(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114611712 B
(45)授权公告日 2022.08.26

(21)申请号 202210501248.0

(22)申请日 2022.05.10

(65)同一申请的已公布的文献号
申请公布号 CN 114611712 A

(43)申请公布日 2022.06.10

(73)专利权人 富算科技(上海)有限公司
地址 200135 上海市浦东新区中国(上海)自由贸易试验区浦东大道1200号2层A区

(72)发明人 尤志强 卞阳

(74)专利代理机构 北京超凡宏宇专利代理事务所(特殊普通合伙) 11463
专利代理师 唐正瑜

(51)Int.Cl.
G06N 20/00(2019.01)
G06F 16/2455(2019.01)
G06F 16/25(2019.01)
G06F 16/28(2019.01)
G06F 16/9535(2019.01)
G06Q 20/40(2012.01)

(56)对比文件
CN 110998516 A,2020.04.10

审查员 张博

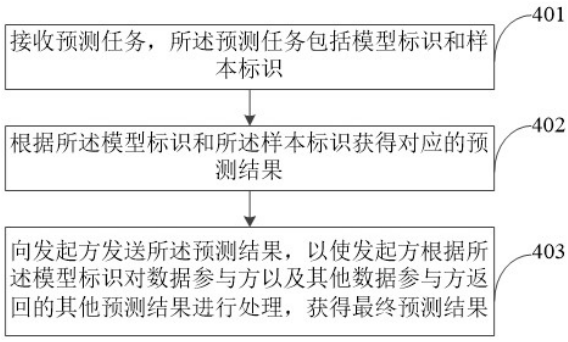
权利要求书2页 说明书11页 附图5页

(54)发明名称

基于异构联邦学习的预测方法、模型生成方法及装置

(57)摘要

本申请提供一种基于异构联邦学习的预测方法、模型生成方法及装置。方法包括:接收预测任务,根据预测任务中的模型标识和样本标识获得预测结果;向发起方发送预测结果,发起方根据模型标识对返回的预测结果进行处理获得最终预测结果;预测结果为利用目标预测模型对样本特征进行处理获得,目标预测模型通过如下方法生成:获取初始预测模型对应的模型文件;初始预测模型采用第一编程语言训练获得;从模型文件中提取初始预测模型的模型参数;将模型参数以数据表的形式存储到Hive数据表中,生成所述目标预测模型,目标预测模型采用第二编程语言实现预测处理;且第一编程语言不同于第二编程语言。本申请可以提高对样本数据预测的效率。



CN 114611712 B

1.一种基于异构联邦学习的预测方法,其特征在于,应用于数据参与方,所述方法包括:

接收预测任务,所述预测任务包括模型标识和样本标识;

根据所述模型标识和所述样本标识获得对应的预测结果;

向发起方发送所述预测结果,以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理,获得最终预测结果;

其中,所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得,所述目标预测模型通过如下方法生成:

获取初始预测模型对应的模型文件;所述初始预测模型采用第一编程语言训练获得;

从所述模型文件中提取所述初始预测模型的模型参数;

将所述模型参数以数据表的形式存储到Hive数据表中,生成所述目标预测模型,所述目标预测模型采用第二编程语言实现预测处理;其中,所述第一编程语言不同于所述第二编程语言。

2.根据权利要求1所述的方法,其特征在于,所述根据所述模型标识和所述样本标识获得对应的预测结果,包括:

根据所述模型标识获取对应的目标预测模型;

根据所述样本标识获取对应的样本数据,并对所述样本数据进行特征提取获得所述样本特征;

根据所述目标预测模型和所述样本特征生成预测结果。

3.根据权利要求1所述的方法,其特征在于,所述根据所述模型标识和所述样本标识获得对应的预测结果,包括:

根据所述模型标识获取对应的目标预测模型,根据所述样本标识获得对应的样本特征;其中,所述样本特征为预先对所述样本标识对应的样本数据进行特征提取后获得的;

根据所述目标预测模型和所述样本特征生成预测结果。

4.根据权利要求1所述的方法,其特征在于,所述根据所述模型标识和所述样本标识获得对应的预测结果,包括:

根据所述模型标识和所述样本标识查询所述预测结果,其中,所述预测结果为预先利用所述模型标识对应的预测模型对所述样本标识对应的样本特征进行处理获得。

5.一种基于异构联邦学习的预测系统,其特征在于,所述预测系统包括发起方和至少一个数据参与方;

所述发起方接收总预测任务,并根据所述总预测任务生成多个预测任务,每个所述预测任务包括模型标识和样本标识;

所述发起方向各所述数据参与方发送对应的所述预测任务;

所述数据参与方用于执行如权利要求1-4任一项所述的方法;

所述发起方在接收到各所述数据参与方返回的所述预测结果后,根据所述模型标识对所述预测结果进行处理,获得最终预测结果。

6.一种基于异构联邦学习的预测装置,其特征在于,包括:

任务接收模块,用于接收预测任务,所述预测任务包括模型标识和样本标识;

结果获得模块,用于根据所述模型标识和所述样本标识获得对应的预测结果;

结果发送模块，用于向发起方发送所述预测结果，以使所述发起方根据所述模型标识对数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；

其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过如下方法生成：

获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；

从所述模型文件中提取所述初始预测模型的模型参数；

将所述模型参数以数据表的形式存储到Hive数据表中，生成所述目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

7.一种电子设备，其特征在于，包括：处理器、存储器和总线，其中，

所述处理器和所述存储器通过所述总线完成相互间的通信；

所述存储器存储有可被所述处理器执行的程序指令，所述处理器调用所述程序指令能够执行如权利要求1-4任一项所述的方法。

8.一种非暂态计算机可读存储介质，其特征在于，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令被计算机运行时，使所述计算机执行如权利要求1-4任一项所述的方法。

基于异构联邦学习的预测方法、模型生成方法及装置

技术领域

[0001] 本申请涉及数据安全技术领域，具体而言，涉及一种基于异构联邦学习的预测方法、模型生成方法及装置。

背景技术

[0002] 联邦学习作为一种数据安全计算技术，特别是联邦学习中的机器学习算法，对于当前金融风控、互联网个性化推荐等业务，具有极高的应用价值。

[0003] 现有的联邦学习系统，一般是使用同种程序语言开发模型训练模块和预测模块，在机器学习、深度学习场景下，python是一种主流的开发语言。虽然python是一种适合快速编程的开发语言，提供较多的第三方功能包以及高级编程接口函数，但由于python是解释型语言，运行速度慢且非常消耗内存。针对数据量较大的情况，其预测效率较低。

发明内容

[0004] 本申请实施例的目的在于提供一种基于异构联邦学习的预测方法、模型生成方法及装置，用以提高对大数据量样本或大流量场景进行预测的效率。

[0005] 第一方面，本申请实施例提供一种基于异构联邦学习的预测方法，应用于数据参与方，所述方法包括：接收预测任务，所述预测任务包括模型标识和样本标识；根据所述模型标识和所述样本标识获得对应的预测结果；向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过如下方法生成：获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；从所述模型文件中提取所述初始预测模型的模型参数；将所述模型参数以数据表的形式存储到Hive数据表中，生成所述目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0006] 本申请实施例通过在基于联邦学习获得的初始预测模型基础上增加转换层，即，将初始预测模型对应的模型参数存储到Hive表中，从而在对样本数据进行预测时，可以不依赖模型训练时所采用的编程语言，采用Hive大数据处理框架可以提高对样本数据预测的效率。

[0007] 在任一实施例中，所述根据所述模型标识和所述样本标识获得对应的预测结果，包括：根据所述模型标识获取对应的目标预测模型；根据所述样本标识获取对应的样本数据，并对所述样本数据进行特征提取获得所述样本特征；根据所述目标预测模型和所述样本特征生成预测结果。本申请实施例可以实现对预测任务的线上实时预测。

[0008] 在任一实施例中，所述根据所述模型标识和所述样本标识获得对应的预测结果，包括：根据所述模型标识获取对应的目标预测模型，根据所述样本标识获得对应的样本特征；其中，所述样本特征为预先对所述样本标识对应的样本数据进行特征提取后获得的；根

据所述目标预测模型和所述样本特征生成预测结果。本申请实施例通过预先对样本数据进行特征处理，当需要进行预测时，直接将样本特征输入模型即可获得预测结果，提高了预测的效率。

[0009] 在任一实施例中，所述根据所述模型标识和所述样本标识获得对应的预测结果，包括：根据所述模型标识和所述样本标识查询所述预测结果，其中，所述预测结果为预先利用所述模型标识对应的预测模型对所述样本标识对应的样本特征进行处理获得。本申请实施例通过预先利用预测模型对样本数据进行预测，获得预测结果，并将预测结果进行存储，在接收到预测任务后，可以直接将预测结果返回，大大提高了返回预测结果的效率。

[0010] 第二方面，本申请实施例提供一种预测模型生成方法，应用于联邦学习架构中的数据参与方，所述方法包括：获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；从所述模型文件中提取所述初始预测模型的模型参数；将所述模型参数以数据表的形式存储到Hive数据表中，生成目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0011] 本申请实施例通过在基于联邦学习获得的初始预测模型基础上增加转换层，即，将初始预测模型对应的模型参数存储到Hive表中，从而在对样本数据进行预测时，可以不依赖模型训练时所采用的编程语言，采用Hive大数据处理框架可以提高对样本数据预测的效率。

[0012] 第三方面，本申请实施例提供一种基于异构联邦学习的预测系统，所述预测系统包括发起方和至少一个数据参与方；所述发起方接收总预测任务，并根据所述总预测任务生成多个预测任务，每个所述预测任务包括模型标识和样本标识；所述发起方向各所述数据参与方发送对应的所述预测任务；所述数据参与方用于执行第一方面所述的方法；所述发起方在接收到各所述数据参与方返回的所述预测结果后，根据所述模型标识对所述预测结果进行处理，获得最终预测结果。

[0013] 第四方面，本申请实施例提供一种基于异构联邦学习的预测装置，包括：任务接收模块，用于接收预测任务，所述预测任务包括模型标识和样本标识；结果获得模块，用于根据所述模型标识和所述样本标识获得对应的预测结果；结果发送模块，用于向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过如下方法生成：获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；从所述模型文件中提取所述初始预测模型的模型参数；将所述模型参数以数据表的形式存储到Hive数据表中，生成所述目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0014] 第五方面，本申请实施例提供一种预测模型生成装置，包括：文件获取模块，用于获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；参数提取模块，用于从所述模型文件中提取所述初始预测模型的模型参数；参数转换模块，用于将所述模型参数转换为数据表的形式，并存储到Hive数据表中，生成目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0015] 第六方面，本申请实施例提供一种电子设备，包括：处理器、存储器和总线，其中，所述处理器和所述存储器通过所述总线完成相互间的通信；所述存储器存储有可被所述处理器执行的程序指令，所述处理器调用所述程序指令能够执行第一方面或第二方面的方法。

[0016] 第七方面，本申请实施例提供一种非暂态计算机可读存储介质，包括：所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令使所述计算机执行第一方面或第二方面的方法。

[0017] 本申请的其他特征和优点将在随后的说明书阐述，并且，部分地从说明书中变得显而易见，或者通过实施本申请实施例了解。本申请的目的和其他优点可通过在所写的说明书、权利要求书、以及附图中所特别指出的结构来实现和获得。

附图说明

[0018] 为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本申请的某些实施例，因此不应被看作是对范围的限定，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他相关的附图。

[0019] 图1为对比实施例提供的同构联邦学习预测服务框架示意图；

[0020] 图2为本申请实施例提供的一种预测模型架构图；

[0021] 图3为本申请实施例提供的一种预测模型生成方法流程示意图；

[0022] 图4为本申请实施例提供的一种基于异构联邦学习的预测方法流程示意图；

[0023] 图5为本申请实施例提供的实时流基本服务架构示意图；

[0024] 图6为本申请实施例提供的实时数据联邦预测系统架构图；

[0025] 图7为本申请实施例提供的一种基于异构联邦学习的预测系统架构示意图；

[0026] 图8为本申请实施例提供的基于异构联邦学习的预测装置结构示意图；

[0027] 图9为本申请实施例提供的预测模型生成装置结构示意图；

[0028] 图10为本申请实施例提供的电子设备实体结构示意图。

具体实施方式

[0029] 联邦机器学习(Federated machine learning/Federated Learning)，又名联邦学习，联合学习，联盟学习。联邦机器学习是一个机器学习框架，能有效帮助多个机构在满足用户隐私保护、数据安全和政府法规的要求下，进行数据使用和机器学习建模。根据参与各方数据源分布的情况不同，联邦学习可以被分为三类：横向联邦学习、纵向联邦学习、联邦迁移学习。本申请实施例主要针对纵向联邦学习进行描述：

[0030] 纵向联邦学习：在两个数据集的用户重叠较多而用户特征重叠较少的情况下，我们把数据集按照纵向(即特征维度)切分，并取出双方用户相同而用户特征不完全相同的那部分数据进行训练。

[0031] 图1为对比实施例提供的同构联邦学习预测服务框架示意图，如图1所示，联邦学习训练模块使用python语言进行开发，模型学习训练结束之后，再由python开发的模型预测服务模块，加载训练好的模型进行预测服务。这种同构的联邦学习服务，由于受到框架及

开发语言限制,导致很难应对海量数据的计算,针对离线处理或者在线推理,性能都大打折扣。

[0032] 针对以上问题,本申请提出一种异构的纵向联邦学习算法预测服务系统,能够支持对所述联邦系统进行改造或者重建,搭建一套高可用的联邦学习应用系统,支持大数据计算,可应对海量实时数据计算,具备高效的算法预测性能,能够落地于工业级层面的企业业务系统中。

[0033] 应当说明的是,前面提到的"同构"是指开发用于训练的模型所使用的编程语言与模型预测服务语言一致,如都是使用python等。而"异构"是指系统中开发模型训练模型的编程语言可以与模型预测服务独立,不耦合,开发训练可以使用python等,而预测服务可以不依赖于相似的框架或者语言,可以采用业内更成熟的大数据处理工具如Hivesql、flinksql、spark、Tensorflow Serving等框架或者技术实现。

[0034] 在介绍本申请的方案之前,先对本申请中所涉及的概念进行介绍:

[0035] Hive:Hive是建立在Hadoop体系架构上的一层SQL抽象,通过SQL语言就可以进行海量数据的处理、分析和统计工作,Hive SQL先被SQL解析器进行解析然后被Hive框架解析成一个MapReduce可执行计划,并按照该计划生成MapReduce任务后交给Hadoop集群处理。Hive可以执行离线任务处理,能够应对几十亿量级的数据处理。

[0036] Flink是一个分布式计算框架,为分布式、高性能、随时可用立即准确的流处理应用程序打造的开源处理框架,快速处理任意规模的数据。能够支持上万亿的Event处理,维护TB级别的处理状态,运行在上千个核心的集群中,用于对无界和有界数据进行有状态计算。Flink的基本数据模型是数据流,及事件(Event)的序列。流可以是无边界的无限流,即所谓的流处理。可以说,有边界的有限流,这样就是批处理。在Flink的流执行模式中,一个事件在一个节点处理完后的输出就可以发到下一个节点立即处理。这样执行引擎并不会引入额外的延迟。

[0037] Flink on Hive:Flink使用HiveCatalog可以通过批或者流的方式来处理Hive中的表。这就意味着Flink既可以作为Hive的一个批处理引擎,也可以通过流处理的方式来读写Hive中的表,从而为实时计算应用和流批一体的落地实践奠定了坚实的基础。Flink支持以批处理(Batch)和流处理(Streaming)的方式写入Hive表。

[0038] 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描述。

[0039] 图2为本申请实施例提供的一种预测模型架构图,如图2所示,在利用第一编程语言对模型进行训练完成后,获得初始预测模型。例如:第一编程语言为python语言,在获得训练好的初始预测模型后,在终端会以pickle的方式序列化,获得.pkl的二进制模型文件。本申请实施例的预测模型架构中增加了转换层,通过转换层将初始预测模型的模型文件转换成数据表的形式,并将数据表存储在Hive数据表中,并且预测模型中集成了Hive-flink流批一体的联邦预测服务模块,使联邦系统具备大数据处理和海量数据实时计算能力。

[0040] 应当说明的是,本申请实施例中对模型训练所使用的第一编程语言不做具体限定,其可以是python语言,也可以是其他不利于进行大数据预测的语言。

[0041] 图3为本申请实施例提供的一种预测模型生成方法流程示意图,如图3所示,应当说明的是,该方法应用于联邦学习系统中的数据参与方,若联邦学习系统中包括多个数据参与方,且每个数据参与方中存储有完整的预测模型中的一部分,则需要在多个数据参与

方中均执行下述步骤。另外，数据参与方可以是终端，也可以是服务器，终端可以为台式电脑、笔记本电脑、平板电脑等智能电子设备。该方法包括：

[0042] 步骤301：获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；

[0043] 步骤302：从所述模型文件中提取所述初始预测模型的模型参数；

[0044] 步骤303：将所述模型参数以数据表的形式存储到Hive数据表中，生成目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0045] 在步骤301中，初始预测模型是预先采用第一编程语言训练获得，其中，第一编程语言可以是python语言，也可以为其他编程语言，本申请实施例对此不作具体限定。对于第一编程语言为python语言的情况，其初始预测模型对应的模型文件为.pkl格式的二进制文件。那么对于不同的第一编程语言，其对应的模型文件的格式不同。

[0046] 假设A方为发起方，且同时也是数据参与方，具有特征X_a和标签y。B方为数据参与方，只具有特征X_b。联邦学习训练结束，A方、B方分别持有完整模型Model的部分模型的参数数据，这里分别称为Model_A, Model_B，其中Model_A, Model_B结合在一起才是完整模型。Model_A, Model_B会被序列化为Model_A.pkl, Model_B.pkl。Model_A.pkl存在A方本地的持久化层，比如mysql、fastdfs等。Model_B.pkl存在B方本地的持久化层。A方和B方分别加载各自本地持久化层中的二进制格式的模型文件。

[0047] 在步骤302中，数据参与方从模型文件中提取训练好的模型参数，可以理解的是，在数据参与方预先配置有提取模型参数的功能。例如：A方获得Model_A 的模型参数para_A, B方获得Model_B的参数para_B。另外，各方可以从模型文件中提取到对应的训练任务Id，在之后该任务id会被作为model_id使用，来唯一区分模型参数。para_A的形式为：Intercept, W_a1, W_a2, ..., W_ak。Para_B的形式为：W_b1, W_b2, ..., W_bm。可以理解的是，Intercept指代广义线性模型中的截距项，其可以在A方，也可以在B方。

[0048] 在步骤303中，数据参与方预先创建Hive数据表，然后将提取到的模型参数存储在Hive数据表中。A方Hive数据表结构如表1所示，B方Hive数据表结构如表2所示：

[0049] 表1

[0050]

Model_id	Intercept	W_a1	W_a2	W_a3	...	W_ak
7rdq2834yuiwqeu9321483243	0.2	0.01	0.03	0.1	...	0.13

[0051] 表2

[0052]

Model_id	W_b1	W_b2	W_b3	...	W_bm
7rdq2834yuiwqeu9321483243	0.012	0.01	0.3	...	0.21

[0053] 数据参与方将模型参数以数据表的形式存储到Hive数据表中后便获得了目标预测模型。可以理解的是，目标预测模型在预测过程中，采用的第二编程语言可以为：Hivesql、Java、Scala等。并且，第二编程语言与第一编程语言不同。

[0054] 本申请实施例通过在基于联邦学习获得的初始预测模型基础上增加转换层，即，将初始预测模型对应的模型参数存储到Hive表中，从而在对样本数据进行预测时，可以不依赖模型训练时所采用的编程语言，采用Hive大数据处理框架可以提高对样本数据预测的效率。

[0055] 图4为本申请实施例提供的一种基于异构联邦学习的预测方法流程示意图，如图4所示，该方法应用于联邦学习系统中的数据参与方，若联邦学习系统中包括多个数据参与方，且每个数据参与方中存储有完整的预测模型中的一部分，则需要在多个数据参与方中均执行下述步骤。另外，数据参与方可以是终端，也可以是服务器，终端可以为台式电脑、笔记本电脑、平板电脑等智能电子设备。该方法包括：

[0056] 步骤401：接收预测任务，所述预测任务包括模型标识和样本标识。

[0057] 其中，数据参与方中可以存储多种预测模型，例如：逻辑回归模型、决策树模型、深度学习模型等，不同的模型对应不同的模型标识。数据参与方中还存储有多种样本数据，不同的样本数据对应不同的样本标识，例如：样本数据可以是银行的客户信息、银行中流水数据、运营商中用户的历史购物数据等。预测任务可以由发起方发送给数据参与方，可以理解的是，对于包括多个数据参与方的情况，数据参与方接收到的预测任务属于总预测任务的一部分，是该数据参与方所需要预测的任务。

[0058] 步骤402：根据所述模型标识和所述样本标识获得对应的预测结果。

[0059] 其中，该预测结果为利用模型标识对应的目标预测模型对样本标识对应的样本数据进行处理后获得的，并且，目标预测模型采用上述预测模型生成方法生成，此处不再赘述。

[0060] 步骤403：向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果。

[0061] 其中，发起方为与数据参与方通信连接的终端或服务器，各个数据参与方在获得预测结果后，将预测结果发送给发起方，由发起方根据模型标识对各个数据参与方返回的预测结果进行综合处理，获得最终预测结果。例如：模型标识对应的模型为逻辑回归模型或决策树二分类模型，则需要将各个数据参与方返回的预测结果进行求和，然后再进行sigmoid转换，深度学习模型往往需要做softmax的计算，而线性回归模型、决策树回归模型则只需要进行求和即可。因此，不同的模型在执行预测值组合转换的时候逻辑上存在差异，因此通过model_id来确定具体的模型类型，进而确定组合的执行逻辑。

[0062] 例如：A方持有的样本特征数据为X_a=[X_a1, X_a2, X_a3, X_a4]，Para_A是由转换层获得的部分模型Model_A的模型权重系数；其中，Para_A=[W_a1,W_a2,W_a3,W_a4]。x_B指的是B方持有的样本特征数据，X_b=[X1_b, X2_b, X3_b]，Para_B是由转换层获得的部分模型Model_B的模型权重系数，且Para_B=[W_b1,W_b2,W_b3]。经过各方本地的预测任务，可以得到对应的部分预测值，比如A方得到predict_value_A = Intercept + X_a1 * W_a1 + X_a2 * W_a2 + X_a3 * W_a3 + X_a4 * W_a4，B方得到predict_value_B= X_b1 * W_b1 + X_b2 * W_b2 + X_b3 * W_b3，而最终的模型预测值为：

[0063] Predict_value = Sigmoid(predict_value_A + predict_value_B)

[0064] 其中，Sigmoid函数形式为：$\sigma(z)=\frac{1}{1+e^{-z}}$。

[0065] 本申请实施例通过在基于联邦学习获得的初始预测模型基础上增加转换层，即，将初始预测模型对应的模型参数存储到Hive表中，从而在对样本数据进行预测时，可以不依赖模型训练时所采用的编程语言，采用Hive大数据处理框架可以提高对样本数据预测的效率。

[0066] 在上述实施例的基础上,针对预测流程,数据参与方可以进行实时在线预测、近实时预测和离线预测三种方式,下面分别针对每种预测方式进行介绍:

[0067] 第一种:实时在线预测,可以理解的是,实时任务一般是秒级或者分钟级执行,当用户产生新的行为数据,短时间内即进行特征提取及预测。实时任务一般用于对用户实时行为严苛的业务场景。

[0068] 第一步:数据参与方根据模型标识从已经存储的目标预测模型中获取对应的目标预测模型;由于目标预测模型的模型参数都是以数据表的形式存储的,因此,可以根据模型标识从Hive数据表中获取对应的模型参数,可以理解的是,获取到模型参数就相当于获取到了目标预测模型。

[0069] 第二步:根据样本标识获取对应的样本数据,并对样本数据进行特征提取获得样本特征;应当说明的是,不同的样本标识提取的样本特征的方法不同,可以预先在数据参与方中配置各样本标识对应的特征提取脚本,通过该脚本可以对样本数据进行特征提取,以获得样本特征。例如:数据参与方中存储了某视频运营商的用户观看视频的历史数据,特征提取脚本中设定了特征提取的规则,例如:从用户观看视频的历史数据中提取观看爱情类视频的次数、观看动作类视频的次数、观看爱情类视频的总时长等等。

[0070] 第三步:根据目标预测模型和样本特征生成预测结果。

[0071] 实时预测方法所对应的基本的服务架构如图5所示:首先收集日志、埋点数据等,将其写入到 Kafka 里面,经过实时计算平台进行处理,将 数据仓库的运营数据存储(Operational Data Store,ODS)层中的明细数据抽取出来,并与DIM层对应的维度关联等操作,执行相应的联邦推理计算,将结果写入到 Redis 等,再通过数据服务提供给业务使用。

[0072] 图6为本申请实施例提供的实时数据联邦预测系统架构图,如图6所示,发起方可以为推荐系统或金融风控系统,还可以为其他系统,本申请实施例对此不作具体限定。业务系统是指数据参与方,业务系统实时从实时消息系统中获取样本数据,业务系统在接收到预测任务后,对样本数据进行Flink分钟级甚至是秒级的联邦推理计算,计算方法如上所述。在获得预测结果后,通过实时数据平台将预测结果返回给发起方。

[0073] 本申请实施例提供的一种基于python与hive-flink流批一体的联邦学习架构,能够支持大规模数据下的离线批处理以及海量实时流处理。

[0074] 第二种:近实时预测

[0075] 第一步:数据参与方对样本数据进行特征提取,获得样本特征。可以理解的是,样本特征的提取方法与上述实施例一致,此处不再赘述。另外,数据参与方可以在接收到样本数据后,便对样本数据进行特征提取,将提取的样本特征进行存储,为后续预测任务做准备。近实时场景下,执行任务的时间间隔以小时计,可以是每小时执行一次特征提取动作。新获取的特征数据将取代老的特征数据,作为样本id对应的最新特征数据。可以理解的是,为了便于对样本数据进行区分,可以为每个样本特征关联样本标识。

[0076] 第二步:数据参与方在接收到预测任务后,根据预测任务中的模型标识从已经存储的目标预测模型中获取对应的目标预测模型;由于目标预测模型的模型参数都是以数据表的形式存储的,因此,可以根据模型标识从Hive数据表中获取对应的模型参数,可以理解的是,获取到模型参数就相当于获取到了目标预测模型。

[0077] 第三步:根据样本标识获取对应的样本特征。可以理解的是,获取样本特征的步骤与获取目标预测模型的步骤没有先后顺序,可以同时进行,也可以是获取样本特征在获取目标预测模型之前或之后。

[0078] 第四步:根据所述目标预测模型和所述样本特征生成预测结果。生成预测结果的方法与上述实施例一致,此处不再赘述。

[0079] 第三种:离线预测

[0080] 第一步:数据参与方对样本数据进行特征提取,获得样本特征。可以理解的是,样本特征的提取方法与上述实施例一致,此处不再赘述。另外,数据参与方可以在接收到样本数据后,便对样本数据进行特征提取,将提取的样本特征进行存储,为后续预测任务做准备。离线批量预测一般是按天进行,可以配置成日任务,每天的某个固定时间段执行特征处理提取。计算得到的特征数据作为样本id对应的特征信息,用于联邦批量预测日任务计算。可以理解的是,为了便于对样本数据进行区分,可以为每个样本特征关联样本标识。

[0081] 第二步:数据参与方利用其内部存储的目标预测模型对样本特征进行处理,获得预测结果,其中,若数据参与方中包括了多个目标预测模型,则可以分别利用每个预测模型均计算出一个预测结果,并将该预测结果进行存储,可以理解的是,在存储时,可以将预测结果、目标预测模型对应的模型标识和样本标识对应存储到预测结果表中。其存储的格式如表3所示:

[0082] 表3

[0083]

Model_id	Sample_id	Predict_value
7rdq2834yuiwqeu9321483243	123456	0.8

[0084] 第三步:数据参与方接收预测任务,根据预测任务中的模型标识和样本标识从存储的预测结果表中获取对应的预测结果。

[0085] 应当说明的是,本申请实施例不限于hive-flink的流批一体服务,还可以是spark、Tensorflow Serving等,只需要在转换层和计算层做相应的适配。

[0086] 图7为本申请实施例提供的一种基于异构联邦学习的预测系统架构示意图,如图7所示,应当说明的是,发起方也可以同时是数据参与方,发起方用于接收总预测任务,并根据总预测任务为每个数据参与方生成对应的预测任务。可以理解的是,为了便于画图,本申请实施例只画出了两个数据参与方,在实际应用中,数据参与方的数量可以是更多或更少,本申请实施例对此不作具体限定。发起方在生成预测任务后,将预测任务发送给各个数据参与方,由数据参与方根据预测任务获得对应的预测结果,可以理解的是,数据参与方获得预测结果的方法可以参见上述各个方法实施例,此处不再赘述。各个数据参与方在获得预测结果后,分别将各自的预测结果返回给发起方,由发起方根据模型标识对预测结果进行处理,获得最终预测结果。可以理解的是,发起方获得最终预测结果的方法参见上述实施例,此处不再赘述。

[0087] 应当说明的是,若发起方同时也是数据参与方,那么发起方也需要根据自己的预测任务进行预测,获得预测结果。

[0088] 本申请实施例提供的基于异构联邦学习的预测系统既能基于python等语言开发训练模型,又能使用大数据技术功能实现大规模数据预测以及海量实时数据推理服务。联邦学习系统不再局限于某种特定语言进行开发应用,有效利用训练和预测场景下更适合的

架构来提供工业级落地应用能力。

[0089] 图8为本申请实施例提供的基于异构联邦学习的预测装置结构示意图，该装置可以是电子设备上的模块、程序段或代码。应理解，该装置与上述图4方法实施例对应，能够执行图4方法实施例涉及的各个步骤，该装置具体的功能可以参见上文中的描述，为避免重复，此处适当省略详细描述。所述装置包括：任务接收模块801、结果获得模块802和结果发送模块803，其中：

[0090] 任务接收模块801用于接收预测任务，所述预测任务包括模型标识和样本标识；

[0091] 结果获得模块802用于根据所述模型标识和所述样本标识获得对应的预测结果；

[0092] 结果发送模块803用于向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；

[0093] 其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过如下方法生成：

[0094] 获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；

[0095] 从所述模型文件中提取所述初始预测模型的模型参数；

[0096] 将所述模型参数以数据表的形式存储到Hive数据表中，生成所述目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0097] 在上述实施例的基础上，结果获得模块802具体用于：

[0098] 根据所述模型标识获取对应的目标预测模型；

[0099] 根据所述样本标识获取对应的样本数据，并对所述样本数据进行特征提取获得所述样本特征；

[0100] 根据所述目标预测模型和所述样本特征生成预测结果。

[0101] 在上述实施例的基础上，结果获得模块802具体用于：

[0102] 根据所述模型标识获取对应的目标预测模型，根据所述样本标识获得对应的样本特征；其中，所述样本特征为预先对所述样本标识对应的样本数据进行特征提取后获得的；

[0103] 根据所述目标预测模型和所述样本特征生成预测结果。

[0104] 在上述实施例的基础上，结果获得模块802具体用于：

[0105] 根据所述模型标识和所述样本标识查询所述预测结果，其中，所述预测结果为预先利用所述模型标识对应的预测模型对所述样本标识对应的样本特征进行处理获得。

[0106] 图9为本申请实施例提供的预测模型生成装置结构示意图，该装置可以是电子设备上的模块、程序段或代码。应理解，该装置与上述图3方法实施例对应，能够执行图3方法实施例涉及的各个步骤，该装置具体的功能可以参见上文中的描述，为避免重复，此处适当省略详细描述。所述装置包括：文件获取模块901、参数提取模块902和参数转换模块903，其中：

[0107] 文件获取模块901用于获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得；

[0108] 参数提取模块902用于从所述模型文件中提取所述初始预测模型的模型参数；

[0109] 参数转换模块903用于将所述模型参数转换为数据表的形式，并存储到Hive数据表中，生成目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言。

[0110] 图10为本申请实施例提供的电子设备实体结构示意图，如图10所示，所述电子设备，包括：处理器(processor)1001、存储器(memory)1002和总线1003；其中，

[0111] 所述处理器1001和存储器1002通过所述总线1003完成相互间的通信；

[0112] 所述处理器1001用于调用所述存储器1002中的程序指令，以执行上述各方法实施例所提供的方法，例如包括：接收预测任务，所述预测任务包括模型标识和样本标识；根据所述模型标识和所述样本标识获得对应的预测结果；向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过上述实施例的方法生成。

[0113] 处理器1001可以是一种集成电路芯片，具有信号处理能力。上述处理器1001可以是通用处理器，包括中央处理器(Central Processing Unit，CPU)、网络处理器(Network Processor，NP)等；还可以是数字信号处理器(DSP)、专用集成电路(ASIC)、现场可编程门阵列(FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。其可以实现或者执行本申请实施例中公开的各种方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0114] 存储器1002可以包括但不限于随机存取存储器(Random Access Memory，RAM)，只读存储器(Read Only Memory，ROM)，可编程只读存储器(Programmable Read-Only Memory，PROM)，可擦除只读存储器(Erasable Programmable Read-Only Memory，EPROM)，电可擦除只读存储器(Electrically Erasable Programmable Read-Only Memory，EEPROM)等。

[0115] 本实施例公开一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，计算机能够执行上述各方法实施例所提供的方法，例如包括：接收预测任务，所述预测任务包括模型标识和样本标识；根据所述模型标识和所述样本标识获得对应的预测结果；向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过上述实施例的方法生成。

[0116] 本实施例提供一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令使所述计算机执行上述各方法实施例所提供的方法，例如包括：接收预测任务，所述预测任务包括模型标识和样本标识；根据所述模型标识和所述样本标识获得对应的预测结果；向发起方发送所述预测结果，以使所述发起方根据所述模型标识对所述数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果；其中，所述预测结果为利用所述模型标识对应的目标预测模型对所述样本标识对应的样本特征进行处理获得，所述目标预测模型通过上述实施例的方法生成。

[0117] 在本申请所提供的实施例中，应该理解到，所揭露装置和方法，可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，又例如，多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0118] 另外，作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0119] 再者，在本申请各个实施例中的各功能模块可以集成在一起形成一个独立的部分，也可以是各个模块单独存在，也可以两个或两个以上模块集成形成一个独立的部分。

[0120] 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。

[0121] 以上所述仅为本申请的实施例而已，并不用于限制本申请的保护范围，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

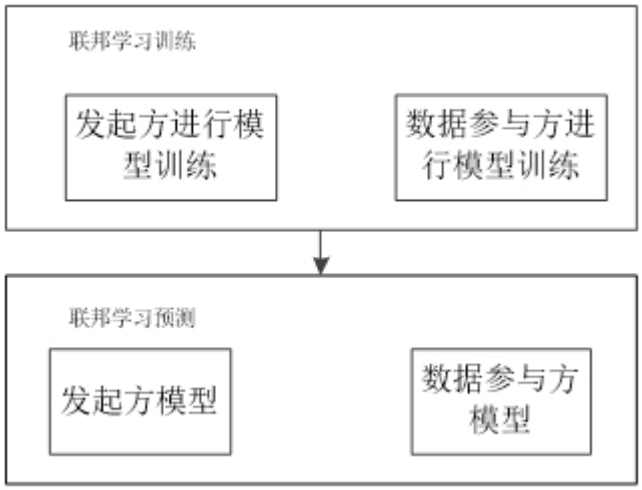

联邦学习训练
发起方进行模型训练
数据参与方进行模型训练
联邦学习预测
发起方模型
数据参与方模型

图1

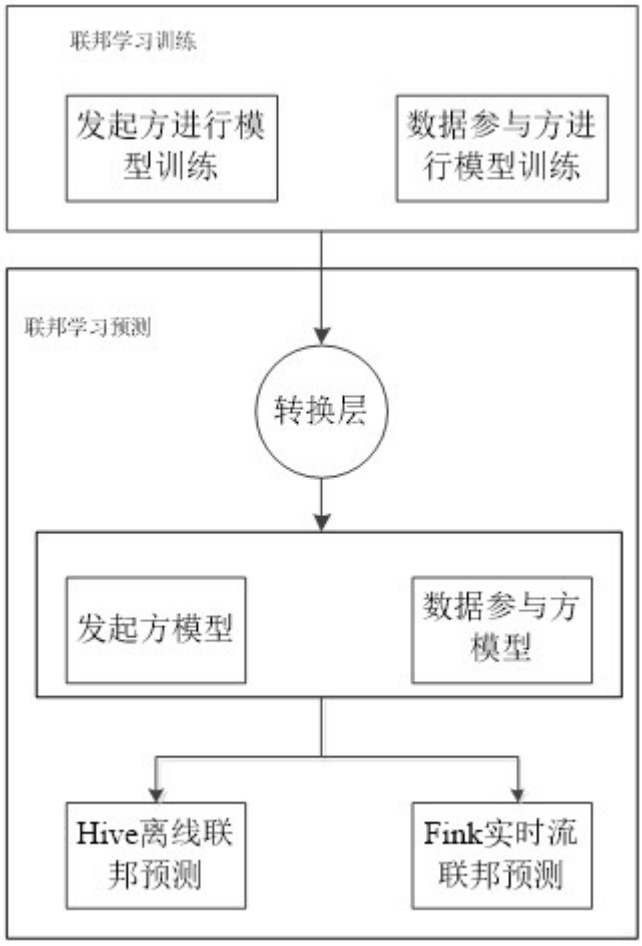

联邦学习训练
发起方进行模型训练
数据参与方进行模型训练
联邦学习预测
转换层
发起方模型
数据参与方模型
Hive离线联邦预测
Fink实时流联邦预测

图2

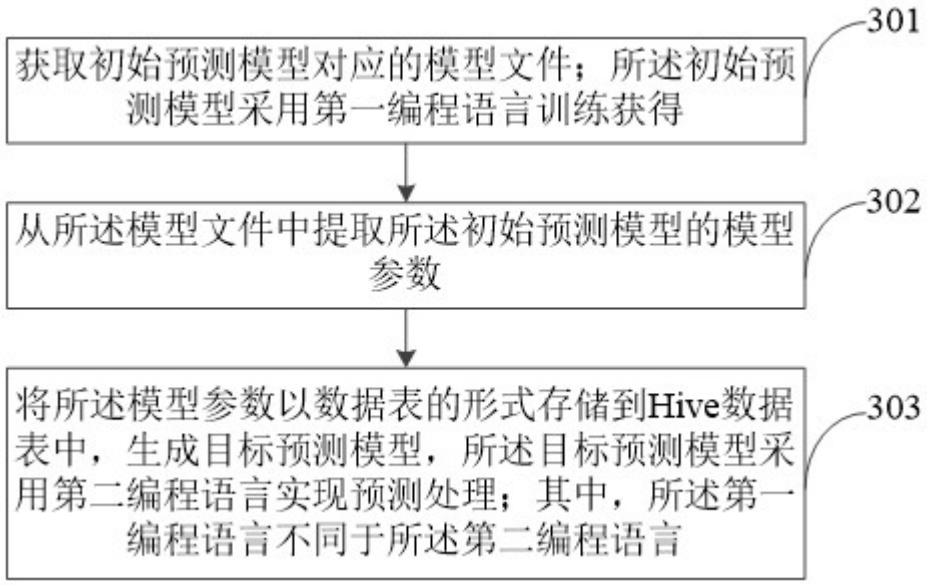

301
获取初始预测模型对应的模型文件；所述初始预测模型采用第一编程语言训练获得
302
从所述模型文件中提取所述初始预测模型的模型参数
303
将所述模型参数以数据表的形式存储到Hive数据表中，生成目标预测模型，所述目标预测模型采用第二编程语言实现预测处理；其中，所述第一编程语言不同于所述第二编程语言

图3

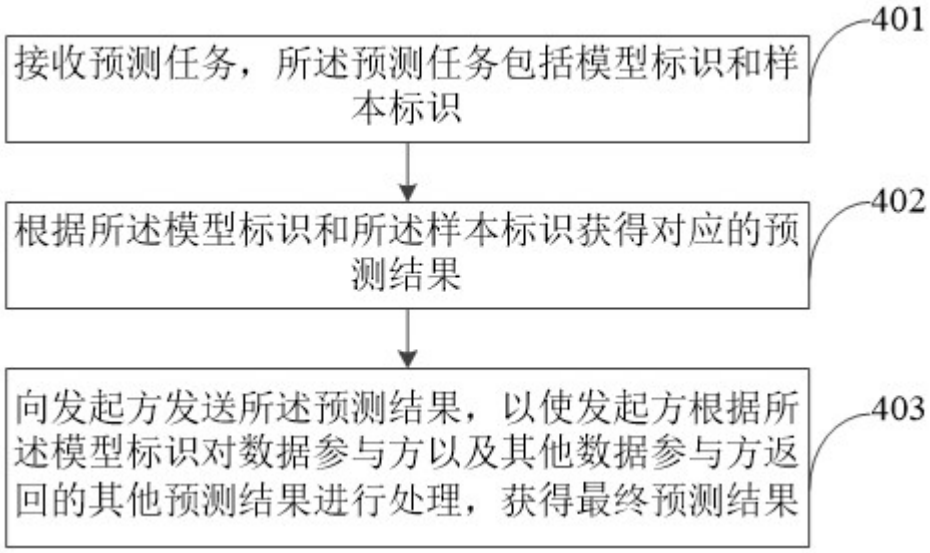

401
接收预测任务，所述预测任务包括模型标识和样本标识
402
根据所述模型标识和所述样本标识获得对应的预测结果
403
向发起方发送所述预测结果，以使发起方根据所述模型标识对数据参与方以及其他数据参与方返回的其他预测结果进行处理，获得最终预测结果

图4

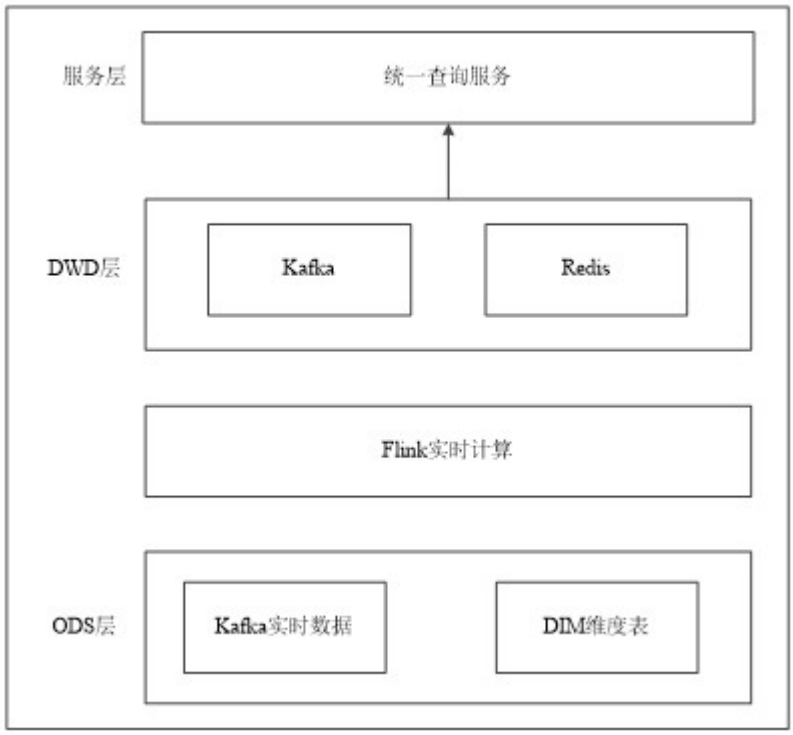

服务层
统一查询服务
DWD层
Kafka
Redis
Flink实时计算
ODS层
Kafka实时数据
DIM维度表

图5

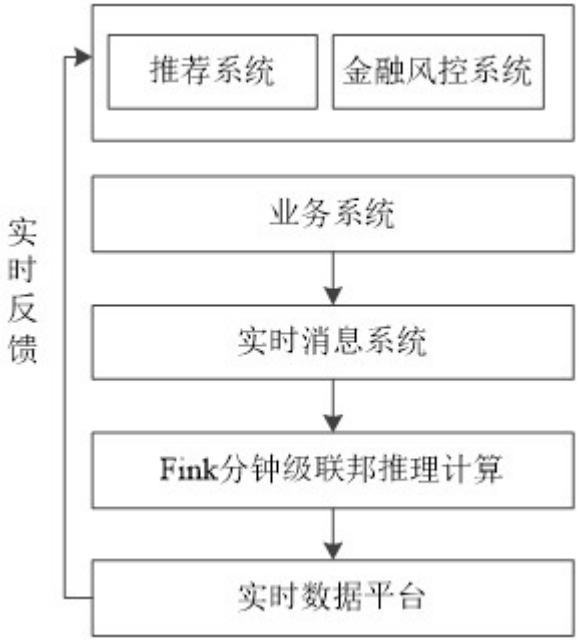

推荐系统
金融风控系统
业务系统
实时消息系统
Fink分钟级联邦推理计算
实时数据平台
实时反馈

图6

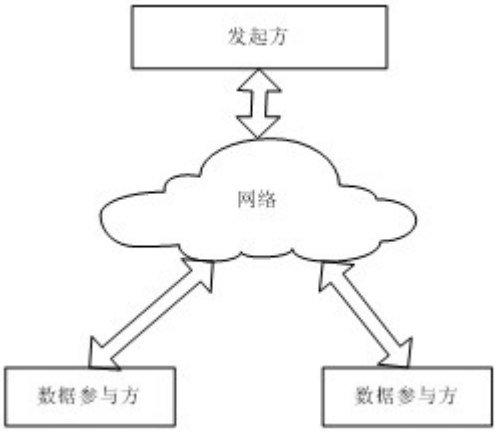

发起方
网络
数据参与方
数据参与方

图7

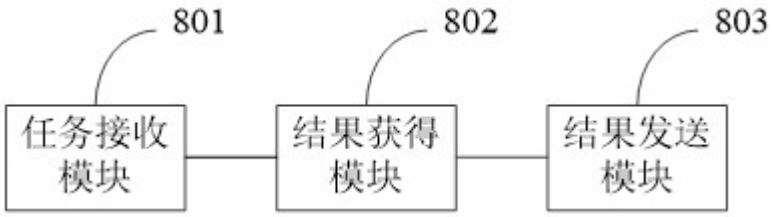

801
802
803
任务接收模块
结果获得模块
结果发送模块

图8

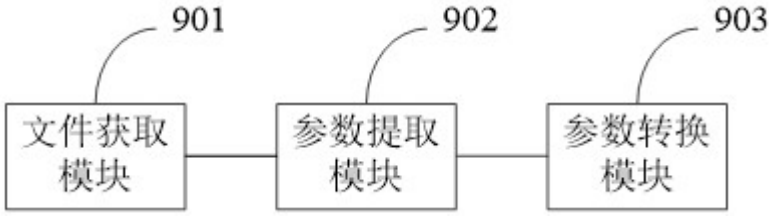

901
902
903
文件获取模块
参数提取模块
参数转换模块

图9

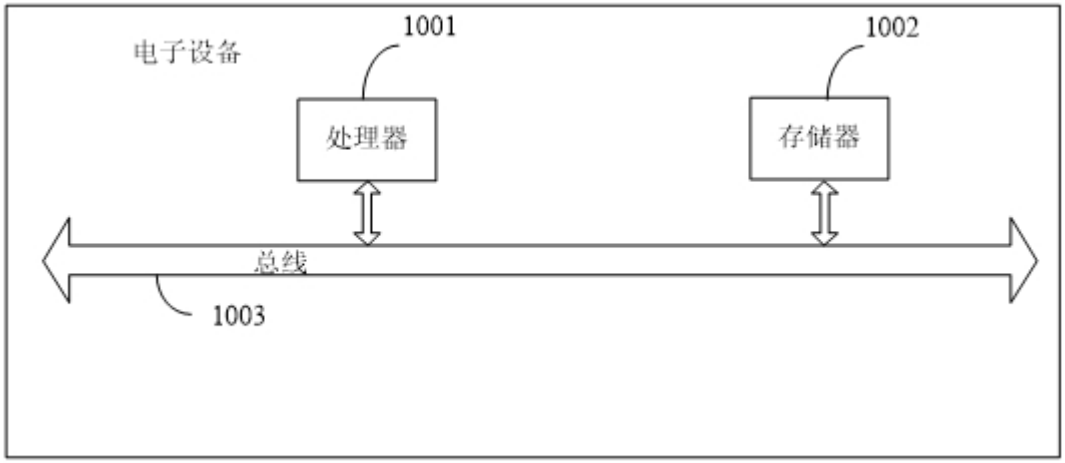


图10