选取至少一条样本作为原始样本，并基于所述原始样本和所述重要特征，构造第二样本集；其中，所述第二样本集包括正第二样本和负第二样本；将所述第二样本集输入所述目标模型，基于所述目标模型输出的分数，确定扰动特征；其中，所述扰动特征包括正扰动特征和负扰动特征，所述扰动特征由所述重要特征中的至少一个组成；使用所述扰动特征构造攻击样本；其中，所述攻击样本用于攻击所述目标模型。

[0006] 本说明书实施例之一提供一种生成攻击样本的系统，所述系统包括：

[0007] 第一获取模块，用于获取第一样本集，将所述第一样本集作为目标模型的测试数据；第二获取模块，用于使用所述第一样本集，获取重要特征；其中，所述重要特征为对目标模型测试结果的影响超过预设影响阈值的特征，所述重要特征包括正重要特征和负重要特征；第一构造模块，用于从所述第一样本集中选取至少一条样本作为原始样本，并基于所述原始样本和所述重要特征，构造第二样本集；其中，所述第二样本集包括正第二样本和负第二样本；扰动特征确定模块，用于将所述第二样本集输入所述目标模型，基于所述目标模型输出的结果，确定扰动特征；其中，所述扰动特征包括正扰动特征和负扰动特征，所述扰动特征由所述重要特征中的至少一个组成；第二构造模块，用于使用所述扰动特征构造攻击样本；其中，所述攻击样本用于攻击所述目标模型。

[0008] 本说明书实施例之一提供一种生成攻击样本的装置，包括：

[0009] 所述装置包括至少一个处理器以及至少一个存储器；所述至少一个存储器用于存

储计算机指令;所述至少一个处理器用于执行所述计算机指令中的至少部分指令以实现生成攻击样本的方法。

附图说明

[0010] 本说明书将以示例性实施例的方式进一步说明,这些示例性实施例将通过附图进行详细描述。这些实施例并非限制性的,在这些实施例中,相同的编号表示相同的结构,其中:

[0011] 图1是根据本说明书一些实施例所示的生成攻击样本的系统的模块图;

[0012] 图2是根据本说明书一些实施例所示的生成攻击样本的方法的示例性流程图;

[0013] 图3是根据本说明书一些实施例所示的生成攻击样本的示例性应用场景图;

[0014] 图4A是根据本说明书一些实施例所示的特征空间和采样序列的示例的示意图;

[0015] 图4B是根据本说明书一些实施例所示的样本1的示例的示意图;

[0016] 图4C是根据本说明书一些实施例所示的使用独热编码表示特征的样本1的示例的示意图;

[0017] 图5是根据本说明书一些实施例所示的正第二样本的示例的示意图;以及

[0018] 图6是根据本说明书一些实施例所示的负第二样本的示例的示意图。

具体实施方式

[0019] 为了更清楚地说明本说明书实施例的技术方案,下面将对实施例描述中所需要使用的附图作简单的介绍。显而易见地,下面描述中的附图仅仅是本说明书的一些示例或实施例,对于本领域的普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图将本说明书应用于其它类似情景。除非从语言环境中显而易见或另做说明,图中相同标号代表相同结构或操作。

[0020] 应当理解,本文使用的“系统”、“装置”、“单元”和/或“模组”是用于区分不同级别的不同组件、元件、部件、部分或装配的一种方法。然而,如果其他词语可实现相同的目的,则可通过其他表达来替换所述词语。

[0021] 如本说明书和权利要求书中所示,除非上下文明确提示例外情形,“一”、“一个”、“一种”和/或“该”等词并非特指单数,也可包括复数。一般说来,术语“包括”与“包含”仅提示包括已明确标识的步骤和元素,而这些步骤和元素不构成一个排它性的罗列,方法或者设备也可能包含其它的步骤或元素。

[0022] 本说明书中使用了流程图用来说明根据本说明书的实施例的系统所执行的操作。应当理解的是,前面或后面操作不一定按照顺序来精确地执行。相反,可以按照倒序或同时处理各个步骤。同时,也可以将其他操作添加到这些过程中,或从这些过程移除某一步或数步操作。

[0023] 图1是根据本说明书一些实施例所示的生成攻击样本的系统的模块图。

[0024] 如图1所示,该生成攻击样本的系统可以包括第一获取模块110、第二获取模块120、第一构造模块130、扰动特征确定模块140和第二构造模块150。

[0025] 第一获取模块110可以用于获取第一样本集,将所述第一样本集作为目标模型的测试数据。关于获取第一样本集,将所述第一样本集作为目标模型的测试数据的详细描述

可以参见图1,在此不再赘述。

[0026] 第二获取模块120可以用于使用所述第一样本集,获取重要特征;其中,所述重要特征为对目标模型测试结果的影响超过预设影响阈值的特征,所述重要特征包括正重要特征和负重要特征。关于使用所述第一样本集,获取重要特征的详细描述可以参见图1,在此不再赘述。

[0027] 第一构造模块130可以用于从所述第一样本集中选取至少一条样本作为原始样本,并基于所述原始样本和所述重要特征,构造第二样本集;其中,所述第二样本集包括正第二样本和负第二样本。关于从所述第一样本集中选取至少一条样本作为原始样本,并基于所述原始样本和所述重要特征,构造第二样本集的详细描述可以参见图1,在此不再赘述。

[0028] 扰动特征确定模块140可以用于将所述第二样本集输入所述目标模型,基于所述目标模型输出的结果,确定扰动特征;其中,所述扰动特征包括正扰动特征和负扰动特征,所述扰动特征由所述重要特征中的至少一个组成。关于将所述第二样本集输入所述目标模型,基于所述目标模型输出的结果,确定扰动特征的详细描述可以参见图1,在此不再赘述。

[0029] 第二构造模块150可以用于使用所述扰动特征构造攻击样本。关于使用所述扰动特征构造攻击样本的详细描述可以参见图1,在此不再赘述。

[0030] 应当理解,图1所示的系统及其模块可以利用各种方式来实现。例如,在一些实施例中,系统及其模块可以通过硬件、软件或者软件和硬件的结合来实现。其中,硬件部分可以利用专用逻辑来实现;软件部分则可以存储在存储器中,由适当的指令执行系统,例如微处理器或者专用设计硬件来执行。本领域技术人员可以理解上述的方法和系统可以使用计算机可执行指令和/或包含在处理器控制代码中来实现,例如在诸如磁盘、CD或DVD-ROM的载体介质、诸如只读存储器(固件)的可编程的存储器或者诸如光学或电子信号载体的数据载体上提供了这样的代码。本说明书的系统及其模块不仅可以有诸如超大规模集成电路或门阵列、诸如逻辑芯片、晶体管等的半导体、或者诸如现场可编程门阵列、可编程逻辑设备等的可编程硬件设备的硬件电路实现,也可以用例如由各种类型的处理器所执行的软件实现,还可以由上述硬件电路和软件的结合(例如,固件)来实现。

[0031] 需要注意的是,以上对于生成攻击样本的系统及其模块的描述,仅为描述方便,并不能把本说明书限制在所举实施例范围之内。可以理解,对于本领域的技术人员来说,在了解该系统的原理后,可能在不背离这一原理的情况下,对各个模块进行任意组合,或者构成子系统与其他模块连接。例如,在一些实施例中,例如,图1中披露的第一获取模块110、第二获取模块120、第一构造模块130、扰动特征确定模块140和第二构造模块150可以是一个系统中的不同模块,也可以是一个模块实现上述的两个或两个以上模块的功能。例如,第一构造模块130、扰动特征确定模块140可以是两个模块,也可以是一个模块同时具有构造样本和确定扰动特征功能。诸如此类的变形,均在本说明书的保护范围之内。

[0032] 图2是根据本说明书一些实施例所示的生成攻击样本的方法的示例性流程图。

[0033] 步骤210,获取第一样本集,将所述第一样本集作为目标模型的测试数据。具体的,该步骤可以由第一获取模块110执行。

[0034] 本说明书的一些实施例实现了针对输入为离散、稀疏特征的模型的黑盒攻击算法。因此,在一些实施例中,可以选择一个输入是离散、稀疏特征的被攻击模型作为目标模

型。例如:目标模型可以为逾期率预估模型,逾期率预估模型的输入可以为贷款申请者的年龄、性别、身高、申请金额以及申请时间等离散特征组成的样本。

[0035] 在一些实施例中,第一获取模块110可以获取第一样本集,将所述第一样本集作为目标模型的测试数据。具体可以包括以下步骤:

[0036] (1)将特征空间随机划分为多个特征序列,将每个特征序列作为一条样本。

[0037] 在一些实施例中,特征空间可以为很多个特征组成的集合。例如:年龄、性别、身高以及星座等特征可以构成一个特征空间。在一些实施例中,样本获取模块110可以从特征空间里随机采样,每次采样多个不同的特征,每次被采样的多个特征作为一个特征序列,经过多次采样从特征空间里获取了多个特征序列,每个特征序列构成一条样本,最终将特征空间中的特征全部采样到。具体的,每个样本的维度与特征空间的维度一致,如果特征空间包含n个特征,则样本可以由表示这n个特征的n个特征向量组成,样本中每个特征向量乘以一个采样位,该采样位有0和1两种取值:0表示该特征没有被当前样本采样到,1表示该特征被当前样本采样到。例如:如图4A所示的特征空间由500个特征组成,从该特征空间采样一个特征序列w01、w03、w10、w11、w12、w21、w30、w491、w493、w500,该特征序列构成了图4B所示的样本1,其中w02没有被采样到,其在样本1中的采样位为0,w03被采样到,其在样本1中的采样位为1。在一些实施例中,特征空间中的特征使用独热(one-hot)编码的特征向量来表示,每个特征向量的维度一致,都是特征空间的大小。独热编码使用N位状态寄存器来对N个状态进行编码,每个状态都有它独立的寄存器位,并且在任意时候,其中只有一位有效,也就是说这N种状态中只有一个状态位值为1,其他状态位都是0。例如:如图4A所示的500维的特征空间,w01的特征向量可以表示为(1,0,0,0,…,0),w02的特征向量可以表示为(0,1,0,0,…,0),…,w500的特征向量可以表示为(0,0,0,0,…,1)。因此图4B所示的样本1可以为图4C中所示的500个特征向量组成的序列:(1,0,0…0)*1,(0,1,0…0)*0,(0,0,1…0)*1,…,(0,0…0,…1)*1。其中,被采样到的特征,例如:w1、w3和w10等,对应的采样位为1,未被采样到的特征,例如:w2、w4和w5等,对应的采样位为0。在一些实施例中,上述从特征空间中获取的多个样本没有标签。

[0038] 在一些实施例中,每次采样的特征的个数在一定的预设范围内,例如:如果一个常用的样本由10个特征组成,那么该预设范围可以为[8,13]。在一些实施例中,每次采样的特征个数可以相同也可以不同。例如:第一次采样8个特征,第2次采样10个特征…。在一些实施例中,特征的编码方式也可以为非独热编码的其他方式,例如:基于分布式词向量模型的word2vec方法等,不受本说明书的表述所限。

[0039] (2)分别将多条样本输入目标模型,将目标模型输出的结果作为样本的标签,将带标签的多条样本作为第一样本集。

[0040] 在一些实施例中,目标模型经过测试可以对输入样本做出正确的判断,因此可以将步骤(1)中获取的多条样本分别输入目标模型,目标模型对输入的样本打分,输出分数表示的结果,可以将该结果作为输入样本的标签。例如:目标模型为逾期率预估模型,可以用于判断申请者的借款是否会逾期还款。将图4c所示的样本1输入目标模型,如果目标模型判断样本1对应的申请者会逾期归还借款,则输出一个大于0.5的分数,例如0.8,样本1的标签可以为0.8。如果目标模型判断样本1对应的申请者不会逾期归还借款,则输出一个低于0.5的分数,例如0.2,样本1的标签可以为0.2。以此类推,获取多条样本对应的标签,将带标签

的多条样本作为第一样本集。

[0041] 步骤220，使用第一样本集，获取重要特征。具体的，该步骤可以由第二获取模块120执行。

[0042] 在一些实施例中，第二获取模块120可以使用步骤210中获取的第一样本集获取重要特征。在一些实施例中，在组成样本的特征中，有些特征对目标模型测试结果的影响相对比较大，有些对目标模型测试结果的影响相对比较小，选取其中对目标模型测试结果的影响相对比较大的特征作为重要特征。在重要特征中，有些特征可以拉高目标模型对于样本的打分，这一部分特征可以称为正重要特征；有些特征可以降低目标模型对于样本的打分，这一部分特征可以称为负重要特征。例如：对于上述示例中的逾期率预估模型，如果申请者的申请金额越大，按时归还借款的概率就越小，则“申请金额”可以作为正重要特征；如果申请者的申请时间越长，按时归还借款的概率越大，则“申请时间”可以作为负重要特征。

[0043] 在一些实施例中，使用第一样本集，获取重要特征可以包括以下步骤：

[0044] (1)使用第一样本集训练线性模型，获取多个权重，多个权重对应于组成第一样本集的多个特征。线性模型可以是一类统计模型的总称，制作方法是用一定的流程将各个环节连接起来，包括线性回归模型和方差分析模型等。在一些实施例中，可以使用第一样本集训练线性回归模型，获取组成第一样本集的多个特征对应的多个权重。具体的，多元线性回归模型可以表示为$Y=\beta1x1+\beta2x2+\beta3x3....+b$，其中，β1为特征x1对应的权重、β2为特征x2对应的权重、β3为特征x3对应的权重、…。例如：图4B所示的样本1，包括w01、w03、w10、…、w500，共计10个特征，通过训练LR模型，可以获取与w01对应的β1、与w03对应的β3、与w10对应的β10、…、与w500对应的β500。第一样本集还包括样本2、样本3和样本4等多个样本，使用上述方法获取每个样本采样的特征对应的权重。在一些实施例中，权重的值可以表示对应特征和标签的相关强度，值越大表示对应特征与标签的相关度越高，对于最终测试结果的影响也越大。权重的符号表示对应特征和标签的相关方向，使用正号“+”和负号“-”来表示。例如：上述步骤示例中的“申请金额”特征可以拉高模型输出的分数，对应的权重符号为“+”；“申请时间”特征可以降低模型输出的分数，对应的权重符号为“-”。

[0045] (2)将多个权重中符号为正的权重按照值由大到小的顺序进行排序，选取前N个权重对应的特征作为正重要特征。为了使用最少的特征构造攻击样本，在一些实施例中，N的取值可以为3。例如：步骤(1)中获取的多个权重中，符号为正的有290个，按照由大到小的顺序将这290个权重排序：β2-80、β4-70.5、β180-69.8、β290-65…，选取β2、β4、β180对应的特征w02、w04、w180作为正重要特征。在一些实施例中，N的取值也可以是其他数值，不受本说明书的表述所限。

[0046] (3)将所述多个权重中符号为负的权重按照值由大到小的顺序进行排序，选取前N个权重对应的特征作为负重要特征。例如：步骤(1)中获取的多个权重中，符号为负的有210个，按照由大到小的顺序将这210个权重排序：β1-(-90.7)、β23-(-89)、β172-(-86.3)、β6-(-84.0)…，选取β1、β23、β172对应的特征w01、w23、w172作为负重要特征。

[0047] 步骤230，从第一样本集中选取至少一条样本作为原始样本，并基于原始样本和重要特征，构造第二样本集。具体的，该步骤可以由构造模块130执行。

[0048] 原始样本可以为没有添加扰动的正常样本。在一些实施例中，第一构造模块130可以从第一样本集中选取至少一条样本作为原始样本。例如：可以选取图4B所示的样本1作为

原始样本。在一些实施例中，也可以通过其他方式获取原始样本，不受本说明书的表述所限。

[0049] 在一些实施例中，可以基于选取的原始样本和步骤220获取的重要特征，构造第二样本集。在一些实施例中，可以从N个正重要特征中任意选取至少一个按照不同的组合添加到原始样本中，构造至少一条正第二样本。下面以原始样本为图4B所示的第一样本集中的样本1，正重要特征为步骤220中获取的w02、w04和w180为例说明：

[0050] 组合一：如图5所示，将w02添加到样本1中，构造一条正第二样本。具体的，将样本1中w02对应的采样位由0变为1，获取样本2_1。

[0051] 组合二：如图5所示，将w02和w04添加到样本1中，构造一条正第二样本。具体的，将样本1中的w02和w04对应的采样位由0变为1，获取样本2_2。

[0052] 组合三：如图5所示，将w02、w04和w180添加到样本1中，构造一条正第二样本。具体的，将样本1中的w02、w04和w180对应的采样位由0变为1，获取样本2_3。

[0053] 在一些实施例中，可以从N个负重要特征中任意选取至少一个按照不同的组合添加到原始样本中，构造至少一条负第二样本。下面以原始样本为图4B所示的第一样本集中的样本1，负重要特征为步骤220中获取的w01、w23和w172为例说明：

[0054] 组合一：如图6所示，将w01添加到样本1中，构造一条负第二样本。具体的，样本1中w01对应的采样位已经为1，因此将样本1作为样本2_4。

[0055] 组合二：如图6所示，将w01和w23添加到样本1中，构造一条负第二样本。具体的，将样本1中的w23对应的采样位由0变为1，获取样本2_5。

[0056] 组合三：如图6所示，将w01、w23和w172添加到样本1中，构造一条如第二样本。具体的，将样本1中的w23和w172对应的采样位由0变为1，获取样本2_6。

[0057] 上述示例构造了3个正第二样本和3个负第二样本，由这6条样本构成第二样本集。在一些实施例中，也可以使用其他组合方式构造正第二样本和负第二样本，不受本说明书的表述所限。

[0058] 步骤240，将第二样本集输入目标模型，基于目标模型输出的分数，确定扰动特征。具体的，该步骤可以由扰动特征确定模块140执行。

[0059] 在一些实施例中，扰动特征可以是在正常样本中额外添加的重要特征，用于改变被攻击模型的输出结果。在一些实施例中，扰动特征可以包括正扰动特征和负扰动特征，正扰动特征添加到正常的样本中后，可以拉高被攻击模型输出的分数，负扰动特征添加到正常的样本中后，可以降低被攻击模型输出的分数。在一些实施例中，扰动特征可以由步骤220中获取的重要特征中的至少一个组成。

[0060] 在一些实施例中，将步骤240中获取的第二样本集中的至少一条正第二样本输入目标模型，获取目标模型输出的至少一个分数。将获取的至少一个分数中最高分数对应的正第二样本中包含的正重要特征作为正扰动特征。以步骤240的示例中获取的第二样本集为例说明：将样本2_1输入目标模型中，目标模型输出分数0.7，将样本2_2输入目标模型中，目标模型输出分数0.9，将样本2_3输入目标模型中，目标模型输出分数0.8。3个分数中的最高分数0.9对应的样本为：样本2_2，样本2_2中包含的正重要特征为：w02和w04，因此将w02和w04作为正扰动特征。

[0061] 将所述第二样本集中的至少一条负第二样本输入所述目标模型，获取所述目标模

型输出的至少一个分数。将所述至少一个分数中最低分数对应的负第二样本中包含的负重要特征作为负扰动特征。以步骤240的示例中获取的第二样本集为例说明:将样本2_4输入目标模型中,目标模型输出分数0.3,将样本2_5输入目标模型中,目标模型输出分数0.5,将样本2_6输入目标模型中,目标模型输出分数0.4。3个分数中的最低分数0.3对应的样本为:样本2_4,样本2_4中包含的负重要特征为:w01,因此将w01作为负扰动特征。

[0062] 步骤250,使用扰动特征构造攻击样本。具体的,该步骤可以由第二构造模块150执行。

[0063] 在一些实施例中,从目标模型的测试数据中选取至少一条样本作为测试样本,将正扰动特征添加到测试样本中,获取正攻击样本。例如:目标模型为逾期率预估模型,选取一条测试样本,将步骤240中获取的正扰动特征w02和w04添加到该测试样本中,构造一条正攻击样本。

[0064] 在一些实施例中,将负扰动特征添加到测试样本中,获取负攻击样本。例如:选取上述示例中的测试样本,将步骤240中获取的负扰动特征w01添加到该测试样本中,构造一条负攻击样本。

[0065] 在一些实施例中,上述构造的攻击样本可以用于攻击目标模型,改变目标模型对于正常样本的判定结果。例如:一条正常样本输入目标模型,模型会输出一个大于0.5的分数,例如0.7,通过给该正常样本添加负扰动特征形成攻击样本,将该攻击样本输入目标模型,模型输出的分数变为0.4,即模型对该正常样本的判断发生了改变。

[0066] 本说明书实施例可能带来的有益效果包括但不限于:本说明书所述的实施例是针对输入为离散稀疏特征的目标模型,利用线性模型学习每个特征的重要性来选取重要特征,可以通过较少次数的尝试获取到具有较强攻击性的扰动特征。通过在正常样本中添加扰动特征构造攻击样本,使用攻击样本攻击目标模型,可以有效的升高/降低模型对于正常样本的打分。如果经过测试目标模型的鲁棒性需要提高,则可以在后续的模型训练过程中,在训练样本中添加扰动特征,从而提高模型的鲁棒性。需要说明的是,不同实施例可能产生的有益效果不同,在不同的实施例里,可能产生的有益效果可以是以上任意一种或几种的组合,也可以是其他任何可能获得的有益效果。

[0067] 应当注意的是,上述有关流程200的描述仅仅是为了示例和说明,而不限定本说明书的适用范围。对于本领域技术人员来说,在本说明书的指导下可以对流程200进行各种修正和改变。然而,这些修正和改变仍在本说明书的范围之内。例如,步骤230可以拆分为两个步骤230_1和230_2,在步骤230_1中从第一样本集中选取至少一条样本作为原始样本,在步骤230_2中构造第二样本集。

[0068] 图3是根据本说明书一些实施例所示的生成攻击样本的方法的示例性应用场景图。

[0069] 如图3所示,在一些实施例中,目标模型可以基于实体对象的数据进行预测,并根据预测结果确定后续的操作。实体对象的数据可以是用户数据和商户数据。其中,用户数据可以包括与用户相关的数据,例如用户输入的文本数据或语音数据等。可以使用目标模型预测用户是否会履行某项操作、用户的信用分数等。商户数据可以包括商户的位置数据、商户在工商登记的数据等。可以使用目标模型预测商户下一个季度的最高营收、客流高峰期等。在一些实施例中,可以使用本说明书中所描述的方法生成攻击样本,用于攻击目标模

型，测试目标模型的鲁棒性是否需要提高。详细测试方法请参见图2，这里不再赘述。

[0070] 本说明书所述的方法还可以应用于其他应用场景，不受本说明书的表述所限。

[0071] 上文已对基本概念做了描述，显然，对于本领域技术人员来说，上述详细披露仅仅作为示例，而并不构成对本说明书的限定。虽然此处并没有明确说明，本领域技术人员可能会对本说明书进行各种修改、改进和修正。该类修改、改进和修正在本说明书中被建议，所以该类修改、改进、修正仍属于本说明书示范实施例的精神和范围。

[0072] 同时，本说明书使用了特定词语来描述本说明书的实施例。如“一个实施例”、“一实施例”、和/或“一些实施例”意指与本说明书至少一个实施例相关的某一特征、结构或特点。因此，应强调并注意的是，本说明书中在不同位置两次或多次提及的“一实施例”或“一个实施例”或“一个替代性实施例”并不一定是指同一实施例。此外，本说明书的一个或多个实施例中的某些特征、结构或特点可以进行适当的组合。

[0073] 此外，本领域技术人员可以理解，本说明书的各方面可以通过若干具有可专利性的种类或情况进行说明和描述，包括任何新的和有用的工序、机器、产品或物质的组合，或对他们的任何新的和有用的改进。相应地，本说明书的各个方面可以完全由硬件执行、可以完全由软件(包括固件、常驻软件、微码等)执行、也可以由硬件和软件组合执行。以上硬件或软件均可被称为“数据块”、“模块”、“引擎”、“单元”、“组件”或“系统”。此外，本说明书的各方面可能表现为位于一个或多个计算机可读介质中的计算机产品，该产品包括计算机可读程序编码。

[0074] 计算机存储介质可能包含一个内含有计算机程序编码的传播数据信号，例如在基带上或作为载波的一部分。该传播信号可能有多种表现形式，包括电磁形式、光形式等，或合适的组合形式。计算机存储介质可以是除计算机可读存储介质之外的任何计算机可读介质，该介质可以通过连接至一个指令执行系统、装置或设备以实现通讯、传播或传输供使用的程序。位于计算机存储介质上的程序编码可以通过任何合适的介质进行传播，包括无线电、电缆、光纤电缆、RF、或类似介质，或任何上述介质的组合。

[0075] 本说明书各部分操作所需的计算机程序编码可以用任意一种或多种程序语言编写，包括面向对象编程语言如Java、Scala、Smalltalk、Eiffel、JADE、Emerald、C++、C#、VB.NET、Python等，常规程序化编程语言如C语言、Visual Basic、Fortran 2003、Perl、COBOL 2002、PHP、ABAP，动态编程语言如Python、Ruby和Groovy，或其他编程语言等。该程序编码可以完全在用户计算机上运行、或作为独立的软件包在用户计算机上运行、或部分在用户计算机上运行部分在远程计算机运行、或完全在远程计算机或服务器上运行。在后种情况下，远程计算机可以通过任何网络形式与用户计算机连接，比如局域网(LAN)或广域网(WAN)，或连接至外部计算机(例如通过因特网)，或在云计算环境中，或作为服务使用如软件即服务(SaaS)。

[0076] 此外，除非权利要求中明确说明，本说明书所述处理元素和序列的顺序、数字字母的使用、或其他名称的使用，并非用于限定本说明书流程和方法的顺序。尽管上述披露中通过各种示例讨论了一些目前认为有用的发明实施例，但应当理解的是，该类细节仅起到说明的目的，附加的权利要求并不仅限于披露的实施例，相反，权利要求旨在覆盖所有符合本说明书实施例实质和范围的修正和等价组合。例如，虽然以上所描述的系统组件可以通过硬件设备实现，但是也可以只通过软件的解决方案得以实现，如在现有的服务器或移动设

备上安装所描述的系统。

[0077] 同理，应当注意的是，为了简化本说明书披露的表述，从而帮助对一个或多个发明实施例的理解，前文对本说明书实施例的描述中，有时会将多种特征归并至一个实施例、附图或对其的描述中。但是，这种披露方法并不意味着本说明书对象所需要的特征比权利要求中提及的特征多。实际上，实施例的特征要少于上述披露的单个实施例的全部特征。

[0078] 一些实施例中使用了描述成分、属性数量的数字，应当理解的是，此类用于实施例描述的数字，在一些示例中使用了修饰词“大约”、“近似”或“大体上”来修饰。除非另外说明，“大约”、“近似”或“大体上”表明所述数字允许有±20％的变化。相应地，在一些实施例中，说明书和权利要求中使用的数值参数均为近似值，该近似值根据个别实施例所需特点可以发生改变。在一些实施例中，数值参数应考虑规定的有效数位并采用一般位数保留的方法。尽管本说明书一些实施例中用于确认其范围广度的数值域和参数为近似值，在具体实施例中，此类数值的设定在可行范围内尽可能精确。

[0079] 针对本说明书引用的每个专利、专利申请、专利申请公开物和其他材料，如文章、书籍、说明书、出版物、文档等，特此将其全部内容并入本说明书作为参考。与本说明书内容不一致或产生冲突的申请历史文件除外，对本说明书权利要求最广范围有限制的文件(当前或之后附加于本说明书中的)也除外。需要说明的是，如果本说明书附属材料中的描述、定义、和/或术语的使用与本说明书所述内容有不一致或冲突的地方，以本说明书的描述、定义和/或术语的使用为准。

[0080] 最后，应当理解的是，本说明书中所述实施例仅用以说明本说明书实施例的原则。其他的变形也可能属于本说明书的范围。因此，作为示例而非限制，本说明书实施例的替代配置可视为与本说明书的教导一致。相应地，本说明书的实施例不仅限于本说明书明确介绍和描述的实施例。

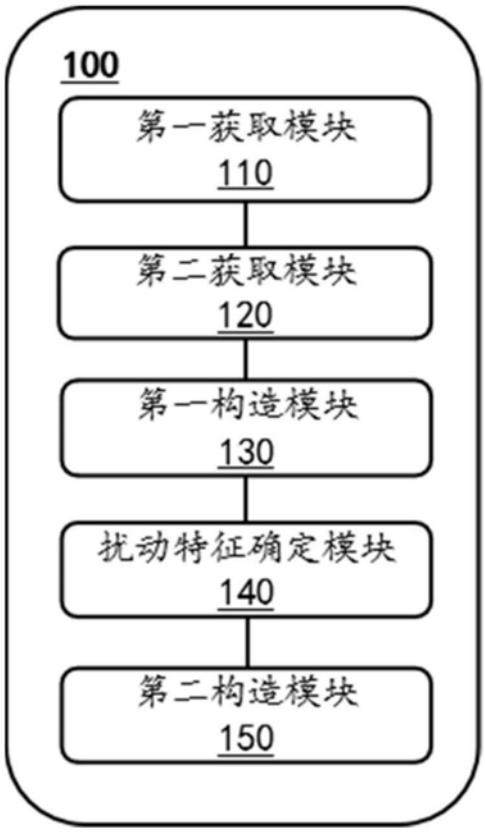

100
第一获取模块
110
第二获取模块
120
第一构造模块
130
扰动特征确定模块
140
第二构造模块
150

图1

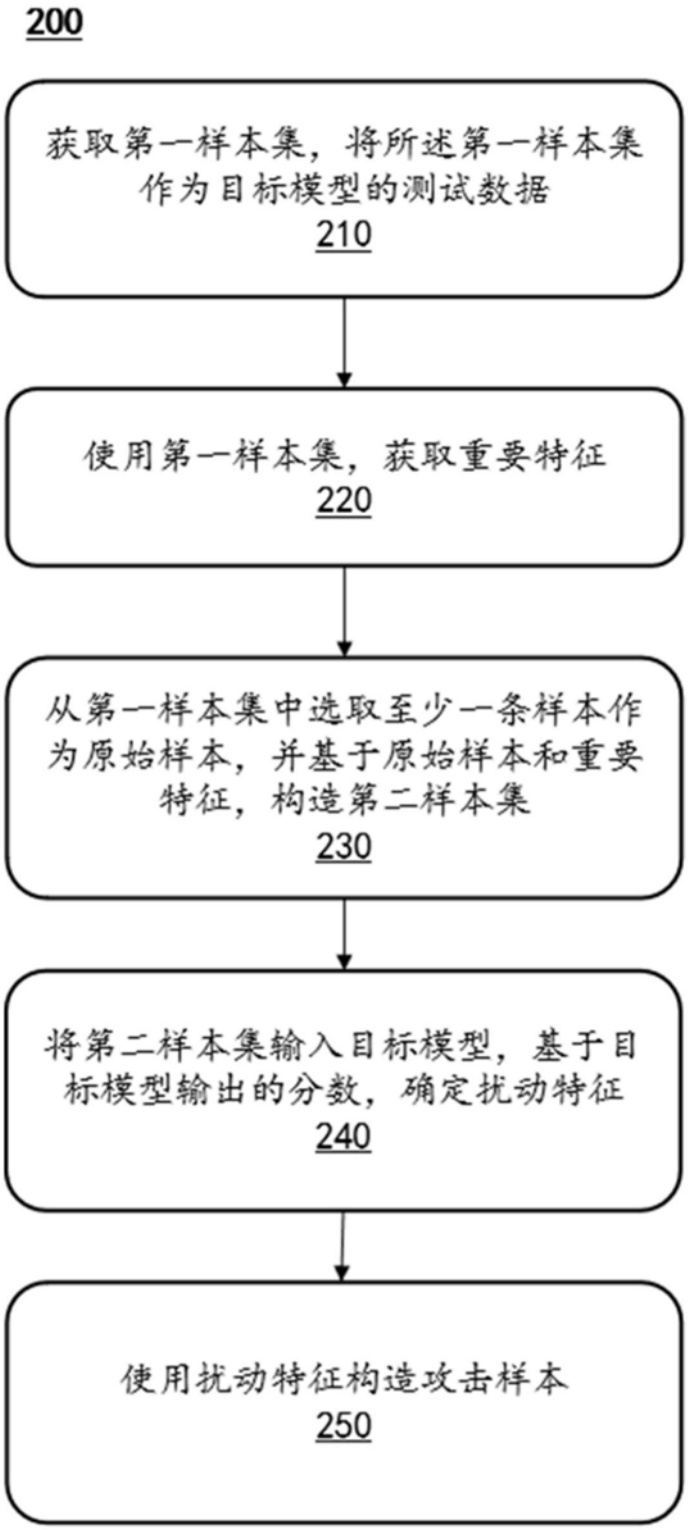

200
获取第一样本集，将所述第一样本集作为目标模型的测试数据
210
使用第一样本集，获取重要特征
220
从第一样本集中选取至少一条样本作为原始样本，并基于原始样本和重要特征，构造第二样本集
230
将第二样本集输入目标模型，基于目标模型输出的分数，确定扰动特征
240
使用扰动特征构造攻击样本
250

图2

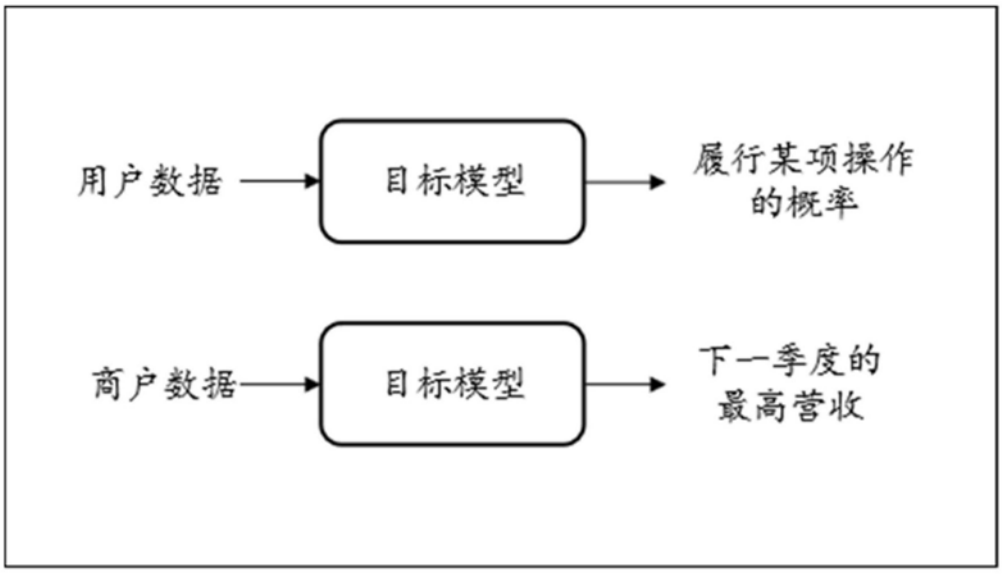


图3

特征空间：

w01	w02	w03	…	w10
w11	w12	w13	…	w20
w21	w22	w23	…	w30
…	…	…	…	…
w491	w492	w493	…	w500

图4A

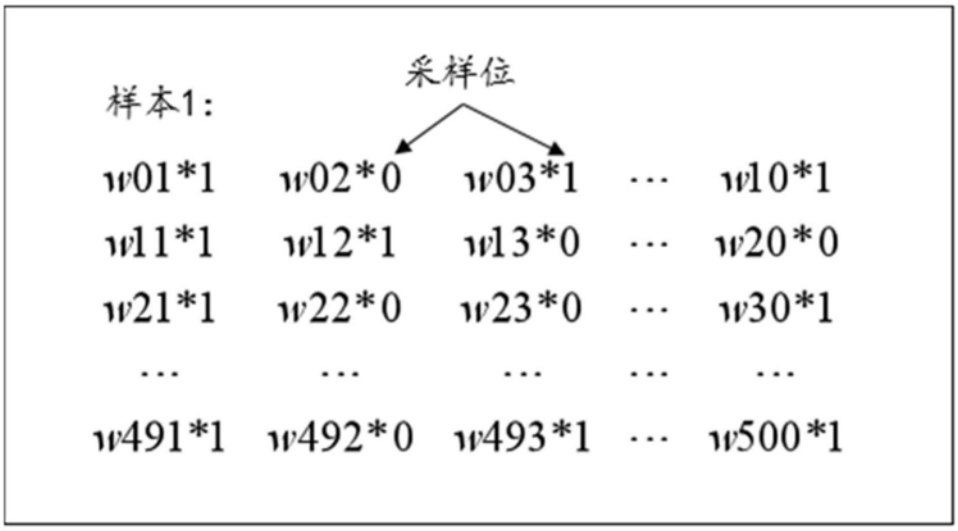


图4B

(1, 0, 0…0)*1, (0, 1, 0…0)*0 (0, 0, 1…0)*1 … (0, 0…0, …1)*1

图4C

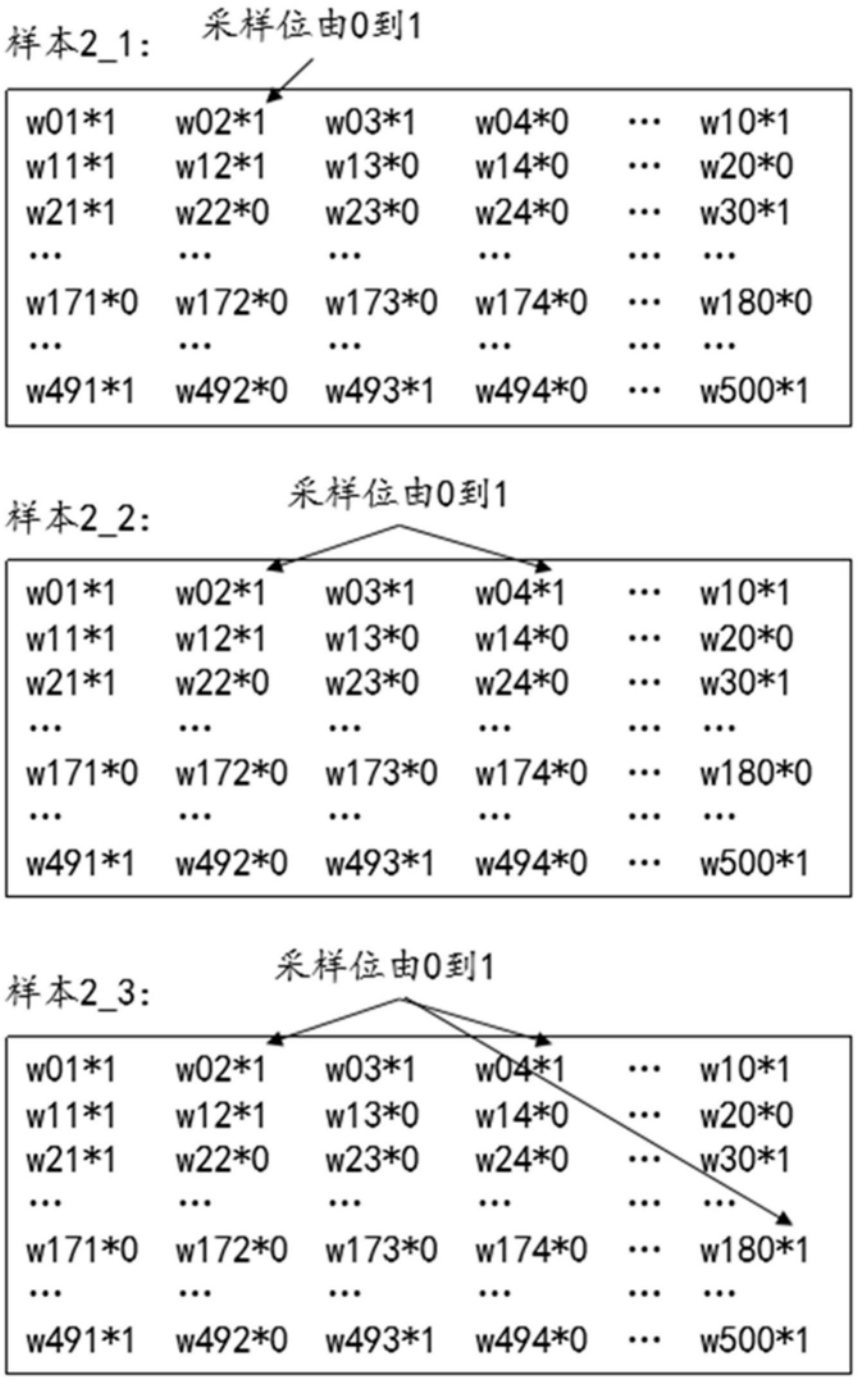

样本2_1:
采样位由0到1
w01*1 w02*1 w03*1 w04*0 … w10*1
w11*1 w12*1 w13*0 w14*0 … w20*0
w21*1 w22*0 w23*0 w24*0 … w30*1
… … … … … …
w171*0 w172*0 w173*0 w174*0 … w180*0
… … … … … …
w491*1 w492*0 w493*1 w494*0 … w500*1
样本2_2:
采样位由0到1
w01*1 w02*1 w03*1 w04*1 … w10*1
w11*1 w12*1 w13*0 w14*0 … w20*0
w21*1 w22*0 w23*0 w24*0 … w30*1
… … … … … …
w171*0 w172*0 w173*0 w174*0 … w180*0
… … … … … …
w491*1 w492*0 w493*1 w494*0 … w500*1
样本2_3:
采样位由0到1
w01*1 w02*1 w03*1 w04*1 … w10*1
w11*1 w12*1 w13*0 w14*0 … w20*0
w21*1 w22*0 w23*0 w24*0 … w30*1
… … … … … …
w171*0 w172*0 w173*0 w174*0 … w180*1
… … … … … …
w491*1 w492*0 w493*1 w494*0 … w500*1

图5

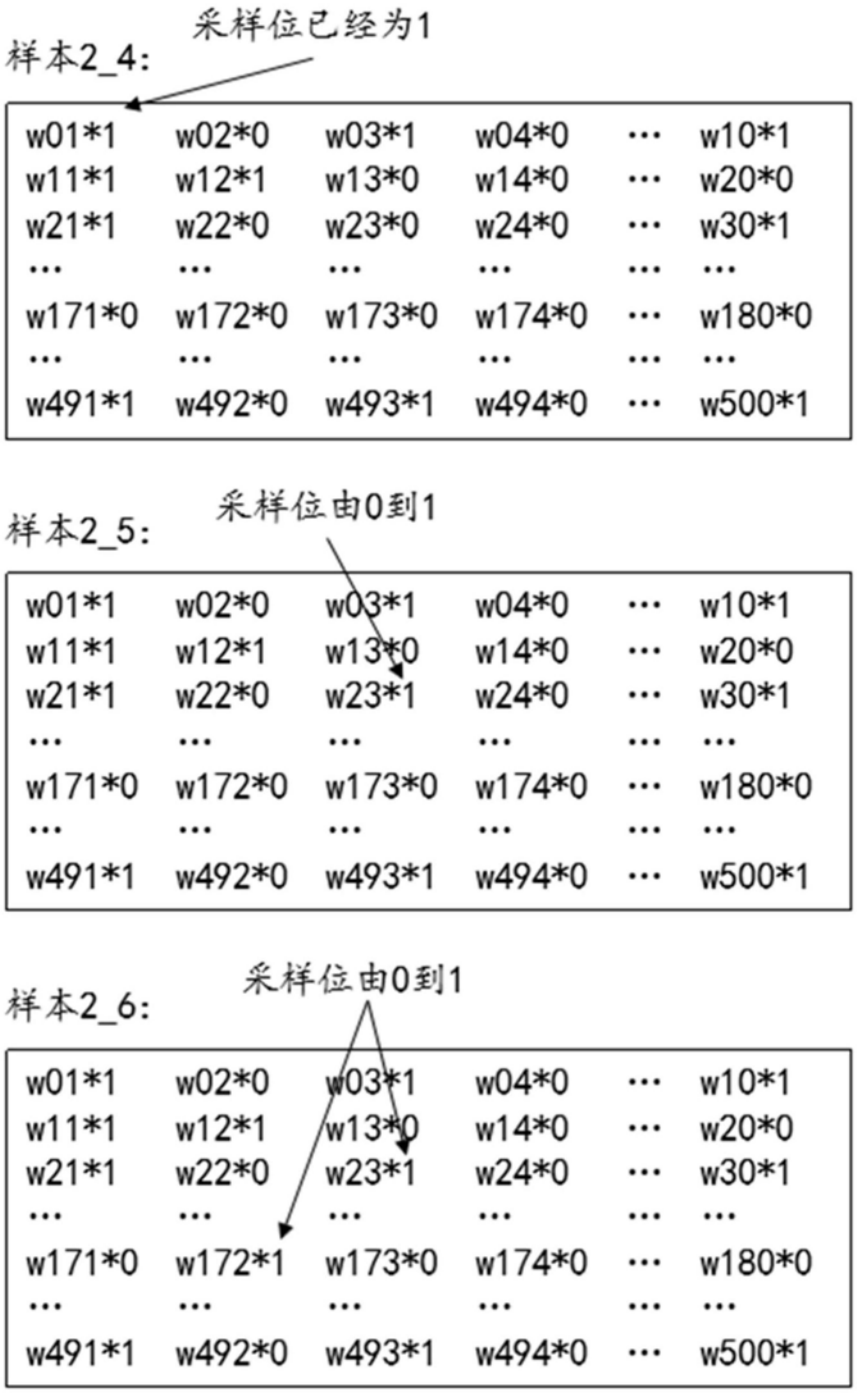

采样位已经为1
样本2_4:
w01*1 w02*0 w03*1 w04*0 … w10*1
w11*1 w12*1 w13*0 w14*0 … w20*0
w21*1 w22*0 w23*0 w24*0 … w30*1
… … … … … …
w171*0 w172*0 w173*0 w174*0 … w180*0
… … … … … …
w491*1 w492*0 w493*1 w494*0 … w500*1
采样位由0到1
样本2_5:
w01*1 w02*0 w03*1 w04*0 … w10*1
w11*1 w12*1 w13*0 w14*0 … w20*0
w21*1 w22*0 w23*1 w24*0 … w30*1
… … … … … …
w171*0 w172*0 w173*0 w174*0 … w180*0
… … … … … …
w491*1 w492*0 w493*1 w494*0 … w500*1
采样位由0到1
样本2_6:
w01*1 w02*0 w03*1 w04*0 … w10*1
w11*1 w12*1 w13*0 w14*0 … w20*0
w21*1 w22*0 w23*1 w24*0 … w30*1
… … … … … …
w171*0 w172*1 w173*0 w174*0 … w180*0
… … … … … …
w491*1 w492*0 w493*1 w494*0 … w500*1

图6