(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112101570 B
(45)授权公告日 2024.09.17

(21)申请号 202011003210.8

(22)申请日 2020.09.22

(65)同一申请的已公布的文献号
申请公布号 CN 112101570 A

(43)申请公布日 2020.12.18

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 付琰 陈亮辉 周洋杰 方军

(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 孟金喆

(51)Int.Cl.
G06N 20/00(2019.01)

(56)对比文件
CN 106055607 A,2016.10.26
CN 111400174 A,2020.07.10

审查员 张博

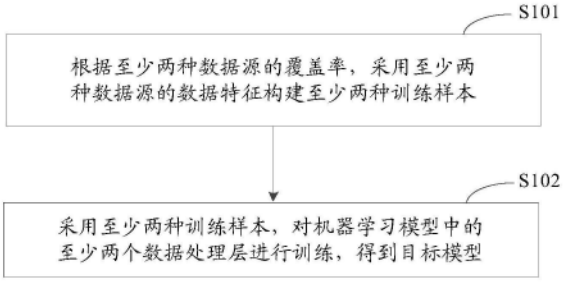
权利要求书3页 说明书15页 附图6页

(54)发明名称

一种模型训练和使用方法、装置、设备及存储介质

(57)摘要

本申请公开了一种模型训练和使用方法、装置、设备及存储介质,涉及人工智能、机器学习和大数据技术领域。其中,模型训练方法的具体实现方案为:根据至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本;其中,所述训练样本关联有至少一种数据源;采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型;其中,不同训练样本训练的数据处理层不同。以提高模型训练效果以及任务预测的准确性。



CN 112101570 B

1.一种模型训练方法,包括:

获取包含至少两种数据源的数据特征集合;其中,所述至少两种数据源是指贴吧数据源、搜索引擎数据源和微博数据源中的至少两种;

根据所述数据特征集合,确定所述至少两种数据源的特征数量和目标特征数量;

根据所述至少两种数据源的特征数量和目标特征数量,确定所述至少两种数据源的覆盖率;

根据所述至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本;其中,所述训练样本关联有至少一种数据源;

采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型;其中,不同训练样本训练的数据处理层不同;

其中,所述根据所述至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本,包括:

根据所述至少两种数据源的覆盖率,将所述至少两种数据源分为至少两组;

从所述至少两组数据源中选择当前组数据源,且将所述当前组数据源的数据特征以及上一训练样本中数据源的数据特征,作为当前训练样本;

其中,所述当前组数据源的覆盖率低于所述上一训练样本中数据源的覆盖率。

2.根据权利要求1所述的方法,其中,采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,包括:

根据训练样本和待训练层数的关联关系,从机器学习模型中的至少两个数据处理层中,确定当前训练样本待训练的数据处理层;

基于所述机器学习模型中已训练的数据处理层,采用当前训练样本对所述当前训练样本待训练的数据处理层进行训练;

其中,所述已训练的数据处理层通过采用在所述当前训练样本之前采用的训练样本训练得到。

3.根据权利要求1所述的方法,其中,采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,包括:

将机器学习模型中位于已训练的数据处理层之后的至少一个数据处理层作为候选训练层,并采用当前训练样本,对所述候选训练层进行训练;

若所述候选训练层的训练结果满足收敛条件,则将所述候选训练层作为当前训练样本训练的数据处理层,且当前训练样本对所述机器学习模型训练结束。

4.根据权利要求1-3中任一项所述的方法,其中,所述机器学习模型为树模型;所述树模型中的至少两个数据处理层为至少两棵决策树。

5.一种模型使用方法,使用权利要求1-4中任一项所述的方法训练的目标模型实现,包括:

根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的至少两种训练样本中,确定所述待预测样本关联的目标训练样本;其中,所述待预测样本关联的数据源包括贴吧数据源、搜索引擎数据源和微博数据源中至少两种;

根据所述目标训练样本,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层;

根据所述待调用数据处理层,对所述待预测样本进行任务预测。

6.根据权利要求5所述的方法,其中,根据所述目标训练样本,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层,包括:

确定所述目标训练样本在模型训练阶段训练的数据处理层;

根据所述目标训练样本训练的数据处理层,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层。

7.根据权利要求5或6所述的方法,其中,所述目标模型为树模型;所述树模型中的至少两个数据处理层为至少两棵决策树。

8.一种模型训练装置,包括:

覆盖率确定模块,具体包括:

特征集合获取单元,用于获取包含至少两种数据源的数据特征集合;其中,所述至少两种数据源是指贴吧数据源、搜索引擎数据源和微博数据源中的至少两种;

特征数量确定单元,用于根据所述数据特征集合,确定所述至少两种数据源的特征数量和目标特征数量;

覆盖率确定单元,用于根据所述至少两种数据源的特征数量和目标特征数量,确定所述至少两种数据源的覆盖率;

训练样本构建模块,用于根据所述至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本;其中,所述训练样本关联有至少一种数据源;

模型训练模块,用于采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型;其中,不同训练样本训练的数据处理层不同;

其中,所述训练样本构建模块包括:

数据源分组单元,用于根据所述至少两种数据源的覆盖率,将所述至少两种数据源分为至少两组;

训练样本确定单元,用于从所述至少两组数据源中选择当前组数据源,且将所述当前组数据源的数据特征以及上一训练样本中数据源的数据特征,作为当前训练样本;

其中,所述当前组数据源的覆盖率低于所述上一训练样本中数据源的覆盖率。

9.根据权利要求8所述的装置,其中,所述模型训练模块具体用于:

根据训练样本和待训练层数的关联关系,从机器学习模型中的至少两个数据处理层中,确定当前训练样本待训练的数据处理层;

基于所述机器学习模型中已训练的数据处理层,采用当前训练样本对所述当前训练样本待训练的数据处理层进行训练;

其中,所述已训练的数据处理层通过采用在所述当前训练样本之前采用的训练样本训练得到。

10.根据权利要求8所述的装置,其中,所述模型训练模块还具体用于:

将机器学习模型中位于已训练的数据处理层之后的至少一个数据处理层作为候选训练层,并采用当前训练样本,对所述候选训练层进行训练;

若所述候选训练层的训练结果满足收敛条件,则将所述候选训练层作为当前训练样本训练的数据处理层,且当前训练样本对所述机器学习模型训练结束。

11.根据权利要求8-10中任一项所述的装置,其中,所述机器学习模型为树模型;所述

树模型中的至少两个数据处理层为至少两棵决策树。

12.一种模型使用装置,使用权利要求1-4中任一项所述的方法训练的目标模型实现,包括:

样本分析模块,用于根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的至少两种训练样本中,确定所述待预测样本关联的目标训练样本;其中,所述待预测样本关联的数据源包括贴吧数据源、搜索引擎数据源和微博数据源中至少两种;

调用数据层确定模块,用于根据所述目标训练样本,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层;

任务预测模块,用于根据所述待调用数据处理层,对所述待预测样本进行任务预测。

13.根据权利要求12所述的装置,其中,所述调用数据层确定模块具体用于:

确定所述目标训练样本在模型训练阶段训练的数据处理层;

根据所述目标训练样本训练的数据处理层,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层。

14.根据权利要求12或13所述的装置,其中,所述目标模型为树模型;所述树模型中的至少两个数据处理层为至少两棵决策树。

15.一种电子设备,其中,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-4中任一项所述的模型训练方法,或执行权利要求5-7中任一项所述的模型使用方法。

16.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行权利要求1-4中任一项所述的模型训练方法,或执行权利要求5-7中任一项所述的模型使用方法。

一种模型训练和使用方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及计算机技术领域,尤其涉及人工智能、机器学习和大数据技术,具体涉及一种模型训练和调用方法。

背景技术

[0002] 目前,机器学习模型在很多领域都发挥着至关重要的作用。其中,模型输入特征的全面性和准确性决定了模型的预测效果。所以为了提升模型的预测效果,研发人员通常会引入不同数据源的数据特征进行模型训练和任务预测。但是不同数据源的覆盖率不一致,对于覆盖率较低的数据源,其包含的数据特征相对薄弱,可能会存在数据特征缺失的情况。所以当使用覆盖率不同的数据源的数据特征进行模型训练或任务预测时,存在训练效果较差,或任务预测准确性低等问题,亟需改进。

发明内容

[0003] 本公开提供了一种模型训练和使用方法、装置、设备及存储介质。

[0004] 根据本公开的一方面,提供了一种模型训练方法,该方法包括:

[0005] 根据至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本;其中,所述训练样本关联有至少一种数据源;

[0006] 采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型;其中,不同训练样本训练的数据处理层不同。

[0007] 根据本公开的第二方面,提供了一种模型使用方法,该方法使用本申请任一实施例的方法训练的目标模型实现,该方法包括:

[0008] 根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的至少两种训练样本中,确定所述待预测样本关联的目标训练样本;

[0009] 根据所述目标训练样本,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层;

[0010] 根据所述待调用数据处理层,对所述待预测样本进行任务预测。

[0011] 根据本公开的第三方面,提供了一种模型训练装置,该装置包括:

[0012] 训练样本构建模块,用于根据至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本;其中,所述训练样本关联有至少一种数据源;

[0013] 模型训练模块,用于采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型;其中,不同训练样本训练的数据处理层不同。

[0014] 根据本公开的第四方面,提供了一种模型使用装置,该装置使用本申请任一实施例的方法训练的目标模型实现,该装置包括:

[0015] 样本分析模块,用于根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的至少两种训练样本中,确定所述待预测样本关联的目标训练样本;

[0016] 调用数据层确定模块,用于根据所述目标训练样本,从所述目标模型的至少两个

数据处理层中,确定待调用数据处理层;

[0017] 任务预测模块,用于根据所述待调用数据处理层,对所述待预测样本进行任务预测。

[0018] 根据本公开的第五方面,提供了一种电子设备,该电子设备包括:

[0019] 至少一个处理器;以及

[0020] 与至少一个处理器通信连接的存储器;其中,

[0021] 存储器存储有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行本申请任一实施例的模型训练方法或模型使用方法。

[0022] 根据本公开的第六方面,提供了一种存储有计算机指令的非瞬时计算机可读存储介质。计算机指令用于使计算机执行本申请任一实施例的模型训练方法或模型使用方法。

[0023] 根据本申请的技术解决了因数据来源的覆盖率不同,导致样本特征缺失,从而影响模型训练效果和模型预测准确性的问题,为模型训练和模型预测提供了一种新思路。

[0024] 应当理解,本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征,也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0025] 附图用于更好地理解本方案,不构成对本申请的限定。其中:

[0026] 图1是根据本申请实施例提供的一种模型训练方法的流程图;

[0027] 图2是根据本申请实施例提供的一种模型训练方法的流程图;

[0028] 图3A是根据本申请实施例提供的一种模型训练方法的流程图;

[0029] 图3B是根据本申请实施例提供的机器学习模型的数据处理层的结构示意图;

[0030] 图4是根据本申请实施例提供的一种模型训练方法的流程图;

[0031] 图5是根据本申请实施例提供的一种模型训练方法的流程图;

[0032] 图6是根据本申请实施例提供的一种模型使用方法的流程图;

[0033] 图7是根据本申请实施例提供的一种模型训练装置的结构示意图;

[0034] 图8是根据本申请实施例提供的一种模型使用装置的结构示意图;

[0035] 图9是用来实现本申请实施例的模型训练或模型使用方法的电子设备的框图。

具体实施方式

[0036] 以下结合附图对本申请的示范性实施例做出说明,其中包括本申请实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本申请的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0037] 图1是根据本申请实施例提供的一种模型训练方法的流程图。本实施例适用于采用不同覆盖率的数据源的数据特征,对机器学习模型进行训练的情况。该实施例可以由电子设备中配置的模型训练装置来执行,该装置可以采用软件和/或硬件来实现。如图1所示,该方法包括:

[0038] S101,根据至少两种数据源的覆盖率,采用至少两种数据源的数据特征构建至少两种训练样本。

[0039] 其中,数据源是指数据特征的获取来源,例如,若需要使用的数据特征为用户的搜索浏览数据的特征,则获取该类特征的数据源可以包括但不限于:贴吧数据源、搜索引擎数据源和微博数据源等。不同数据源因注册用户数量的不同,对应的覆盖率也不同,注册用户数量越多的数据源对应的覆盖率越高。例如,使用贴吧的用户少于使用搜索引擎的用户,所以贴吧数据源的覆盖率就低于搜索引擎数据源的覆盖率。需要说明的是,在本申请实施例中,各种数据源的覆盖率可以是预先计算好的,也可以是训练时根据所需要使用到的数据特征所属的数据源,实时进行计算。具体的计算方法将在后续实施例进行详细介绍。

[0040] 训练样本可以是指对机器学习模型进行模型训练时使用的样本数据,本申请实施例对机器学习模型的训练分为至少两个阶段,相应的,所需构建的训练样本的数量为至少两种。需要说明的是,本申请实施例中,每种训练样本关联有至少一种数据源,即每种训练样本中包含从至少一种数据源中获取的数据特征。另外,不同训练样本关联的数据源相互交叠,例如,第一训练样本关联的数据源为搜索引擎数据源,第二训练样本关联的数据源为搜索引擎数据源和微博数据源;第三训练样本关联的数据源为搜索引擎数据源、微博数据源和贴吧数据源,这三种训练样本中,任意两种训练样本所关联的数据源相互之间都存在交叠。

[0041] 可选的,本申请实施例对于不同覆盖率的数据源对应的数据特征,并不是直接将其作为一个训练样本,而是将这些数据特征进行重新组合,得到至少两种训练样本。具体的,由于本申请实施例是根据数据特征所属数据源的覆盖率的不同,对机器学习模型进行多阶段的训练,所以可以是根据模型的工作原理和/或数据特征所属数据源的种类等,先确定需要对机器学习模型分几个阶段进行训练,也就是说确定本步骤需要构建几种训练样本(即每个训练阶段对应一种训练样本),进而再确定每种训练样本关联哪些数据源的数据特征。可选的,在确定每种训练样本所关联的数据源时,可以随着训练样本关联训练次数的递增,训练样本关联的数据源的种类也随之增加,如可以是当前训练样本在上一训练样本关联的数据源的基础上,增加至少一种新的数据源,且新增加的数据源的覆盖率要低于上一训练样本关联的数据源的覆盖率。

[0042] S102,采用至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型。

[0043] 其中,本申请实施例中的机器学习模型的种类有很多,该类机器学习模型需要具备的特征为:包含有至少两个数据处理层,且每一数据处理层在其前一数据处理层的处理结果上进行迭代计算。例如,满足该特征的机器学习模型可以包括但不限于:基于梯度提升决策树(Gradient Boosting Decision Tree,GBDT)算法构建的树模型、逻辑回归(Logistic Regression,LR)模型、深度学习模型等。可选的,当所述机器学习模型为树模型时,所述树模型中的至少两个数据处理层为至少两棵决策树。

[0044] 可选的,由于S101已经构建了至少两种训练样本,且每种训练样本对应机器学习模型训练的一个阶段,所以本步骤可以是为不同训练阶段确定需要训练的数据处理层,进而将该训练样本输入到机器学习模型中,对其需要训练的数据处理层进行参数训练。需要说明的是,本申请实施例中,不同训练样本训练的是同一个机器学习模型,只是每种训练样本训练该机器学习模型的数据处理层不同。也就是说,在本申请实施例中,采用一种训练样本训练后的数据处理层,不在基于其他种类的训练样本进行重复训练,以保证后续任务预

测时预测结果的准确性。

[0045] 可选的,在本申请实施例中,每种训练样本对应机器学习模型中需要训练的数据处理层可以是预先设置好的,例如,假设机器学习模型有10层,训练样本有3种,可以预先设置好第一训练样本训练机器学习模型的第1-5层,第二训练样本训练机器学习模型的第6-8层,第三训练样本训练机器学习模型的第9-10层。还可以是对于每种训练样本预先不知道其需要训练的数据处理层,而是在上一训练样本训练的数据处理层之后,对后续未训练过的数据处理层接着进行训练,具体的训练数量可以在本阶段训练的过程中确定,例如,假设上一训练阶段训练的是机器学习模型的第1-5层,则当前阶段可以是采用当前阶段对应的训练样本从机器学习模型的第6层开始训练,若该训练样本训练到模型收敛时训练到了第8层,则可以确定当前训练样本训练的数据处理层为机器学习模型的第6-8层。

[0046] 本申请实施例的技术方案,对于数据特征所属数据源的覆盖率不一致,存在缺失数据的情况下,为了保证训练的机器学习模型的准确性,不是简单的进行缺失特征补充(其中,简单的进行缺失特征补充,无法保证补充特征的准确),也不是针对不同覆盖率的数据源训练多个机器学习模型来进行任务预测(其中,训练多个机器学习模型成本高,其占用资源多)。而是将数据特征按照所属数据源的覆盖率,分为多种训练样本,每种训练样本对应一个训练阶段,分阶段对一个机器学习模型的不同数据处理层进行训练。从而实现无需补充缺失特征,只训练一个机器学习模型,后续根据待预测数据特征所属数据源的覆盖率,选择调用不同的数据处理层,即可准确进行任务预测,提高模型训练效果,以及后续任务预测准确性的同时,节约了资源,降低了模型训练的功耗,为采用不同覆盖率的数据源的数据特征进行模型训练提供了一种新思路。

[0047] 图2是根据本申请实施例提供的一种模型训练方法的流程图,本实施例在上述实施例的基础上,给出了根据至少两种数据源的覆盖率,采用至少两种数据源的数据特征构建至少两种训练样本的具体情况介绍,如图2所示,该方法包括:

[0048] S201,根据至少两种数据源的覆盖率,将至少两种数据源分为至少两组。

[0049] 可选的,在本申请实施例中,可以是先根据模型的工作原理和/或数据特征所属数据源的种类等,确定需要分几个阶段对机器学习模型进行训练,本步骤具体将至少两种数据源划分为几种,取决于需要对机器学习模型进行几个阶段的训练。例如,若已经确定需要对机器学习模型分两个阶段进行训练,则此时可以将至少两种数据源分为两组。

[0050] 可选的,在确定了至少两种数据待划分的组数之后,具体如何对至少两种数据源进行划分的方法有很多,对此本实施例不进行限定。方式一、可以是根据需要划分的组数,确定需要使用的覆盖率阈值的数量和数值。进而按照确定的覆盖率阈值的数量(其中,该数量可以比需要划分的组数少1)和数值,对至少两种数据源划分为至少两组。例如,假设本步骤需要将至少两种数据源划分的组数为2组,则此时需要使用的覆盖率阈值的数量为1个,若覆盖率范围为1-100,则此时可以设置覆盖率阈值的数值为50。进而将至少两种数据源中,覆盖率大于或等于50的划分为一组,作为高覆盖率组,将覆盖率小于50的划分为一组,作为低覆盖率组。方式二、可以将至少两种不同覆盖率的数据源按照覆盖率高低顺序进行排序,然后根据待划分的组数,将至少两种不同覆盖率的数据源进行划分,例如,假设本步骤需要将4种数据源划分为两组,此时可以是将覆盖率较高的两种数据源划分为一组,即高覆盖率组,将另外两种划分为一组,即低覆盖率组。

[0051] S202,从至少两组数据源中选择当前组数据源,且将当前组数据源的数据特征以及上一训练样本中数据源的数据特征,作为当前训练样本。

[0052] 其中,当前组数据源的覆盖率低于上一训练样本中数据源的覆盖率。

[0053] 可选的,本申请实施例可以是将S201划分的各组数据源,按照数据源覆盖率从高到底的顺序,依次将每一组数据源作为当前组数据源。若当前组数据源是第一组数据源时,可以是直接将该第一组数据源的数据特征作为当前训练样本,即当前训练样本是第一训练样本,其之前没有上一训练样本;若当前组数据源不是第一组数据源,则可以是将当前组数据源的数据特征和上一训练样本中所包含的所有数据源的数据特征一并作为当前训练样本的数据特征。例如,S201将数据源划分为两组,即高覆盖率组和低覆盖率组,则本步骤可以是将高覆盖率组的数据源的数据特征作为第一训练样本,将低覆盖率组的数据源的数据特征,和第一训练样本包含的数据特征(即高覆盖率组的数据源的数据特征)一并作为第二训练样本。

[0054] S203,采用至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型。

[0055] 其中,不同训练样本训练的数据处理层不同。

[0056] 本申请实施例的技术方案,根据至少两种数据源的覆盖率先将至少两组数据源分为至少两组,然后针对每一组,将该组数据源和上一训练样本中的数据源的数据特征一并作为当前训练样本,来分阶段为机器学习模型中的不同数据处理层进行训练。本实施例的方案确定每阶段的训练样本时,当前阶段的训练样本的覆盖率低于上一阶段,且当前阶段的训练样本中包含上一阶段的训练样本中的数据特征,保证了每阶段训练样本的全面性和准确性,为训练样本的构建提供了一种优选方案,进而保证了后续分阶段训练机器学习模型的训练效果。

[0057] 图3A是根据本申请实施例提供的一种模型训练方法的流程图;图3B是根据本申请实施例提供的机器学习模型的数据处理层的结构示意图。本实施例在上述实施例的基础上,给出了一种采用至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练的具体情况介绍,如图3A-3B所示,该方法包括:

[0058] S301,根据至少两种数据源的覆盖率,采用至少两种数据源的数据特征构建至少两种训练样本。

[0059] 其中,本申请实施例中的训练样本关联有至少一种数据源。

[0060] S302,根据训练样本和待训练层数的关联关系,从机器学习模型中的至少两个数据处理层中,确定当前训练样本待训练的数据处理层。

[0061] 其中,所谓训练样本和待训练层数的关联关系可以是预先设置好的每种训练样本与需要训练的数据处理层的层数之间的关系。

[0062] 可选的,在本申请实施例中,该关联关系可以是预先设置好每一种训练样本需要训练机器学习模型的哪几层,此时可以依次将至少两种训练样本中的每一训练样本作为当前训练样本,然后根据该当前训练样本在关联关系中记录的需要训练机器学习模型的哪几层,从机器学习模型中选出这几层作为当前训练样本待训练的数据处理层。例如,待训练的机器学习模型为图3B所示的机器学习模型,假设训练样本和待训练层数的关联关系为第一训练样本训练该机器学习模型的第1-5层,第二训练样本训练该机器学习模型的第6-8层,

第三训练样本训练该机器学习模型的第9-10层,则可以基于该关联关系,直接将该机器学习模型的第1-5层作为第一训练样本待训练的数据处理层;将该机器学习模型的第6-8层作为第二训练样本待训练的数据处理层;将该机器学习模型的第9-10层作为第三训练样本待训练的数据处理层。可选的,该关联关系还可以是预先设置好每种训练样本需要训练的机器学习模型的数据处理层的总数量,此时可以依次将至少两种训练样本中的每一训练样本作为当前训练样本,然后根据该当前训练样本在关联关系中记录的需要训练机器学习模型的数据处理层的总数量,结合上一训练样本待训练的数据处理层,从机器学习模型中选出当前训练样本待训练的数据处理层。例如,待训练的机器学习模型为图3B所示的机器学习模型,假设训练样本和待训练层数的关联关系为第一训练样本训练机器学习模型的5层,第二训练样本训练机器学习模型的3层,第三训练样本训练机器学习模型的2层,则可以将机器学习模型的第1-5层作为第一训练样本待训练的数据处理层;将第一训练样本训练的数据处理层之后的第6-8层作为第二训练样本待训练的数据处理层;将第二训练样本训练的数据处理层之后的第9-10层作为第三训练样本待训练的数据处理层。

[0063] S303,基于机器学习模型中已训练的数据处理层,采用当前训练样本对当前训练样本待训练的数据处理层进行训练。

[0064] 其中,所述已训练的数据处理层通过采用在当前训练样本之前采用的训练样本训练得到。例如,待训练的机器学习模型为图3B所示的机器学习模型,假设S301构建了3种训练样本(即第一训练样本、第二训练样本和第三训练样本),且当前训练阶段为第三训练阶段,即采用的当前训练样本为第三训练样本,则本申请实施例的已训练的数据处理层为采用第一训练样本和第二训练样本在第一训练阶段和第二训练阶段已经训练好的机器学习模型的第1-8层。

[0065] 可选的,在本申请实施例中,依次将至少两种训练样本中的每一训练样本作为当前训练样本,基于当前训练样本之前的所有训练样本已经训练的数据处理层,采用当前训练样本对其待训练的数据处理层进行训练,具体的,可以是将当前训练样本输入到机器学习模型中,先通过已训的数据处理层对当前训练样本进行处理,然后根据已训练的数据处理层的处理结果,进一步训练当前训练样本待训练的数据处理层的参数。例如,如图3B所示,假设当前训练样本待训练的是机器学习模型的第9-10层;当前训练样本之前的训练样本已经训练好了机器学习模型的第1-8层,此时可以是基于已经训练好的机器学习模型的第1-8层,采用当前训练样本,对机器学习模型的第9-10层进行训练,更新该机器学习模型的第9-10层的参数值。需要说明的是,在本申请实施例中,针对每一个训练样本都执行完上述S303的操作之后,该机器学习模型即训练完成,得到后续进行任务预测的目标模型。

[0066] 本申请实施例的技术方案,对于至少两种数据源的数据特征,根据数据源的覆盖率的不同,构建至少两种训练样本,然后依次将每种训练样本作为当前训练样本,根据训练样本和待训练层数之间的关联关系,确定当前训练样本待训练的数据处理层,进而基于已经训练的数据处理层,采用当前训练样本对该当前训练样本的待训练的数据处理层进行训练,所有训练样本都执行完训练操作后,得到目标模型。本申请实施例的方案,研发人员可以结合机器学习模型的各数据处理层的特性,以及训练效果等,预先灵活设置各种训练样本与其待训练数据处理层之间的关联关系。实现快速且灵活的确定各训练样本待训练的数据处理层,在提高模型训练效果的同时,保证了模型训练的灵活性和高效性。

[0067] 图4是根据本申请实施例提供的一种模型训练方法的流程图。本实施例在上述实施例的基础上,给出了另一种采用至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练的具体情况介绍。需要说明的是,本实施例的方案所适用的机器学习模型为树模型,例如,可以是基于GBDT算法构建的树模型,相应的,该树模型中的至少两个数据处理层为至少两棵决策树。

[0068] 接下来先对GBDT算法进行简单介绍:GBDT算法可以是由k棵基础决策树组成的一个加法运算式:$\hat{y}=\sum_{k=1}^{K} f_k(x)$,其中$f_k$表示第k棵决策树的输出值,每一棵决策树的训练目标都是拟合真实的y值和目前的$\hat{y}$之间的残差。在训练GBDT模型时,首先训练得到第一棵决策树,把这棵决策树的输出值记为$f_1(x)$。那么第二决策棵树的训练目标就是$y-f_1(x)$,第三棵树的训练目标是$y-f_1(x)-f_2(x)$,以此类推。目前很多流行的机器学习库,例如优化的分布式梯度增强库(xgboost)、梯度提升算法库(catboost)等都是GBDT算法的一种实现。

[0069] 如图4所示,本实施例训练树模型的方法包括:

[0070] S401,根据至少两种数据源的覆盖率,采用至少两种数据源的数据特征构建至少两种训练样本。

[0071] 其中,本申请实施例的训练样本关联有至少一种数据源。

[0072] S402,将机器学习模型中位于已训练的数据处理层之后的至少一个数据处理层作为候选训练层,并采用当前训练样本,对候选训练层进行训练。

[0073] 可选的,所谓候选训练层可以是机器学习模型中,位于已训练数据处理层之后的各数据处理层,需要说明的是,本申请实施例中,可以是先以已训练的数据处理层之后的第一个数据处理层作为候选训练层,然后采用当前训练样本对该数据处理层进行训练,并判断训练结果与训练目标相比,残差是否满足收敛条件,如残差值是否不在减小或残差值是否在可接受误差范围内,若是,则说明该候选训练层选择的准确,训练结果满足收敛条件,可执行S403的操作,若否,则说明该候选训练层还需要继续增加,如将该数据处理层之后的下一数据处理层也添加到该候选训练层中,继续采用当前训练样本对新加入的该候选训练层进行训练,得到训练结果,并判断该训练结果与训练目标相比,残差是否满足收敛条件,若满足则执行S403的操作,否则按照上述介绍的方案继续增加候选训练层的数量,直到对候选训练层训练的结果与训练目标的残差满足收敛条件为止。

[0074] 例如,若图3B所示的机器学习模型为树模型,且第一训练阶段已经结束,训练了该树模型的第1-5棵决策树,此时当前训练阶段(即第二训练阶段)可以是先将第6棵决策树作为当前训练样本的候选训练层,采用当前训练样本计算第6层的输出结果与训练目标相比,残差值是否不在减小,若是,则说明模型满足收敛条件,若否,则说明模型还没有收敛,此时需要将第7层也添加到候选训练层中,继续计算第7层的输出结果与训练目标之间的残差值是否不在减小,依次类推,直到残差值不在减小,则认为模型训练到满足收敛条件,方可执行后续S403的操作。

[0075] S403,若候选训练层的训练结果满足收敛条件,则将候选训练层作为当前训练样本训练的数据处理层,且当前训练样本对机器学习模型训练结束。

[0076] 可选的,在本申请实施例中,若S403对采用当前训练样本,对候选训练层进行训练的训练结果满足收敛条件,则此时当前训练样本对机器学习模型的当前阶段训练结束,且

满足收敛条件时对应的候选训练层即为当前训练样本训练的数据处理层。

[0077] 需要说明的是,本申请实施例可以是依次将S401构建的每一训练样本作为当前训练样本,来执行S402和S403的操作,直到所有训练样本都执行完上述S402和S403的操作,则此时机器学习模型训练结束,得到目标模型。

[0078] 本申请实施例的技术方案,对于至少两种数据源的数据特征,根据数据源的覆盖率的不同,构建至少两种训练样本,然后依次将每种训练样本作为当前训练样本,将机器学习模型中,已训练的数据处理层之后的至少一个数据处理层作为候选训练层,采用当前训练样本对候选训练层进行训练,若训练结果满足收敛条件,则当前训练样本对机器学习模型训练结束,且将该候选训练层作为当前训练样本训练的数据处理层。本申请实施例的方案,每一训练样本需要训练的数据处理层的层数是在模型训练过程中,根据收敛条件来确定,并非人工预先设置的,提高了各训练样本训练的数据处理层的准确性,进而提高了模型训练的准确性。

[0079] 图5是根据本申请实施例提供的一种模型训练方法的流程图。本实施例在上述实施例的基础上,给出了一种确定数据源的覆盖率的方法介绍,如图5所示,该方法包括:

[0080] S501,获取包含至少两种数据源的数据特征集合。

[0081] 其中,数据特征集合可以是包含了多次从至少两种数据源中获取的数据特征的集合。

[0082] 可选的,在本申请实施例中,可以是分批次(如执行5次)从至少两种数据源中执行获取数据特征的操作,并将每次获取的数据特征作为一个子集放在数据特征集合中。需要说明的是,虽然每次都从至少两种数据源中获取特征数据,但是由于不同数据的覆盖率不同,所以每次并不一定从各数据源中都获取到了特征数据。例如,若两种数据源分别为搜索引擎数据源和贴吧数据源,由于搜索引擎数据源的注册用户明显高于贴吧数据源的注册用户,即搜索引擎数据源的覆盖率高于贴吧数据源的覆盖率。假设五次分别获取的是用户1、用户2、用户3、用户4和用户5在这两种数据源中的搜索浏览特征,且只有用户1和用户3同时使用搜索引擎和贴吧,用户2、用户4和用户5只用搜索引擎,则此时虽然对两个数据源都执行了5次获取数据特征的操作,但实际并不是每次都同时获取到搜索引擎的搜索流量特征1和贴吧的搜索流量特征2。此时得到的数据特征集合为:

$$\begin{bmatrix} \text{搜索浏览特征1} & \text{搜索流量特征2} \\ \text{搜索浏览特征1} & 0 \\ \text{搜索流量特征1} & \text{搜索流量特征2} \\ \text{搜索浏览特征1} & 0 \\ \text{搜索浏览特征1} & 0 \end{bmatrix},$$

其中该数据特征集合的一行数据代表一次获取的特征子集。

[0083] S502,根据数据特征集合,确定至少两种数据源的特征数量和目标特征数量。

[0084] 其中,本申请实施例中,目标特征数量是指获取数据特征集时,若每次从各数据源获取数据特征时都能获取到,则执行多次获取操作后,针对一个数据源获取到的数据特征的数量,也就是说,该目标特征数量为获取数据特征集合时,执行获取操作的次数。

[0085] 可选的,本申请实施例中,要确定每种数据源的覆盖率,就需要先确定数据特征集合中至少两种数据源的特征数量和目标特征数量,具体的,确定至少两种数据源的特征数量时,可以是针对每种数据源,统计数据特征集合中包含的该种数据源的数据特征的特征

数量,也就是说,针对每种数据源,统计多次获取操作获取到该种数据源的数据特征的次数;确定目标特征数量时,可以是统计获取操作的总次数。例如,针对S501示出的数据特征集合,可以是统计包含各种数据源的数据特征的总行数,作为各种数据源的特征数量,如包含搜索引擎数据源的搜索浏览特征1的总行数为5行,则搜索引擎数据源的特征数量为5,包含贴吧数据源的搜索浏览特征2的总行数为2行,则贴吧数据源的特征数量为2。可以统计数据特征集合的总行数5,作为目标特征数量。

[0086] S503,根据至少两种数据源的特征数量和目标特征数量,确定至少两种数据源的覆盖率。

[0087] 可选的,本申请实施例可以是针对每一种数据源,将该种数据源的特征数量占目标特征数量的比例,作为该种数据源的覆盖率。例如,若搜索引擎数据源的特征数量为5,贴吧数据源的特征数量为2,目标特征数量为5,则搜索引擎数据源的覆盖率为5/5=1;贴吧数据源的覆盖率为2/5=0.4。显然搜索引擎数据源的覆盖率高于贴吧数据源的覆盖率。

[0088] S504,根据至少两种数据源的覆盖率,采用至少两种数据源的数据特征构建至少两种训练样本。

[0089] 其中,训练样本关联有至少一种数据源;

[0090] S505,采用至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型。

[0091] 其中,不同训练样本训练的数据处理层不同。

[0092] 本申请实施例的技术方案,根据获取的包含至少两种数据源的数据特征集合,确定各数据源的数据特征数量和目标特征数量,并将各数据源的特征数量占目标特征数量的比值作为各种数据源的覆盖率。进而对于至少两种数据源的数据特征,根据其数据源的覆盖率的不同,构建至少两种训练样本,依次采用各训练样本分阶段为机器学习模型中的不同数据处理层进行训练。本申请实施例的方案根据多次获取的包含至少两种数据源的数据特征集合,来计算各数据源的覆盖率,时效性和准确性更高,为后续依据该覆盖率进行模型训练提供了保证。

[0093] 图6是根据本申请实施例提供的一种模型使用方法的流程图。本实施例适用于基于上述各实施例训练的目标模型,执行任务预测的情况。该实施例可以由电子设备中配置的模型使用装置来执行,该装置可以采用软件和/或硬件来实现。如图6所示,该方法包括:

[0094] S601,根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的至少两种训练样本中,确定待预测样本关联的目标训练样本。

[0095] 其中,待预测样本可以是执行预测操作时,需要输入到训练好的目标模型中的输入数据。可选的,本申请实施例中,待预测样本中包含的数据特征属于至少一种数据源。本实施例中所谓的模型训练阶段可以是指上述任意实施例的模型训练方法中,对机器学习模型进行训练得到目标模型的阶段。

[0096] 可选的,在本申请实施例中,可以是先确定待预测样本中包含的数据特征属于的至少一种数据源,然后实时计算各数据源的覆盖率,或者直接参考模型训练阶段计算出的各数据源的覆盖率,由于在模型训练阶段,对模型进行训练前执行过采用至少两种数据源的数据特征构建至少两种训练样本的操作,所以此时本步骤可以是判断待预测样本关联的数据源的覆盖率与模型训练阶段构建的哪种训练样本关联的数据源的覆盖率一致,则将该

训练样本作为待预测样本关联的目标训练样本。

[0097] 例如,假设模型训练阶构建了两种训练样本,即将搜索引擎数据源的浏览搜索特征1作为第一训练样本,将搜索引擎数据源的浏览搜索特征1和贴吧搜索数据源的浏览搜索特征2作为第二训练样本。若本实施例待预测样本中只包含浏览搜索特征1,则该待预测样本与第一训练样本关联的数据源的覆盖率一致,即都只包含高覆盖率的搜索引擎数据源,则该待预测样本关联的目标训练样本为模型训练阶段构建的第一训练样本。同理,若待训练样本中同时包含浏览搜索特征1和数据特征2,或者只包含数据特征2(该情况出现的概率极低),则此时该待预测样本关联的目标训练样本为模型训练阶段构建的第二训练样本。

[0098] S602,根据目标训练样本,从目标模型的至少两个数据处理层中,确定待调用数据处理层。

[0099] 其中,待调用数据处理层可以是执行本次预测任务,需要从目标模型中调用的数据处理层。目标模型可以是采用上述任意实施例的方法对机器学习模型进行训练得到的。该目标模型需要具备的特征为:包含至少两个数据处理层,且每一数据处理层在上一数据处理层的处理结果上进行迭代计算。例如,满足该特征的目标模型可以包括但不限于:基于梯度提升决策树(Gradient Boosting Decision Tree,GBDT)算法构建的树模型、逻辑回归(Logistic Regression,LR)模型、深度学习模型等。可选的,当所述目标模型为树模型时,所述树模型中的至少两个数据处理层为至少两棵决策树。

[0100] 可选的,本申请实施例中,待预测样本关联的目标训练样本不同,执行本次预测任务,从目标模型中调用的数据处理层也就不同。可以是模型训练好后,记录模型训练阶段构建的各种训练样本与其对应的待调用数据处理层之间的映射关系,此时可以通过该映射关系来确定目标训练样本对应的待调用数据处理层。还可以是查找目标训练样本在模型训练阶段训所训练的数据处理层,并将该数据处理层和位于该数据处理层之前的各数据处理层作为目标训练样本对应的待调用数据处理层。例如,若目标训练样本在模型训练阶段所训练的是机器学习模型的第6-8层,则此时可以是将训练好的目标模型的第1-8层作为目标训练样本的待调用数据处理层。

[0101] S603,根据待调用数据处理层,对待预测样本进行任务预测。

[0102] 可选的,本申请实施例在确定了待调用数据处理层之后,可以是将待预测样本输入到训练好的目标模型中,调用S602确定的待调用数据处理层对输入的预测样本进行处理,得到预测结果。

[0103] 需要说明的是,在本申请实施例中,模型训练阶段,采用一种训练样本训练后的数据处理层,不再基于其他种类的训练样本进行重复训练,例如,采用只关联高覆盖率数据源的第一训练样本对机器学习模型的第1-5层训练后,采用同时包含高低覆盖率数据源的第二训练样本就只对机器学习模型的第6-8层进行训练,机器学习模型的第1-5层不重复训练。本申请实施例这样设置的好处是,保证模型训练结果的准确性。因为在模型使用阶段,若待预测训练样本关联的目标训练样本为第一训练样本时,此时需要调用训练后的目标模型的第1-5层作为待调用数据处理层。若第1-5层没有经过重复训练,其就是通过只关联高覆盖率数据源的训练样本进行训练的,则此时可以基于目标模型的第1-5层精准对只关联高覆盖率数据源的待预测样本进行预测,若第1-5层采用第二训练样本进行了重复训练,因为第二样本中还关联低覆盖率数据源,此时重复训练后的第1-5层就无法准确预测只关联

高覆盖率数据源的待预测样本,影响预测结果的准确性。

[0104] 本申请实施例的技术方案,在分阶段训练得到目标模型之后,根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的各训练样本中,确定该待预测样本关联的目标训练样本,进而根据该目标训练样本确定待调用数据处理层,采用待预测样本,调用目标模型的待调用数据处理层进行任务预测。本申请实施例的方案,对关联不同覆盖率数据源的待预测样本,选择调用同一目标模型的不同数据处理层来执行预测任务,无需部署多个目标模型,在提高任务预测准确性的同时,节约了资源,降低了模型训练的功耗,为模型的使用提供了一种新思路。

[0105] 进一步的,在本申请实施例中,根据所述目标训练样本,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层的过程还可以包括:确定目标训练样本在模型训练阶段训练的数据处理层;根据目标训练样本训练的数据处理层,从目标模型的至少两个数据处理层中,确定待调用数据处理层。具体的,目标训练样本属于模型训练阶段构建的至少两个训练样本中的一个,在模型训练阶段,每种训练样本都对应训练机器学习模型中的一部分数据处理层,所以此时,可以将该目标训练样本在模型训练阶段训练的数据处理层,以及该数据处理层之前的各数据处理层一并作为本次预测的待调用数据处理层。这样设置的好处是结合模型训练阶段来确定本次待调用的数据处理层,保证了本次调用的数据处理层的准确性,为了准确进行任务预测提供了保障。

[0106] 图7是根据本申请实施例提供的一种模型训练装置的结构示意图。本实施例适用于采用不同覆盖率的数据源的数据特征,对机器学习模型进行训练的情况。该装置可实现本申请任意实施例的模型训练方法。该装置700具体包括如下:

[0107] 训练样本构建模块701,用于根据至少两种数据源的覆盖率,采用所述至少两种数据源的数据特征构建至少两种训练样本;其中,所述训练样本关联有至少一种数据源;

[0108] 模型训练模块702,用于采用所述至少两种训练样本,对机器学习模型中的至少两个数据处理层进行训练,得到目标模型;其中,不同训练样本训练的数据处理层不同。

[0109] 本申请实施例的技术方案,对于数据特征所属数据源的覆盖率不一致,存在缺失数据的情况下,为了保证训练的机器学习模型的准确性,不是简单的进行缺失特征补充(其中,简单的进行缺失特征补充,无法保证补充特征的准确),也不是针对不同覆盖率的数据源训练多个机器学习模型来进行任务预测(其中,训练多个机器学习模型成本高,其占用资源多)。而是将数据特征按照所属数据源的覆盖率,分为多种训练样本,每种训练样本对应一个训练阶段,分阶段对一个机器学习模型的不同数据处理层进行训练。从而实现无需补充缺失特征,只训练一个机器学习模型,后续根据待预测数据特征所属数据源的覆盖率,选择调用不同的数据处理层,即可准确进行任务预测,提高模型训练效果,以及后续任务预测准确性的同时,节约了资源,降低了模型训练的功耗,为采用不同覆盖率的数据源的数据特征进行模型训练提供了一种新思路。

[0110] 进一步的,所述训练样本构建模块701包括:

[0111] 数据源分组单元,用于根据至少两种数据源的覆盖率,将所述至少两种数据源分为至少两组;

[0112] 训练样本确定单元,用于从所述至少两组数据源中选择当前组数据源,且将所述当前组数据源的数据特征以及上一训练样本中数据源的数据特征,作为当前训练样本;

[0113] 其中,所述当前组数据源的覆盖率低于所述上一训练样本中数据源的覆盖率。

[0114] 进一步的,所述模型训练模块702具体用于:

[0115] 根据训练样本和待训练层数的关联关系,从机器学习模型中的至少两个数据处理层中,确定当前训练样本待训练的数据处理层;

[0116] 基于机器学习模型中已训练的数据处理层,采用当前训练样本对所述当前训练样本待训练的数据处理层进行训练;

[0117] 其中,所述已训练的数据处理层通过采用在所述当前训练样本之前采用的训练样本训练得到。

[0118] 进一步的,所述模型训练模块702还具体用于:

[0119] 将机器学习模型中位于已训练的数据处理层之后的至少一个数据处理层作为候选训练层,并采用当前训练样本,对所述候选训练层进行训练;

[0120] 若所述候选训练层的训练结果满足收敛条件,则将所述候选训练层作为当前训练样本训练的数据处理层,且当前训练样本对所述机器学习模型训练结束。

[0121] 进一步的,所述装置还包括覆盖率确定模块,所述覆盖率确定模块具体包括:

[0122] 特征集合获取单元,用于获取包含所述至少两种数据源的数据特征集合;

[0123] 特征数量确定单元,用于根据所述数据特征集合,确定所述至少两种数据源的特征数量和目标特征数量;

[0124] 覆盖率确定单元,用于根据所述至少两种数据源的特征数量和目标特征数量,确定所述至少两种数据源的覆盖率。

[0125] 进一步的,所述机器学习模型为树模型;所述树模型中的至少两个数据处理层为至少两棵决策树。

[0126] 图8是根据本申请实施例提供的一种模型使用装置的结构示意图;本实施例适用于基于上述各实施例训练的目标模型,执行任务预测的情况。该装置800具体包括如下:

[0127] 样本分析模块801,用于根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的至少两种训练样本中,确定所述待预测样本关联的目标训练样本;

[0128] 调用数据层确定模块802,用于根据所述目标训练样本,从所述目标模型的至少两个数据处理层中,确定待调用数据处理层;

[0129] 任务预测模块803,用于根据所述待调用数据处理层,对所述待预测样本进行任务预测。

[0130] 本申请实施例的技术方案,在分阶段训练得到目标模型之后,根据待预测样本关联的数据源的覆盖率,从模型训练阶段构建的各训练样本中,确定该待预测样本关联的目标训练样本,进而根据该目标训练样本确定待调用数据处理层,采用待预测样本,调用目标模型的待调用数据处理层进行任务预测。本申请实施例的方案,对关联不同覆盖率数据源的待预测样本,选择调用同一目标模型的不同数据处理层来执行预测任务,无需部署多个目标模型,在提高任务预测准确性的同时,节约了资源,降低了模型训练的功耗,为模型的使用提供了一种新思路。

[0131] 进一步的,所述调用数据层确定模块802具体用于:

[0132] 确定所述目标训练样本在模型训练阶段训练的数据处理层;

[0133] 根据所述目标训练样本训练的数据处理层,从所述目标模型的至少两个数据处理

层中,确定待调用数据处理层。

[0134] 进一步的,所述目标模型为树模型;所述树模型中的至少两个数据处理层为至少两棵决策树。

[0135] 根据本申请的实施例,本申请还提供了一种电子设备和一种可读存储介质。

[0136] 如图9所示,是根据本申请实施例的实现模型训练或模型使用方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0137] 如图9所示,该电子设备包括:一个或多个处理器901、存储器902,以及用于连接各部件的接口,包括高速接口和低速接口。各个部件利用不同的总线互相连接,并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理,包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如,耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中,若需要,可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样,可以连接多个电子设备,各个设备提供部分必要的操作(例如,作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图9中以一个处理器901为例。

[0138] 存储器902即为本申请所提供的非瞬时计算机可读存储介质。其中,所述存储器存储有可由至少一个处理器执行的指令,以使所述至少一个处理器执行本申请所提供的模型训练或模型使用方法。本申请的非瞬时计算机可读存储介质存储计算机指令,该计算机指令用于使计算机执行本申请所提供的模型训练或模型使用方法。

[0139] 存储器902作为一种非瞬时计算机可读存储介质,可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块,如本申请实施例中的模型训练或模型使用方法对应的程序指令/模块(例如,附图7所示的训练样本构建模块701和模型训练模块702,或附图8所示的样本分析模块801、调用数据层确定模块802和任务预测模块803)。处理器901通过运行存储在存储器902中的非瞬时软件程序、指令以及模块,从而执行服务器的各种功能应用以及数据处理,即实现上述方法实施例中的模型训练或模型使用方法。

[0140] 存储器902可以包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需要的应用程序;存储数据区可存储根据实现模型训练或模型使用方法的电子设备的使用所创建的数据等。此外,存储器902可以包括高速随机存取存储器,还可以包括非瞬时存储器,例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中,存储器902可选包括相对于处理器901远程设置的存储器,这些远程存储器可以通过网络连接至实现模型训练或模型使用方法的电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0141] 实现模型训练或模型使用方法的电子设备还可以包括:输入装置903和输出装置904。处理器901、存储器902、输入装置903和输出装置904可以通过总线或者其他方式连接,图9中以通过总线连接为例。

[0142] 输入装置903可接收输入的数字或字符信息,以及产生与实现模型训练或模型使

用方法的电子设备的用户设置以及功能控制有关的键信号输入,例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置904可以包括显示设备、辅助照明装置(例如,LED)和触觉反馈装置(例如,振动电机)等。该显示设备可以包括但不限于,液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中,显示设备可以是触摸屏。

[0143] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0144] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0145] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0146] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0147] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务中,存在的管理难度大,业务扩展性弱的缺陷。

[0148] 根据本申请实施例的技术方案,对于数据特征所属数据源的覆盖率不一致,存在缺失数据的情况下,为了保证训练的机器学习模型的准确性,不是简单的进行缺失特征补

充(其中,简单的进行缺失特征补充,无法保证补充特征的准确),也不是针对不同覆盖率的数据源训练多个机器学习模型来进行任务预测(其中,训练多个机器学习模型成本高,其占用资源多)。而是将数据特征按照所属数据源的覆盖率,分为多种训练样本,每种训练样本对应一个训练阶段,分阶段对一个机器学习模型的不同数据处理层进行训练。从而实现无需补充缺失特征,只训练一个机器学习模型,后续根据待预测数据特征所属数据源的覆盖率,选择调用不同的数据处理层,即可准确进行任务预测,提高模型训练效果,以及后续任务预测准确性的同时,节约了资源,降低了模型训练的功耗,为模型的训练和使用提供了一种新思路。

[0149] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本申请公开的技术方案所期望的结果,本文在此不进行限制。

[0150] 上述具体实施方式,并不构成对本申请保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等,均应包含在本申请保护范围之内。

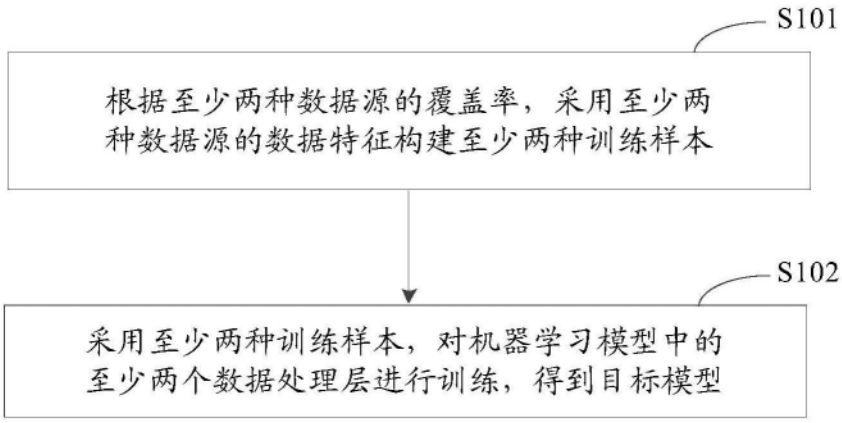

S101
根据至少两种数据源的覆盖率，采用至少两种数据源的数据特征构建至少两种训练样本
S102
采用至少两种训练样本，对机器学习模型中的至少两个数据处理层进行训练，得到目标模型

图1


S201
根据至少两种数据源的覆盖率，将至少两种数据源分为至少两组
S202
从至少两组数据源中选择当前组数据源，且将当前组数据源的数据特征以及上一训练样本中数据源的数据特征，作为当前训练样本
S203
采用至少两种训练样本，对机器学习模型中的至少两个数据处理层进行训练，得到目标模型

图2

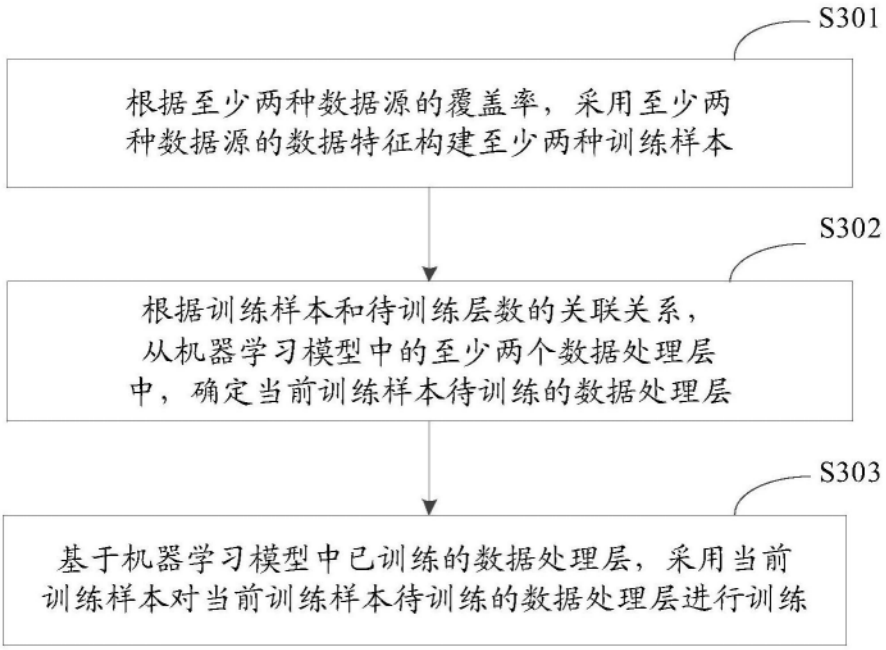

S301
根据至少两种数据源的覆盖率，采用至少两种数据源的数据特征构建至少两种训练样本
S302
根据训练样本和待训练层数的关联关系，从机器学习模型中的至少两个数据处理层中，确定当前训练样本待训练的数据处理层
S303
基于机器学习模型中已训练的数据处理层，采用当前训练样本对当前训练样本待训练的数据处理层进行训练

图3A

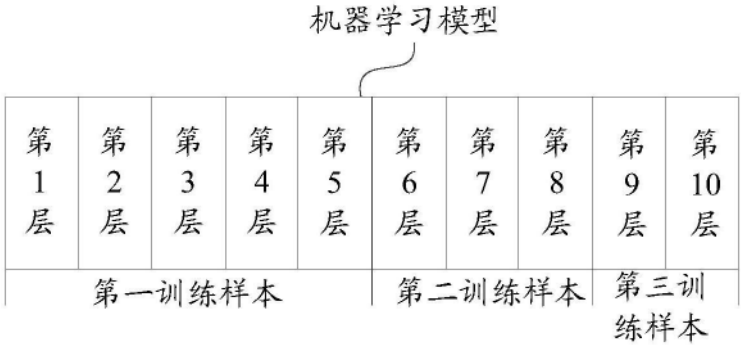

机器学习模型
第1层
第2层
第3层
第4层
第5层
第6层
第7层
第8层
第9层
第10层
第一训练样本
第二训练样本
第三训练样本

图3B

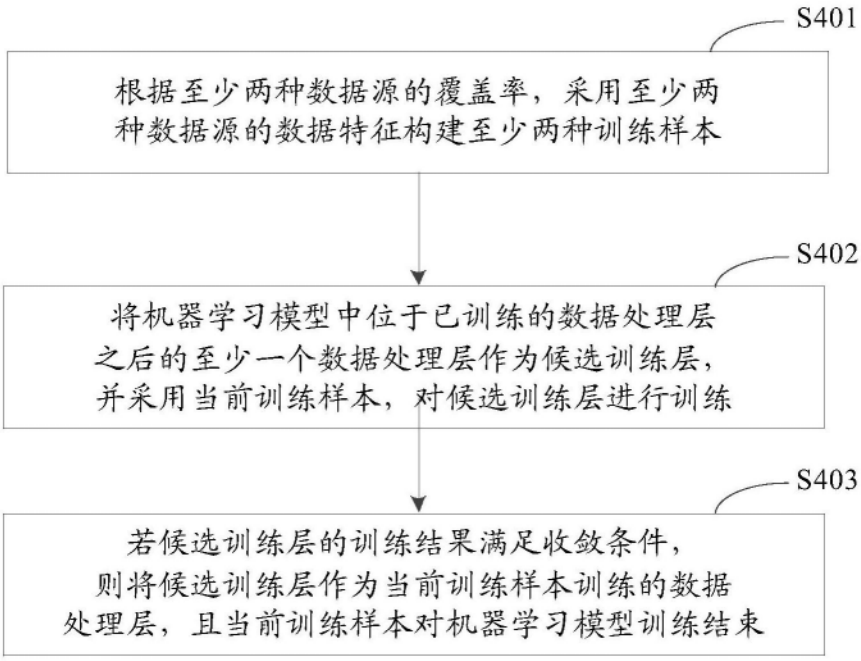

S401
根据至少两种数据源的覆盖率，采用至少两种数据源的数据特征构建至少两种训练样本
S402
将机器学习模型中位于已训练的数据处理层之后的至少一个数据处理层作为候选训练层，并采用当前训练样本，对候选训练层进行训练
S403
若候选训练层的训练结果满足收敛条件，则将候选训练层作为当前训练样本训练的数据处理层，且当前训练样本对机器学习模型训练结束

图4

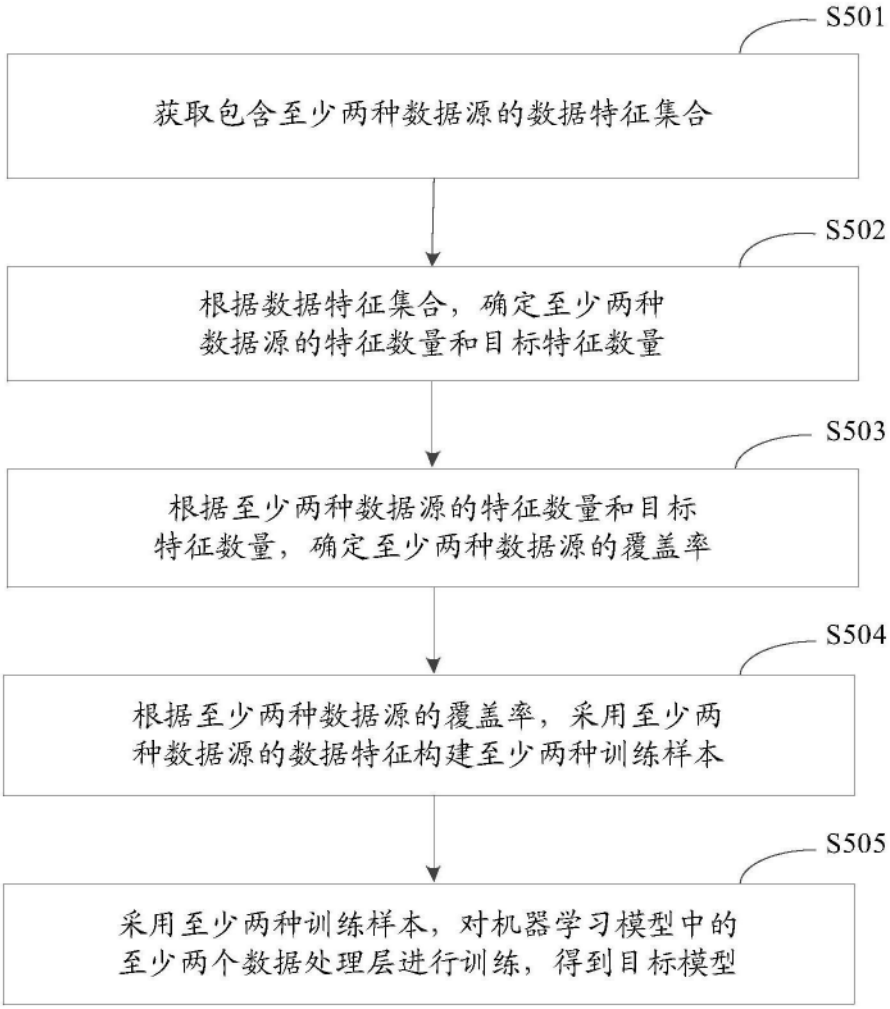

S501
获取包含至少两种数据源的数据特征集合
S502
根据数据特征集合，确定至少两种数据源的特征数量和目标特征数量
S503
根据至少两种数据源的特征数量和目标特征数量，确定至少两种数据源的覆盖率
S504
根据至少两种数据源的覆盖率，采用至少两种数据源的数据特征构建至少两种训练样本
S505
采用至少两种训练样本，对机器学习模型中的至少两个数据处理层进行训练，得到目标模型

图5

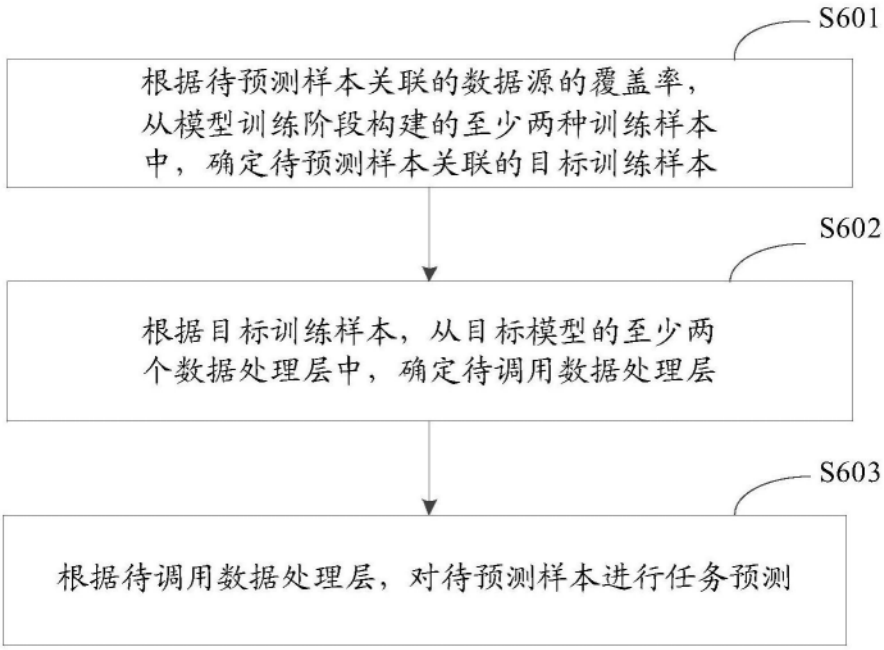

S601
根据待预测样本关联的数据源的覆盖率，从模型训练阶段构建的至少两种训练样本中，确定待预测样本关联的目标训练样本
S602
根据目标训练样本，从目标模型的至少两个数据处理层中，确定待调用数据处理层
S603
根据待调用数据处理层，对待预测样本进行任务预测

图6

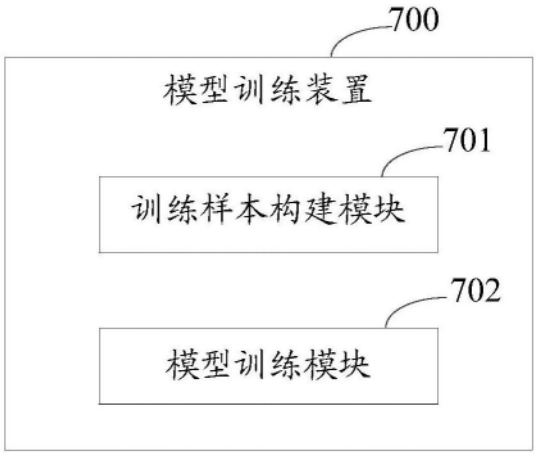

700
模型训练装置
701
训练样本构建模块
702
模型训练模块

图7

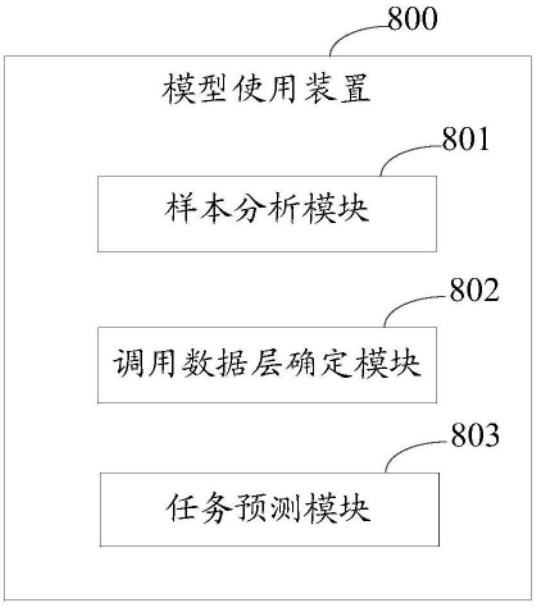

800
模型使用装置
801
样本分析模块
802
调用数据层确定模块
803
任务预测模块

图8

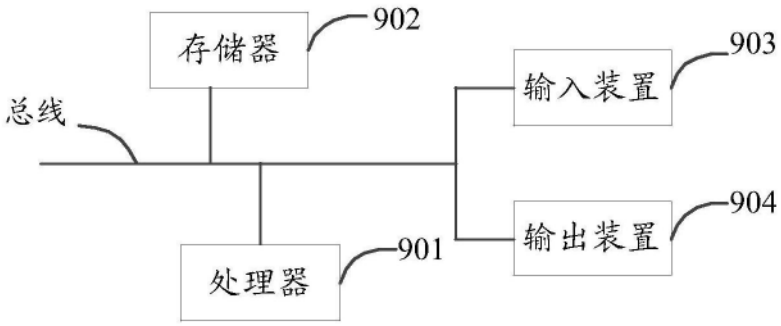

存储器
902
总线
输入装置
903
输出装置
904
处理器
901

图9