

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 5 月 14 日 (14.05.2020)



(10) 国际公布号
WO 2020/093718 A1

(51) 国际专利分类号:
G06F 16/26 (2019.01)

(21) 国际申请号: PCT/CN2019/094741

(22) 国际申请日: 2019 年 7 月 4 日 (04.07.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811327417.3 2018 年 11 月 8 日 (08.11.2018) CN

(71) 申请人: 北京字节跳动网络技术有限公司 (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市石景山区实兴大街 30 号院 3 号楼 2 层 B-0035 房间, Beijing 100041 (CN)。

(72) 发明人: 李伟健(LI, Weijian); 中国北京市海淀区知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100190 (CN)。王长虎(WANG, Changhu); 中国北京市海淀区知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100190 (CN)。

(74) 代理人: 北京市立方律师事务所 (LIFANG & PARTNERS); 中国北京市东城区东四十条甲 22 号南新仓商务大厦 A 座 1105 室, Beijing 100088 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: TRAINING DATA RE-SAMPLING METHOD AND APPARATUS, AND STORAGE MEDIUM AND ELECTRONIC DEVICE

(54) 发明名称: 训练数据重采样方法、装置、存储介质及电子设备

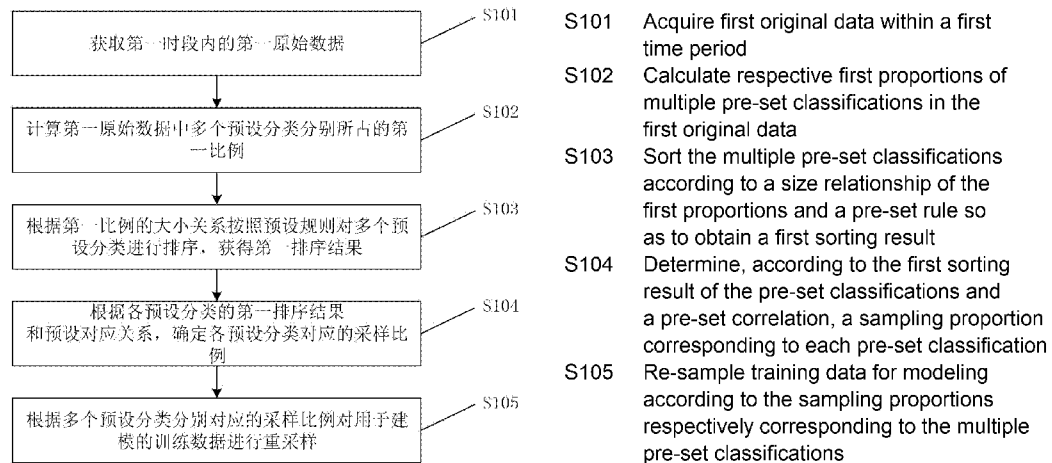


图 1

(57) Abstract: Disclosed are a training data re-sampling method and apparatus, and a storage medium and an electronic device. The method comprises: acquiring first original data within a first time period (S101); calculating respective first proportions of multiple pre-set classifications in the first original data (S102); sorting the multiple pre-set classifications according to a size relationship of the first proportions and a pre-set rule so as to obtain a first sorting result (S103); determining, according to the first sorting result of the pre-set classifications and a pre-set correlation, a sampling proportion corresponding to each pre-set classification (S104), wherein the pre-set correlation is a correlation between the first sorting result and the sampling proportion; and re-sampling training data for

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

— 发明人资格 (细则4.17(iv))

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

modeling according to the sampling proportions respectively corresponding to the multiple pre-set classifications (S105). The problem of a classification model being unfriendly to a small category can be solved, and the classification accuracy of a classification model, obtained through training by means of training data, for different applications can be improved, thereby improving user experience.

(57) 摘要: 一种训练数据重采样方法、装置、存储介质及电子设备。该方法包括: 获取第一时段内的第一原始数据(S101); 计算第一原始数据中多个预设分类分别所占的第一比例(S102); 根据第一比例的大小关系按照预设规则对多个预设分类进行排序, 获得第一排序结果(S103); 根据各预设分类的第一排序结果和预设对应关系, 确定各预设分类对应的采样比例(S104), 所述预设对应关系为所述第一排序结果与所述采样比例之间的对应关系; 根据多个预设分类分别对应的采样比例对用于建模的训练数据进行重采样(S105)。可以解决分类模型对小类别不友好的问题, 提高通过该训练数据训练得到的分类模型针对不同应用的分类准确性, 从而提高用户体验。

训练数据重采样方法、装置、存储介质及电子设备

相关申请的交叉引用

5 本申请要求 2018 年 11 月 8 日在中国知识产权局提交的中国专利申请 No. 201811327417.3 的优先权，通过引用将该中国专利申请公开的全部内容并入本文。

技术领域

10 本公开涉及数据挖掘领域，具体地，涉及一种训练数据重采样方法、装置、存储介质及电子设备。

背景技术

在机器学习中，训练数据中的针对分类模型中不同分类的样本数目可能
15 出现相差巨大的情况，例如，在 N 个训练数据中，属于第一类的样本数目可能与属于第二类的样本数目以及属于第三类的样本数目等都相差巨大（例如属于第一类的样本数目可能占到 N 个训练数据中的 90%，属于第二类 and 第三类的样本数目可能一共才占到 N 个训练数据中的 10%），这样直接用样本数目不均衡的训练数据去对分类模型进行训练时，机器学习算法倾向于产生不太
20 令人满意的分类模型，例如可能会造成分类模型对于训练数据中样本数目较少的分类欠拟合、对训练模型中样本数量较多的分类过拟合的情况，实际上，如果不均衡比例超过 4:1，分类模型就会偏向于大的类别而忽略小的类别。因此，用没有处理过的不均衡的训练数据训练出的分类模型可能对实际数据的分类效果并不理想。目前，针对训练数据不均衡的情况，通常会采用对训练数
25 据进行重采样的方法。

发明内容

本公开的目的是提供一种训练数据重采样方法、装置、存储介质及电子设备

备，能够针对训练数据不均衡的情况，根据实际的原始数据中不同分类所占的比例来对训练数据进行重采样处理，从而解决分类模型对小类别不友好的问题。

为了实现上述目的，本公开提供一种训练数据重采样方法，所述方法包括：

5 获取第一时段内的第一原始数据；

计算所述第一原始数据中多个预设分类分别所占的第一比例；

根据所述第一比例的大小关系按照预设规则对所述多个预设分类进行排序，获得第一排序结果；

10 根据各预设分类的第一排序结果和预设对应关系，确定各预设分类对应的采样比例，所述预设对应关系为所述第一排序结果与所述采样比例之间的对应关系；

根据所述多个预设分类分别对应的所述采样比例对用于建模的训练数据进行重采样。

可选地，所述方法还包括：在获得所述第一排序结果之后，

15 获取第二时段内的第二原始数据；

计算所述第二原始数据中所述多个预设分类分别所占的第二比例；

根据所述第二比例的大小关系按照所述预设规则对所述多个预设分类进行排序，获得第二排序结果；

20 在所述第一排序结果和所述第二排序结果一致时，执行所述确定各预设分类对应的采样比例的步骤。

可选地，在所述第一排序结果和所述第二排序结果不一致时，重新确定所述第二时段，并将所述第二排序结果确定为第一排序结果；返回所述获取第二时段内的第二原始数据的步骤。

25 可选地，当至少两个预设分类的比例相同时，按照所述至少两个预设分类的优先级确定所述至少两个预设分类的排序。

本公开还提供一种训练数据重采样装置，所述装置包括：

第一获取模块，用于获取第一时段内的第一原始数据；

第一计算模块，用于计算所述第一原始数据中多个预设分类分别所占的第一比例；

第一排序模块，用于根据所述第一比例的大小关系按照预设规则对所述多个预设分类进行排序，获得第一排序结果；

比例获取模块，用于根据各预设分类的所述第一排序结果和预设对应关系，确定各预设分类对应的采样比例，所述预设对应关系为所述第一排序结果与
5 与所述采样比例之间的对应关系；

重采样模块，用于根据所述多个预设分类分别对应的所述采样比例对用于建模的训练数据进行重采样。

可选地，所述装置还包括：

第二获取模块，用于在获得所述第一排序结果之后，获取第二时段内的第
10 二原始数据；

第二计算模块，用于计算所述第二原始数据中所述多个预设分类分别所占的第二比例；

第二排序模块，用于根据所述第二比例的大小关系按照所述预设规则对所述多个预设分类进行排序，获得第二排序结果；

15 排名比较模块，用于在所述第一排序结果和所述第二排序结果一致时，触发所述比例获取模块确定各预设分类对应的所述采样比例。

可选地，所述排名比较模块还被配置成：

在所述第一排序结果和所述第二排序结果不一致时，重新确定所述第二时段，并将所述第二排序结果确定为第一排序结果；

20 触发所述第二获取模块获取第二时段内的第二原始数据。

可选地，当至少两个预设分类的比例相同时，按照所述至少两个预设分类的优先级确定所述至少两个预设分类的排序。

本公开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上述方法。

25 本公开还提供一种电子设备，其包括：

存储器，其上存储有计算机程序；

处理器，用于执行所述存储器中的所述计算机程序，以实现上述方法。

根据本公开的实施例，能够针对训练数据不均衡的情况，根据实际的原始数据中不同分类所占的比例来对训练数据进行重采样处理，从而解决分类模

型对小类别不友好的问题，提高通过该训练数据训练得到的分类模型针对不同应用的分类准确性，从而提高用户体验。

附图说明

5 附图是用来帮助对本公开的进一步理解，并且构成说明书的一部分，与下面的具体实施方式一起用于解释本公开，但并不构成对本公开的限制。在附图中：

图 1 是根据本公开的示例性实施例的一种训练数据重采样方法的流程图。

图 2 是根据本公开的示例性实施例的又一训练数据重采样方法的流程图。

10 图 3 是根据本公开的示例性实施例的一种训练数据重采样装置的结构框图。

图 4 是根据本公开的示例性实施例的又一训练数据重采样装置的结构框图。

图 5 是根据本公开的示例性实施例的一种电子设备的框图。

15

具体实施方式

以下结合附图对本公开的具体实施方式进行详细说明。应当理解的是，此处所描述的具体实施方式仅用于说明和解释本公开，并不用于限制本公开。

20 由于在建立分类模型时经常会遇到训练数据不均衡的情况，而根据分类不均衡的训练数据训练出来的分类模型经常会偏向占比较多的类别，对占比较多的类别可能会出现过拟合，而对占比较少的类别可能会出现欠拟合的情况，因此，本领域技术人员在建立分类模型时，通常会制定诸如提升分类算法或平衡训练数据的类（数据预处理）的策略，其中，后者因为应用范围广泛而更常使用，即对不均衡的训练数据进行重采样的处理。

25 根据重采样的原理，通常将重采样分为过采样（over-sampling）和欠采样（under-sampling），过采样即通过增加训练数据中占比较少的类别的训练数据的数量，从而增加训练数据中少数类的代表性，欠采样即通过减少训练数据中占比较多的类别的训练数据的数量，从而减少训练数据中多数类的代表性，以此，来达到平衡训练数据中不同比例的类别的数据数量，解决训练数据不平衡

的问题。重采样是采用的方法通常会为最邻近法、双线性内插法以及三次卷积内插法这三种。

由此可见，根据不同领域对训练数据重采样需求的不同，相同的重采样方法对训练数据处理的效果也不尽相同。因此，如何根据不同领域的分类分布的实际情况来确定如何对不均衡训练数据进行重采样，是亟需解决的重要问题。

因此，本公开提出了一种训练数据重采样方法。

图1是根据本公开的示例性实施例的一种训练数据重采样方法的流程图，如图1所示，该方法可以包括步骤S101至步骤S105。

在步骤S101中，获取第一时段内的第一原始数据。原始数据是指分类模型所适用的实际应用中产生的实际数据，例如，对在线小视频进行分类的分类模型所应用的在线小视频应用中实际产生的数据，第一时段内的第一原始数据可以为例如在2018年10月15日0:00到2018年10月16日0:00之间用户实际上传的小视频数据。

在步骤S102中，计算第一原始数据中多个预设分类分别所占的第一比例。在获取到该第一原始数据之后，对该第一原始数据中多个预设分类分别所占的第一比例进行计算，该预设分类即为人为设定好的不同的分类，例如风景类、宠物类、舞蹈类、技术类、搞笑类等等。步骤S102中即可以得到在该第一时段内获取到的第一原始数据中不同预设分类分别所占的第一比例。

在步骤S103中，根据第一比例的大小关系按照预设规则对多个预设分类进行排序，获得第一排序结果（即，获得各个预设分类的排名）。

在步骤S104中，根据各预设分类的排名和预设对应关系，确定各预设分类对应的采样比例，该预设对应关系为排名与采样比例之间的对应关系。预设规则可以为从大到小，或者从小到大等，该预设规则应该与该预设对应关系互相配合，从而获取能够反映该第一原始数据中各个预设分类实际分布情况的采样比例。例如，在步骤S103中可以按照第一比例从大到小的顺序对该多个预设分类进行排序，该预设对应关系可以为例如表1所示：

表1

排名	1	2	3	4	5
采样比例	50%	20%	10%	10%	10%

其中，该预设分类不限于如表 1 中所示的 5 类，可以为其他任意个数。只要每个预设分类所对应的采样比例的总和为 100% 即可。由此，如表 1 所示，5 个预设分类就能够分别得到与之相对应的采样比例。

5 在步骤 S105 中，根据多个预设分类分别对应的采样比例对用于建模的训练数据进行重采样。在通过步骤 S101 至步骤 S104 得到第一时段的第一原始数据中不同预设分类分别对应的采样比例之后，就可以根据该采样比例对训练数据进行重采样了，本公开中对重采样采用的上采样还是下采样，以及具体采用何种采样方法不做限制，只要是根据图 1 中所示的步骤 S101 至步骤 S104
10 而得到的各个预设分类的采样比例进行重采样即可。

根据本公开的实施例，能够针对训练数据不均衡的情况，根据实际的原始数据中不同分类所占的比例来对训练数据进行重采样处理，从而解决分类模型对小类别不友好的问题，提高通过该训练数据训练得到的分类模型针对不同应用的分类准确性，从而提高用户体验。

15 图 2 是根据本公开的示例性实施例的又一训练数据重采样方法的流程图。如图 2 所示，该方法除了包括图 1 中所示的步骤 S101 至步骤 S105 之外，还可以包括步骤 S201 至步骤 S205。

在步骤 S201 中，获取第二时段内的第二原始数据。本公开中该第二时段与第一时段不相同，但允许第二时段和第一时段之间有重叠的时间，而且没有
20 固定的先后顺序关系，即第一时段在第二时段之前或之后都可，但在先的时段的结束时间与在后的时段的开始时间之间的间隔需要小于预设阈值。该预设阈值可以为例如 24 小时、48 小时等，这样能够避免第一时段与第二时段选取的间隔过大，所对应的第一原始数据和第二原始数据的数据内容变化过大，从而导致影响其中的各个预设分类所占比例的问题。其中，该第二原始数据中对
25 原始数据的定义与上述第一原始数据中对原始数据的定义相同。

在步骤 S202 中，计算第二原始数据中多个预设分类分别所占的第二比例。

在步骤 S203 中，根据第二比例的大小关系按照预设规则对多个预设分类进行排序，获得第二排序结果。上述步骤 S202 和步骤 S203 与图 1 中所示的步骤 S102 和步骤 S103 中的步骤相似，步骤 S102 和步骤 S103 是根据在第一

原始数据中的各个预设分类所占的比例确定各个预设分类所对应的采样比例，步骤 S202 和步骤 S203 是根据在第二原始数据中的各个预设分类所占的比例确定各个预设分类所对应的采样比例，分别得到了第一排序结果和第二排序结果。

5 在步骤 S204 中，对在步骤 S103 和步骤 S203 中分别得到的第一排序结果和第二排序结果进行比较。如果二者一致，则表示该第一排序结果和第二排序结果是准确的，可以用于确定各个预设分类的不同采样比例，因此转至步骤 S104 来确定个预设分类对应的采样比例，最后执行步骤 S105 来根据各个预设分类对应的采样比例来对用于建模的训练数据进行重采样，以解决由于训
10 练数据不均衡导致的分类模型效果的问题。而如果二者不一致，则转至步骤 S205。

 在步骤 S205 中，重新确定第二时段，并将第二排序结果确定为第一排序结果。其中，重新确定的第二时段，与在步骤 S201 中使用的第二时段不相同，重新确定的第二时段与在步骤 S201 中使用的第二时段之间的关系限定，与
15 在步骤 S201 中使用的第二时段与第一时段的关系相似，都允许二者之间有重叠的时间，而且没有固定的先后顺序关系，另外，重新确定的第二时段也不能与第一时段以及之前所有曾经被确定为第二时段的时段相同，这样，保证了每一次获取原始数据进行比例计算以及排序时，都能保证选取的内容能够真实反映实际应用中不同预设分类所占的比例，且不重复。在重新确定了第二时段
20 之后，将上述的第二排序结果确定为第一排序结果，并返回步骤 S201 重新获取第二排序结果，进而对新获取的第二排序结果与上一次获取的曾经是第二排序结果的第一排序结果进行比较，从而来决定是否能够根据一致的采样比例对训练数据重采样。

 综上，在本实施例中，先对第一时段的第一原始数据中各个预设分类所占
25 的比例得到第一排序结果，然后对第二时段的第二原始数据中的各个预设分类所占的比例得到第二排序结果，并对第一排序结果和第二排序结果进行比较。如果一致，则可以表明该排序结果无误差，可以直接根据该排名确定各个预设分类的采样比例，进而根据该采样比例对训练数据重采样；如果不一致，则表明二者中有至少一者是有误差的，从而需要重新选取新的时段中的新的

原始数据，并重新根据新的原始数据得到一个新的排序结果，并与最近一次获取到的排序结果进行比较，直到比较结果一致为止。当该比较结果经过多次比较之后达到一致时，此时就根据该一致的排序结果来确定各个预设分类对应的采样比例，并根据该采样比例对训练数据进行重采样。

5 根据本公开的实施例，通过至少两次对不同时段的不同原始数据中的各个预设分类所占的比例排名进行比较，就能够确定出一个能够真正反映原始数据中不同预设分类所占比例的排名，避免了由于所选取的时段比较特殊或者偶发性事件导致的所确定的原始数据中各个预设分类所占比例不能很好的反映实际原始数据中各个预设分类所占比例的情况，进而保证了根据该排名
10 确定的采样比例进行重采样时的准确性，从而使得根据重采样后的训练数据训练得到的分类模型的效果进一步提升。

在一种可能的实施方式中，当至少两个预设分类的比例相同时，按照至少两个预设分类的优先级确定至少两个预设分类的排序。在图 1 中所示的步骤 S103 和图 2 所示的步骤 S203 中根据各个原始数据中多个预设分类分别所占
15 的比例的大小关系按照预设规则进行排序时，如果出现两个预设分类所占的比例相同时，可以根据比例相同的预设分类的优先级来确定这些相同比例的预设分类的先后顺序。例如，令预设分类 A、预设分类 B、预设分类 C 在一个时段中的原始数据中所占的比例都为 2%，预设分类 A、预设分类 B、预设分类 C 的优先级为预设分类 A>预设分类 B>预设分类 C，那么在排序时，
20 如果按照从大到小的预设规则，那么应该按照预设分类 A 排在三者之首，其次是预设分类 B，最后是预设分类 C，如果是按照从小到大的规则，则与之正好相反，预设分类 C 应该排在最前，其次是预设分类 B，最后是预设分类 A。

图 3 是根据本公开的示例性实施例的一种训练数据重采样装置的结构框图，如图 3 所示，该装置可以包括：第一获取模块 10，用于获取第一时段内的
25 的第一原始数据；第一计算模块 20，用于计算第一原始数据中多个预设分类分别所占的第一比例；第一排序模块 30，用于根据第一比例的大小关系按照预设规则对多个预设分类进行排序，获得第一排序结果；比例获取模块 40，用于根据各预设分类的第一排序结果和预设对应关系，确定各预设分类对应的采样比例，该预设对应关系为第一排序结果与采样比例之间的对应关系；重

采样模块 50，用于根据多个预设分类分别对应的采样比例对用于建模的训练数据进行重采样。

图 4 是根据本公开的示例性实施例的又一训练数据重采样装置的结构框图。与图 3 所示的训练数据重采样装置相比，图 4 所示的装置还可以包括：

- 5 第二获取模块 60，用于在第一排序模块 30 获得第一排序结果之后，获取第二时段内的第二原始数据；第二计算模块 70，用于计算第二原始数据中多个预设分类分别所占的第二比例；第二排序模块 80，用于根据第二比例的大小关系按照预设规则对多个预设分类进行排序，获得第二排序结果；排名比较模块 90，用于在第一排序结果和第二排序结果一致时，触发比例获取模块 40 确定
10 各预设分类对应的采样比例。

在一种可能的实施方式中，排名比较模块 90 还可以被配置成：在第一排序结果和第二排序结果不一致时，重新确定第二时段，并将第二排序结果确定为第一排序结果；触发第二获取模块 60 获取第二时段内的第二原始数据。

- 在一种可能的实施方式中，当至少两个预设分类的比例相同时，按照至少
15 两个预设分类的优先级确定至少两个预设分类的排序

关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

- 图 5 是根据本公开的示例性实施例的一种电子设备 500 的框图。例如，
电子设备 500 可以被提供为服务器。参照图 5，电子设备 500 包括处理器 522，
20 其数量可以为一个或多个，以及存储器 532，用于存储可由处理器 522 执行的
计算机程序。存储器 532 中存储的计算机程序可以包括一个或一个以上的每
一个对应于一组指令的模块。此外，处理器 522 可以被配置为执行该计算机
程序，以执行上述的训练数据重采样方法。

- 另外，电子设备 500 还可以包括电源组件 526 和通信组件 550，该电源组
25 件 526 可以被配置为执行电子设备 500 的电源管理，该通信组件 550 可以被
配置为实现电子设备 500 的通信，例如，有线或无线通信。此外，该电子设备
500 还可以包括输入/输出（I/O）接口 558。电子设备 500 可以操作基于存储在存储器 532 的操作系统，例如 Windows Server™、Mac OS X™、Unix™、
Linux™ 等等。

在另一示例性实施例中，还提供了一种包括程序指令的计算机可读存储介质，该程序指令被处理器执行时实现上述的训练数据重采样方法。例如，该计算机可读存储介质可以为上述包括程序指令的存储器 532，上述程序指令可由电子设备 500 的处理器 522 执行以完成上述的训练数据重采样方法。

5 以上结合附图详细描述了本公开的示例性实施方式，但是，本公开并不限于上述实施方式中的具体细节，在本公开的技术构思范围内，可以对本公开的技术方案进行多种简单变型，这些简单变型均属于本公开的范围。

10 另外需要说明的是，在上述具体实施方式中所描述的各个具体技术特征，在不矛盾的情况下，可以通过任何合适的方式进行组合。为了避免不必要的重复，本公开对各种可能的组合方式不再另行说明。

此外，本公开的各种不同的实施方式之间也可以进行任意组合，只要其不违背本公开的构思，就应被视为在本公开的范围內。

权 利 要 求 书

1. 一种训练数据重采样方法，所述方法包括：

获取第一时段内的第一原始数据；

计算所述第一原始数据中多个预设分类分别所占的第一比例；

5 根据所述第一比例的大小关系按照预设规则对所述多个预设分类进行排序，获得第一排序结果；

根据各预设分类的所述第一排序结果和预设对应关系，确定各预设分类对应的采样比例，所述预设对应关系为所述第一排序结果与所述采样比例之间的对应关系；

10 根据所述多个预设分类分别对应的所述采样比例，对用于建模的训练数据进行重采样。

2. 根据权利要求 1 所述的方法，所述方法还包括：在获得所述第一排序结果之后，

15 获取第二时段内的第二原始数据；

计算所述第二原始数据中所述多个预设分类分别所占的第二比例；

根据所述第二比例的大小关系按照所述预设规则对所述多个预设分类进行排序，获得第二排序结果；

20 在所述第一排序结果和所述第二排序结果一致时，执行所述确定各预设分类对应的采样比例的步骤。

3. 根据权利要求 2 所述的方法，其中，

在所述第一排序结果和所述第二排序结果不一致时，重新确定所述第二时段，并将所述第二排序结果确定为第一排序结果；

25 返回所述获取第二时段内的第二原始数据的步骤。

4. 根据权利要求 1-3 中任一项所述的方法，其中，当至少两个预设分类的比例相同时，按照所述至少两个预设分类的优先级确定所述至少两个预设

分类的排序。

5. 一种训练数据重采样装置，所述装置包括：

第一获取模块，用于获取第一时段内的第一原始数据；

5 第一计算模块，用于计算所述第一原始数据中多个预设分类分别所占的第一比例；

第一排序模块，用于根据所述第一比例的大小关系按照预设规则对所述多个预设分类进行排序，获得第一排序结果；

10 比例获取模块，用于根据各预设分类的所述第一排序结果和预设对应关系，确定各预设分类对应的采样比例，所述预设对应关系为所述第一排序结果与所述采样比例之间的对应关系；

重采样模块，用于根据所述多个预设分类分别对应的所述采样比例对用于建模的训练数据进行重采样。

15 6. 根据权利要求 5 所述的装置，所述装置还包括：

第二获取模块，用于在所述第一排序模块获得所述第一排序结果之后，获取第二时段内的第二原始数据；

第二计算模块，用于计算所述第二原始数据中所述多个预设分类分别所占的第二比例；

20 第二排序模块，用于根据所述第二比例的大小关系按照所述预设规则对所述多个预设分类进行排序，获得第二排序结果；

排名比较模块，用于在所述第一排序结果和所述第二排序结果一致时，触发所述比例获取模块确定各预设分类对应的所述采样比例。

25 7. 根据权利要求 6 所述的装置，其中，所述排名比较模块还被配置成：

在所述第一排序结果和所述第二排序结果不一致时，重新确定所述第二时段，并将所述第二排序结果确定为第一排序结果；

触发所述第二获取模块获取第二时段内的第二原始数据。

8. 根据权利要求 5-7 中任一项所述的装置，其中，当至少两个预设分类的比例相同时，按照所述至少两个预设分类的优先级确定所述至少两个预设分类的排序。

5 9. 一种计算机可读存储介质，其上存储有计算机程序，所述程序被处理器执行时实现权利要求 1-4 中任一项所述的方法。

10. 一种电子设备，包括：

存储器，其上存储有计算机程序；

10 处理器，用于执行所述存储器中的所述计算机程序，以实现权利要求 1-4 中任一项所述的方法。

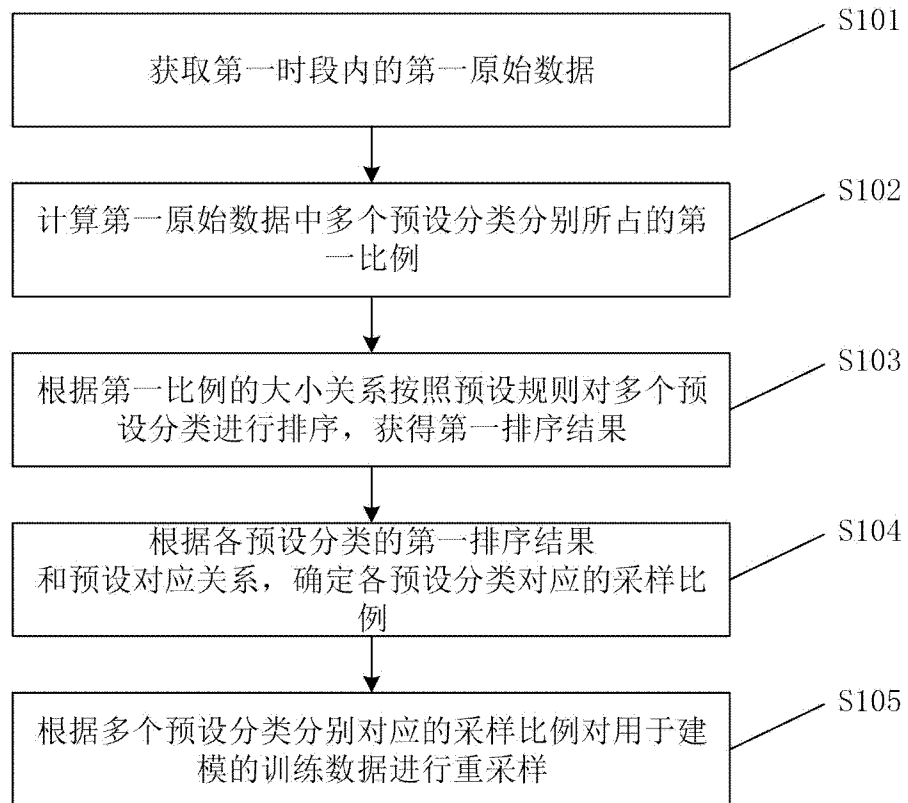


图 1

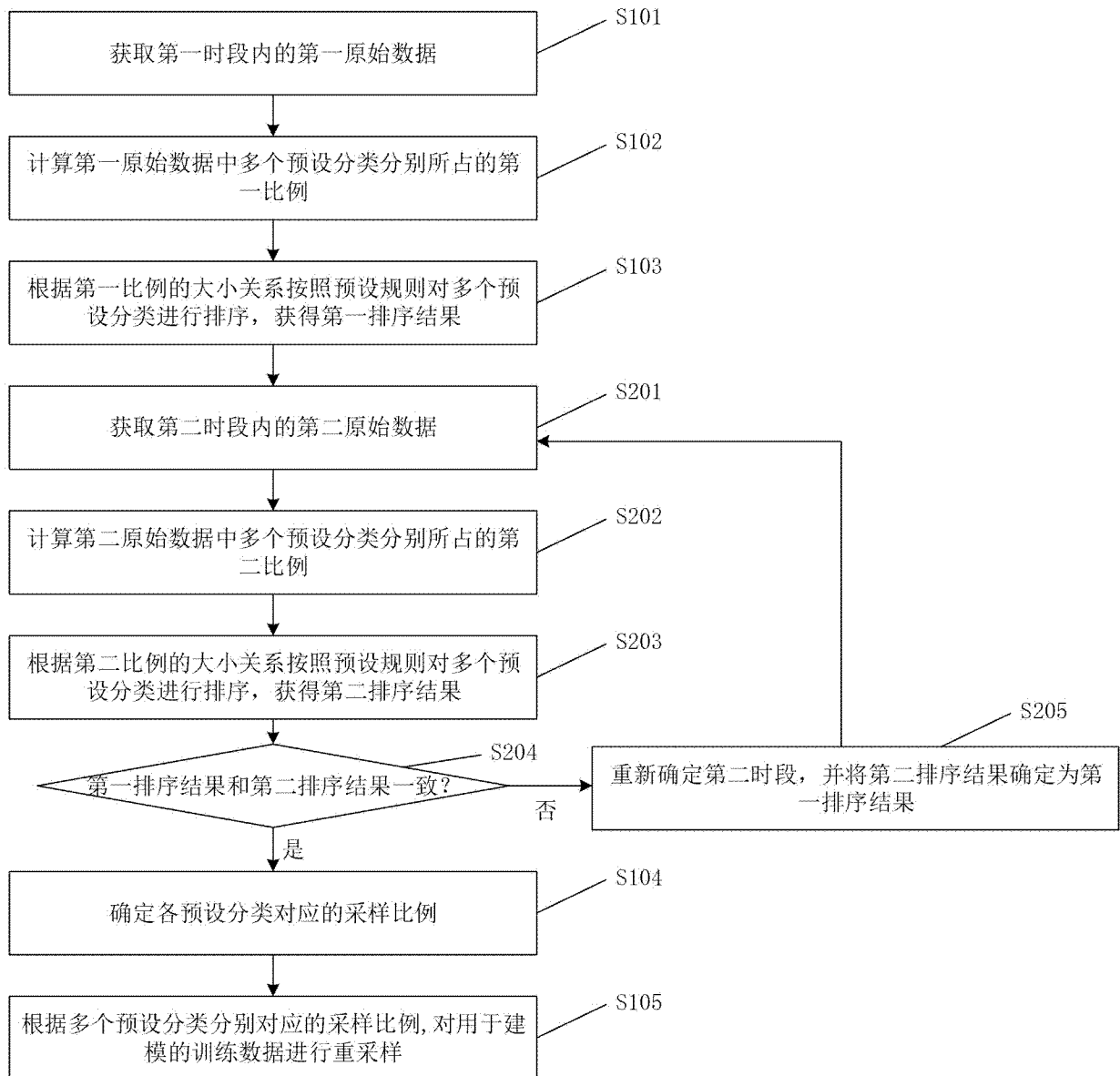


图 2

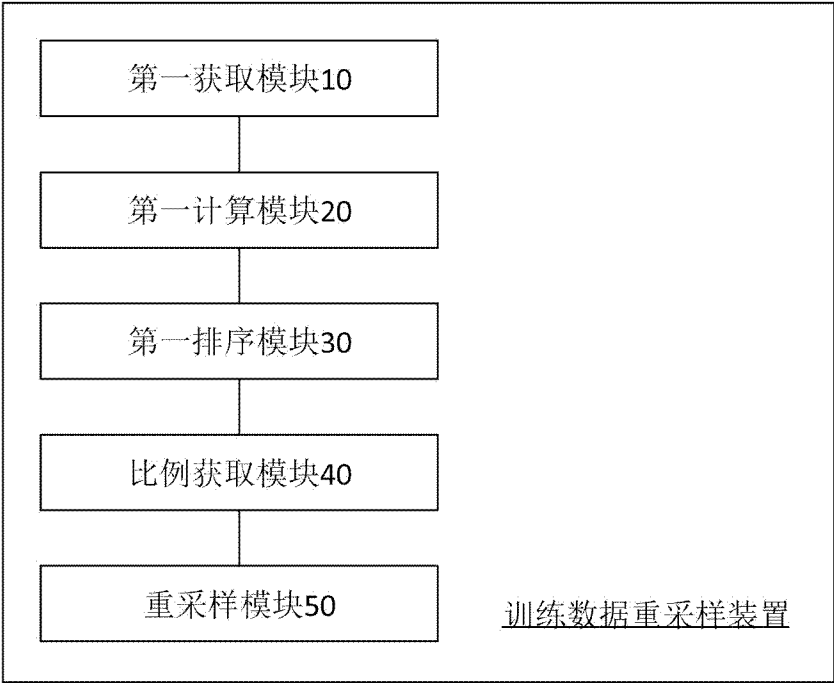


图 3

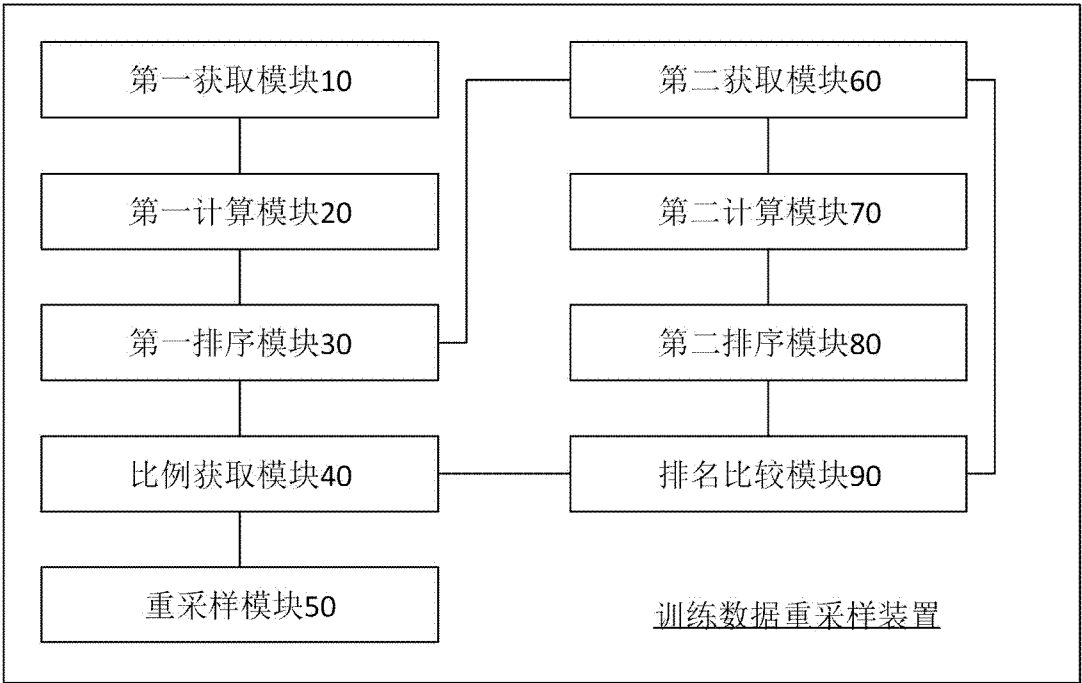


图 4

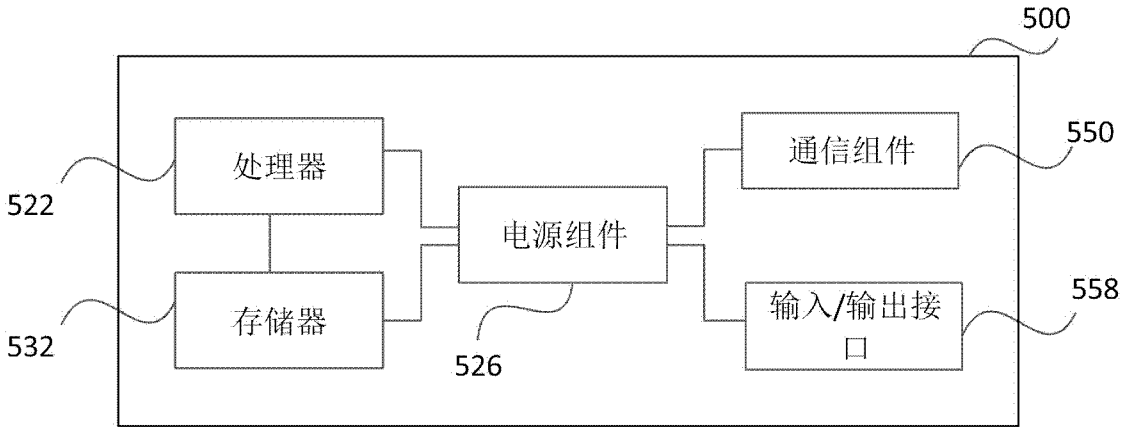


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/094741

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/26(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNKI, CNPAT, IEEE: 采样, 训练, 数据, 分类, 类别, 比例, 重新, 重采样, 建模, 排序, train+, data, classif+, categor+, sampl+, percent, re+, model, query

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109635034 A (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) 16 April 2019 (2019-04-16) claims 1-10	1-10
A	CN 108717547 A (SIC YOUE DATA CO., LTD.) 30 October 2018 (2018-10-30) claims 1 and 8	1-10
A	CN 107563435 A (HARBIN INSTITUTE OF TECHNOLOGY SHENZHEN GRADUATE SCHOOL) 09 January 2018 (2018-01-09) entire document	1-10
A	US 2013097103 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 18 April 2013 (2013-04-18) entire document	1-10
A	CN 107545275 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 05 January 2018 (2018-01-05) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

24 September 2019

Date of mailing of the international search report

09 October 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/094741

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109635034	A	16 April 2019	None			
CN	108717547	A	30 October 2018	None			
CN	107563435	A	09 January 2018	WO	2019041629	A1	07 March 2019
US	2013097103	A1	18 April 2013	None			
CN	107545275	A	05 January 2018	None			

国际检索报告

国际申请号

PCT/CN2019/094741

A. 主题的分类 G06F 16/26 (2019.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																				
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) WPI, EPDOC, CNKI, CNPAT, IEEE: 采样, 训练, 数据, 分类, 类别, 比例, 重新, 重采样, 建模, 排序, train+, data, classif+, categor+, sampl+, percent, re+, model, query																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 109635034 A (北京字节跳动网络技术有限公司) 2019年 4月 16日 (2019 - 04 - 16) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 108717547 A (国信优易数据有限公司) 2018年 10月 30日 (2018 - 10 - 30) 权利要求1, 8</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 107563435 A (哈尔滨工业大学深圳研究生院) 2018年 1月 9日 (2018 - 01 - 09) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>US 2013097103 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2013年 4月 18日 (2013 - 04 - 18) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 107545275 A (华南理工大学) 2018年 1月 5日 (2018 - 01 - 05) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 109635034 A (北京字节跳动网络技术有限公司) 2019年 4月 16日 (2019 - 04 - 16) 权利要求1-10	1-10	A	CN 108717547 A (国信优易数据有限公司) 2018年 10月 30日 (2018 - 10 - 30) 权利要求1, 8	1-10	A	CN 107563435 A (哈尔滨工业大学深圳研究生院) 2018年 1月 9日 (2018 - 01 - 09) 全文	1-10	A	US 2013097103 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2013年 4月 18日 (2013 - 04 - 18) 全文	1-10	A	CN 107545275 A (华南理工大学) 2018年 1月 5日 (2018 - 01 - 05) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 109635034 A (北京字节跳动网络技术有限公司) 2019年 4月 16日 (2019 - 04 - 16) 权利要求1-10	1-10																		
A	CN 108717547 A (国信优易数据有限公司) 2018年 10月 30日 (2018 - 10 - 30) 权利要求1, 8	1-10																		
A	CN 107563435 A (哈尔滨工业大学深圳研究生院) 2018年 1月 9日 (2018 - 01 - 09) 全文	1-10																		
A	US 2013097103 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2013年 4月 18日 (2013 - 04 - 18) 全文	1-10																		
A	CN 107545275 A (华南理工大学) 2018年 1月 5日 (2018 - 01 - 05) 全文	1-10																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期 2019年 9月 24日		国际检索报告邮寄日期 2019年 10月 9日																		
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 姚楠 电话号码 86-(10)-53961521																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/094741

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	109635034	A	2019年 4月 16日	无	
CN	108717547	A	2018年 10月 30日	无	
CN	107563435	A	2018年 1月 9日	W0 2019041629 A1	2019年 3月 7日
US	2013097103	A1	2013年 4月 18日	无	
CN	107545275	A	2018年 1月 5日	无	