

一种基于梯度提升决策树的模型训练方法及装置

技术领域

[01] 本说明书实施例涉及信息技术领域，尤其涉及一种基于梯度提升决策树的模型训练方法及装置。

5 背景技术

[02] 众所周知，当需要训练应用于某个业务场景的预测模型时，通常需要从该业务场景的数据域获取大量数据进行标注，作为已标注样本，进行模型训练。如果已标注样本的数量较少，则通常无法得到效果合格的模型。需要说明的是，某个业务场景的数据域，实际上是基于该业务场景所产生的业务数据的集合。

- 10 [03] 然而，实践中，某些特殊业务场景下积累的数据较少。这导致当需要训练应用于某个特殊业务场景的模型时，无法从该特殊业务场景的数据域获取足够的已标注样本，从而无法得到效果合格的模型。

发明内容

- 15 [04] 为了解决某些特殊业务场景下积累的数据较少导致无法训练出效果合格的模型的问题，本说明书实施例提供一种基于梯度提升决策树的模型训练方法及装置，技术方案如下：

[05] 根据本说明书实施例的第1方面，提供一种基于梯度提升决策树的模型训练方法，用于训练应用于目标业务场景的目标模型，所述方法包括：

- 20 [06] 获取第一样本集合；所述第一样本集合是从源业务场景的数据域获取的已标注样本的集合；所述源业务场景是与所述目标业务场景相近的业务场景；

[07] 使用所述第一样本集合，执行梯度提升决策树 GBDT 算法流程，依次训练至少一个决策树，直至满足预设的训练暂停条件；

[08] 根据使用所述第一样本集合训练出的决策树，确定训练残差；

- 25 [09] 获取第二样本集合；所述第二样本集合是从所述目标业务场景的数据域获取的已标注样本的集合；

- [10] 使用所述第二样本集合, 基于所述训练残差继续执行 GBDT 算法流程, 依次训练至少一个决策树, 直至满足预设的训练停止条件;
- [11] 其中, 所述目标模型是由已训练出的决策树集成得到的。
- [12] 根据本说明书实施例的第 2 方面, 提供一种预测方法, 包括:
- 5 [13] 从目标业务场景的数据域获取待预测数据;
- [14] 根据所述待预测数据, 确定所述待预测数据对应的模型输入特征;
- [15] 将所述模型输入特征输入到应用于所述目标业务场景的预测模型, 以输出预测结果; 所述预测模型是根据上述第 1 方面的方法得到的。
- [16] 根据本说明书实施例的第 3 方面, 提供一种基于梯度提升决策树的模型训练装置,
- 10 [17] 第一获取模块, 获取第一样本集合; 所述第一样本集合是从源业务场景的数据域获取的已标注样本的集合; 所述源业务场景是与所述目标业务场景相近的业务场景;
- [18] 第一训练模块, 使用所述第一样本集合, 执行梯度提升决策树 GBDT 算法流程, 依次训练至少一个决策树, 直至满足预设的训练暂停条件;
- 15 [19] 计算模块, 根据使用所述第一样本集合训练出的决策树, 确定训练残差;
- [20] 第二获取模块, 获取第二样本集合; 所述第二样本集合是从所述目标业务场景的数据域获取的已标注样本的集合;
- [21] 第二训练模块, 使用所述第二样本集合, 基于所述训练残差继续执行 GBDT 算法流程, 依次训练至少一个决策树, 直至满足预设的训练停止条件;
- 20 [22] 其中, 所述目标模型是由已训练出的决策树集成得到的。
- [23] 根据本说明书实施例的第 4 方面, 提供一种预测装置, 包括:
- [24] 获取模块, 从目标业务场景的数据域获取待预测数据;
- [25] 确定模块, 根据所述待预测数据, 确定所述待预测数据对应的模型输入特征;
- [26] 输入模块, 将所述模型输入特征输入到应用于所述目标业务场景的预测模型, 以
- 25 [27] 输出预测结果; 所述预测模型是根据上述第 1 方面的方法得到的。
- [27] 本说明书实施例所提供的技术方案, 将一个 GBDT 算法流程划分为两个阶段, 在前一阶段, 从与目标业务场景相近的业务场景的数据域获取已标注样本依次训练若干决

策树，并确定经过前一阶段训练后产生的训练残差；在下一阶段，从目标业务场景的数据域获取已标注样本，并基于所述训练残差，继续训练若干决策树。最终，应用于目标业务场景的模型实际上是由前一阶段训练出的决策树与后一阶段训练出的决策树集成得到的。通过本说明书实施例，虽然目标业务场景下积累的数据不足，但是，可以借助
5 与目标业务场景相近的业务场景的数据，训练应用于目标业务场景的模型。经过测试，可以得到效果合格的模型。

[28] 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和解释性的，并不能限制本说明书实施例。

[29] 此外，本说明书实施例中的任一实施例并不需要达到上述的全部效果。

10 附图说明

[30] 为了更清楚地说明本说明书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说明书实施例中记载的一些实施例，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。

15 [31] 图 1 是本说明书实施例提供的一种基于梯度提升决策树的模型训练方法的流程示意图；

[32] 图 2 是本说明书实施例提供的方案架构示意图；

[33] 图 3 是本说明书实施例提供的一种预测方法的流程示意图；

20 [34] 图 4 是本说明书实施例提供的一种基于梯度提升决策树的模型训练装置的结构示意图；

[35] 图 5 是本说明书实施例提供的一种预测装置的结构示意图；

[36] 图 6 是用于配置本说明书实施例方法的一种设备的结构示意图。

具体实施方式

25 [37] 本发明借鉴了机器学习技术领域的迁移学习思想。在面对训练应用于目标业务场景的模型的需求时，如果目标业务场景下积累的数据不足，那么可以利用与目标业务场景相近的业务场景下积累的数据进行模型训练。

[38] 具体地,本发明将迁移学习思想与梯度提升决策树(Gradient Boosting Decision Tree, GBDT)算法相结合,对GBDT算法流程进行了改进。在本说明书实施例中,针对一个GBDT算法流程,先使用与目标业务场景相近的业务场景下产生的数据进行训练,满足一定的训练暂停条件之后,暂停训练,并计算当前的训练残差;随后,使用目标业务场景下产生的数据,基于所述训练残差继续训练,直到满足一定的训练停止条件。如此,将训练得到的GBDT模型应用于目标业务场景,可以取得较好的预测效果。

[39] 需要说明的是,在本文中,与目标业务场景相近的业务场景,实际上是与目标业务场景相类似或相关联的业务场景。本文将与目标业务场景相近的业务场景称为源业务场景。

10 [40] 举例来说,假设目标业务场景是男性商品推荐场景,为了更好的根据男性用户的年龄进行商品推荐,需要训练用于预测男性用户年龄的模型。然而,由于男性商品推荐功能上线不久,积累的男性用户购买记录较少(购买记录中记载了购买者的各种特征信息以及购买者的年龄),因为无法获得足够的已标注样本进行训练。于是,可以以女性商品推荐场景为目标业务场景对应的源业务场景。由于女性商品推荐功能早已上线,已经积累了大量女性用户购买记录,因此,在本说明书实施例中,可以借助于积累的大量女性用户购买记录,使用少量的男性用户购买记录训练出效果合格的,用于预测男性用户年龄的模型。

20 [41] 为了使本领域技术人员更好地理解本说明书实施例中的技术方案,下面将结合本说明书实施例中的附图,对本说明书实施例中的技术方案进行详细地描述,显然,所描述的实施例仅仅是本说明书的一部分实施例,而不是全部的实施例。基于本说明书中的实施例,本领域普通技术人员所获得的所有其他实施例,都应当属于保护的范围。

[42] 以下结合附图,详细说明本说明书各实施例提供的技术方案。

[43] 图1是本说明书实施例提供的一种基于梯度提升决策树的模型训练方法的流程图,包括以下步骤:

25 [44] S100: 获取第一样本集合。

[45] 本方法的目的是训练应用于目标业务场景的目标模型。

[46] 在本说明书实施例中,目标业务场景对应的源业务场景的数据域中积累的数据较多,可以从源业务场景的数据域中获取大量业务数据进行标注,得到足量的已标注样本。步骤S100中的第一样本集合是从所述源业务场景的数据域获取的已标注样本的集合。

[47] 沿用前文所述的例子，目标业务场景是男性商品推荐场景，源业务场景是女性商品推荐场景。从源业务场景的数据域获取若干已标注样本，具体可以从女性商品推荐场景积累的购买记录中，获取若干购买记录，然后，针对获取的每个购买记录，从该购买记录中提取购买者的学历、收入、工作类型、身高等特征信息，构建该购买记录对应的用户特征向量，然后，从该购买记录中提取购买者的年龄，作为该用户特征向量的标注值，如此，得到一个已标注样本。

[48] S102: 使用所述第一样本集合，执行梯度提升决策树 GBDT 算法流程，依次训练至少一个决策树，直至满足预设的训练暂停条件。

[49] 此处先对 GBDT 算法的原理进行介绍。

10 [50] GBDT 算法是一种典型的集成学习算法，在 GBDT 算法流程中，使用一些已标注样本依次训练出两个以上的决策树，然后将训练出的各决策树集成为一个模型，作为训练结果。

[51] 其中，训练的第一个决策树实际上用于拟合各已标注样本的标注值。训练出第一个决策树之后，可以计算得到当前的训练残差。所述训练残差用于表征，截至当前的训练进度，对各已标注样本进行预测得到的预测值与各已标注样本的标注值之间的差距。15 可见，在一个 GBDT 算法流程中，每训练出一个决策树，所述训练残差会更新一次。

[52] 在 GBDT 算法流程中，训练第一个决策树之后，会继续训练下一个决策树。针对除第一个决策树之外的每个其他决策树，该其他决策树用于拟合根据该其他决策树之前所有决策树计算出的训练残差。如此，随着 GBDT 算法流程的推进，越来越多的决策树20 被依次训练出来，所述训练残差会越来越小。当训练残差足够小时，说明当前的模型的模型参数对各已标注样本的标注值的拟合效果达标，此时，便可以结束训练。

[53] 还需要说明的是，在 GBDT 算法中，拟合有两层含义：

[54] 其一，训练的第一个决策树用于拟合样本的标注值。此处的拟合实际上是指，针对第一个决策树，以样本的特征为决策树输入，以样本的标注值为决策树输出，训练决策树参数，尽量使得决策树参数、决策树输入与决策树输出相匹配。25

[55] 其二，针对后续训练的每个决策树，该决策树用于拟合根据之前所有决策树计算得到的训练残差。此处的拟合实际上是指，针对后续训练的每个决策树，以样本的特征为决策树输入，以训练残差为决策树输出，训练决策树参数，尽量使得决策树参数、决策树输入与决策树输出相匹配。

- [56] 而在本说明书实施例 5 中，将 GDBT 算法流程拆分为两个阶段（前一阶段与后一阶段）。步骤 S102 是在前一阶段所执行的步骤。具体地，在步骤 S102 中，使用第一样本集合，执行 GDBT 算法流程，依次训练至少一个决策树，当满足预设的训练暂停条件时，暂停训练，也就是完成前一阶段的训练。
- 5 [57] 需要说明的是，所述训练暂停条件可以根据实际需要指定。例如，所述训练暂停条件可以是，使用所述第一样本集合训练出的决策树的数量达到第一指定数量。在实际应用中，采用对决策树数量进行限定的方式决定暂停第一阶段训练的时机，较为直观，方便执行。
- [58] 又如，所述训练暂停条件可以是，基于已训练出的各决策树计算得到的训练残差
- 10 落入第一指定阈值区间。
- [59] S104：根据使用所述第一样本集合训练出的决策树，确定训练残差。
- [60] 当满足所述训练暂停条件时，暂停训练，并根据使用所述第一样本集合训练出的决策树，计算截至当前，训练过程产生的训练残差。
- [61] 也就是说，在步骤 S104 中，也就是根据前一阶段训练出的各决策树计算当前的训
- 15 练残差，这也意味着完成了对前一阶段的阶段性训练结果的整合。
- [62] S106：获取第二样本集合。
- [63] 其中，所述第二样本集合是从所述目标业务场景的数据域获取的已标注样本的集合。
- [64] 需要说明的是，一般而言，第一样本集合中的已标注样本的数量显著大于第二样
- 20 本集合中的已标注样本的数量。
- [65] 还需要说明的，在本说明书实施例 5 中，并不对执行步骤 S106 的时机进行限制。实际上，可以在步骤 S100~S104 中任一步骤之前执行步骤 S106。
- [66] S108：使用所述第二样本集合，基于所述训练残差继续执行所述 GDBT 算法流程，依次训练至少一个决策树，直至满足预设的训练停止条件。
- 25 [67] 在前一阶段之后，需要将所述训练残差迁移到后一阶段。步骤 S108 是在后一阶段所执行的步骤。在后一阶段，使用第二样本集合，基于所述训练残差，继续执行 GDBT 算法流程，当满足预设的训练停止条件时，停止训练，也就完成了后一阶段的训练。
- [68] 在步骤 S108 中，实际上是使用第二样本集合，继承前一阶段的阶段性训练结果，

继续训练决策树。

[69] 其中，使用所述第二样本集合训练出的第一个决策树，用于拟合前一阶段产生的训练残差。针对使用所述第二样本集合训练出的第一个决策树之后的每个决策树，该决策树用于拟合根据该决策树之前的所有决策树（包括使用所述第一样本集合训练出的决策树和使用所述第二样本集合训练出的决策树）计算出的训练残差。

[70] 需要说明的是，当满足预设的训练停止条件时，所述后一阶段才会结束（即停止训练）。所述训练停止条件可以根据实际需要指定。

[71] 例如，所述训练停止条件可以是，使用所述第二样本集合训练出的决策树的数量达到第二指定数量。又如，所述训练停止条件可以是，基于已训练的各决策树计算得到的训练残差落入第二指定阈值区间。通常，所述第二阈值区间的右端点的取值小于所述第一阈值区间的左端点的取值。

[72] 通过步骤 S100~S108，可以训练出若干决策树。于是，可以将训练出的各决策树进行集成，得到的应用于所述目标业务场景的目标模型。具体地，所述目标模型可以是对各决策树，按产生顺序，由先到后进行排序得到的决策树序列。

[73] 通过图 1 所示的模型训练方法，将一个 GBDT 算法流程划分为两个阶段，在前一阶段，从与目标业务场景相近的业务场景的数据域获取已标注样本依次训练若干决策树，并确定经过前一阶段训练后产生的训练残差；在后一阶段，从目标业务场景的数据域获取已标注样本，并基于所述训练残差，继续训练若干决策树。最终，应用于目标业务场景的模型实际上是由第一阶段训练出的决策树与第二阶段训练出的决策树集成得到的。

通过本说明书实施例，虽然目标业务场景下积累的数据不足，但是，可以借助与目标业务场景相近的业务场景的数据，训练应用于目标业务场景的模型。经过测试，可以得到效果合格的模型。

[74] 此外，在本说明书实施例中，可以从不止一个源业务场景的数据域获取已标注样本，进行所述前一阶段的训练。例如，假设目标业务场景是男性商品推荐场景，那么可以获取的源业务场景为女性商品推荐场景和儿童商品推荐场景，依次使用这两个源业务场景下产生的数据进行 GBDT 算法流程中前一阶段的训练。

[75] 具体地，在步骤 S108 之前，可以获取第三样本集合。所述第三样本集合是从其他源业务场景的数据域获取的已标注样本的集合。

[76] 接着，在步骤 S108 之前，使用所述第三样本集合，基于所述训练残差继续执行

GBDT 算法流程,依次训练至少一个决策树,直至满足预设的训练再暂停条件,并根据使用所述第一样本集合训练出的决策树和使用所述第三样本集合训练出的决策树,重新确定所述训练残差。

[77] 其中,所述训练再暂停条件可以根据实际需要指定。例如,所述训练暂停条件可以是,使用所述第三样本集合训练出的决策树的数量达到第三指定数量。又如,所述训练暂停条件可以是,基于已训练出的各决策树计算得到的训练残差落入第三指定阈值区间。

[78] 也就是说,在 GBDT 算法流程的前一阶段,可以先从一个源业务场景的数据域获取已标注样本的集合,并开始执行 GBDT 算法流程。当满足训练暂停条件时,先暂停训练,并计算当前的训练残差。随后,更换使用另一个源业务场景对应的已标注样本的结合,基于所述训练残差,继续执行 GBDT 算法流程。当满足训练再暂停条件时,再次暂停,并重新计算当前的训练残差。

[79] 之后,还可以再次更换其他源业务场景,继续进行前一阶段的训练。总之,本领域技术人员应当理解,如下技术方案应在本发明所要求的保护范围之内:

[80] 将 GBDT 算法流程分为两个阶段,在前一阶段,依次使用至少两个源业务场景下产生的数据执行 GBDT 算法流程,在下一阶段,使用目标业务场景下产生的数据继续执行 GBDT 算法流程。

[81] 图 2 是本说明书实施例提供的方案架构示意图。如图 2 所示,使用不止一个(图中以 3 个为例)源业务场景的数据域的数据执行 GBDT 算法流程中前一阶段的训练,然后,使用目标业务场景的数据域的数据开始下一阶段的训练,以拟合前一阶段产生的训练残差,最终得到模型。

[82] 如图 2 所示,使用源业务场景 A 积累的数据开始执行 GBDT 算法流程,当使用源业务场景 A 积累的数据所训练出的决策树的数量达到 N_1 时,计算截至当前的训练残差,记为训练残差 1。随后,使用源业务场景 B 积累的数据,基于训练残差 1,继续执行 GBDT 算法流程,当使用源业务场景 B 积累的数据所训练出的决策树的数量达到 N_2 时,计算截至当前的训练残差,记为训练残差 2。随后,使用源业务场景 C 积累的数据,基于训练残差 2,继续执行 GBDT 算法流程,当使用源业务场景 C 积累的数据所训练出的决策树的数量达到 N_3 时,计算截至当前的训练残差,记为训练残差 3。至此,前一阶段结束。可见,在图 2 中,训练残差 3 实际上就是下一阶段所要继承的训练残差(即步骤 S108

中所述训练残差)。

[83] 在下一阶段,使用目标业务场景积累的数据,基于训练残差 3 继续执行 GBDT 算法流程,当使用目标业务场景积累的数据所训练出的决策树的数量达到 N_4 时,停止训练。

5 [84] 图 3 是本说明书实施例提供的一种预测方法的流程示意图,包括如下步骤:

[85] S300: 从目标业务场景的数据域获取待预测数据。

[86] S302: 根据所述待预测数据,确定所述待预测数据对应的模型输入特征。

[87] S304: 将所述模型输入特征输入到应用于所述目标业务场景的预测模型,以输出预测结果。

10 [88] 其中,所述预测模型是根据图 1 所示的训练方法得到的。

[89] 基于图 1 所示的模型训练方法,本说明书实施例还对应提供了一种基于梯度提升决策树的模型训练装置,用于训练应用于目标业务场景的目标模型,如图 4 所示,所述装置包括:

[90] 第一获取模块 401,获取第一样本集合;所述第一样本集合是从源业务场景的数据域获取的已标注样本的集合;所述源业务场景是与所述目标业务场景相近的业务场景;

[91] 第一训练模块 402,使用所述第一样本集合,执行梯度提升决策树 GBDT 算法流程,依次训练至少一个前决策树,直至满足预设的训练暂停条件;

[92] 计算模块 403,根据使用所述第一样本集合训练出的决策树,确定训练残差;

[93] 第二获取模块 404,获取第二样本集合;所述第二样本集合是从所述目标业务场景的数据域获取的已标注样本的集合;

[94] 第二训练模块 405,使用所述第二样本集合,基于所述训练残差继续执行 GBDT 算法流程,依次训练至少一个后决策树,直至满足预设的训练停止条件;

[95] 其中,所述目标模型是由已训练出的决策树集成得到的。

[96] 所述训练暂停条件,具体包括:

25 [97] 使用所述第一样本集合训练出的决策树的数量达到第一指定数量。

[98] 所述训练停止条件,具体包括:

[99] 使用所述第一样本集合训练出的决策树的数量达到第二指定数量。

[100] 所述装置还包括:

[101] 再处理模块 406, 在使用所述第二样本集合, 基于所述训练残差继续执行 GBDT 算法流程之前, 获取第三样本集合; 所述第三样本集合是从其他源业务场景的数据域获取的已标注样本的集合; 使用所述第三样本集合, 基于所述训练残差继续执行 GBDT 算法流程, 依次训练至少一个决策树, 直至满足预设的训练再暂停条件; 根据使用所述第一样本集合训练出的决策树和使用所述第三样本集合训练出的决策树, 重新确定所述训练残差。

[102] 所述训练再暂停条件, 具体包括:

[103] 使用所述第三样本集合训练出的决策树的数量达到第三指定数量。

10 [104] 基于图 3 所示的预测方法, 本说明书实施例还对应提供了一种预测装置, 如图 5 所示, 包括:

[105] 获取模块 501, 从目标业务场景的数据域获取待预测数据;

[106] 确定模块 502, 根据所述待预测数据, 确定所述待预测数据对应的模型输入特征;

15 [107] 输入模块 503, 将所述模型输入特征输入到应用于所述目标业务场景的预测模型, 以输出预测结果; 所述预测模型是根据图 1 所示的方法得到的。

[108] 本说明书实施例还提供一种计算机设备, 其至少包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序, 其中, 处理器执行所述程序时实现图 1 所示的方法。

20 [109] 图 6 示出了本说明书实施例所提供的一种更为具体的计算设备硬件结构示意图, 该设备可以包括: 处理器 1010、存储器 1020、输入/输出接口 1030、通信接口 1040 和总线 1050。其中处理器 1010、存储器 1020、输入/输出接口 1030 和通信接口 1040 通过总线 1050 实现彼此之间在设备内部的通信连接。

25 [110] 处理器 1010 可以采用通用的 CPU (Central Processing Unit, 中央处理器)、微处理器、应用专用集成电路 (Application Specific Integrated Circuit, ASIC)、或者一个或多个集成电路等方式实现, 用于执行相关程序, 以实现本说明书实施例所提供的技术方案。

[111] 存储器 1020 可以采用 ROM (Read Only Memory, 只读存储器)、RAM (Random Access Memory, 随机存取存储器)、静态存储设备, 动态存储设备等形式实现。存储

器 1020 可以存储操作系统和其他应用程序，在通过软件或者固件来实现本说明书实施例所提供的技术方案时，相关的程序代码保存在存储器 1020 中，并由处理器 1010 来调用执行。

5 [112] 输入/输出接口 1030 用于连接输入/输出模块，以实现信息输入及输出。输入输出/模块可以作为组件配置在设备中（图中未示出），也可以外接于设备以提供相应功能。其中输入设备可以包括键盘、鼠标、触摸屏、麦克风、各类传感器等，输出设备可以包括显示器、扬声器、振动器、指示灯等。

10 [113] 通信接口 1040 用于连接通信模块（图中未示出），以实现本设备与其他设备的通信交互。其中通信模块可以通过有线方式（例如 USB、网线等）实现通信，也可以通过无线方式（例如移动网络、WIFI、蓝牙等）实现通信。

[114] 总线 1050 包括一通路，在设备的各个组件（例如处理器 1010、存储器 1020、输入/输出接口 1030 和通信接口 1040）之间传输信息。

15 [115] 需要说明的是，尽管上述设备仅示出了处理器 1010、存储器 1020、输入/输出接口 1030、通信接口 1040 以及总线 1050，但是在具体实施过程中，该设备还可以包括实现正常运行所必需的其他组件。此外，本领域的技术人员可以理解的是，上述设备中也可以仅包含实现本说明书实施例方案所必需的组件，而不必包含图中所示的全部组件。

[116] 本说明书实施例还提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现图 1 所示的方法。

20 [117] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于
25 存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

[118] 通过以上的实施方式的描述可知，本领域的技术人员可以清楚地了解到本说明书实施例可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解，本说明书实

施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在存储介质中，如 ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本说明书实施例各个实施例或者实施例的某些部分所述的方法。

- 5 [119] 上述实施例阐明的系统、方法、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机，计算机的具体形式可以是个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件收发设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任意几种设备的组合。
- 10 [120] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于方法实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。以上所描述的方法实施例仅仅是示意性的，其中所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，在实施本说明书实施例方案时
- 15 可以把各模块的功能在同一个或多个软件和/或硬件中实现。也可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。
- [121] 以上所述仅是本说明书实施例的具体实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本说明书实施例原理的前提下，还可以做出若干改进和润饰，
- 20 这些改进和润饰也应视为本说明书实施例的保护范围。

权利要求书

1、一种基于梯度提升决策树的模型训练方法，用于训练应用于目标业务场景的目标模型，所述方法包括：

5 获取第一样本集合；所述第一样本集合是从源业务场景的数据域获取的已标注样本的集合；所述源业务场景是与所述目标业务场景相近的业务场景；

使用所述第一样本集合，执行梯度提升决策树 GBDT 算法流程，依次训练至少一个决策树，直至满足预设的训练暂停条件；

根据使用所述第一样本集合训练出的决策树，确定训练残差；

10 获取第二样本集合；所述第二样本集合是从所述目标业务场景的数据域获取的已标注样本的集合；

使用所述第二样本集合，基于所述训练残差继续执行 GBDT 算法流程，依次训练至少一个决策树，直至满足预设的训练停止条件；

其中，所述目标模型是由已训练出的决策树集成得到的。

2、如权利要求 1 所述的方法，所述训练暂停条件，具体包括：

15 使用所述第一样本集合训练出的决策树的数量达到第一指定数量。

3、如权利要求 1 所述的方法，所述训练停止条件，具体包括：

使用所述第二样本集合训练出的决策树的数量达到第二指定数量。

4、如权利要求 1 所述的方法，在使用所述第二样本集合，基于所述训练残差继续执行 GBDT 算法流程之前，所述方法还包括：

20 获取第三样本集合；所述第三样本集合是从其他源业务场景的数据域获取的已标注样本的集合；

使用所述第三样本集合，基于所述训练残差继续执行 GBDT 算法流程，依次训练至少一个决策树，直至满足预设的训练再暂停条件；

25 根据使用所述第一样本集合训练出的决策树和使用所述第三样本集合训练出的决策树，重新确定所述训练残差。

5、如权利要求 4 所述的方法，所述训练再暂停条件，具体包括：

使用所述第三样本集合训练出的决策树的数量达到第三指定数量。

6、一种预测方法，包括：

从目标业务场景的数据域获取待预测数据；

30 根据所述待预测数据，确定所述待预测数据对应的模型输入特征；

将所述模型输入特征输入到应用于所述目标业务场景的预测模型，以输出预测结果；

所述预测模型是根据权利要求 1~5 任一项所述的方法得到的。

7、一种基于梯度提升决策树的模型训练装置，用于训练应用于目标业务场景的目标模型，所述装置包括：

5 第一获取模块，获取第一样本集合；所述第一样本集合是从源业务场景的数据域获取的已标注样本的集合；所述源业务场景是与所述目标业务场景相近的业务场景；

第一训练模块，使用所述第一样本集合，执行梯度提升决策树 GBDT 算法流程，依次训练至少一个前决策树，直至满足预设的训练暂停条件；

计算模块，根据使用所述第一样本集合训练出的决策树，确定训练残差；

10 第二获取模块，获取第二样本集合；所述第二样本集合是从所述目标业务场景的数据域获取的已标注样本的集合；

第二训练模块，使用所述第二样本集合，基于所述训练残差继续执行 GBDT 算法流程，依次训练至少一个后决策树，直至满足预设的训练停止条件；

其中，所述目标模型是由已训练出的决策树集成得到的。

8、如权利要求 7 所述的装置，所述训练暂停条件，具体包括：
15 使用所述第一样本集合训练出的决策树的数量达到第一指定数量。

9、如权利要求 7 所述的装置，所述训练停止条件，具体包括：
使用所述第一样本集合训练出的决策树的数量达到第二指定数量。

10、如权利要求 7 所述的装置，所述装置还包括：

再处理模块，在使用所述第二样本集合，基于所述训练残差继续执行 GBDT 算法流
20 程之前，获取第三样本集合；所述第三样本集合是从其他源业务场景的数据域获取的已标注样本的集合；使用所述第三样本集合，基于所述训练残差继续执行 GBDT 算法流程，依次训练至少一个决策树，直至满足预设的训练再暂停条件；根据使用所述第一样本集合训练出的决策树和使用所述第三样本集合训练出的决策树，重新确定所述训练残差。

11、如权利要求 10 所述的装置，所述训练再暂停条件，具体包括：
25 使用所述第三样本集合训练出的决策树的数量达到第三指定数量。

12、一种预测装置，包括：

获取模块，从目标业务场景的数据域获取待预测数据；

确定模块，根据所述待预测数据，确定所述待预测数据对应的模型输入特征；

30 输入模块，将所述模型输入特征输入到应用于所述目标业务场景的预测模型，以输出预测结果；所述预测模型是根据权利要求 1~5 任一项所述的方法得到的。

13、一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行

的计算机程序，其中，所述处理器执行所述程序时实现如权利要求 1 至 5 任一项所述的方法。

14、一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其中，所述处理器执行所述程序时实现如权利要求 6 所述的方法。