(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 122132312 A
(43)申请公布日 2026.06.02

(21)申请号 202610243080.6

(22)申请日 2026.02.28

(71)申请人 启元实验室
地址 100084 北京市海淀区紫雀路55号院8号楼-1至6层101

(72)发明人 裴华鑫 许伟超 杨敬轩 褚栖桐 赵千川

(74)专利代理机构 北京律和信知识产权代理事务所(普通合伙) 11446
专利代理师 武玉琴

(51)Int.Cl.
G06F 11/3668(2025.01)
G06N 3/0499(2023.01)
G06N 3/084(2023.01)
G01M 17/00(2006.01)
B64F 5/60(2017.01)
B63B 71/00(2020.01)

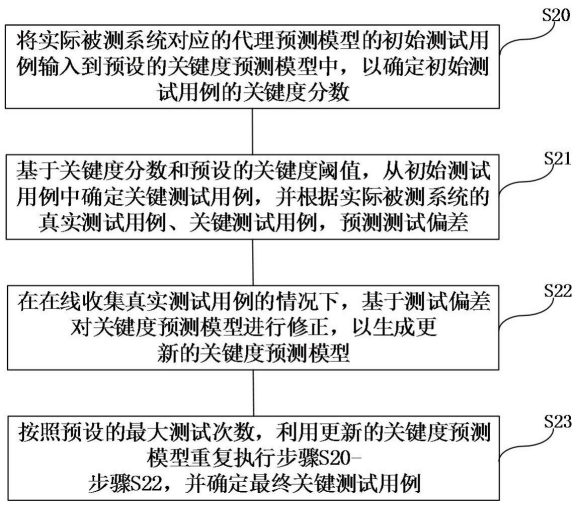
权利要求书2页 说明书13页 附图5页

(54)发明名称

一种面向无人系统差异性的弹性测试方法及装置

(57)摘要

本申请提供一种面向无人系统差异性的弹性测试方法及装置,涉及人工智能技术领域。步骤1:将实际被测系统对应的代理预测模型的初始测试用例输入到关键度预测模型中,以确定初始测试用例的关键度分数;步骤2:基于关键度分数和关键度阈值从初始测试用例中确定关键测试用例,并根据实际被测系统的真实测试用例、关键测试用例预测测试偏差;步骤3:在在线收集真实测试用例的情况下基于测试偏差对关键度预测模型进行修正以生成更新的关键度预测模型;步骤4:按照预设的最大测试次数,利用更新的关键度预测模型重复执行步骤1-3并确定最终关键测试用例。提升了智能测试技术在不同型号无人系统测试中的弹性能力,实现测试用例库的快速优化与适配。



CN 122132312 A

1.一种面向无人系统差异性的弹性测试方法,其特征在于,包括:

步骤1:将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中,以确定所述初始测试用例的关键度分数;

步骤2:基于所述关键度分数和预设的关键度阈值,从所述初始测试用例中确定关键测试用例,并根据所述实际被测系统的真实测试用例、所述关键测试用例,预测测试偏差;

步骤3:在在线收集所述真实测试用例的情况下,基于所述测试偏差对所述关键度预测模型进行修正,以生成更新的关键度预测模型;

步骤4:按照预设的最大测试次数,利用所述更新的关键度预测模型重复执行步骤1-3,并确定最终关键测试用例。

2.根据权利要求1所述的方法,其特征在于,所述基于所述关键度分数和预设的关键度阈值,从所述初始测试用例中确定关键测试用例,并根据所述实际被测系统的真实测试用例、所述关键测试用例,预测测试偏差,包括:

在所述关键度分数大于所述关键度阈值的情况下,将所述关键度分数对应的初始测试用例确定为关键测试用例;

根据所述实际被测系统的真实测试用例、所述关键测试用例,预测测试偏差。

3.根据权利要求2所述的方法,其特征在于,所述根据所述实际被测系统的真实测试用例、所述关键测试用例,预测测试偏差,包括:

在获取到所述实际被测系统的离线真实数据集的情况下,从所述离线真实数据集中提取所述真实测试用例;

利用混淆矩阵,根据所述关键测试用例中的预测标签和所述真实测试用例中的真实标签,确定测试过程的二分类指标;

基于所述二分类指标,预测所述测试偏差。

4.根据权利要求2所述的方法,其特征在于,所述根据所述实际被测系统的真实测试用例、所述关键测试用例,预测测试偏差,包括:

在未获取到所述实际被测系统的离线真实数据集的情况下,

基于所述代理预测模型,利用所述关键测试用例对所述实际被测系统进行测试,以生成对应的真实测试用例;

基于预设的、基于神经网络的关键度预测函数、所述关键测试用例的所述关键度分数和所述真实测试用例的真实标签,确定交叉熵损失;

基于所述交叉熵损失,预测所述测试偏差。

5.根据权利要求4所述的方法,其特征在于,所述在在线收集所述真实测试用例的情况下,基于所述测试偏差对所述关键度预测模型进行修正,以生成更新的关键度预测模型,包括:

在在线收集所述真实测试用例的情况下,利用预设训练策略,以所述交叉熵损失为损失函数,在训练数据集中进行预设的微调轮次的训练,以生成所述更新的代理预测模型,其中,所述预设训练策略为与修正前的所述代理预测模型使用的训练方式相同的类别均衡策略。

6.根据权利要求2所述的方法,其特征在于,所述在所述关键度分数大于所述关键度阈值的情况下,将所述关键度分数对应的初始测试用例确定为关键测试用例,包括:

在所述关键度分数大于所述关键度阈值的情况下,按照预设采样策略中采样概率,从自然分布中进行采样,以从所述初始测试用例中确定所述关键测试用例。

7.根据权利要求6所述的方法,其特征在于,所述在所述关键度分数大于所述关键度阈值的情况下,按照预设采样策略中采样概率,从自然分布中进行采样,以从所述初始测试用例中确定所述关键测试用例,包括:

在所述关键度分数大于所述关键度阈值的情况下,按照所述采样概率,从所述自然分布中进行采样,以从所述初始测试用例中确定预备测试用例;

从均匀分布中抽取随机数,并在所述随机数小于所述采样概率的情况下,将所述预备测试用例确定为所述关键测试用例;

在所述随机数大于或者等于所述采样概率的情况下,持续从自然分布中采样得到更新的预备测试用例,并将所述更新的预备测试用例的关键度分数与所述关键度阈值进行对比。

8.根据权利要求6所述的方法,其特征在于,还包括:

在所述关键度分数小于或者等于所述关键度阈值的情况下,按照剩余采样概率,从自然分布中进行重采样得到新的初始测试用例,并比较所述新的初始测试用例的关键度分数和所述关键度阈值,其中,所述剩余采样概率为1与所述采样概率的差值。

9.一种面向无人系统差异性的弹性测试装置,其特征在于,包括:

分数确定模块,用于将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中,以确定所述初始测试用例的关键度分数;

测试偏差确定模块,用于基于所述关键度分数和预设的关键度阈值,从所述初始测试用例中确定关键测试用例,并根据所述实际被测系统的真实测试用例、所述关键测试用例,预测测试偏差;

模型修正模块,用于在在线收集所述真实测试用例的情况下,基于所述测试偏差对所述关键度预测模型进行修正,以生成更新的关键度预测模型;

循环测试修正模块,用于按照预设的最大测试次数,利用所述更新的关键度预测模型重复调用所述分数确定模块、所述测试偏差确定模块和所述模型修正模块,并确定最终关键测试用例。

10.一种电子设备,其特征在于,包括:

处理器;

存储器,存储有计算机程序,当所述计算机程序被所述处理器执行时,使得所述处理器执行如权利要求1-8任一项所述的方法。

一种面向无人系统差异性的弹性测试方法及装置

技术领域

[0001] 本申请涉及人工智能技术领域,例如涉及一种面向无人系统差异性的弹性测试方法及装置。

背景技术

[0002] 无人系统测试是指对无人系统(如:无人车、无人机、无人船等)的功能与性能进行测试验证,以确定其在特定任务下是否具备脱离人类操控并自主完成任务的能力。有别于传统系统,智能无人系统具有应用范围广、需求差异大、承担任务复杂、行为机理不明确等特性。无人系统的测试面临系统结构未知、工作状态难以获取、测试用例空间维数高等问题,采用传统的测试技术将导致测试效率低、测试成本高、能力边界难以触及,严重滞缓了无人系统的迭代优化与应用进程。无人系统智能测试是指引入人工智能技术生成测试用例库,实现在超高维测试用例空间中高效搜索出具有代表性的关键测试用例库,以加速摸清无人系统的能力边界,最终推动解决传统测评手段难以满足智能无人系统测试需求的难题。实际上,智能测试的核心思想是“AI(Artificial Intelligence,人工智能)测AI”,一方面,被测对象(即:智能无人系统)是智能化水平较高的系统;另一方面,测试手段同样具有较高的智能化水平。“AI测AI”技术在智能无人系统测试验证方面已初步展现出优势与潜力,为提升智能测试的实际应用价值,相关基础理论与关键技术仍有待深入研究。作为无人系统从研发到实际应用中的关键环节,智能无人系统的测试面临系统结构未知、工作状态难以获取、测试用例空间维数高等问题,采用传统的测试技术将导致测试效率低、测试成本高、关键性能考核不充分等问题,严重滞缓了无人系统的迭代优化与应用进程。以“AI测AI”为核心思想的智能测试技术已初显优势,在加速摸清无人系统的能力边界、降低测试成本方面展现出了极大的潜力。

[0003] 相关技术中,智能测试高度依赖训练数据的规模与质量,使得采用无人系统的代理预测模型生成测试用例库是实施智能测试的重要手段。然而,由于代理预测模型与实际被测系统之间存在差异,导致离线生成的测试用例库通用性差,难以较好地适用于机理不同、性能不同的无人系统的测试验证,严重限制了智能测试技术的快速推广与应用。因此,迫切需要提升智能测试技术在不同型号无人系统测试中的弹性能力,实现测试用例库的快速优化与适配。

发明内容

[0004] 本申请旨在提供一种面向无人系统差异性的弹性测试方法及装置。

[0005] 根据本申请的一方面,提出一种面向无人系统差异性的弹性测试方法,包括:步骤1:将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中,以确定初始测试用例的关键度分数;步骤2:基于关键度分数和预设的关键度阈值,从初始测试用例中确定关键测试用例,并根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差;步骤3:在在线收集真实测试用例的情况下,基于测试偏差对关键度预测模型进

行修正,以生成更新的关键度预测模型;步骤4:按照预设的最大测试次数,利用更新的关键度预测模型重复执行步骤1-3,并确定最终关键测试用例。

[0006] 根据本申请的一方面,提出一种面向无人系统差异性的弹性测试装置,包括:

[0007] 分数确定模块,用于将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中,以确定初始测试用例的关键度分数;

测试偏差确定模块,用于基于关键度分数和预设的关键度阈值,从初始测试用例中确定关键测试用例,并根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差;

模型修正模块,用于在在线收集真实测试用例的情况下,基于测试偏差对关键度预测模型进行修正,以生成更新的关键度预测模型;

循环测试修正模块,用于按照预设的最大测试次数,利用更新的关键度预测模型重复调用分数确定模块、测试偏差确定模块和模型修正模块,并确定最终关键测试用例。

[0008] 根据本申请的一方面,提出一种电子设备,该电子设备包括:处理器;存储器,存储有计算机程序,当计算机程序被处理器执行时,使得处理器执行如上文的方法。

[0009] 根据本申请的一方面,提出一种非瞬时性计算机可读介质,其上存储有可读指令,当指令被处理器执行时,使得处理器执行如上文中的方法。

[0010] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性的,并不能限制本申请。

[0011] 有益效果:

通过本申请所提供的上述实施例,引入智能弹性测试机制,通过构建代理预测模型与关键度预测模型相结合的动态评估体系,有效量化了无人系统代理模型与实际系统之间的差异性,为解决测试用例通用性差的难题提供了可衡量的技术路径。基于主动学习的自适应修正框架,能够在在线收集真实测试用例的过程中实时检测并修正测试偏差,显著提升了测试用例库在资源受限条件下的快速优化与适配能力,实现了测试策略从静态预设到动态弹性的跨越。针对无人系统高维参数空间导致的维数灾难问题,创新融合深度学习技术,通过关键度预测模型对海量测试用例进行智能筛选与优先级划分,突破了传统方法在复杂场景下处理高维数据的局限性,大幅提升了测试效率与覆盖率。构建了“AI测AI”的闭环优化体系,通过多次迭代的模型修正与偏差收敛机制,实现了代理模型与实际系统差异性的持续缩减,进而得到更加准确的测试用例库。

附图说明

[0012] 为了更清楚地说明本申请实施例中的技术方案,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,还可以根据这些附图获得其他的附图,而并不超出本申请要求保护的范围。

[0013] 图1为本申请实施例提供的智能弹性测试的框架的示意图;

图2为本申请实施例提供的面向无人系统差异性的弹性测试方法的流程图;

图3为本申请实施例提供的代理预测模型的多层感知机的结构组成示意图;

图4为本申请实施例提供的面向无人系统差异性的弹性测试装置的框图;

图5为本申请实施例提供的电子设备的结构示意图。

具体实施方式

[0014] 现在将参考附图更全面地描述示例实施例。然而,示例实施例能够以多种形式实施,且不应被理解为限于在此阐述的实施例;相反,提供这些实施例使得本申请将全面和完整,并将示例实施例的构思全面地传达给本领域的技术人员。在图中相同的附图标记表示相同或类似的部分,因而将省略对它们的重复描述。

[0015] 此外,所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中,提供许多具体细节从而给出对本申请的实施例的充分理解。然而,本领域技术人员将意识到,可以实践本申请的技术方案而没有特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知方法、装置、实现或者操作以避免模糊本申请的各方面。

[0016] 附图中所示的方框图仅仅是功能实体,不一定必须与物理上独立的实体相对应。即,可以采用软件形式来实现这些功能实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0017] 附图中所示的流程图仅是示例性说明,不是必须包括所有的内容和操作/步骤,也不是必须按所描述的顺序执行。例如,有的操作/步骤还可以分解,而有的操作/步骤可以合并或部分合并,因此实际执行的顺序有可能根据实际情况改变。

[0018] 应理解,虽然本文中可能使用术语第一、第二、第三等来描述各种组件,但这些组件不应受这些术语限制。这些术语乃用以区分一组件与另一组件。因此,下文论述的第一组件可称为第二组件而不偏离本申请概念的教示。如本文中所使用,术语“及/或”包括相关联的列出项目中的任一个及一或多者的所有组合。

[0019] 本申请提出的智能测试方法是一种弹性测试方法,旨在提升智能测试技术在不同型号无人系统测试中的弹性能力,推动解决无人系统智能测试技术通用性差的问题。具体而言,基于人工智能的测试技术高度依赖训练数据的规模与质量,使得采用无人系统代理预测模型生成测试用例库是实施“AI测AI”的重要手段。然而,由于代理预测模型与实际被测系统之间存在差异,导致采用代理预测模型生成的测试用例库通用性差,难以较好地适用于机理不同、性能不同的无人系统的测试验证,严重限制了智能测试技术的快速推广与应用。在采用代理预测模型离线构建的测试用例库的基础上,智能弹性测试技术旨在根据真实被测对象的特性实现测试用例库的在线优化与适配,大幅提升智能测试技术的实际应用价值。

[0020] 基于此,本申请提供了一种面向无人系统差异性的弹性测试方法及装置。

[0021] 图1为本申请实施例提供的智能弹性测试的框架的示意图。本申请中可以利用代理模型(代理预测模型)对测试用例关键度预测模型(后文称为关键度预测模型)进行训练。利用采样策略和关键度预测模型输出的关键度分数来确定合适的测试用例,即关键测试用例,并在实际被测系统上进行关键测试用例的执行测试,得到测试数据用于预测测试用例的关键度的预测偏差,进而在在线状态下对关键度预测模型进行模型修正,实现在线优化。

[0022] 具体的实现方式可以参考以下实施例。

[0023] 图2为本申请实施例提供的面向无人系统差异性的弹性测试方法的流程图。如图2

所示,该方法包括:步骤S20、步骤S21、步骤S22和步骤S23。

[0024] 在步骤S20中,将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中,以确定初始测试用例的关键度分数。

[0025] 本申请中,代理预测模型一般是与实际被测系统具有相似或相同工作环境的其他不同模型。可以是“替身”或“双胞胎”,也可以是相同工作环境中其他算法模型(偏差更大)。无论是何种,都面临与实际被测对象的偏差带来的影响。

[0026] 初始测试用例可以为代理预测模型经过了多次实验测试后的数据对,包含测试参数和最终测试任务的成功或者失败的标签,参数和标签整体作为一个初始测试用例。可以预先构建关键度预测模型,将初始测试用例输入该模型,可以直接输出每个初始测试用例的关键度分数。

[0027] 在一些实现方式中,代理预测模型可以通过深度学习的数据驱动方式获得。代理预测模型可以理解为一个将高维参数空间S映射到区间[0,1]的关键度预测函数。然而由于应用环境复杂,被测对象内部机理不明,被测对象在某一测试用例下的执行情况难以预测,导致关键度预测函数获得困难,本申请具体可以使用基于深度学习的方式训练以人工神经网络为基础的代理预测模型。

[0028] 利用梯度反向传播训练所搭建的人工神经网络。多层感知机(Multi-Layer Perceptron,MLP)是人工神经网络的一类基础算法,其通过多个全连接层和非线性激活函数可以实现对函数的拟合。利用反向传播算法,多层感知机能够从足够数量的输入与输出数据中学习到数据的信息,并拟合出输入到输出的映射函数。本申请采用的代理预测模型由多层多层感知机组成,并使用ReLU函数为中间层激活函数,Sigmoid函数将输出层输出标准化到[0,1]之间,代理预测模型的多层感知机的结构组成可以如图3所示。

[0029] 训练损失使用二分类交叉熵损失,并使用Adam(Adaptive Moment Estimation Optimizer,自适应矩估计优化器)进行梯度下降。由于关键测试用例的稀疏性,导致失败的测试用例占比极小,由此导致的类别不平衡问题会严重影响多层感知机的训练,导致多层感知机更加倾向于将测试用例预测为负样本,而无法学习到输入到输出的映射关系,本专利采用类别均衡采样策略。具体而言,在多层感知机训练时需要多个训练样本组合成一个训练批次统一计算损失函数并进行梯度传播,在构建训练批次时,调整正样本与负样本的采样概率,从而使得每个批次正负样本比例接近1:1,具体而言,假设正负样本数量比为a:b,则每个正负样本的采样概率比例为b:a。

[0030] 在步骤S21中,基于关键度分数和预设的关键度阈值,从初始测试用例中确定关键测试用例,并根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差。

[0031] 本申请中,可以预先设置受到关键度分数和关键度阈值的对比情况影响的关键测试用例的选取方式。基于选取方式和当前的关键度分数、预设的关键度阈值,从初始测试用例中选取出关键测试用例。关键度阈值可以设定为0.5,但是在一些实现方式中,根据深度学习二分类的思想,阈值可以随意设定,越高代表越严格,可能会过滤掉过多数据,但预测的准确率也一般更高。

[0032] 本申请还可以预先设置测试偏差的预测方式。首先获取实际被测系统的真实测试用例,即实际被测系统按照关键测试用例对应的测试参数运行并产生对应的标签得到的。在一些实现方式中,测试偏差可以用其他特征来表征,例如召回率、精准率等。本申请可以

按照预测方式,代入真实测试用例、关键测试用例,得到用于预测测试偏差的参数,可以预先设置这些参数与测试偏差之间的对应关系。

[0033] 在步骤S22中,在在线收集真实测试用例的情况下,基于测试偏差对关键度预测模型进行修正,以生成更新的关键度预测模型。

[0034] 本申请中,真实测试用例可能是在线收集的,也可能是离线收集的。针对在线收集的情况下,可以利用测试偏差来选择修正参数或者函数,对关键度预测模型进行修正,得到更新的关键度预测模型。

[0035] 在一些实现方式中,离线状态在此方法里不考虑修正,即离线可以认为是整个测试过程中不会根据实时测试反馈调整的关键度预测模型。

[0036] 在步骤S23中,按照预设的最大测试次数,利用更新的关键度预测模型重复执行步骤S20-步骤S22,并确定最终关键测试用例。

[0037] 本申请可以根据实际要求预先设置最大测试次数。将初始测试用例输入到新的关键度预测模型,输出对应的关键度分数,然后再次预测测试偏差,直到循环了最大测试次数后停止,得到最终的关键测试用例。

[0038] 本申请引入智能弹性测试机制,通过构建代理预测模型与关键度预测模型相结合的动态评估体系,有效量化了无人系统代理模型与实际系统之间的差异性,为解决测试用例通用性差的难题提供了可衡量的技术路径。基于主动学习的自适应修正框架,能够在在线收集真实测试用例的过程中实时检测并修正测试偏差,显著提升了测试用例库在资源受限条件下的快速优化与适配能力,实现了测试策略从静态预设到动态弹性的跨越。针对无人系统高维参数空间导致的维数灾难问题,创新融合深度学习技术,通过关键度预测模型对海量测试用例进行智能筛选与优先级划分,突破了传统方法在复杂场景下处理高维数据的局限性,大幅提升了测试效率与覆盖率。构建了“AI测AI”的闭环优化体系,通过多次迭代的模型修正与偏差收敛机制,实现了代理模型与实际系统差异性的持续缩减,进而得到更加准确的测试用例库。

[0039] 根据一些实施例,在确定关键测试用例的过程中,具体可以在关键度分数大于关键度阈值的情况下,将关键度分数对应的初始测试用例确定为关键测试用例;根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差。

[0040] 本申请中,可以将关键度分数和关键度阈值进行对比,如果关键度分数大于关键度阈值,则可以将关键度分数对应的初始测试用例作为关键测试用例,其中,关键测试用例可以用于表征任务执行的结果为失败标签的任务对应的测试用例。

[0041] 如果关键度分数小于或者等于关键度阈值,那么剔除对应的初始测试用例,实现测试用例的筛选。然后可以根据真实测试用例、关键测试用例来预测测试偏差。

[0042] 本申请通过设定关键度阈值,自动从海量初始测试用例中筛选出具有高关键性的测试用例,有效解决了在庞大参数空间和有限测试资源下“测什么”的决策难题,使测试活动能够精准聚焦于最可能暴露系统差异性的关键场景,极大提升了测试效率。基于筛选出的关键测试用例与实际系统的真实测试结果进行比对,能够直接、量化地预测和度量代理模型与实际被测系统之间的测试偏差。这为客观评估代理模型的保真度、以及诊断模型与实物差异的具体表现提供了关键技术手段。

[0043] 根据一些实施例,如果存在离线真实数据集,预测测试偏差的过程中,具体可以从

离线真实数据集中提取真实测试用例;利用混淆矩阵,根据关键测试用例中的预测标签和真实测试用例中的真实标签,确定测试过程的二分类指标;基于二分类指标,预测测试偏差。

[0044] 本申请中,虽然不考虑离线状态下的模型修正,但是可以计算离线状态下某一关键用例的测试偏差,作为最终组成的合适的测试用例库的基础。

[0045] 在具备离线真实数据集的情况下,可以通过二分类问题的预测标签与真实标签的相符与差异的情况来衡量二分类的准确率。一般使用混淆矩阵来统计预测结果,根据在真实数据集中统计预测标签与真实标签是否一致,可以分为4种情况:

真实标签和预测标签都为正(即真实测试用例和关键测试用例中任务的执行结果都为触发了失败事件),表征正确的预测,用TP表示,其真值为正;

真实标签和预测标签都为负(即真实测试用例和关键测试用例中任务的执行结果都为触发了成功事件),表征正确的预测,用TN表示,其真值为负;

真实标签为正但预测标签为负,表征错误的预测,用FN表示;

真实标签为负但预测标签为正,表征错误的预测,用FP表示。

[0046] 利用混淆矩阵和上述的预测标签来计算二分类指标,可以包括精确率(Precision)、召回率(Precision-Recall,PR)和平均精度(Average Precision,AP)等,具体计算方式如下:

$$\text{Precision} = \frac{TP}{TP+FP},$$

$$\text{PR} = \frac{TP}{TP+FN}。$$

[0047] 精确率定义在所有被预测为正类的样本中实际为正类的比例,用于衡量模型预测正类时的准确性,而召回率描述了实际为正类的样本中,有多少被模型正确识别为正类,反映了模型对正类样本的敏感性。在复杂问题中,精确率与召回率一般无法同时做得最优,例如提高分类阈值会带来精确率提升,但降低召回率。平均精度则是一个更加全面的性能评价标准。平均精度通过计算精确率-召回率(Precision-Recall,PR)曲线下的面积来衡量模型在不同阈值下的整体性能。理论上,平均精度的定义为:

$$\text{AP} = \int_0^1 P(r)dr$$

[0048] 其中$P(r)$是PR曲线中的精确率与召回率的函数关系。实际测试过程中,由于曲线是离散的,平均精度通常通过插值计算得到。平均精度可以综合反映模型对分类任务的整体能力,尤其适合用于模型间的性能比较。

[0049] 以上指标通过在真实数据集中对代理预测模型进行评估,可以得到代理预测模型的二分类能力,这个二分类能力也反映了将代理模型用于测试用例关键度预测时,预测偏差的大小。

[0050] 本申请通过明确从离线真实数据集中提取真实测试用例,实现了对实际被测系统历史或基准行为的高效、标准化利用。这为在无法进行大规模实时实测的场景下,仍能启动并实施智能弹性测试流程提供了关键的数据基础和技术路径。通过引入混淆矩阵这一经典机器学习评估工具,将代理模型在“关键测试用例”上的预测结果与实际系统的“真实标签”

进行系统化比对,实现了对二者差异性的结构化、精细化度量。这标志着对无人系统“测试偏差”的评估从定性或单一指标描述,转向了基于多维度二分类指标的定量、可解释分析。基于混淆矩阵计算出的二分类指标,能够从不同侧面(如模型的精确性、完备性、综合性能等)预测和刻画测试偏差。这种多维度的偏差描述,相比单一误差值,更能精准定位代理模型与实际系统在哪些方面存在差异。

[0051] 根据一些实施例,针对在线获取真实数据集的情况下,即在未获取到实际被测系统的离线真实数据集的情况下,预测测试偏差具体可以基于代理预测模型,利用关键测试用例对实际被测系统进行测试,以生成对应的真实测试用例;基于预设的基于神经网络的关键度预测函数、关键测试用例的关键度分数和真实测试用例的真实标签,确定交叉熵损失;基于交叉熵损失,预测测试偏差。

[0052] 本申请中,可以直接利用关键测试用例的测试参数来测试实际被测系统,在线采集到真实测试用例。

[0053] 关键度预测偏差(简称为预测偏差)估计函数可以用交叉熵损失来表示,即:

$$L_f = -\sum_{i=1}^{N}\left[y_i \log\left(f(s_i)\right) + (1-y_i)\log\left(1-f(s_i)\right)\right],$$

其中,s_i用于表征第i个关键测试用例,$f(s_i)$为预测的第i个测试用例的关键度分数,$f(\cdot)$用于表征基于神经网络的关键度预测函数,y_i是第i个真实测试用例的表征任务成功或者失败的标签,为0或1。L_f即为估计的测试偏差。

[0054] 本申请针对在线采集数据集的情况,提出了一种自举式解决方案。通过直接利用代理预测模型生成的关键测试用例对实际系统进行在线测试,能够“从零开始”动态生成所需的真实测试用例集。这突破了智能弹性测试对预先完备数据的依赖,极大地扩展了该框架的适用范围和部署灵活性。该方法将测试执行(在线获取真实标签)与偏差预测(计算交叉熵损失)无缝衔接,形成了一个完整的自动化数据采集与评估循环。无需人工介入数据分析,系统即可自主完成从触发测试到量化偏差的全过程,显著提升了测试流程的自主化与智能化水平。通过预设基于神经网络的关键度预测函数,并计算其关于关键测试用例的预测分数(关键度分数)与真实标签之间的交叉熵损失,为测试偏差提供了一个具有坚实信息论基础的、连续且可微的量化指标。该指标不仅能反映预测的正确与否,更能敏感地捕捉预测置信度与真实结果之间的概率分布差异,提供了比传统分类指标更精细的偏差洞察。

[0055] 根据一些实施例,在线收集真实测试用例的情况下,修正关键度预测模型的具体过程为利用预设训练策略,以交叉熵损失为损失函数,在训练数据集中进行预设的微调轮次的训练,以生成更新的代理预测模型,其中,预设训练策略为与修正前的代理预测模型使用的训练方式相同的类别均衡策略。

[0056] 本申请中,偏差估计函数L_f的形式是交叉熵损失的形式,因此,可以利用梯度下降方法对代理预测模型的多层感知机参数进行微调以提升预测模型的精准率和召回率。

[0057] 在一些实现方式中,在线测试过程中,每个关键测试用例都能获得被测对象实际的执行结果,得到一个“测试用例参数-标签”数据对并形成训练数据集。在资源充足条件下,智能弹性测试算法设定每一轮采集预设数量个数据后暂停测试,在训练数据集中对关键度预测模型进行微调后成为更新的代理预测模型,具体微调方法如下:

设定微调轮次,以交叉熵损失为损失函数,在训练数据集中训练微调轮次对应的

轮数。为防止在小数据集中微调过多导致神经网络遗忘在过去学习到的信息,微调轮次应该较小,如取2。同样地,为了防止类别不均衡导致训练效果不佳,在微调过程中也使用与训练代理预测模型时相同的类别均衡策略。

[0058] 本申请通过将预测得到的“交叉熵损失”直接作为损失函数,实现了“测试-评估-修正”的自动化闭环。这使得关键度预测模型的优化不再是离线、静态的任务,而是成为在线测试流程中一个可根据实时反馈(测试偏差)进行自我调整的动态环节,显著提升了测试系统的智能性与自主进化能力。通过限定使用微调轮次进行训练,避免了在在线修正过程中因过度拟合少量新数据而导致模型性能剧烈波动或灾难性遗忘的风险,确保了模型更新的平稳性和可控性。采用“与修正前的代理预测模型使用的训练方式相同的类别均衡策略”作为预设训练策略,确保了模型修正过程在方法论上与初始模型训练保持一致。这不仅能有效缓解在线数据流可能带来的类别不平衡问题,还可以保证了新旧模型在数据理解和决策逻辑上的内在一致性,避免了因修正策略不一致而引入新的、不可控的系统偏差。

[0059] 根据一些实施例,在关键度分数大于关键度阈值的情况下,确定关键测试用例的过程中,具体可以按照预设采样策略中采样概率,从自然分布中进行采样,以从初始测试用例中确定关键测试用例。

[0060] 本申请中,为了兼顾探索与开发,设定一个较小的随机的采样概率ϵ,该采样概率可以在0.05-0.2之间取值。对于每一测试,将有ϵ的概率直接从自然分布$\mathrm{p(s)}$中采样测试用例。自然分布即随机分布,例如定义为在参数最低和最高值之间的均匀分布。采样时就是随机数生成,不需要一个巨大的测试用例集合,然后让关键度预测模型判断是否为关键测试用例,筛选结束后用来测试,记关键度分数大于关键度阈值的测试用例的集合为:$S_c = \{s \in S, f_s > f_{th}\}$。

[0061] 其中,f_{th}为关键度阈值,S用于表征预先设置或者采集的测试用例集合,s用于表征某一测试用例,S_c用于表征关键度分数大于关键度阈值的测试用例集合。

[0062] 则本申请的测试用例的采样分布可以表示为:

$$\mathrm{q(s)} = \epsilon \mathrm{p(s)} + (1-\epsilon)\frac{II_{S_c}(S)\mathrm{p(s)}}{IE_P[II_{S_c}(S)]},$$

其中,II_{S_c}为示性函数,当$s \in S$时$II_{S_c} = 1$,否则$II_{S_c} = 0$。$IE_P[II_{S_c}(S)]$代表示性函数作用于S后的期望,可视为自然分布中选择到集合S_c的概率。

[0063] 本申请通过引入基于预设采样策略和采样概率的随机采样机制,将关键测试用例的确定过程从简单的阈值硬筛选,升级为一种结合确定性(关键度分数)与随机性(概率采样)的智能化决策,避免了仅依赖分数排序可能导致的测试场景选择僵化或过拟合,增强了测试探索的多样性与鲁棒性。本申请并非机械地选择分数最高的用例,而是依据概率从高分区域(关键度分数大于阈值)中进行采样。这实质上是将探索-利用权衡引入测试用例生成过程:既倾向于选择模型认为关键(高不确定性或高影响力)的用例(即“探索-利用权衡”中的利用权衡),又保留了对该区域内不同场景进行探索的可能性,有助于发现那些分数相近但特性迥异的关键缺陷,提升了测试对未知风险的挖掘能力。

[0064] 根据一些实施例,在关键度分数大于关键度阈值的情况下,从初始测试用例中确定关键测试用例具体可以按照采样概率,从自然分布中进行采样,以从初始测试用例中确

定预备测试用例;从均匀分布中抽取随机数,并在随机数小于采样概率的情况下,将预备测试用例确定为关键测试用例;在随机数大于或者等于采样概率的情况下,持续从自然分布中采样得到更新的预备测试用例,并将更新的预备测试用例的关键度分数与关键度阈值进行对比。

[0065] 本申请中,首先从自然分布$p(s)$中采样得到测试用例s,然后从均匀分布$U(0,1)$中抽取随机数(随机数生成函数实现,如python的random库。),若随机数小于ϵ,则直接使用该测试用例作为关键测试用例,否则持续地从中采样测试用例s',将s'的关键度分数与关键度阈值进行对比,直至测试用例$s' \in S_c$后使用s'来测试实际被测系统。

[0066] 关键度代表着十几倍测系统在当前测试用例下触发失败事件的概率,如果关键度预测偏差越小,则中测试用例越容易失败,使用采样分布得到的测试失败率越高。

[0067] 本申请的采样测试方法一方面结合关键度预测模型输出的测试用例的关键度分数对测试用例进行筛选,提高了触发失败事件的概率,另一方面引入了随机采样概率,防止只在自然分布中采样导致对测试空间探索不全面。通过引入随机采样概率,在测试次数足够的情况下,理论上可实现对全测试空间100%的覆盖率。

[0068] 根据一些实施例,在关键度分数小于或者等于关键度阈值的情况下,本申请还可以按照剩余采样概率,从自然分布中进行重采样得到新的初始测试用例,并比较新的初始测试用例的关键度分数和关键度阈值,其中,剩余采样概率为1与采样概率的差值。

[0069] 本申请中,如果关键度分数小于或者等于关键度阈值,则按照1-ϵ的采样概率重新采样得到新的初始测试用例,进而更新关键度分数进行重新的计算。

[0070] 本申请通过将总的采样概率明确划分为两部分(采样概率用于选择高关键度用例,剩余采样概率用于对低关键度区域进行重采样),建立了一套覆盖全部参数空间的、结构化的概率分配策略。这确保了测试资源不会完全局限于当前模型识别出的高关键度区域,而是有策略地保留了一部分资源用于持续探索整个参数空间,实现了“重点验证”与“广泛探索”之间的动态平衡与协同。

[0071] 本申请涉及3个典型场景:

1.协同搜索场景

环境中包含1个蓝方智能体、3个红方智能体组成的群系统、2个固定位置障碍物。

[0072] 在此环境下,智能群系统的任务通过智能体之间的协作,完成对蓝方智能体的搜索与打击。设红方群系统中任一智能体在T个时间步内实现对蓝方智能体的追击则表示群系统在当前回合完成任务。此外,蓝方智能体的躲避策略是预先设置的启发式规则,本场景被测无人系统为红方群系统。

[0073] 具体实施时,将坐标原点设在环境中心,横向和纵向分别为x轴和y轴,环境的x坐标和y坐标范围设置为$[-1,1]$。环境中可以配置的变量为3个红方智能体、1个蓝方智能体和2个障碍物的2维坐标位置,共计12个参数。

[0074] 2.协同导航场景

环境中包含3个红方智能体,以及3个需要到达的目标地点。本场景用于测试红方多智能体协同导航并占领目标地点的能力。在此场景中,本场景被测无人系统为3个智能体组成的群系统,它们需要无碰撞地在给定时间T内没有遗漏地达到3个目的地,如果中途发

生碰撞或者未在给定时间内到达3个目的地,则视为任务失败。

[0075] 场景2地图与场景1一致,将环境中心设置为坐标原点,并将横向和纵向范围设置为$[-1,1]$。环境中可以配置的变量为3个红方智能体以及3个目的地的2维坐标位置,共计12个参数。

[0076] 3.双足机器人场景

环境中存在人形双足机器人以及由各种不同地形组成的跑道,用于测试双足机器人的运动控制能力。

[0077] 在此场景中,双足机器人沿着直线向前行走,在行进过程中会遇到不同地形阻碍。地形类型包括树桩、陷阱、带有起伏的平地、上下坡等。被测无人系统为双足机器人,其任务是不断向前行走,并在给定时间步T内保持行走状态而不摔倒。本场景在当机器人摔倒时判定该系统任务失败。

[0078] 本阶段已采取的实验中,固定了双足机器人行走路线上遇到的地形种类,通过配置各地形的长度来改变地形,从而生成测试用例。

[0079] 在一些实现方式中,考虑到同类型无人系统在不同任务下的测试用例具有本质上的区别,因此,在本申请中,若无人系统类型相同,但搭载不同智能算法、执行不同类型任务,则将其视为不同的被测对象。例如,无人战车群体系统,在协同围捕场景和协同导航场景中执行的任务完全不一样,所采用的决策算法也有本质区别,导致测试用例完全不一样,因此,视为两种被测对象。

[0080] 本申请中,3个场景测试用例由一个高维向量来表示。具体地,在场景1和场景2中,为了描述环境运行过程中的动态信息,将前5个时间步中场景实体的位置拼接成一个高维向量来代表整个测试用例。由于场景1和2中可变参数是12个,因此每个测试用例由$12\times5=60$维高维向量表示。在场景3中,由于总时间步T的限制,过于靠后的地形在每次测试中无法被无人系统遇到,因此只考虑前10个地形的长度作为参数,于是场景3测试用例状态由10维高维向量表示。

[0081] 由于高维参数空间及其连续性,无法遍历测试空间且难以捕捉高维参数与测试结果之间的关系。为此,本申请通过训练4层$\mathrm{MLP}(d,256,256,256,1)$来获得代理预测模型,输入层的输入数据的特征维度为d,是一个变量,代表具体的数值,取决于具体的任务,例如,输入的事一个包含10个特征的向量,那么d为10。该代理模型有3个隐藏层,每个隐藏层有256个神经元,输出层的最终输出的维度为1。其中多层感知机的输入为表示每个测试用例的状态维度,在3个场景中分别为60、60、10。对于3个场景,实验分别收集了包含代理模型测试结果的较大规模的训练数据集,并按照3:1:1的比例划分为训练集、验证集、测试集。如前述描述,使用交叉熵损失和Adam优化器训练多层感知机。

[0082] 3个实验场景被测无人系统的底层实现算法均为深度学习算法。在试验中关于代理模型的选取方式解释如下:

对于场景1和场景2,使用相同底层算法、但不同版本模型参数的无人系统作为代理模型;对于场景3,使用不同底层算法的无人系统作为代理模型。通过这些代理模型收集足够的数据用于代理预测模型的训练。

[0083] 在具体的实现过程中,资源充足条件下,使用1个代理模型通过不断在线测试并自适应微调预测模型实现智能弹性测试。实验微调间隔为5000,且每次微调利用已有数据训

练2轮。智能弹性测试终止条件设定为达到一个足够大的测试次数,对于3个场景分别为200000、200000和100000。仅仅依靠测试次数以及被测对象的任务失败率难以全面体现算法性能,为了客观评价智能弹性测试过程中代理预测模型的表现,客观地描述测试用例关键度预测偏差随智能弹性测试过程的变化,在每个场景均提前收集了被测对象实际测试的200000组测试数据构建成“测试数据集”。根据关键度预测偏差的离线估计方法,将讨论代理预测模型在测试数据集中精确率、召回率、平均精度等指标。

[0084] 下面描述本申请的装置实施例,其可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节,可参照本申请方法实施例。

[0085] 图4为本申请实施例提供的面向无人系统差异性的弹性测试装置的框图。如图4所示,400包括分数确定模块401、测试偏差确定模块402、模型修正模块403以及循环测试修正模块404。

[0086] 分数确定模块401,用于将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中,以确定初始测试用例的关键度分数;

测试偏差确定模块402,用于基于关键度分数和预设的关键度阈值,从初始测试用例中确定关键测试用例,并根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差;

模型修正模块403,用于在在线收集真实测试用例的情况下,基于测试偏差对关键度预测模型进行修正,以生成更新的关键度预测模型;

循环测试修正模块404,用于按照预设的最大测试次数,利用更新的关键度预测模型重复调用分数确定模块401、测试偏差确定模块402和模型修正模块403,并确定最终关键测试用例。

[0087] 可选地,测试偏差确定模块402具体用于:

在关键度分数大于关键度阈值的情况下,将关键度分数对应的初始测试用例确定为关键测试用例;

根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差。

[0088] 可选地,测试偏差确定模块402在根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差的情况下,具体用于:

在获取到实际被测系统的离线真实数据集的情况下,从离线真实数据集中提取真实测试用例;

利用混淆矩阵,根据关键测试用例中的预测标签和真实测试用例中的真实标签,确定测试过程的二分类指标;

基于二分类指标,预测测试偏差。

[0089] 可选地,测试偏差确定模块402在根据实际被测系统的真实测试用例、关键测试用例,预测测试偏差的情况下,具体用于:

在未获取到实际被测系统的离线真实数据集的情况下,

基于代理预测模型,利用关键测试用例对实际被测系统进行测试,以生成对应的真实测试用例;

基于预设的基于神经网络的关键度预测函数、关键测试用例的关键度分数和真实测试用例的真实标签,确定交叉熵损失;

基于交叉熵损失,预测测试偏差。

[0090] 可选地,模型修正模块403具体用于:

在在线收集真实测试用例的情况下,利用预设训练策略,以交叉熵损失为损失函数,在训练数据集中进行预设的微调轮次的训练,以生成更新的代理预测模型,其中,预设训练策略为与修正前的代理预测模型使用的训练方式相同的类别均衡策略。

[0091] 可选地,测试偏差确定模块402在在关键度分数大于关键度阈值的情况下,将关键度分数对应的初始测试用例确定为关键测试用例,具体用于:

在关键度分数大于关键度阈值的情况下,按照预设采样策略中采样概率,从自然分布中进行采样,以从初始测试用例中确定关键测试用例。

[0092] 可选地,测试偏差确定模块402在关键度分数大于关键度阈值的情况下,按照预设采样策略中采样概率,从自然分布中进行采样,以从初始测试用例中确定关键测试用例,具体用于:

在关键度分数大于关键度阈值的情况下,按照采样概率,从自然分布中进行采样,以从初始测试用例中确定预备测试用例;

从均匀分布中抽取随机数,并在随机数小于采样概率的情况下,将预备测试用例确定为关键测试用例;

在随机数大于或者等于采样概率的情况下,持续从自然分布中采样得到更新的预备测试用例,并将更新的预备测试用例的关键度分数关键度阈值与进行对比。

[0093] 可选地,面向无人系统差异性的弹性测试装置400还包括重采样模块405,用于:

在关键度分数小于或者等于关键度阈值的情况下,按照剩余采样概率,从自然分布中进行重采样得到新的初始测试用例,并比较新的初始测试用例的关键度分数和关键度阈值,其中,剩余采样概率为1与采样概率的差值。

[0094] 装置执行与前面提供的方法类似的功能,其他功能可参见前面的描述,此处不再赘述。

[0095] 图5为本申请实施例提供的电子设备的结构示意图,如图5所示,本实施例的电子设备500可以包括:存储器501和处理器502。

[0096] 存储器501上存储有计算机程序,当计算机程序被处理器502执行时,使得前述处理器502执行上述实施例中的方法。

[0097] 其中,处理器502和存储器501相连,如通过总线相连。

[0098] 可选地,电子设备500还可以包括收发器。需要说明的是,实际应用中收发器不限于一个,该电子设备500的结构并不构成对本申请实施例的限定。

[0099] 处理器502可以是CPU(Central Processing Unit,中央处理器),通用处理器,DSP(Digital Signal Processor,数据信号处理器),ASIC(Application Specific Integrated Circuit,专用集成电路),FPGA(Field Programmable Gate Array,现场可编程门阵列)或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结合本申请公开内容所描述的各种示例性的逻辑方框,模块和电路。处理器502也可以是实现计算功能的组合,例如包含一个或多个微处理器组合,DSP和微处理器的组合等。

[0100] 总线可包括一通路,在上述组件之间传送信息。总线可以是PCI(Peripheral

Component Interconnect,外设部件互连标准)总线或EISA(Extended Industry Standard Architecture,扩展工业标准结构)总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示,图中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0101] 存储器501可以是ROM(Read Only Memory,只读存储器)或可存储静态信息和指令的其他类型的静态存储设备,RAM(Random Access Memory,随机存取存储器)或者可存储信息和指令的其他类型的动态存储设备,也可以是EEPROM(Electrically Erasable Programmable Read Only Memory,电可擦可编程只读存储器)、CD-ROM(Compact Disc Read Only Memory,只读光盘)或其他光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。

[0102] 存储器501用于存储执行本申请方案的应用程序代码,并由处理器502来控制执行。处理器502用于执行存储器501中存储的应用程序代码,以实现前述方法实施例所示的内容。

[0103] 其中,电子设备包括但不限于:移动电话、笔记本电脑、数字广播接收器、PDA(个人数字助理)、PAD(平板电脑)、PMP(便携式多媒体播放器)、车载终端(例如车载导航终端)等等的移动终端以及诸如数字TV、台式计算机等等的固定终端。还可以为服务器等。图5示出的电子设备仅仅是一个示例,不应对本申请实施例的功能和使用范围带来任何限制。

[0104] 本实施例的电子设备,可以用于执行上述任一实施例的方法,其实现原理和技术效果类似,此处不再赘述。

[0105] 本申请还提供一种非瞬时性计算机可读存储介质,其上存储有计算机可读指令,当前述指令被处理器执行时,使得处理器执行如上实施例中的方法。

[0106] 本领域普通技术人员可以理解:实现上述各方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成。前述的程序可以存储于一种非瞬时性计算机可读取存储介质中。该程序在执行时,执行包括上述各方法实施例的步骤;而前述的存储介质包括:ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0107] 以上对本申请实施例进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明仅用于帮助理解本申请的方法及其核心思想。同时,本领域技术人员依据本申请的思想,基于本申请的具体实施方式及应用范围上做出的改变或变形之处,都属于本申请保护的范围。综上所述,本说明书内容不应理解为对本申请的限制。

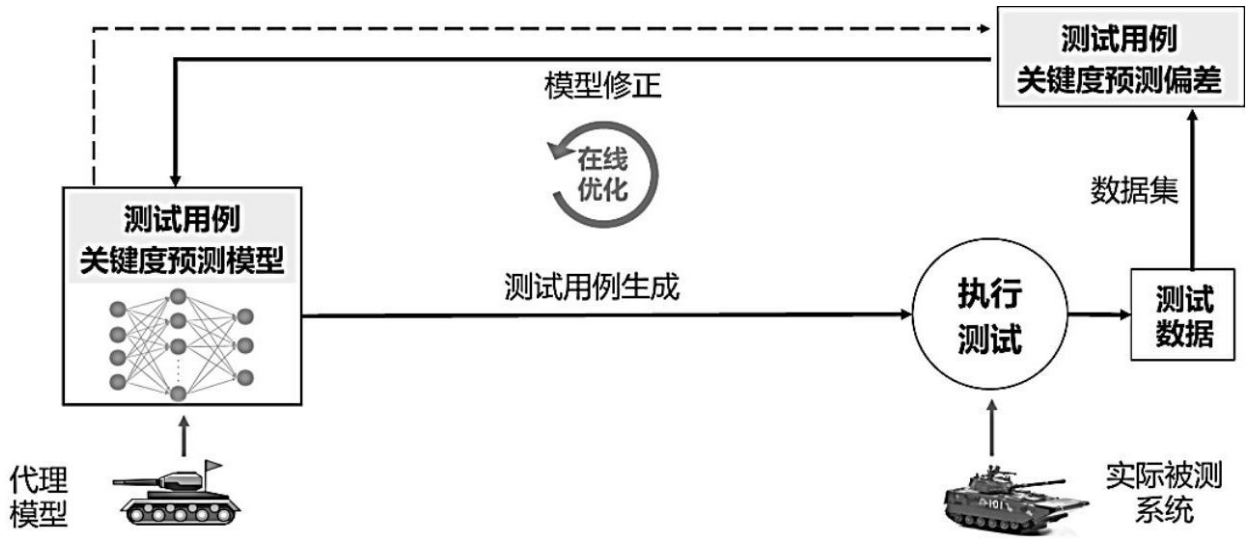

测试用例
关键度预测偏差
模型修正
在线
优化
数据集
测试用例
关键度预测模型
测试用例生成
执行
测试
测试
数据
代理
模型
实际被测
系统

图1

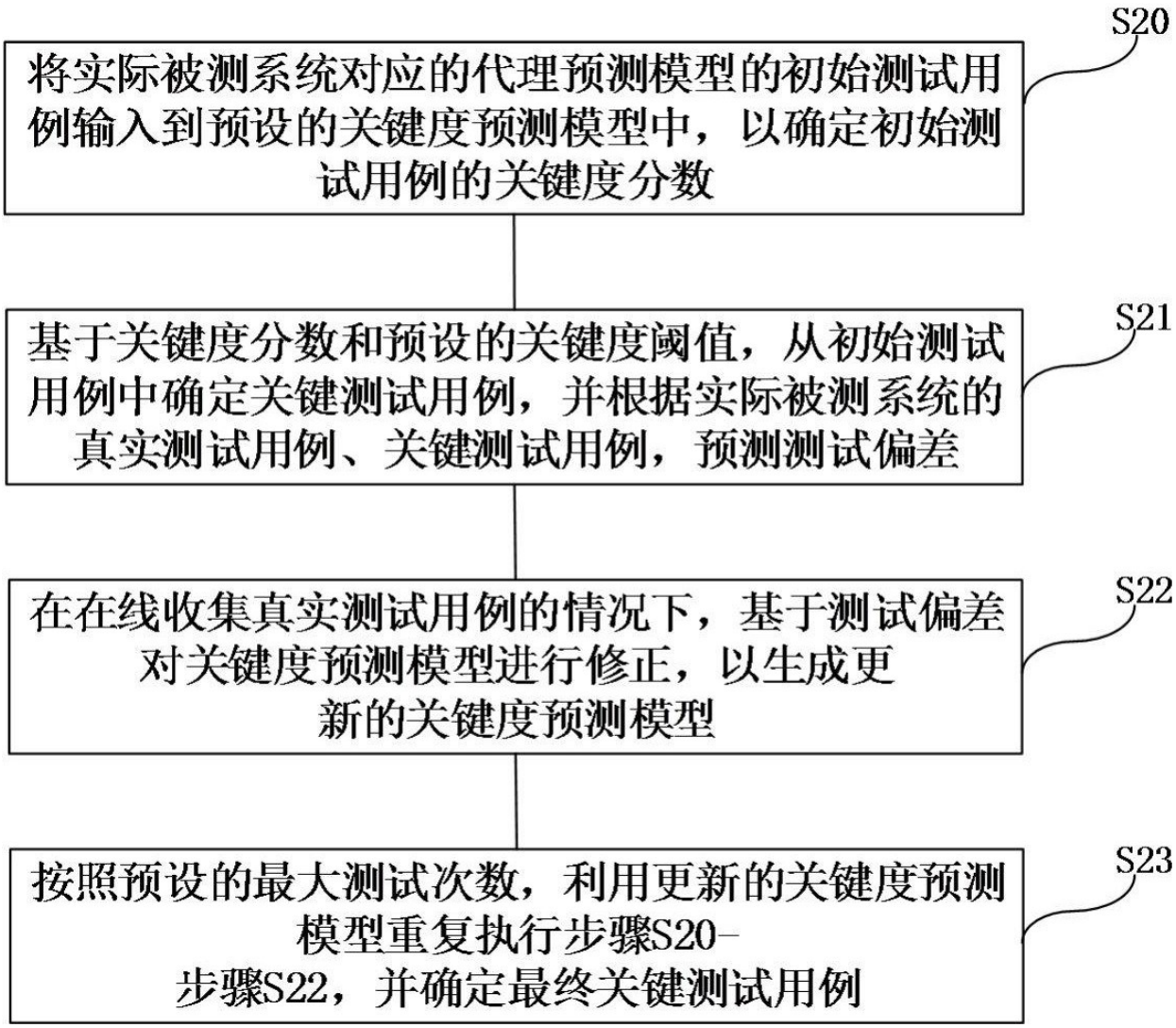

S20
将实际被测系统对应的代理预测模型的初始测试用例输入到预设的关键度预测模型中，以确定初始测试用例的关键度分数
S21
基于关键度分数和预设的关键度阈值，从初始测试用例中确定关键测试用例，并根据实际被测系统的真实测试用例、关键测试用例，预测测试偏差
S22
在在线收集真实测试用例的情况下，基于测试偏差对关键度预测模型进行修正，以生成更新的关键度预测模型
S23
按照预设的最大测试次数，利用更新的关键度预测模型重复执行步骤S20-步骤S22，并确定最终关键测试用例

图2

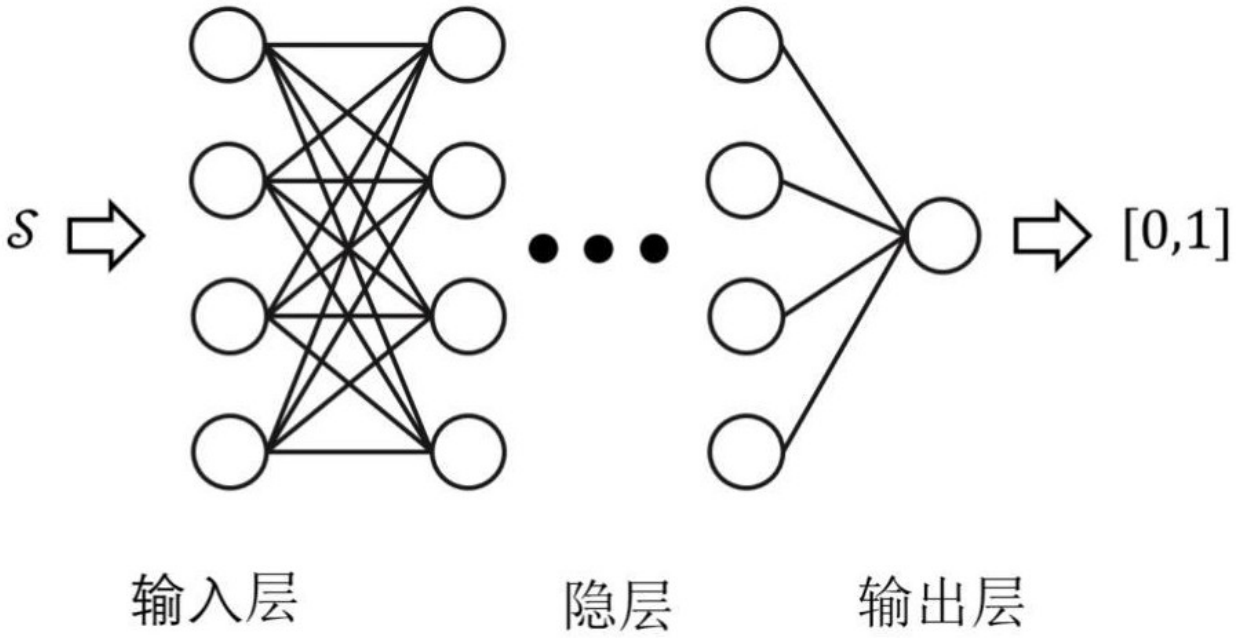


图3

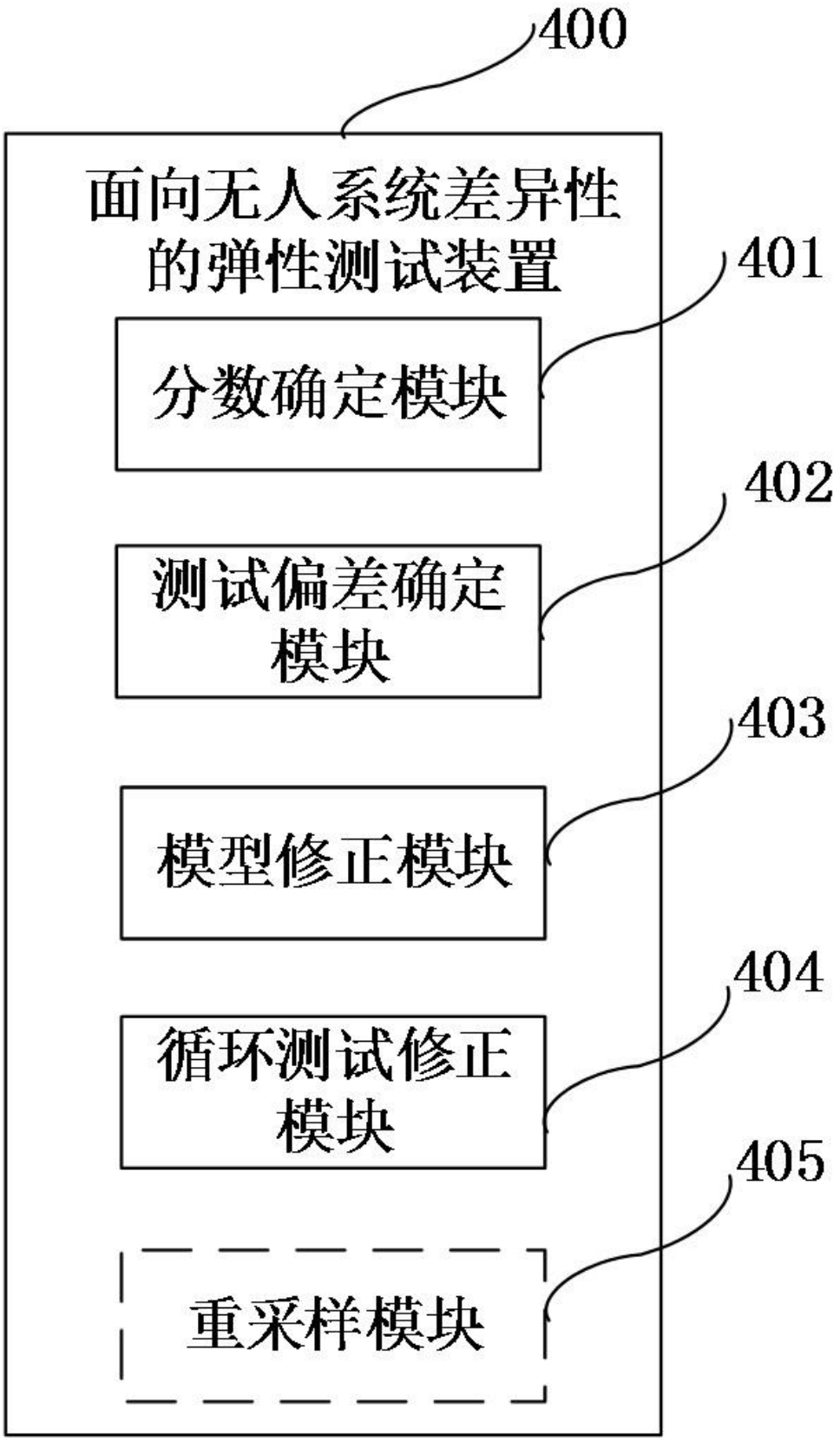

400
面向无人系统差异性的弹性测试装置
401
分数确定模块
402
测试偏差确定模块
403
模型修正模块
404
循环测试修正模块
405
重采样模块

图4

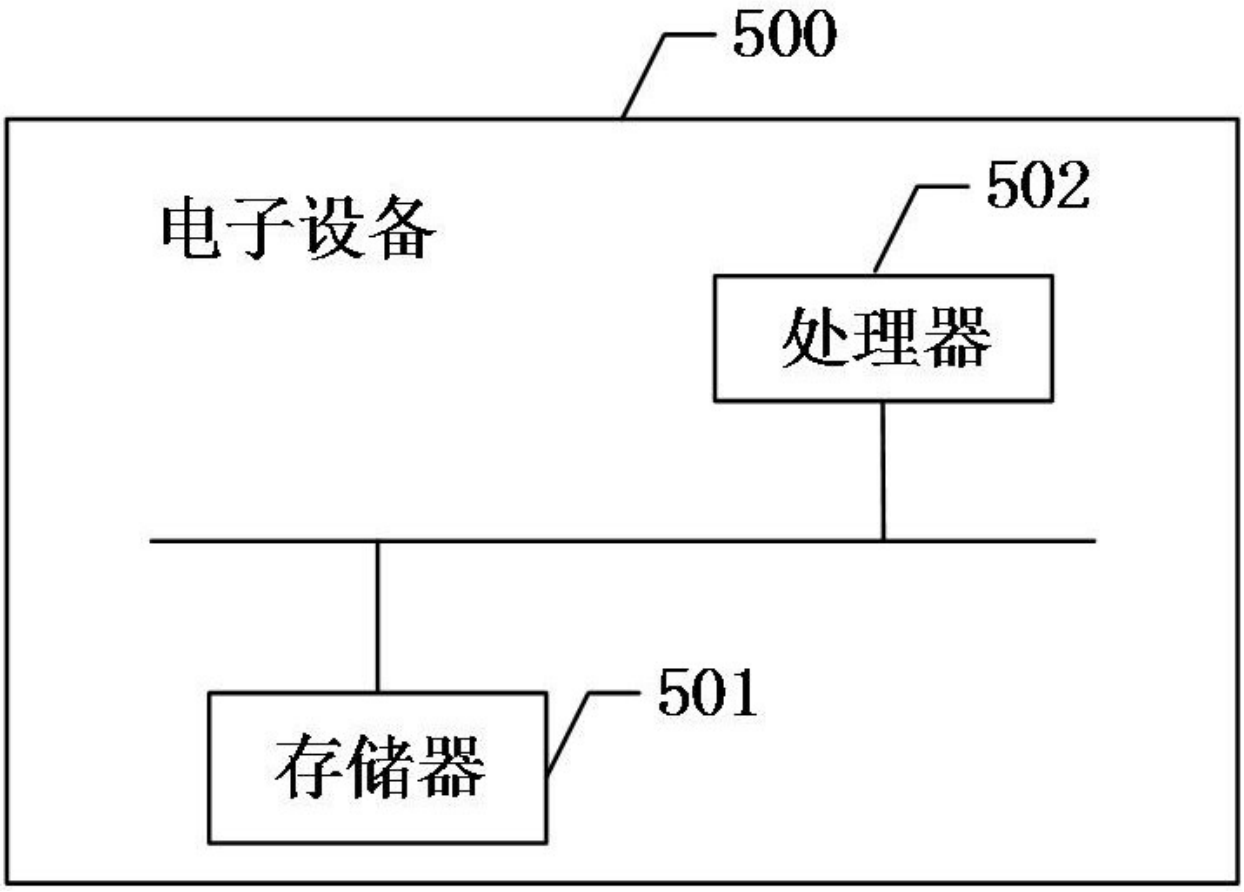


图5