(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 108021931 A
(43)申请公布日 2018.05.11

(21)申请号 201711160012.0

(22)申请日 2017.11.20

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 陈凡 齐翔 王德胜 王韩彬 郭棋林

(74)专利代理机构 北京博思佳知识产权代理有限公司 11415
代理人 林祥

(51)Int.Cl.
G06K 9/62(2006.01)

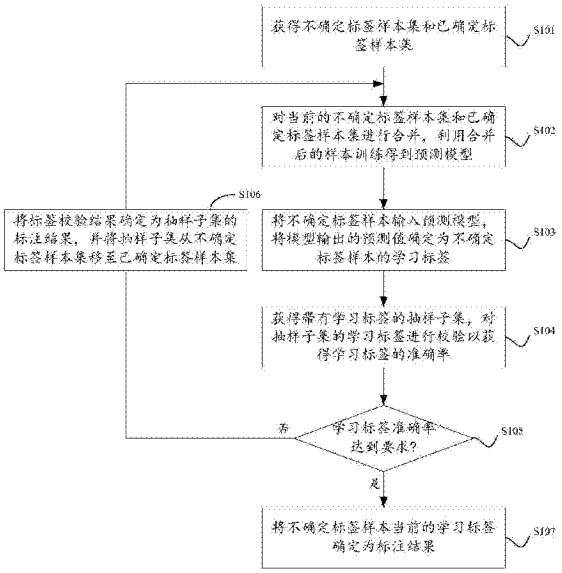
权利要求书3页 说明书11页 附图4页

(54)发明名称

一种数据样本标签处理方法及装置

(57)摘要

公开了一种数据样本标签处理方法及装置。一种数据样本标签处理方法包括：获得不确定标签样本集和已确定标签样本集；利用以下步骤进行迭代处理，直到学习标签的准确率达到预设的要求：对当前的不确定标签样本集和已确定标签样本集进行合并，训练得到预测模型；将不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；迭代结束后，将不确定标签样本当前的学习标签确定为标注结果。



CN 108021931 A

1.一种数据样本标签处理方法，该方法包括：

获得不确定标签样本集和已确定标签样本集；利用以下步骤进行迭代处理，直到学习标签的准确率达到预设的要求：

对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；

将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

迭代结束后，将不确定标签样本当前的学习标签确定为标注结果。

2.根据权利要求1所述的方法，所述不确定标签样本集中的样本均为有标签样本；所述利用合并后的样本训练得到预测模型包括：

利用有监督学习算法对合并后的样本进行训练，得到预测模型。

3.根据权利要求1所述的方法，所述不确定标签样本集中的样本均为无标签样本、或仅部分样本带有标签；所述利用合并后的样本训练得到预测模型包括：

利用半监督学习算法对合并后的样本进行训练，得到预测模型。

4.根据权利要求1所述的方法，所述获得不确定标签样本集，包括：

在初始获得的待处理样本集中样本数量未达到预设需求的情况下，将该待处理样本集输入生成式对抗网络，得到与待处理样本集同分布的生成样本集；

将待处理样本集与生成样本集合并，得到不确定标签样本集。

5.根据权利要求4所述的方法，所述将不确定标签样本集中的不确定标签样本输入所述预测模型，包括：

将不确定标签样本集中，属于待处理样本集的部分输入所述预测模型。

6.根据权利要求4所述的方法，若所述待处理样本集中的样本均为有标签样本，则所述将该待处理样本集输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，包括：

将该待处理样本集的特征部分和标签部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，所述生成样本集中的样本均为有标签样本；

或者

将该待处理样本集的特征部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，所述生成样本集中的样本均为无标签样本。

7.根据权利要求4所述的方法，若所述待处理样本集中的样本均中的样本均为无标签样本、或仅部分样本带有标签，则所述将该待处理样本集输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，包括：

将该待处理样本集的特征部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，所述生成样本集中的样本均为无标签样本。

8.一种数据模型训练方法，该方法包括：

获得不确定标签样本集和已确定标签样本集；利用以下步骤进行迭代处理，直到学习

标签的准确率达到预设的要求：

对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；

将所述不确定标签样本集中的不确定标签样本输入所述预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

对带有学习标签的不确定标签样本集进行抽样得到抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移动至已确定标签样本集；

迭代结束后，将当前的预测模型确定为适用于所述不确定标签样本集的预测模型。

9.一种数据样本标签处理装置，该装置包括：

输入模块，用于获得不确定标签样本集和已确定标签样本集；

学习模块，用于对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

校验模块，用于根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

所述学习模块和所述校验模块相互配合实现迭代处理，直到学习标签的准确率达到预设的要求；

输出模块，用于在迭代结束后，将不确定标签样本当前的学习标签确定为标注结果。

10.根据权利要求9所述的装置，所述不确定标签样本集中的样本均为有标签样本；所述学习模块具体用于：

利用有监督学习算法对合并后的样本进行训练，得到预测模型。

11.根据权利要求9所述的装置，所述不确定标签样本集中的样本均为无标签样本、或仅部分样本带有标签；所述学习模块具体用于：

利用半监督学习算法对合并后的样本进行训练，得到预测模型。

12.根据权利要求9所述的装置，所述输入模块包括：

生成子模块，用于在初始获得的待处理样本集中样本数量未达到预设需求的情况下，将该待处理样本集输入生成式对抗网络，得到与待处理样本集同分布的生成样本集；

合并子模块，用于将待处理样本集与生成样本集合并，得到不确定标签样本集。

13.根据权利要求12所述的装置，所述学习模块具体用于：

将不确定标签样本集中，属于待处理样本集的部分输入所述预测模型。

14.根据权利要求12所述的装置，若所述待处理样本集中的样本均为有标签样本，则所述生成子模块具体用于：

将该待处理样本集的特征部分和标签部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，所述生成样本集中的样本均为有标签样本；

或者

将该待处理样本集的特征部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，所述生成样本集中的样本均为无标签样本。

15.根据权利要求12所述的装置，若所述待处理样本集中的样本均中的样本均为无标签样本、或仅部分样本带有标签，则生成子模块具体用于：

将该待处理样本集的特征部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，所述生成样本集中的样本均为无标签样本。

16.一种数据模型训练装置，该装置包括：

输入模块，用于获得不确定标签样本集和已确定标签样本集；

学习模块，用于对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

校验模块，用于根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

所述学习模块和所述校验模块相互配合实现迭代处理，直到学习标签的准确率达到预设的要求；

输出模块，用于在迭代结束后，将当前的预测模型确定为适用于所述不确定标签样本集的预测模型。

17.一种计算机设备，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8任一项所述的方法。

一种数据样本标签处理方法及装置

技术领域

[0001] 本说明书实施例涉及数据分析技术领域，尤其涉及一种数据样本标签处理方法及装置。

背景技术

[0002] 在机器学习领域，利用大量的数据样本进行训练，可以获得各种形式的数据模型以解决实际问题。机器学习可分为有监督学习和无监督学习，其中监督学习是利用一组已知类别的数据样本来调整预测模型的参数、使其达到性能要求的过程。监督学习使用的训练样本均为已标记样本，即每条样本同时包含“特征值”和“标签值”。

[0003] 有监督学习和无监督学习分别可以适用于一定的需求场景，然而在实际应用中，经常会遇到需要采用有监督学习解决问题、但是数据样本标签不准确甚至无标签的情况。理论上虽然可以采用人工的方式对分别对每条数据样本的标签进行纠正或重新标注，然而在大数据的应用场景下，这种纯人工的处理方式是不现实的。因此，如何对不确定标签样本实现高效、准确的标注，已经成为行业内备受关注的问题。

发明内容

[0004] 针对上述技术问题，本说明书实施例提供一种数据样本标签处理方法及装置，技术方案如下：

[0005] 根据本说明书实施例的第一方面，提供一种数据样本标签处理方法，该方法包括：

[0006] 获得不确定标签样本集和已确定标签样本集；利用以下步骤进行迭代处理，直到学习标签的准确率达到预设的要求：

[0007] 对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；

[0008] 将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

[0009] 根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

[0010] 迭代结束后，将不确定标签样本当前的学习标签确定为标注结果。

[0011] 根据本说明书实施例的第二方面，提供一种数据模型训练方法，该方法包括：

[0012] 获得不确定标签样本集和已确定标签样本集；利用以下步骤进行迭代处理，直到学习标签的准确率达到预设的要求：

[0013] 对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；

[0014] 将所述不确定标签样本集中的不确定标签样本输入所述预测模型，将模型输出的

预测值确定为不确定标签样本的学习标签；

[0015] 对带有学习标签的不确定标签样本集进行抽样得到抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移动至已确定标签样本集；

[0016] 迭代结束后，将当前的预测模型确定为适用于所述不确定标签样本集的预测模型。

[0017] 根据本说明书实施例的第三方面，提供一种数据样本标签处理装置，该装置包括：

[0018] 输入模块，用于获得不确定标签样本集和已确定标签样本集；

[0019] 学习模块，用于对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

[0020] 校验模块，用于根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

[0021] 所述学习模块和所述校验模块相互配合实现迭代处理，直到学习标签的准确率达到预设的要求；

[0022] 输出模块，用于在迭代结束后，将不确定标签样本当前的学习标签确定为标注结果。

[0023] 根据本说明书实施例的第四方面，提供一种数据模型训练装置，该装置包括：

[0024] 输入模块，用于获得不确定标签样本集和已确定标签样本集；

[0025] 学习模块，用于对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

[0026] 校验模块，用于根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

[0027] 所述学习模块和所述校验模块相互配合实现迭代处理，直到学习标签的准确率达到预设的要求；

[0028] 输出模块，用于在迭代结束后，将当前的预测模型确定为适用于所述不确定标签样本集的预测模型。

[0029] 本说明书实施例所提供的技术方案，在已拥有大量已确定标签样本集的情况下，首先通过对已知标签样本信息的学习，对不确定标签样本的标签进行初步标注，以得到不确定标签样本的学习标签，然后对学习标签进行抽样校验，并将校验后的结果反馈至学习阶段，使其利用校验后的结果重新进行学习。通过上述方式来不断改善学习结果，直到满足需求。应用上述方案，不仅可以在仅付出少量校验成本的情况下，实现对不确定标签的标注或纠正，还能够针对不确定标签样本的自有特征，得到可适用于不确定标签样本的预测模

型。

[0030] 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。

[0031] 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。

附图说明

[0032] 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。

[0033] 图1是本说明书实施例的数据样本标签处理方法的流程示意图；

[0034] 图2是本说明书实施例的整体设计架构示意图；

[0035] 图3是本说明书实施例的不确定标签样本生成示意图；

[0036] 图4是本说明书实施例的数据模型训练方法的流程示意图；

[0037] 图5是本说明书实施例的标签处理装置及训练装置的结构示意图；

[0038] 图6是用于配置本说明书实施例装置的一种设备的结构示意图。

具体实施方式

[0039] 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。

[0040] 在对数据进行分类预测等应用场景(例如垃圾信息识别、欺诈行为识别等)，理想的情况是所有的数据样本都带有标签，从而通过有监督学习得到预测模型。虽然对数据样本进行标记的代价较高，但是为了得到性能较好的预测模型，在实现需求的初期也不可避免地要投入成本，以得到数量足够的有标签样本。

[0041] 当训练得到的模型性能达到需求后，就可以投入使用以对未知数据进行分类预测了。但是实际应用中，可能会出现的情况是：由于业务领域、应用场景等方面的差别，导致未知数据与训练模型时所使用数据样本的特征并不完全相同，因此使用已有的模型对这些未知数据进行预测时，经常达不到期望的效果。例如，对于“识别垃圾信息”这一需求，根据电商平台的交易评论内容训练出的识别模型，如果用来识别视频网站评论内容中的垃圾信息，很有可能会出现大量的识别错误。

[0042] 另一种类似的问题是：给定一批“新类型”的数据，希望能够据此训练出适用于这批数据的识别模型。然而这批数据可能是完全不带有任何标签，也可能是全部或部分带有标签、但是无法保证这些标签的准确性(例如可能是粗糙标注等原因导致)。例如，给定一批视频网站中的评论内容，希望训练出适用于网站中垃圾评论内容识别的模型。那么，根据现有技术，如果要满足模型训练的要求，又需要花费大量成本对这批新类型数据进行标注。

[0043] 针对上述需求，本说明书实施例提供一种数据样本标签处理方案，该方案的设计架构如图1所示，具体架构说明如下：

[0044] 1)输入部分：

[0045] 以“不确定标签样本集”和“已确定标签样本集”作为输入信息。

[0046] 其中“不确定标签样本集”中的样本可能是完全不带有任何标签，也可能是全部或部分带有标签、但是无法保证这些标签的准确性。

[0047] “已确定标签样本集”指当前已拥有的、携带可靠标签的样本集合。具体的标签来源这里不做限定，总之可以将这部分视为已知信息。

[0048] 实际情况中，“不确定标签样本集”和“已确定标签样本集”的整体使用需求相同，但又具有不同的数据特征。例如前面例子中的“视频网站评论内容”和“交易平台评论内容”，都是用于识别垃圾评论内容，但是不同领域的评论内容又各有特色。则前者相当于“未确定标签样本集”，后者相当于“已确定标签样本集”。

[0049] 为便于描述，后续将分别以符号“U”和“L”表示“不确定标签样本集”和“已确定标签样本集”。

[0050] 2)输出部分：

[0051] 以“U的标注结果”和“适用于U的预测模型”作为输出信息，当然在实际应用中，可能只需要上述两种输出信息中的一种。

[0052] 3)学习模块：

[0053] 学习模块又可以进一步分为训练和标注两个子模块(图中未示出)：

[0054] a训练子模块：

[0055] 以U和L的合并结果作为依据，通过训练得到预测模型。其中根据U的具体情况不同，将采用不同训练方式：

[0056] 如果U中样本均带有标签，则对U和L进行合并后，使用有监督学习的方式训练得到预测模型。

[0057] 如果U中样本完全不带有任何标签、或者仅部分样本带有标签，则对U和L进行合并后，使用半监督学习的方式训练得到预测模型。

[0058] 可见，无论采用哪种训练方式，由于训练数据覆盖了两类数据的不同特征，因此训练出的模型都能够适用于两种不同数据类型的预测。

[0059] b标注子模块：

[0060] 利用训练子模块训练出的模型，对U中的数据进行预测，将预测结果定义为“学习标签”。

[0061] 4)校验模块：

[0062] 由于学习模块的训练过程中，使用了大量了“不确定标签样本”，因此初期训练得到的模型效果很一般是不理想的(除非U中有大量标注结果正确的样本、而且这个结论是预先已知的，但是这种情况也就没有必要使用合并的样本进行训练了)，因此需要对学习标签进行校验。

[0063] 为保证校验结果的可靠性，这里可以采用人工参与的方式进行校验，校验模块可提供样本数据、标注结果等信息的显示功能，并且提供标注或纠错等操作接口，以方便相关人员进行校验。当然，在能够保证校验结果可靠性的前提下，还也可以采用其他方式实现校验，本申请对此并不进行限定。

[0064] 值得说明的是，由于校验模块的功能仅是从整体上评估标注结果是否理想，因此

这里并不需要对所有的学习标签都进行校验，只需对少量抽样数据进行校验即可，从而实现校验代价的节省。

[0065] 如果评估结果不理想，则需要触发新一轮的标签学习。另一方面，从“不确定标签样本”中抽样出的数据，经过标签校验后，就可以当作“确定标签样本”使用，因此将校验结果反馈给标签学习系统后，能够令每次标签学习的准确率不断趋于优化。两个模块通过上述方式配合，可以进行多次再学习，直到校验结果满足需求。

[0066] 基于上述设计方案，本说明书进一步提供相应的数据样本标签处理方法，参见图2所示，该方法可以包括以下步骤：

[0067] S101，获得不确定标签样本集和已确定标签样本集；

[0068] 为描述方便，在本实施例中，仍以符号“U”和“L”表示“不确定标签样本集”和“已确定标签样本集”。

[0069] 如前所述，U中的样本可能是完全不带有任何标签，也可能是全部或部分带有标签、但是无法保证这些标签的准确性。而L则指代当前已拥有的、携带可靠标签的样本集合。U和L的整体使用需求相同，但又具有不同的数据特征。

[0070] 根据前面的描述可知，训练模型时，采用U和L的合并结果作为训练样本，而模型训练的一个重要需求是：使得模型能够适用于U和L两种不同的数据类型的预测，这就要求U和L都要达到一定的样本数量，而且U和L的比例相差不能过于悬殊。由于单独利用L已经能够单独训练出性能满足需求的模型，因此这里可以认为L中样本的绝对数量是足够的；但是U的样本数量则具有很大的不确定性，如果U中的样本数量过少，则在无法在训练过程中提供足够的U的数据特征，进而导致训练出的模型对无法更好地适应对U类数据的预测。

[0071] 如果U中的样本数量不足，则可以使用GAN(Generative Adversarial Networks，生成式对抗网络)，模拟U的情况再生成一部分样本。

[0072] GAN是一种可以根据已有的真实样本构建出新样本的技术，GAN由生成模型(generative model)和判别模型(discriminative model)组成。生成模型的功能是捕捉已有样本数据的分布，用服从某一分布(例如均匀分布，高斯分布等)的噪声生成类似真实样本据的新样本，追求效果是越像真实样本越好；判别模型是一个二分类器，用于判断一个样本是真实样本还是生成样本。

[0073] 在GAN的训练过程中固定一方，更新另一方的网络权重，交替迭代，在这个过程中，生成模型和判别模型双方都极力优化自己的网络，从而形成竞争对抗，直到双方达到一个动态的平衡，此时生成模型恢复了训练数据的分布(造出了和真实样本一模一样的样本)，判别模型也无法再判断出是真实样本还是生成样本。

[0074] 因此，假设初始给定的待处理样本(本说明书中以U_0表示)数量无法满足训练需求，则可以将U_0输入GAN，由GAN输出与U_0同分布的生成样本集(本说明书中以U_G表示)；然后将U_0与U_G进行合并，如图3所示，即有以下关系：

[0075] $U=U_0+U_G$

[0076] 可以理解的是，本说明书中的“同分布”，并不是严格数学意义上的同分布，而是GAN所模拟出的同分布。

[0077] 由于GAN既可以生成有标签样本，也可以生成无标签样本，那么可以根据U_0的不同情况，采用不同的样本生成方式：

[0078] 如果U_0中的样本均为有标签样本，则可以有两种处理方式：

[0079] 1)将U_0的特征部分和标签部分均输入GAN，得到带有标签的U_G，这种情况下，U中的样本也均为有标签样本。

[0080] 2)仅将U_0的特征部分输入GAN，得到不带标签的U_G，这种情况下，U中仅部分样本带有标签。

[0081] 如果U_0中的样本中的样本均为无标签样本、或仅部分样本带有标签，则可以将U_0的特征部分输入GAN，得到不带标签的U_G，这种情况下，U中样本的标签携带情况与U_0一致。

[0082] 需要生成U_G的样本数量可以根据训练需求确定，这里的训练既包括对样本绝对数量的需求、也包括对样本相对数量的需求。一般而言，希望U与L的比例不低于1:4，当然U也可以比L更大，该比例需求可以根据实际情况设计，本申请对此不需要进行限定。

[0083] S102，对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；

[0084] 使用合并后的样本集S(其中S=U+L)进行模型训练，由于L中的样本都是带有标签的，因此根据U中样本的标签携带情况将决定S中样本的标签携带情况。

[0085] 如果U中样本均带有标签，则S中样本也均带有标签，此时训练需求转化为有监督学习问题。

[0086] 如果U中样本均不带有标签、或部分带有标签，则S中样本也是部分带有标签的，此时训练需求转化为半监督学习问题。半监督学习是也一种利用部分有标签样本以及部分无标签样本实现模型训练的技术，值得注意的是，现有技术中，半监督学习所针对的有标签样本和无标签样本是同一类型的数据样本，而本申请中U和L是两套并不完全一致的样本，因此严格意义上讲与半监督学习的传统应用场景有所区别。由于其整体需求相同，因此在算法层面仍然可以使用半监督学习算法，但是其训练结果需要配合后续的校验步骤多次调整才能满足应用需求。

[0087] 根据具体的应用场景不同，可以选用不同形式的模型以及相应的学习算法，对于本说明书对此并不限定。例如，对于文本识别应用，可以通过构建基于RNN(Recurrent neural Network，循环神经网络)深度学习模型训练文本数据。

[0088] S103，将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

[0089] 对于S102所产出的预测模型，可以将U中的样本输入该模型，在本说明书中，将模型输出的预测值称为样本的“学习标签”，值得注意的是，该学习标签与U中样本是否带有标签或标签是否准确并无必然关联。

[0090] 需要说明的是，这里的“不确定标签样本”既可以是U中的全部样本，也可用是U中样本的一部分。

[0091] 例如，如果在S101采用了GAN生成新样本，则本步骤中可以仅将当前U中属于U_0的那部分样本输入预测模型。这样处理的原因是，相对于U_G而言，U_0才是真实的数据，后续对这部分数据进行校验的意义更大，而且从“标注”需求而言，也只有U_0才是真正需要进行标注处理的对象。这里需要明确的是“属于U_0的那部分样本”并不等同于U_0，这是因为随着整个方案的迭代，U的规模是逐步缩减的，相应地“属于U_0的那部分样本”也会逐步变小。

[0092] 另外，在本步骤中也可以对U(或U_0)进行抽样，仅将抽样结果输入预测模型，从而

得到抽样结果的学习标签。抽样的目的是降低校验的代价，可以在本步骤实现，也可以在后续步骤中实现。

[0093] S104，根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率；

[0094] 如果在S103中未作抽样处理，则在本步骤中，对所有带有学习标签的样本进行抽样，得到抽样子集；如果在S103中已作抽样处理，则在本步骤中可以直接使用S103输出的结果作为抽样子集。

[0095] 抽样的数量可以根据实际需求确定，一般综合考虑以下几个因素：

[0096] 1)是否有足够的代表性：抽样数量越大，代表性越好。

[0097] 2)对校验代价的影响：抽样数量越小，校验代价越低。

[0098] 3)对迭代速度的影响：抽样数量越大，则每次校验后反馈给下一次学习的有用信息越多，相应也会提高整体方案的迭代速度。

[0099] 在实际应用中，也可以在迭代过程中使用动态的抽样率，例如随着迭代的进行，模型性能逐渐趋于稳定，可以逐步降低抽样率。当然，本领域技术人员可以根据实际需求设计抽样方案，本说明书对此不做限定。

[0100] S105，判断学习标签的准确率是否达到预设的要求，如果是则继续执行S107，否则执行S106后返回S102；

[0101] S106，将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

[0102] S107，将不确定标签样本当前的学习标签确定为标注结果。

[0103] 假设以U_s表示抽样子集，如果U_s的学习标签准确率达到某个指标(例如95％)，则认为整个U(或U_0)的学习标签是可信的，可以直接作为标注结果输出，否则需要触发下一轮学习。

[0104] 经过校验的U_s相当于已经具备了可靠的标签，因此在下一轮学习中，U_s可以作为“确定标签样本”使用，即在每次迭代之前，对U和L做以下更新：

[0105] $U=U-U_s$

[0106] $L=L+U_s$

[0107] 通过S102～S106的迭代处理，由于具备可靠标签的样本逐步增加，而且这些样本是具有“新类型数据”特征的样本，因此能够令每次标签学习的准确率不断趋于优化，并且逐步适应“新类型数据”的预测，最终达到应用需求。

[0108] 可以理解的是，迭代停止后的U已经和初始的U不同，因此最终的输的标注结果应包括“当前U中样本最新的学习标签”以及“历次迭代过程中已经过校验确认可靠的标签”

[0109] 此外，迭代停止后，最终的预测模型也可以作为另一项输出信息，如图4的S108所示(其他步骤与图2所示一致，这里不再重复说明)，该模型都能够适用于U和L两种不同数据类型。

[0110] 应用本说明书所提供的方案，对于已拥有海量数据以及较成熟预测模型的企业而言，能够有效将自身的预测能力向外部输出，为客户或合作伙伴提供技术服务，同时也能够不断丰富自身模型的预测能力。

[0111] 例如在电商平台及支付平台的风控系统中，已经具有强大的文本识别能力，可以

从用户生成内容中识别出灌水、广告、暴恐政和黄赌毒等信息。一些其他行业的外部商户也具有类似的需求，例如微博、视频、直播等UGC(User Generated Content，用户生成内容)相关领域，如果这些外部商户没有能力对用户生成内容样本进行准确标注，则可以基于本说明书所提供的技术方案，结合电商平台及支付平台自身已有的垃圾文本数据以及识别模型，对其他行业提供的样本数据进行学习。相对于完全人工对外部样本进行标注或纠正的方式而言效率更高，更容易实现规模化。

[0112] 相应于上述方法实施例，本说明书实施例还提供一种数据样本标签处理装置或数据模型训练装置，参见图5所示，该装置可以包括：

[0113] 输入模块110，用于获得不确定标签样本集和已确定标签样本集；

[0114] 学习模块120，用于对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

[0115] 校验模块130，用于根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

[0116] 学习模块120和校验模块130相互配合实现迭代处理，直到学习标签的准确率达到预设的要求；

[0117] 输出模块140，可以用于在迭代结束后，将不确定标签样本当前的学习标签确定为标注结果。也可以用于在迭代结束后，将当前的预测模型确定为适用于不确定标签样本集的预测模型。

[0118] 在本说明书提供的一种具体实施方式中，若不确定标签样本集中的样本均为有标签样本；则学习模块120可以具体用于：利用有监督学习算法对合并后的样本进行训练，得到预测模型。

[0119] 若不确定标签样本集中的样本均为无标签样本、或仅部分样本带有标签；则学习模块120可以具体用于：利用半监督学习算法对合并后的样本进行训练，得到预测模型。

[0120] 在本说明书提供的一种具体实施方式中，输入模块110可以包括：

[0121] 生成子模块，用于在初始获得的待处理样本集中样本数量未达到预设需求的情况下，将该待处理样本集输入生成式对抗网络，得到与待处理样本集同分布的生成样本集；

[0122] 合并子模块，用于将待处理样本集与生成样本集合并，得到不确定标签样本集。

[0123] 在本说明书提供的一种具体实施方式中，学习模块120可以具体用于：

[0124] 将不确定标签样本集中，属于待处理样本集的部分输入预测模型。

[0125] 在本说明书提供的一种具体实施方式中，若待处理样本集中的样本均为有标签样本，则生成子模块130可以具体用于：

[0126] 将该待处理样本集的特征部分和标签部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，生成样本集中的样本均为有标签样本；

[0127] 或者

[0128] 将该待处理样本集的特征部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，生成样本集中的样本均为无标签样本。

[0129] 若待处理样本集中的样本均中的样本均为无标签样本、或仅部分样本带有标签，则生成子模块130可以具体用于：

[0130] 将该待处理样本集的特征部分输入生成式对抗网络，得到与待处理样本集同分布的生成样本集，生成样本集中的样本均为无标签样本。

[0131] 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前述的数据样本标签处理方法数据模型训练方法。该方法至少包括：

[0132] 获得不确定标签样本集和已确定标签样本集；利用以下步骤进行迭代处理，直到学习标签的准确率达到预设的要求：

[0133] 对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型；

[0134] 将不确定标签样本集中的不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签；

[0135] 根据当前的不确定标签样本集，获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率，如果学习标签的准确率未达到预设的要求，则将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集；

[0136] 迭代结束后，将不确定标签样本当前的学习标签确定为标注结果、或者将当前的预测模型确定为适用于所述不确定标签样本集的预测模型。

[0137] 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。

[0138] 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。

[0139] 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。

[0140] 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。

[0141] 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。

[0142] 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。

[0143] 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。

[0144] 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前述的数据样本标签处理方法数据模型训练方法。该方法至少包括:

[0145] 获得不确定标签样本集和已确定标签样本集;利用以下步骤进行迭代处理,直到学习标签的准确率达到预设的要求:

[0146] 对当前的不确定标签样本集和已确定标签样本集进行合并,利用合并后的样本训练得到预测模型;

[0147] 将不确定标签样本集中的不确定标签样本输入预测模型,将模型输出的预测值确定为不确定标签样本的学习标签;

[0148] 根据当前的不确定标签样本集,获得带有学习标签的抽样子集,对抽样子集的学习标签进行校验以获得学习标签的准确率,如果学习标签的准确率未达到预设的要求,则将标签校验结果确定为抽样子集的标注结果,并将抽样子集从不确定标签样本集移至已确定标签样本集;

[0149] 迭代结束后,将不确定标签样本当前的学习标签确定为标注结果、或者将当前的预测模型确定为适用于所述不确定标签样本集的预测模型。

[0150] 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。

[0151] 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。

[0152] 上述实施例阐明的系统、装置、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。

[0153] 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，对于装置实施例而言，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。以上所描述的装置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。

[0154] 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，这些改进和润饰也应视为本说明书实施例的保护范围。

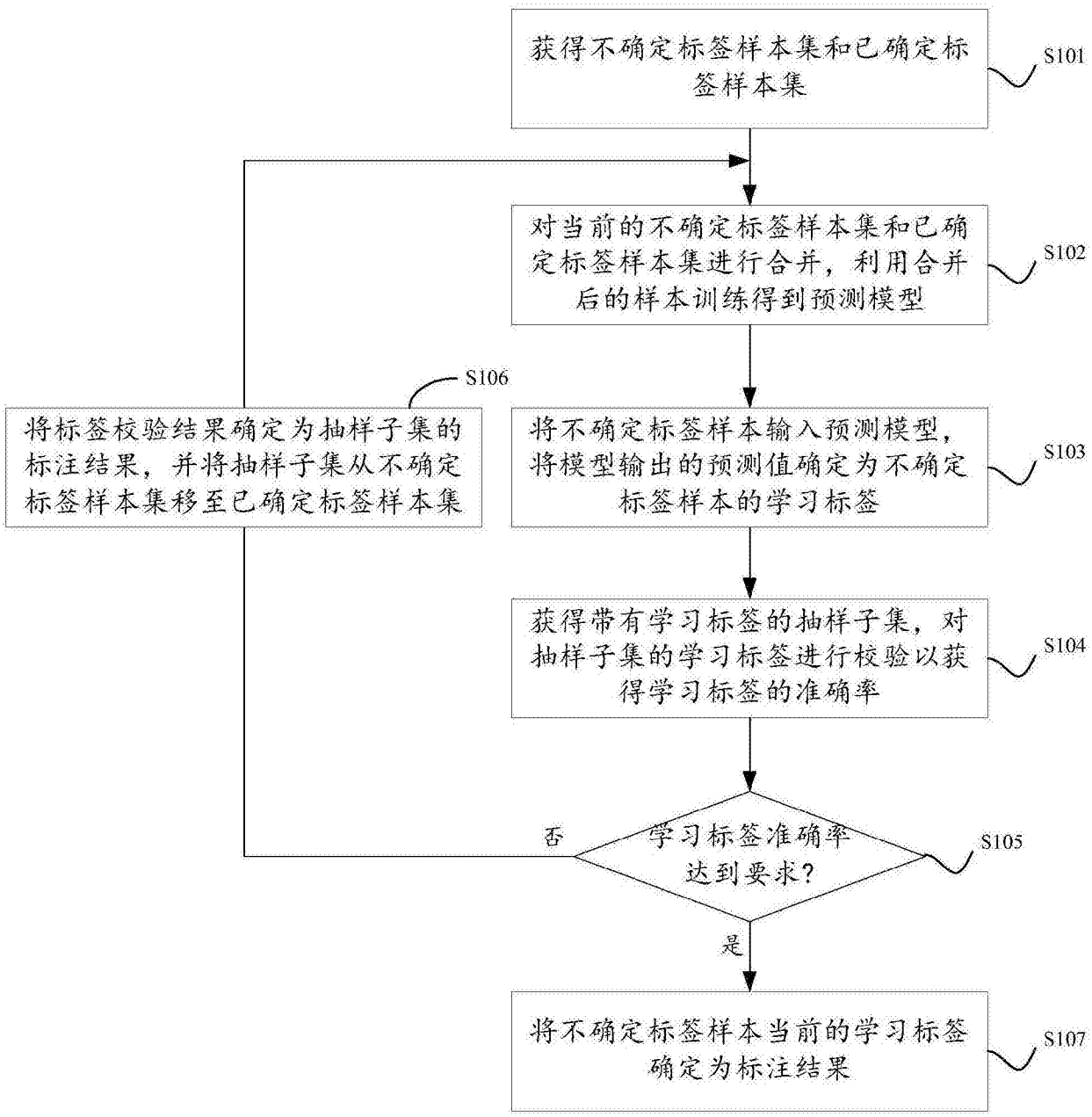

获得不确定标签样本集和已确定标签样本集
S101
对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型
S102
将不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签
S103
获得带有学习标签的抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率
S104
学习标签准确率达到要求?
S105
否
是
将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集
S106
将不确定标签样本当前的学习标签确定为标注结果
S107

图1

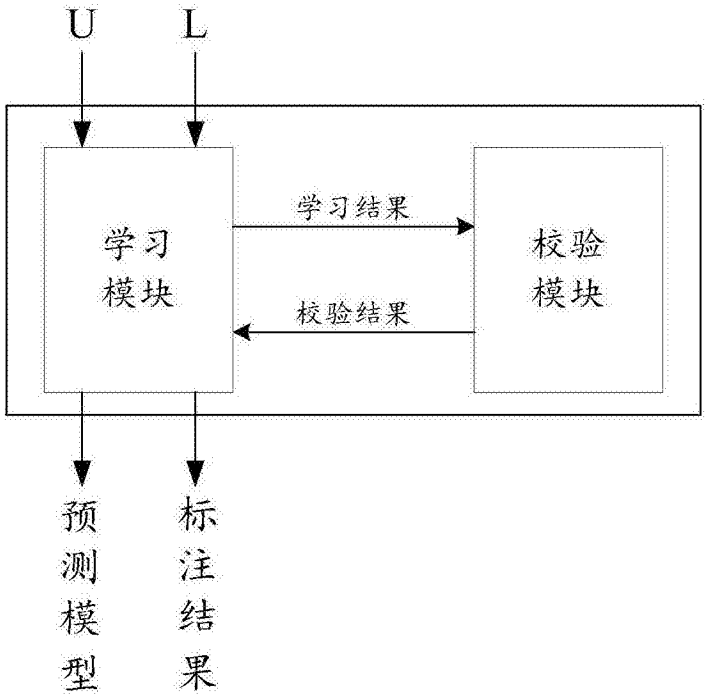


图2

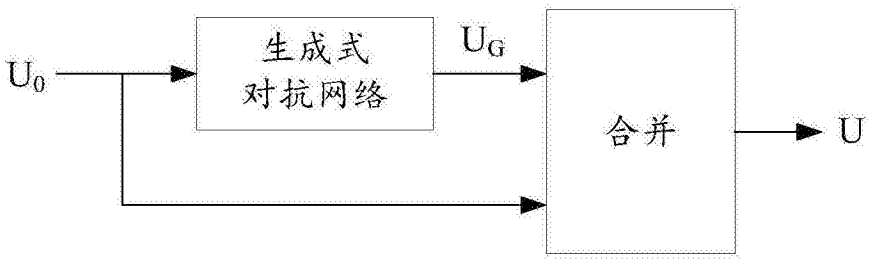


图3

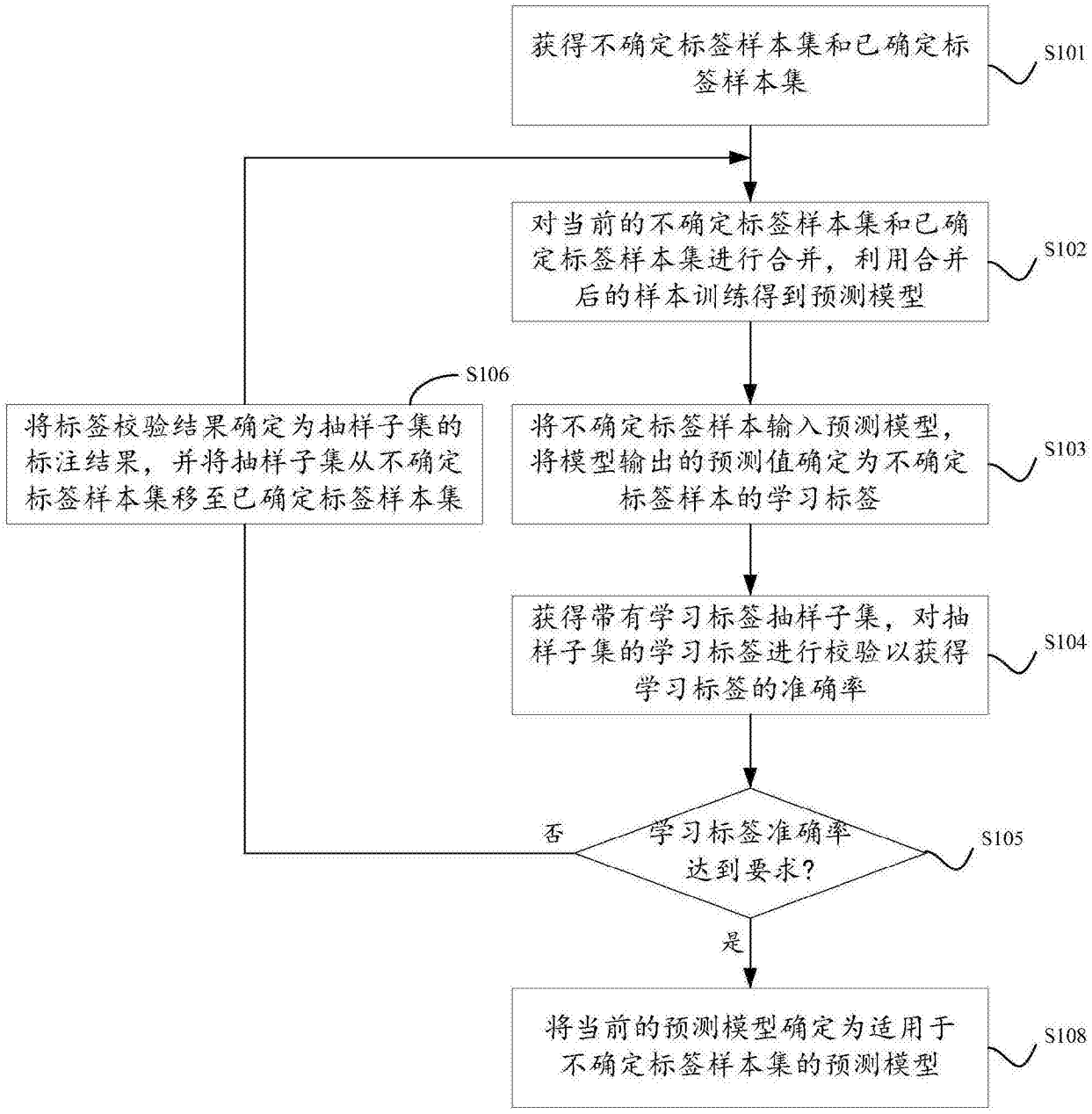

获得不确定标签样本集和已确定标签样本集
S101
对当前的不确定标签样本集和已确定标签样本集进行合并，利用合并后的样本训练得到预测模型
S102
将不确定标签样本输入预测模型，将模型输出的预测值确定为不确定标签样本的学习标签
S103
获得带有学习标签抽样子集，对抽样子集的学习标签进行校验以获得学习标签的准确率
S104
学习标签准确率达到要求?
S105
否
是
将标签校验结果确定为抽样子集的标注结果，并将抽样子集从不确定标签样本集移至已确定标签样本集
S106
将当前的预测模型确定为适用于不确定标签样本集的预测模型
S108

图4

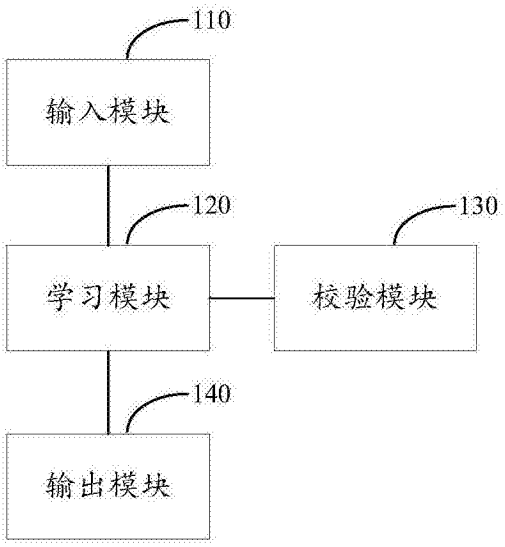

110
输入模块
120
130
学习模块
校验模块
140
输出模块

图5

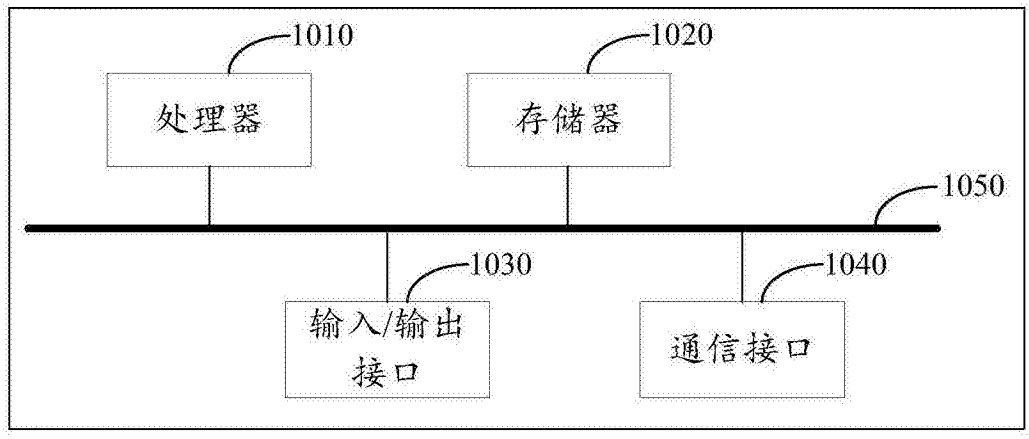

1010
处理器
1020
存储器
1050
1030
输入/输出
接口
1040
通信接口

图6