

(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
2017 年 9 月 8 日 (08.09.2017)



(10) 国际公布号
WO 2017/148266 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2017/073719
- (22) 国际申请日: 2017 年 2 月 16 日 (16.02.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610113716.1 2016 年 2 月 29 日 (29.02.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 周俊 (ZHOU, Jun); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。
- (74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (CN)。

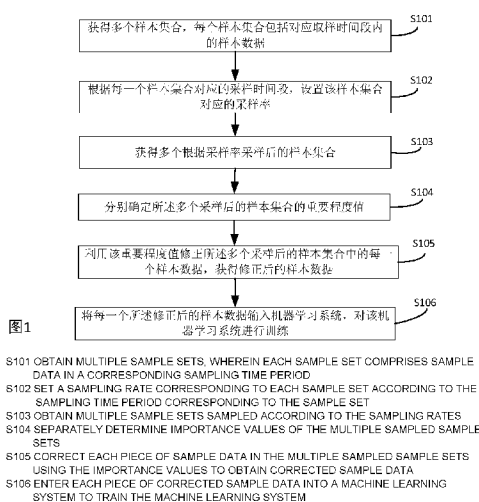
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA,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

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) Title: METHOD AND SYSTEM FOR TRAINING MACHINE LEARNING SYSTEM

(54) 发明名称: 一种机器学习系统的训练方法和训练系统



(57) Abstract: A method and system for training a machine learning system. Multiple pieces of sample data are used to train the machine learning system. The method comprises: obtaining multiple sample sets, wherein each sample set comprises sample data in a corresponding sampling time period (S101); setting a sampling rate corresponding to each sample set according to the sampling time period corresponding to the sample set (S102); obtaining multiple sample sets sampled according to the sampling rates (S103); separately determining importance values of the multiple sampled sample sets (S104); correcting each piece of sample data in the multiple sampled sample sets using the importance values to obtain corrected sample data (S105); and entering each piece of corrected sample data into the machine learning system to train the machine learning system (S106). Processing sample data before entering the sample data into a machine learning system can reduce requirements for the memory resource of the machine as well as reducing the influence on the learning effect of the machine learning system as far as possible.

(57) 摘要:

[见续页]



WO 2017/148266 A1



一种机器学习系统的训练方法和训练系统，利用多个样本数据对机器学习系统进行训练，该方法包括：获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据(S101)；根据每一个样本集合对应的取样时间段，设置该样本集合对应的采样率(S102)；获得多个根据采样率采样后的样本集合(S103)；分别确定所述多个采样后的样本集合的重要程度值(S104)；利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据(S105)；将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练(S106)。将样本数据输入机器学习系统之前对样本数据进行处理，在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

一种机器学习系统的训练方法和训练系统

本申请要求 2016 年 02 月 29 日递交的申请号为 201610113716.1、发明名称为“一种机器学习系统的训练方法和训练系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及大数据处理领域，尤其涉及一种机器学习系统的训练方法和训练系统。

背景技术

10 在如今的大数据时代，互联网公司获取超大规模数据已非常容易。据不完全统计，谷歌 2012 年每天 30 亿 query/300 亿广告，脸书用户 2013 年每天分享 43 亿内容，阿里巴巴 2015 双十一当天就有超过 7 亿笔交易。这些公司通过机器学习系统，去挖掘数据里面的金矿，包括用户兴趣/行为/习惯等等。

机器学习系统设计为模仿人脑的神经网络，用于预测用户的行为。在机器学习系统
15 上线之前，需要通过大规模的数据进行训练。然而在训练过程中，大规模的数据必然要求大规模的机器资源才能有效处理，例如腾讯的广告数据，都是 PB 级别，必然要用到千台机器以上，这对大部分公司来说，都是个巨大的成本。

为了降低成本，提高机器学习系统的效率，通常的处理方式是通过随机样本采样的手段减少机器学习系统处理的数据量。随机样本采样就是以一定概率丢弃样本，例如对
20 每一个样本随机生成 1 个 0-1 范围内的浮点数，当浮点数大于阈值时则直接丢弃该样本。然而，随机丢弃样本的方式会丢弃大量的有用数据，损害机器学习系统的训练效果，降低预测的精度。

发明内容

25 鉴于上述问题，提出了本申请实施例以便提供一种克服上述问题或者至少部分地解决上述问题的机器学习系统的训练方法和训练系统。

为解决上述问题，本申请一实施例公开一种机器学习系统的训练方法，利用多个样本数据对机器学习系统进行训练，所述训练方法包括：

获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

30 根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

获得多个根据采样率采样后的样本集合；

分别确定所述多个采样后的样本集合的重要程度值；

利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

5 将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

本申请另一实施例公开一种机器学习系统的训练系统，利用多个样本数据对机器学习系统进行训练，所述训练系统包括：

第一获取模块，用于获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

10 采样率设置模块，用于根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

第二获取模块，用于获得多个根据采样率采样后的样本集合；

重要程度值确定模块，用于分别设置所述多个采样后的样本集合的重要程度值；

15 样本数据修正模块，用于利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

训练模块，用于将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

本申请实施例至少具有以下优点：本申请实施例公开一种机器学习系统的训练方法和训练系统，在将样本数据输入机器学习系统之前对样本数据进行处理，包括获取根据
20 取样时间段划分的样本集合、根据取样时间段设置每个样本集合的采样率、根据采样率进行采样、确定采样后样本集合的重要程度值以及利用该重要程度值将样本数据进行修正，并将样本数据输入机器学习系统进行训练，在降低机器学习系统处理的数据量的同时保证了重要数据的采用率和利用程度，在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

25

附图说明

图 1 是本申请第一实施例的机器学习系统的训练方法的流程图。

图 2 是本申请第二实施例的机器学习系统的训练方法的流程图。

图 3 是本申请第三实施例的机器学习系统的训练方法的流程图。

30 图 4 是本申请第四实施例的机器学习系统的训练系统的方框图。

图 5 是本申请第五实施例的机器学习系统的训练系统的方框图。

图 6 是本申请第六实施例的机器学习系统的训练系统的方框图。

具体实施方式

5 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员所获得的所有其他实施例，都属于本申请保护的范围。

10 本申请的核心思想之一在于，提出一种机器学习系统的训练方法和训练系统，利用多个样本数据对机器学习系统进行训练，包括根据样本数据的取样时间段将样本数据划分为多个样本集合；根据取样时间段设置每一个样本集合的采样率；根据采样率对每一个样本集合采样，并修改每一个采样后的样本集合对应的重要程度值；利用重要程度值修正每一个样本数据，并将该修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

15

第一实施例

本申请第一实施例提出一种机器学习系统的训练方法，如图 1 所示为本申请一实施例的机器学习系统的训练方法的流程图，本实施例提出的机器学习系统的训练方法包括如下步骤：

20 S101，获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

在这一步骤中，每一个样本数据例如为一个向量，该向量中的其中一个维度例如为该样本数据的取样时间。在本步骤中可以将所有样本数据的取样时间划分为多个取样时间段，并将多个样本数据根据取样时间段划分为多个样本集合，每一个样本集合对应一个取样时间段。

25 例如，所有样本数据的取样时间是从 1 月 24 日至 1 月 29 日，则可以将这一取样时间划分为多个取样时间段，例如 1 月 29 日、1 月 27 日至 1 月 28 日、1 月 24 日至 1 月 26 日三个取样时间段。按照上述三个取样时间段，将样本数据划分为 1 月 29 日取样的样本集合、1 月 27 日至 1 月 28 日取样的样本数据集合、1 月 24 日至 1 月 26 日取样的样本集合。因此，每一个样本集合对应一个取样时间段。

30 值得注意的是，上述取样时间段可以是依据开发者或使用者设定的规则划分，可以

平均分布或者不均匀分布，本申请并不以此为限。

S102，根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

在这一步骤中，可以根据取样时间段设置对应的每一个取样集合的样本率。例如，可以按照取样时间段越新的样本集合对应的采样率越高的原则设置采样率。即，所述样本集合的采样率随着该样本集合对应的取样时间段从旧到新而增加。例如在上述示例中，
5 可以将 1 月 29 日取样的样本数据对应的样本集合的采样率设置为 1.0，将 1 月 27 日至 1 月 28 日取样的样本数据对应的样本集合的采样率设置为 0.5，将 1 月 24 日至 1 月 26 日取样的样本数据对应的样本集合的采样率设置为 0.1。

S103，获得多个根据采样率采样后的样本集合；

10 在这一步骤中，可以根据上一步骤中设置的采样率，对每一个样本集合内的样本进行采样。例如某一个样本集合中包含的样本数据为 1000 个，采样率为 0.1，则采样后该样本集合中包含的样本数据的个数为 $1000 \times 0.1 = 100$ 个。通过采样后，样本集合中的样本数据为 100 个，这 100 个样本数据对应的集合可以称为采样后的样本集合。

S104，分别确定所述多个采样后的样本集合的重要程度值；

15 在一实施例中，重要程度值可以是人为或者机器算法设定的系数，每一个采样后的样本集合对应的重要程度值分别可以人为设定或通过机器以一定规则设定。在上述步骤中，可以在该样本集合原重要程度值的基础上，设置新的重要程度值。

S105，利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

20 在这一步骤中，可以利用该重要程度值修正多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

利用该重要程度值修正每一个样本数据，可以是将每一个向量的每一个特征维度与重要程度值相乘，使该向量等比例放大，获得修正后的样本数据。

例如，该样本集合原有的或者默认的重要程度值为 1，在这一步骤中可以修正为 2，
25 因此某个原来为 $a(1, 1, 1, 2, \dots \dots n)$ 的样本数据在这一步骤中可以修正为 $a(2, 2, 2, 4, \dots \dots 2n)$ ，即为修正后的样本数据。

然而，正如本领域技术人员可以得知的，重要程度值并不限于人为或者机器算法设定的系数，在其他实施例，还可以有多种方法，例如对样本数据 $a(1, 1, 1, 2, \dots \dots n)$ 进行数学运算， $a1=f(a)$ 等等，这里的函数 f 可以为等比相乘函数，或者类似指数运算等等各
30 种数学函数，也可以对样本进行修正。

S106, 将每一个所述修正后的样本数据输入机器学习系统, 对该机器学习系统进行训练。

在这一步骤中, 可以将修正后的样本数据输入机器学习系统, 对该机器学习系统进行训练。在训练中, 首先对损失函数求导, 计算出梯度, 再结合初始的权重以及设置的步长, 根据公式“新的权重=旧的权重+步长*梯度”通过迭代的方式计算出接近最优解的权重值。

综上所述, 本申请第一实施例公开一种机器学习系统的训练方法, 在将样本数据输入机器学习系统之前对样本数据进行处理, 降低数据量的同时保证了重要数据的采用率和利用程度, 在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

第二实施例

本申请第二实施例提出一种机器学习系统的训练方法, 如图 2 所示为本申请第二实施例的机器学习系统的训练方法的流程图, 本实施例提出的机器学习系统的训练方法包括如下步骤:

S201, 获得多个样本集合, 每个样本集合包括对应取样时间段内的样本数据;

S202, 根据每一个样本集合对应的采样时间段, 设置该样本集合对应的采样率;

S203, 获得多个根据采样率采样后的样本集合;

上述三个步骤与第一实施例中的步骤 S101、S102、S103 相同或相似, 在此不再赘述。

S204, 分别确定所述多个采样后的样本集合的重要程度值;

步骤 S204 例如可以包括:

子步骤 S204a: 基于对应的采样率对所述采样后的样本集合的初始重要程度值进行修正, 得到所述采样后的样本集合的重要程度值;

所述重要程度值和初始重要程度值为正比关系, 和所述采样后的样本集合的采样率为反比关系。

在子步骤 S204a 中, 例如可以通过该样本集合原先对应的重要程度值与采样率的比值, 计算新的重要程度值。例如, 可以按照下述公式初次设置每一个样本集合的重要程度值:

$$Y1=Y/a;$$

其中 $Y1$ 为对应于该样本集合的设置后的重要程度值；

Y 为对应于该样本集合的原始的重要程度值；

a 为所述样本集合的采样率。

5 举例来说，在第一实施例所提供的示例中，如果针对 1 月 24 日至 1 月 26 日这一取样时间段的采样率为 0.1，并且该集合对应的重要程度值设为 0.2；针对 1 月 29 日这一取样时间段的采样率为 0.5，并且该集合对应的重要程度值设为 1；针对 1 月 27 日至 1 月 28 日这一取样时间段的采样率为 1，并且该集合对应的重要程度值设为 5，则根据 $Y1=Y/a$ ，可以得出按照取样时间段由旧到新排列的这三个集合的重要程度值分别为 2、2、5。

10 步骤 S204 例如还可以包括：

子步骤 S204b，按照预置规则，提高最新的取样时间段对应的样本集合的重要程度值。

在子步骤 S204b 中，这一预置规则例如可以包括：

15 提高后的最新的取样时间段对应的样本集合的重要程度值与提高前的最新的取样时间段对应的样本集合的重要程度值成正比，并与样本集合的总个数成正比。

在这一子步骤中，例如可以按照下述公式再次设置最新的取样时间段对应的样本集合的重要程度值：

$$Z1=Z*b;$$

其中 $Z1$ 为对应于该样本集合的再次修改后的重要程度值；

20 Z 为对应于该样本集合的初次修改后的重要程度值；

b 为样本集合的总个数。

举例来说，根据步骤 S204b 得出的按照取样时间段由旧到新排列的三个样本集合对应的重要程度值分别为 2、2、5，在这一步骤中，可以针对取样时间段最新的采样后的样本集合，即第三个样本集合，再次提升其重要程度值。

25 例如，可以按照下述公式再次设置最新的取样时间段对应的样本集合的重要程度值：

$$Z1=Z*b;$$

其中 $Z1$ 为对应于该样本集合的再次设置后的重要程度值；

Z 为对应于该样本集合的初次设置后的重要程度值；

b 为样本集合的总个数。

30 举例来说，在子步骤 S204a 中获得的取样时间段最新的样本集合对应的初次设置后

的重要程度值为 5，在这一子步骤中，可以通过 $Z1=Z*b$ 的公式，获取再次设置后的重要程度值为 $5*3=15$ 。

值得注意的是，子步骤 S204b 可以在子步骤 S204a 之前或之后执行，或者是单独执行。即，子步骤 S204b 相对于子步骤 S204a 是独立的，并不依赖于子步骤 S204a。

5 S205，利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

在这一步骤例如可以包括如下子步骤：

S205a，将每一个所述重要程度值与对应的采样后的样本集合中的每一个样本数据相乘，获得修正后的样本数据。

10 S206，将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

这一步骤可以与第一实施例中的步骤 S106 相同或相似，在此不再赘述。

综上所述，本申请第二实施例公开一种机器学习系统的训练方法，在将样本数据输入机器学习系统之前对样本数据进行处理，通过对不同样本集合的重要程度值的设置，
15 在降低数据量的同时保证了重要数据的采用率和利用程度，在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

第三实施例

本申请第三实施例提出一种机器学习系统的训练方法，如图 3 所示为本申请第三实
20 施例的机器学习系统的训练方法的流程图，本实施例提出的机器学习系统的训练方法包括如下步骤：

S301，获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

S302，根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

S303，获得多个根据采样率采样后的样本集合；

25 S304，分别确定所述多个采样后的样本集合的重要程度值；

S305，利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

上述步骤 S301 至 S305 可以与第一实施例公开的步骤 S101 至 S105 相同或相似，也可以与第二实施例公开的步骤 S201 至 S205 相同或相似，在此不再赘述。

30 本实施例还可以包括如下步骤：

S306, 将每一个所述修正后的样本数据输入机器学习系统, 对该机器学习系统进行训练。

在这一步骤中, 可以将修正后的样本数据输入机器学习系统, 对该机器学习系统进行训练。在训练中, 首先对损失函数求导, 计算出梯度, 再结合初始的权重以及设置的步长, 根据公式“新的权重=旧的权重+步长*梯度”通过迭代的方式计算出接近最优解的权重值。

这一步骤可以包括如下子步骤:

S306a, 计算出每一个所述修正后的样本数据的梯度;

S306b, 降低每一个所述样本数据的梯度的精度;

10 S306c, 将降低精度后的梯度输入所述机器学习系统, 对该机器模型进行训练。

在步骤 S306a 中, 可以首先计算每一个修正后的样本数据的梯度, 该梯度为损失函数的导数, 通过对损失函数求导, 可以获得梯度。

在步骤 S306b 中, 机器学习系统的训练一般采用梯度下降法, 每一台机器都需要计算出梯度。如果储存 1 个梯度需要 8 byte (字节), 则 100 亿个梯度需要
15 10000000000*8/1024/1024/1024 = 74.5G 的存储空间。如果将储存一个梯度的字节数压缩至 4 byte, 则 100 亿的梯度仅仅需要 32.25G 内存。

在步骤 S306b 中, 可以使用下述公式, 减少每一个本数据的梯度的存储字节, 以实现降低精度:

$$X1 = \text{floor}(c * X + (\text{rand}()) / d) / c$$

20 其中 floor 为向下取整; rand() 为产生 0-d 之间的浮点数; X1 为低精度浮点数, 例如为计算机存储需要 4 个字节的 float, 在这里表示减少后每一个所述样本数据的梯度的存储字节; X 为高精度浮点数, 例如为计算机存储需要 8 个字节的 double, 为减少前每一个所述样本数据的梯度的存储字节。

另外, 通过利用 rand 函数引入随机因素, 来尽量降低浮点数的累计误差。例如, 利用 $(c * X + (\text{rand}()) / d)$ 的算法, 让 X 乘以一个固定的数, 然后加上一个在 0-1 范围内的浮点数, 目的在于在引入随机因素。C 的值是个经验值, 例如可以为 536870912。D 例如可以为 $2^{32}-1$, 即 2147483647, 是 rand 函数所能产生的上限。

通过上述公式, 可以实现将一个高精度的浮点数, 转成一个低精度的浮点数, 并且尽可能减低累计误差。

30 综上所述, 本申请第三实施例公开一种机器学习系统的训练方法, 在将样本数据输

入机器学习系统之前对样本数据进行处理，通过对不同样本集合的重要程度值的设置，以及在降低梯度精度时的处理，在降低数据量的同时保证了重要数据的采用率和利用程度，在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

5 第四实施例

本申请第四实施例提出一种机器学习系统的训练系统，如图 4 所示为本申请第四实施例的机器学习系统的训练系统的方框图，本实施例提出的机器学习系统的训练系统利用多个样本数据对机器学习系统进行训练，所述训练系统 400 包括：

10 第一获取模块 401，用于获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

采样率设置模块 402，用于根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

第二获取模块 403，用于获得多个根据采样率采样后的样本集合；

重要程度值确定模块 404，用于分别设置所述多个采样后的样本集合的重要程度值；

15 样本数据修正模块 405，用于利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

训练模块 406，用于将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

20 优选地，在本实施例中，所述样本集合的采样率随着该样本集合对应的取样时间段从旧到新而增加。

综上所述，本申请第四实施例公开一种机器学习系统的训练系统，在将样本数据输入机器学习系统之前对样本数据进行处理，降低数据量的同时保证了重要数据的采用率和利用程度，在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

25

第五实施例

本申请第五实施例提出一种机器学习系统的训练系统，如图 5 所示为本申请第五实施例的机器学习系统的训练系统的方框图，本实施例提出的机器学习系统的训练系统利用多个样本数据对机器学习系统进行训练，所述训练系统 500 包括：

30 第一获取模块 501，用于获得多个样本集合，每个样本集合包括对应取样时间段内

的样本数据；

采样率设置模块 502，用于根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

第二获取模块 503，用于获得多个根据采样率采样后的样本集合；

5 重要程度值确定模块 504，用于分别设置所述多个采样后的样本集合的重要程度值；

样本数据修正模块 505，用于利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

训练模块 506，用于将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

10 在本实施例中，所述样本数据修正模块 505 用于：

将每一个所述重要程度值与对应的采样后的样本集合中的每一个样本数据相乘，获得修正后的样本数据。

在本实施例中，所述重要程度值确定模块 504 包括：

15 初次修正子模块 504a，用于基于对应的采样率对所述采样后的样本集合的初始重要程度值进行修正，得到所述采样后的样本集合的重要程度值；

所述重要程度值和初始重要程度值为正比关系，和所述采样后的样本集合的采样率为反比关系。

例如，所述初次修正子模块可以按照下述公式初次设置每一个所述样本集合的重要程度值：

20 $Y1=Y/a;$

其中 Y1 为对应于该样本集合的设置后的重要程度值；

Y 为对应于该样本集合的原始的重要程度值；

a 为所述样本集合的采样率。

在本实施例中，所述重要程度值确定模块 504 还可以包括：

25 二次修正子模块 504b，用于按照预置规则，提高最新的取样时间段对应的样本集合的重要程度值。

优选地，所述预置规则包括：

提高后的最新的取样时间段对应的样本集合的重要程度值与提高前的最新的取样时间段对应的样本集合的重要程度值成正比，并与样本集合的总个数成正比。

30 例如，可以按照下述公式再次设置最新的取样时间段对应的样本集合的重要程度值：

$$Z1=Z*b;$$

其中 Z1 为对应于该样本集合的再次设置后的重要程度值;

Z 为对应于该样本集合的初次设置后的重要程度值;

b 为样本集合的总个数。

5 在本实施例中, 所述样本集合的采样率随着该样本集合对应的取样时间段从旧到新而增加。

综上所述, 本申请第五实施例公开一种机器学习系统的训练系统, 在将样本数据输入机器学习系统之前对样本数据进行处理, 通过对不同样本集合的重要程度值的设置, 在降低数据量的同时保证了重要数据的采用率和利用程度, 在减轻机器的内存资源需求
10 的同时尽量降低对机器学习系统的学习效果的影响。

第六实施例

本申请第六实施例提出一种机器学习系统的训练系统, 如图 6 所示为本申请第六实施例的机器学习系统的训练系统的方框图, 本实施例提出的机器学习系统的训练系统利用多个样本数据对机器学习系统进行训练, 所述训练系统 600 包括:
15

第一获取模块 601, 用于获得多个样本集合, 每个样本集合包括对应取样时间段内的样本数据;

采样率设置模块 602, 用于根据每一个样本集合对应的采样时间段, 设置该样本集合对应的采样率;

20 第二获取模块 603, 用于获得多个根据采样率采样后的样本集合;

重要程度值确定模块 604, 用于分别设置所述多个采样后的样本集合的重要程度值;

样本数据修正模块 605, 用于利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据, 获得修正后的样本数据;

训练模块 606, 用于将每一个所述修正后的样本数据输入机器学习系统, 对该机器学习系统进行训练。
25

在本实施例中, 所述训练模块 606 包括:

计算子模块 606a, 用于计算出每一个所述修正后的样本数据的梯度;

精度降低子模块 606b, 用于降低每一个所述梯度的精度;

训练子模块 606c, 用于将降低精度后的梯度输入所述机器学习系统, 对该机器学习模型
30 进行训练。

在本实施例中，所述精度降低子模块 606b 用于：

利用下述公式，减少每一个梯度的存储字节，以实现降低精度：

$$X1 = \text{floor}(c * X + (\text{rand}()) / d) / c$$

其中 floor 为向下取整；rand（）为产生 0-d 之间的浮点数；X1 为减少后的存储字节

5 数；X 为减少前的存储字节数。

综上所述，本申请第六实施例公开一种机器学习系统的训练系统，在将样本数据输入机器学习系统之前对样本数据进行处理，通过对不同样本集合的重要程度值的设置，以及在降低梯度精度时的处理，在降低数据量的同时保证了重要数据的采用率和利用程度，在减轻机器的内存资源需求的同时尽量降低对机器学习系统的学习效果的影响。

10

对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

15 本领域内的技术人员应明白，本申请实施例的实施例可提供为方法、装置、或计算机程序产品。因此，本申请实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

20 在一个典型的配置中，所述计算机设备包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM) 和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。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号存储。信号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
25 于存储可以被计算设备访问的信号。按照本文中的界定，计算机可读介质不包括非持续
30

性的电脑可读媒体(transitory media)，如调制的数据信号和载波。

本申请实施例是参照根据本申请实施例的方法、终端设备(系统)、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

尽管已描述了本申请实施例的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请实施例范围的所有变更和修改。

最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

以上对本申请所提供的一种针对混淆脚本语言的定位方法和系统，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，以上实施例的说明只是

用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权利要求书

1. 一种机器学习系统的训练方法，利用多个样本数据对机器学习系统进行训练，其特征在于，所述训练方法包括：

获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

5 根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

获得多个根据采样率采样后的样本集合；

分别确定所述多个采样后的样本集合的重要程度值；

利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

10 将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

2. 如权利要求 1 所述的机器学习系统的训练方法，其特征在于，所述利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据的步骤包括：

15 将每一个所述重要程度值与对应的采样后的样本集合中的每一个样本数据相乘，获得修正后的样本数据。

3. 如权利要求 1 所述的机器学习系统的训练方法，其特征在于，所述将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练的步骤包括：

计算出每一个所述修正后的样本数据的梯度；

降低每一个所述梯度的精度；

20 将降低精度后的梯度输入所述机器学习系统，对该机器学习模型进行训练。

4. 如权利要求 3 所述的机器学习系统的训练方法，其特征在于，所述降低每一个所述梯度的精度的步骤包括：

利用下述公式，减少每一个梯度的存储字节，以实现降低精度：

$$X1 = \text{floor}(c * X + (\text{rand}()) / d) / c$$

25 其中 floor 为向下取整；rand（）为产生 0-d 之间的浮点数；X1 为减少后的存储字节数；X 为减少前的存储字节数。

5. 如权利要求 1 所述的机器学习系统的训练方法，其特征在于，所述分别确定所述多个采样后的样本集合的重要程度值步骤包括：

30 基于对应的采样率对所述采样后的样本集合的初始重要程度值进行修正，得到所述采样后的样本集合的重要程度值；

所述重要程度值和初始重要程度值为正比关系，和所述采样后的样本集合的采样率为反比关系。

6. 如权利要求 5 所述的机器学习系统的训练方法，其特征在于，所述分别设置所述多个采样后的样本集合的重要程度值步骤还包括：

5 按照预置规则，提高最新的取样时间段对应的样本集合的重要程度值。

7. 如权利要求 6 所述的机器学习系统的训练方法，其特征在于，所述预置规则包括：提高后的最新的取样时间段对应的样本集合的重要程度值与提高前的最新的取样时间段对应的样本集合的重要程度值成正比，并与样本集合的总个数成正比。

8. 如权利要求 1 所述的机器学习系统的训练方法，其特征在于，在根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率的步骤中，所述样本集合的采样率随着该样本集合对应的取样时间段从旧到新而增加。

9. 一种机器学习系统的训练系统，利用多个样本数据对机器学习系统进行训练，其特征在于，所述训练系统包括：

15 第一获取模块，用于获得多个样本集合，每个样本集合包括对应取样时间段内的样本数据；

采样率设置模块，用于根据每一个样本集合对应的采样时间段，设置该样本集合对应的采样率；

第二获取模块，用于获得多个根据采样率采样后的样本集合；

重要程度值确定模块，用于分别设置所述多个采样后的样本集合的重要程度值；

20 样本数据修正模块，用于利用该重要程度值修正所述多个采样后的样本集合中的每一个样本数据，获得修正后的样本数据；

训练模块，用于将每一个所述修正后的样本数据输入机器学习系统，对该机器学习系统进行训练。

10. 如权利要求 9 所述的机器学习系统的训练系统，其特征在于，所述样本数据修正模块用于：

将每一个所述重要程度值与对应的采样后的样本集合中的每一个样本数据相乘，获得修正后的样本数据。

11. 如权利要求 9 所述的机器学习系统的训练系统，其特征在于，所述训练模块包括：

30 计算子模块，用于计算出每一个所述修正后的样本数据的梯度；

精度降低子模块，用于降低每一个所述梯度的精度；

训练子模块，用于将降低精度后的梯度输入所述机器学习系统，对该机器模型进行训练。

12. 如权利要求 11 所述的机器学习系统的训练系统，其特征在于，所述精度降低子
5 模块用于：

利用下述公式，减少每一个梯度的存储字节，以实现降低精度：

$$X1 = \text{floor}(c * X + (\text{rand}()) / d) / c$$

其中 floor 为向下取整；rand（）为产生 0-d 之间的浮点数；X1 为减少后的存储字节数；X 为减少前的存储字节数。

10 13. 如权利要求 9 所述的机器学习系统的训练系统，其特征在于，所述重要程度值确定模块包括：

初次修正子模块，用于基于对应的采样率对所述采样后的样本集合的初始重要程度值进行修正，得到所述采样后的样本集合的重要程度值；

15 所述重要程度值和初始重要程度值为正比关系，和所述采样后的样本集合的采样率为反比关系。

14. 如权利要求 13 所述的机器学习系统的训练系统，其特征在于，所述重要程度值确定模块还包括：

二次修正子模块，用于按照预置规则，提高最新的取样时间段对应的样本集合的重要程度值。

20 15. 如权利要求 14 所述的机器学习系统的训练系统，其特征在于，所述预置规则包括：

提高后的最新的取样时间段对应的样本集合的重要程度值与提高前的最新的取样时间段对应的样本集合的重要程度值成正比，并与样本集合的总个数成正比。

25 16. 如权利要求 9 所述的机器学习系统的训练系统，其特征在于，所述采样率设置模块用于，将所述样本集合的采样率设置为随着该样本集合对应的取样时间段从旧到新而增加。

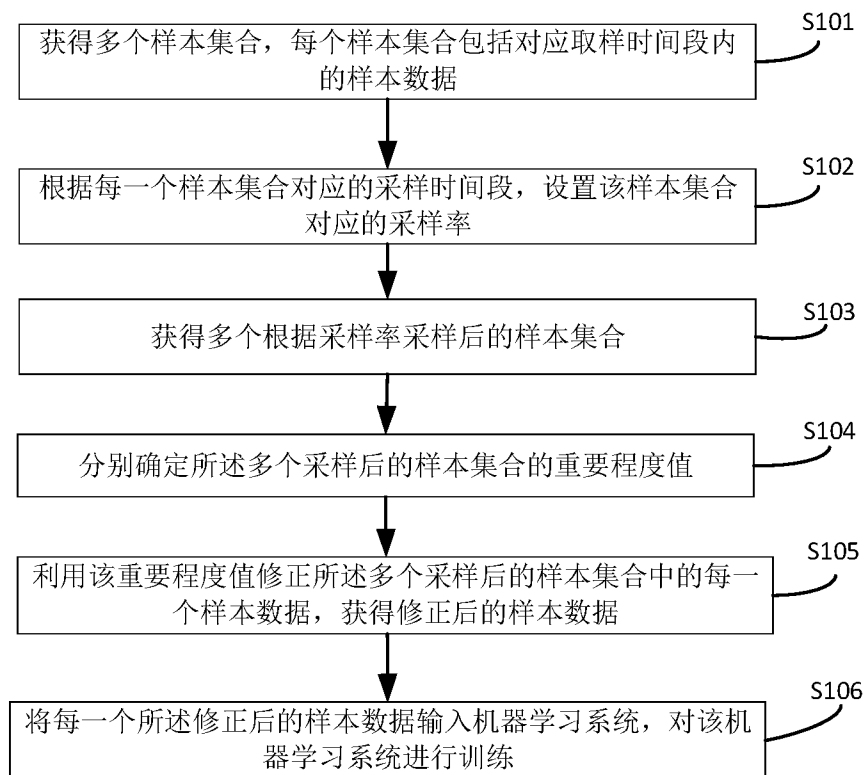


图1

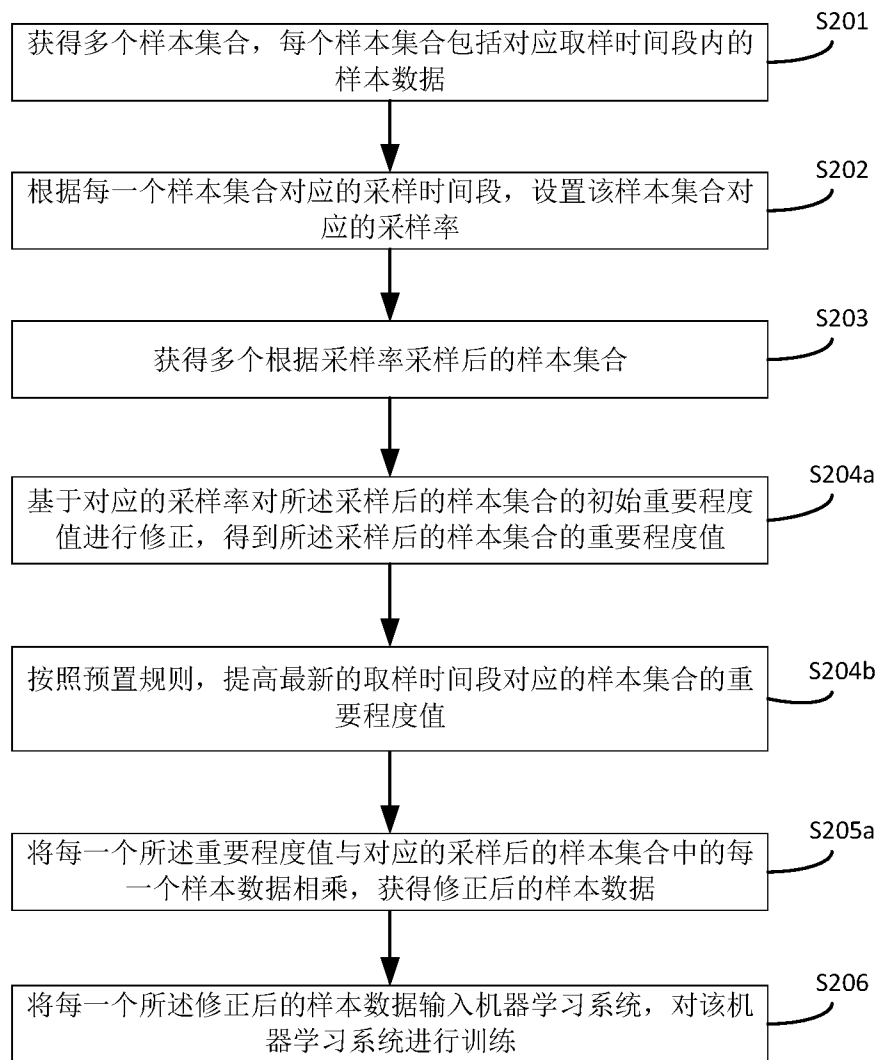


图2

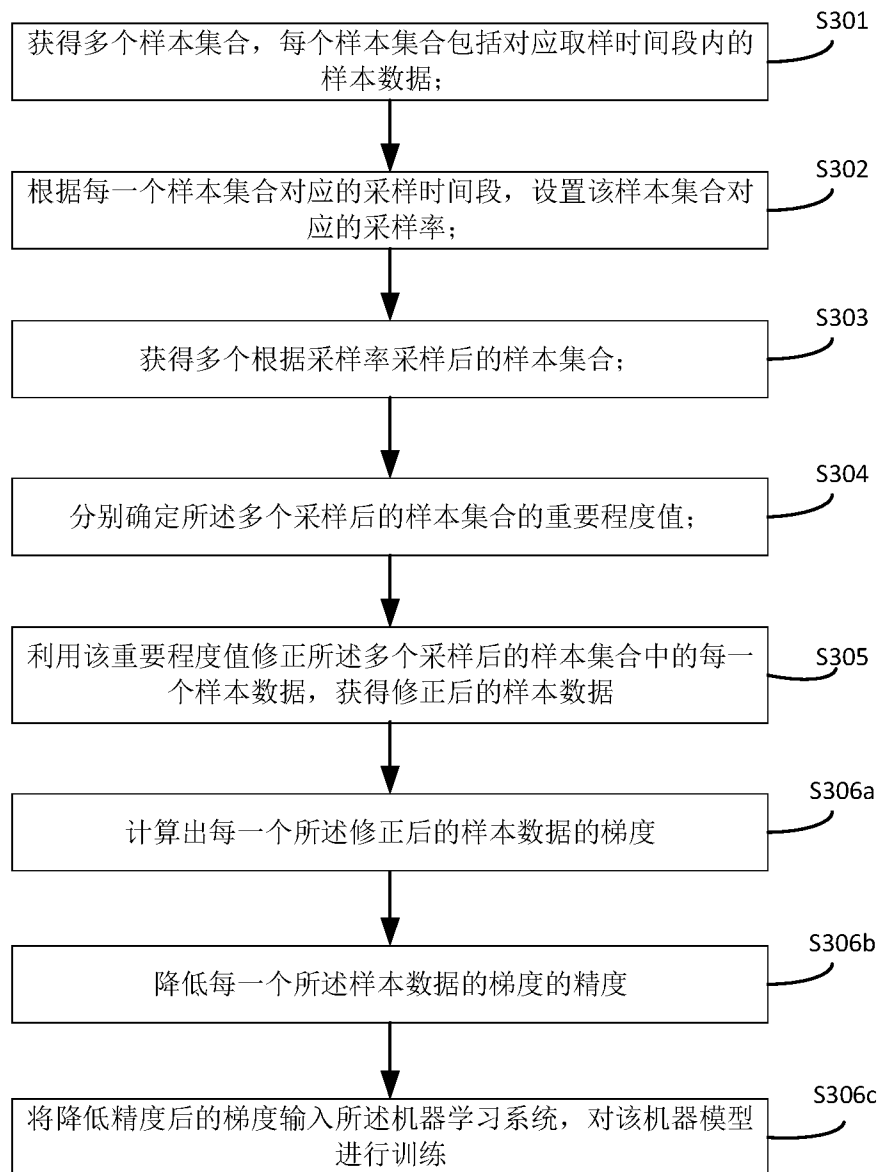


图3

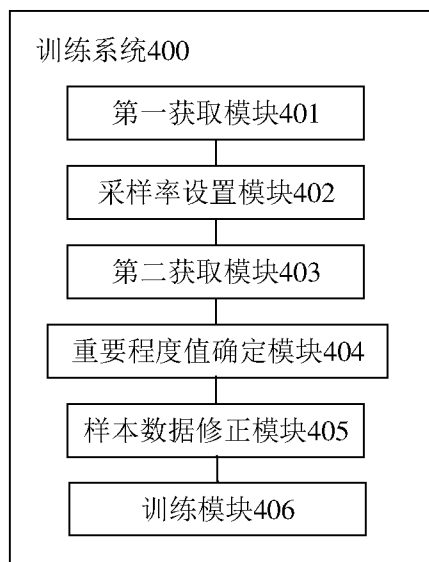


图4

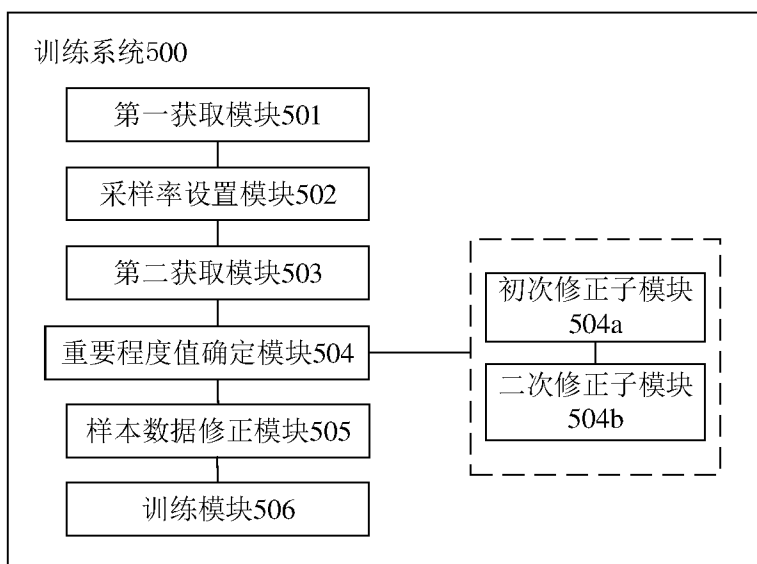


图5

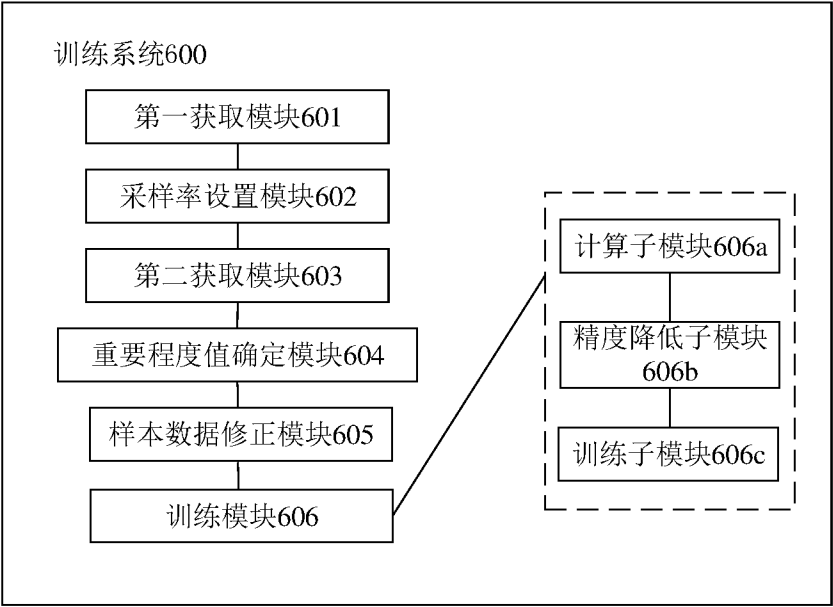


图6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/073719

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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)

CNPAT, CNKI, WPI, EPODOC, IEEE: multiple, learning, training, sampl+, weight, optimiz+, classify, modify

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102156907 A (INST. OF COMPUTING TECH., CAS) 17 August 2011 (17.08.2011) description, paragraphs [0040]-[0059], [0066]-[0074] and [0087]-[0099]	1, 2, 9, 10
Y	CN 103136361 A (CHEN, Yifei) 05 June 2013 (05.06.2013) description, paragraphs [0034]-[0038]	1, 2, 9, 10
A	CN 104166668 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 26 November 2014 (26.11.2014) the whole document	1-16
A	CN 101968851 A (XIDIAN UNIVERSITY) 09 February 2011 (09.02.2011) the whole document	1-16
A	US 2012209612 A1 (INTONOW) 16 August 2012 (16.08.2012) the whole document	1-16

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 05 April 2017	Date of mailing of the international search report 26 April 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LI, Nan Telephone No. (86-10) 62413690

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/073719

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102156907 A	17 August 2011	None	
CN 103136361 A	05 June 2013	None	
CN 104166668 A	26 November 2014	None	
CN 101968851 A	09 February 2011	None	
US 2012209612 A1	16 August 2012	TW 201246183 A	16 November 2012
		TW 201447871 A	16 December 2014
		WO 2012108975 A2	16 August 2012
		CN 103403710 A	20 November 2013
		JP 2014506686 A	17 March 2014
		EP 2673775 A2	18 December 2013

A. 主题的分类 G06F 17/30 (2006.01) 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)) CNPAT, CNKI, WPI, EPODOC, IEEE: 学习, 训练, 样本, 采样, 权重, 加权, 优化, 分类, 修正, 权值, 倍数, learning, training, sampl+, weight, optimiz+, classify, modify																				
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>Y</td> <td>CN 102156907 A (中国科学院计算技术研究所) 2011年 8月 17日 (2011 - 08 - 17) 说明书第0040-0059, 0066-0074, 0087-0099段</td> <td>1, 2, 9, 10</td> </tr> <tr> <td>Y</td> <td>CN 103136361 A (陈一飞) 2013年 6月 5日 (2013 - 06 - 05) 说明书第0034-0038段</td> <td>1, 2, 9, 10</td> </tr> <tr> <td>A</td> <td>CN 104166668 A (南京邮电大学) 2014年 11月 26日 (2014 - 11 - 26) 全文</td> <td>1-16</td> </tr> <tr> <td>A</td> <td>CN 101968851 A (西安电子科技大学) 2011年 2月 9日 (2011 - 02 - 09) 全文</td> <td>1-16</td> </tr> <tr> <td>A</td> <td>US 2012209612 A1 (INTONOW) 2012年 8月 16日 (2012 - 08 - 16) 全文</td> <td>1-16</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	Y	CN 102156907 A (中国科学院计算技术研究所) 2011年 8月 17日 (2011 - 08 - 17) 说明书第0040-0059, 0066-0074, 0087-0099段	1, 2, 9, 10	Y	CN 103136361 A (陈一飞) 2013年 6月 5日 (2013 - 06 - 05) 说明书第0034-0038段	1, 2, 9, 10	A	CN 104166668 A (南京邮电大学) 2014年 11月 26日 (2014 - 11 - 26) 全文	1-16	A	CN 101968851 A (西安电子科技大学) 2011年 2月 9日 (2011 - 02 - 09) 全文	1-16	A	US 2012209612 A1 (INTONOW) 2012年 8月 16日 (2012 - 08 - 16) 全文	1-16
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
Y	CN 102156907 A (中国科学院计算技术研究所) 2011年 8月 17日 (2011 - 08 - 17) 说明书第0040-0059, 0066-0074, 0087-0099段	1, 2, 9, 10																		
Y	CN 103136361 A (陈一飞) 2013年 6月 5日 (2013 - 06 - 05) 说明书第0034-0038段	1, 2, 9, 10																		
A	CN 104166668 A (南京邮电大学) 2014年 11月 26日 (2014 - 11 - 26) 全文	1-16																		
A	CN 101968851 A (西安电子科技大学) 2011年 2月 9日 (2011 - 02 - 09) 全文	1-16																		
A	US 2012209612 A1 (INTONOW) 2012年 8月 16日 (2012 - 08 - 16) 全文	1-16																		
☐ 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 2017年 4月 5日		国际检索报告邮寄日期 2017年 4月 26日																		
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 李楠 电话号码 (86-10) 62413690																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/073719

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	102156907	A	2011年 8月 17日	无			
CN	103136361	A	2013年 6月 5日	无			
CN	104166668	A	2014年 11月 26日	无			
CN	101968851	A	2011年 2月 9日	无			
US	2012209612	A1	2012年 8月 16日	TW	201246183	A	2012年 11月 16日
				TW	201447871	A	2014年 12月 16日
				WO	2012108975	A2	2012年 8月 16日
				CN	103403710	A	2013年 11月 20日
				JP	2014506686	A	2014年 3月 17日
				EP	2673775	A2	2013年 12月 18日