

通信方法、装置及系统

相关申请的交叉引用

本申请要求在 2020 年 09 月 25 日提交中国专利局、申请号为 PCT/CN2020/117940、
申请名称为“通信方法、装置及系统”的中国专利申请的优先权，其全部内容通过引用结
5 合在本申请中。

技术领域

本申请涉及通信技术领域，尤其涉及通信方法、装置及系统。

10 背景技术

机器学习模型的训练通常是通过学习某一组输入特征与输出目标之间的映射来进行的，通过优化某些损失函数来使得机器学习模型的输出结果（即预测值）与实际结果（即
标签值/真实值）之间的误差最小化。当训练出最优模型后，利用该模型的输出对未来的情
况进行预测。在理想情况下，假定未来将使用到的数据与模型训练期间所使用的数据类似，
15 具体来说，可能假设训练时的输入特征和预测时的输入特征的分布保持恒定。但是实际中
这种假设通常不成立，数据的特征会随着时间的推移由于网络部署变化、应用层业务要求
变化、网络实际用户分布变化等而变化，因此，模型的性能（即泛化能力）会随时间推
移逐渐下降。具体的表现可能是，模型的准确率降低，即模型的预测值与真实值之间的误
差变大。

20 以数据分析网元的训练功能和推断功能分离的场景为例，支持训练功能的数据分析网
元（简称训练数据分析网元）无法感知支持推断功能的数据分析网元（简称推断数据分
析网元）中的模型使用效果，且推断数据分析网元没有能力进行模型训练。因此，当模型性
能下降时，若推断数据分析网元继续使用性能下降的模型进行数据分析，会导致数据分
析结果不准确。

25

发明内容

本申请提供通信方法、装置及系统，用以实现在模型性能下降时及时对模型进行再次
训练，从而保证模型性能。

30 第一方面，本申请实施例提供一种通信方法，包括：第一数据分析网元接收来自第二
数据分析网元的第一信息，所述第一信息包括模型的性能报告，所述模型的性能报告用于
指示所述模型的性能的评估结果，或者，所述模型的性能报告用于指示所述模型的性能的
评估结果不满足所述模型的性能指标的要求；所述第一数据分析网元根据所述模型的性能
报告更新所述模型的第一模型信息，获得所述模型的第二模型信息；所述第一数据分析网
元向所述第二数据分析网元发送第二信息，所述第二信息包括所述第二模型信息。

35 基于上述方案，当第二数据分析网元无法完成模型训练时，则第二数据分析网元可以
向第一数据分析网元发送模型的性能报告，从而第一数据分析网元可以根据该模型的性能
报告对模型进行更新，得到模型的第二模型信息，并将第二模型信息发送给第二数据分
析网元，使得第二数据分析网元可以基于第二模型信息更新模型，从而可以实现在模型性能
下降时及时对模型进行训练，进而可以保证模型性能。

在一种可能的实现方法中，所述第一数据分析网元向所述第二数据分析网元发送第三信息，所述第三信息包括所述模型的性能指标，所述模型的性能指标用于获取所述模型的性能的评估结果。

5 基于上述方案，第一数据分析网元可以预先向第二数据分析网元发送模型的性能指标，从而使得第二数据分析网元基于该模型的性能指标生成模型的性能报告，有助于帮助第一数据分析网元判断是否启动模型训练，并提升模型训练后的模型性能。

在一种可能的实现方法中，所述第一数据分析网元向第三数据分析网元发送所述第二信息。

10 基于上述方案，第一数据分析网元不仅可以向第二数据分析网元发送第二信息，还可以向其它使用该模型的网元，如第三数据分析网元发送该第二信息，使得第三数据分析网元也可以利用第二模型信息更新模型，提升模型使用效果。

在一种可能的实现方法中，所述第一数据分析网元接收来自第二数据分析网元的第一信息，包括：所述第一数据分析网元通过网络存储网元接收来自所述第二数据分析网元的所述第一信息。所述第一数据分析网元向所述第二数据分析网元发送第二信息，包括：所述
15 所述第一数据分析网元通过网络存储网元向所述第二数据分析网元发送所述第二信息。

基于上述方案，可以通过网络存储网元作为中间网元，实现第一数据分析网元与第二数据分析网元之间的模型更新交互，这可以适用于第一数据分析网元与第二数据分析网元之间没有接口的场景。

20 第二方面，本申请实施例提供一种通信方法，包括：第二数据分析网元向第一数据分析网元发送第一信息，所述第一信息包括模型的性能报告，所述模型的性能报告用于指示所述模型的性能的评估结果，或者，所述模型的性能报告用于指示所述模型的性能的评估结果不满足所述模型的性能指标的要求；所述第二数据分析网元接收来自所述第一数据分析网元的第二信息，所述第二信息包括所述模型的第二模型信息，所述第二模型信息是根据所述模型的性能报告更新所述模型的第一模型信息得到的；所述第二数据分析网元根据
25 所述第二模型信息，更新所述模型。

基于上述方案，当第二数据分析网元无法完成模型训练时，则第二数据分析网元可以向第一数据分析网元发送模型的性能报告，从而第一数据分析网元可以根据该模型的性能报告对模型进行更新，得到模型的第二模型信息，并将第二模型信息发送给第二数据分析网元，使得第二数据分析网元可以基于第二模型信息更新模型，从而可以实现在模型性能
30 下降时及时对模型进行再次训练，进而可以保证模型性能。

在一种可能的实现方法中，所述第二数据分析网元接收来自所述第一数据分析网元的第三信息，所述第三信息包括所述模型的性能指标，所述模型的性能指标用于获取所述模型的性能的评估结果。

35 基于上述方案，第一数据分析网元可以预先向第二数据分析网元发送模型的性能指标，从而使得第二数据分析网元基于该模型的性能指标生成模型的性能报告，有助于帮助第一数据分析网元判断是否启动模型训练，并提升模型训练的准确性。

在一种可能的实现方法中，所述第二数据分析网元向第一数据分析网元发送第一信息，包括：所述第二数据分析网元通过网络存储网元向所述第一数据分析网元发送所述第一信息；所述第二数据分析网元接收来自所述第一数据分析网元的第二信息，包括：所述第二
40 数据分析网元通过网络存储网元接收来自所述第一数据分析网元的第二信息。

基于上述方案，可以通过网络存储网元作为中间网元，实现第一数据分析网元与第二数据分析网元之间的模型更新交互，这可以适用于第一数据分析网元与第二数据分析网元之间没有接口的场景。

5 基于上述第一方面、或第一方面的任意实现方法、或第二方面、或第二方面的任意实现方法：

在一种可能的实现方法中，所述模型性能指标包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性。

10 在一种可能的实现方法中，所述第三信息还包括以下一项或多项：分析类型标识、所述模型的标识、子模型的标识，所述分析类型标识用于指示所述模型的分析类型。

在一种可能的实现方法中，所述第三信息还包括以下一项或多项：上报周期、门限信息，所述上报周期用于指示上报所述模型的性能报告的时间，所述门限信息用于指示上报所述模型的性能报告的条件。

15 基于上述方案，第一数据分析网元可以指示第二数据分析网元上报模型的性能报告的时间和/或条件，从而实现有条件的上报，可以节约资源开销。

在一种可能的实现方法中，所述第一信息还包括所述模型的性能报告对应的以下一项或多项信息：时间、区域、切片。

基于上述方案，当第一信息还包括模型的性能报告对应的时间、区域或切片，则可以提升第一数据分析网元执行模型再次训练后的模型性能。

20 在一种可能的实现方法中，所述第二信息还包括以下一项或多项：所述模型的标识、子模型的标识、所述模型的性能评估结果、所述模型的性能评估结果对应的硬件能力信息、所述模型的大小、所述模型的推理时长。

25 基于上述方案，第一数据分析网元将模型的性能评估结果、模型的性能评估结果对应的硬件能力信息、模型的大小或模型的推理时长中的一个或多个发送给第二数据分析网元，有助于第二数据分析网元确定是否使用该模型，进而可以减少资源开销的浪费。

30 第三方面，本申请实施例提供一种通信方法，包括：第一数据分析网元将模型的第一信息更新为所述模型的第二信息；所述第一数据分析网元确定所述模型的第二信息的索引信息，所述第二信息的索引信息包括第一标识信息，所述第一标识信息用于指示所述模型的第二信息；所述第一数据分析网元向所述第二数据分析网元发送所述第二信息的索引信息，所述第二信息的索引信息用于所述模型的第二信息的获取。其中，模型的第二信息的索引信息，也可以称为第二信息对应的模型索引信息。

35 基于上述方案，第一数据分析网元对模型进行更新，得到模型的第二信息后，可以将该第二信息的索引信息发送给第二数据分析网元，从而第二数据分析网元可以根据该索引信息获取到新的模型信息，即该第二信息，进而第二数据分析网元可以根据新的模型信息更新模型，实现模型性能提升。

在一种可能的实现方法中，所述第二信息的索引信息还包括以下一项或多项：所述模型对应的分析类型标识、所述模型的标识、所述模型的第二信息的版本信息。

40 在一种可能的实现方法中，所述第一数据分析网元从所述第二数据分析网元接收所述模型的第一信息的索引信息，所述第一信息的索引信息包括第二标识信息，所述第二标识信息用于指示所述模型的第一信息；所述第一数据分析网元根据所述第一信息的索引信息

获取所述模型的第一信息。

在一种可能的实现方法中，所述第一信息的索引信息还包括以下一项或多项：所述模型对应的分析类型标识、所述模型的标识、所述模型的第一信息的版本信息。

在一种可能的实现方法中，第一数据分析网元将模型的第一信息更新为所述模型的第二信息，包括：所述第一数据分析网元从所述第二数据分析网元获取第一请求，所述第一请求用于更新所述模型的第一信息，所述第一请求包括所述模型的第一信息的索引信息；所述第一数据分析网元根据所述第一信息的索引信息获取所述模型的第一信息；所述第一数据分析网元更新所述模型的第一信息，得到所述模型的第二信息。

在一种可能的实现方法中，第一数据分析网元从第二数据分析网元接收模型的第一信息的索引信息，包括：所述第一数据分析网元向所述第二数据分析网元发送第二请求，所述第二请求用于请求所述模型的第一信息的索引信息，所述第二请求包括所述模型对应的分析类型标识；所述第一数据分析网元从所述第二数据分析网元接收第二响应，所述第二响应包括所述模型的第一信息的索引信息。

在一种可能的实现方法中，所述第一数据分析网元通过网络存储网元从第二数据分析网元接收所述模型的第一信息的索引信息。

在一种可能的实现方法中，所述第一数据分析网元通过网络存储网元向第二数据分析网元发送所述模型的第二信息的索引信息。

在一种可能的实现方法中，所述第一数据分析网元为分布式学习中的客户端数据分析网元，所述第二数据分析网元为分布式学习中的服务端数据分析网元。

在一种可能的实现方法中，所述分布式学习为联邦学习。

在一种可能的实现方法中，所述第一数据分析网元为支持推理功能的数据分析网元，所述第二数据分析网元为支持训练功能的数据分析网元。

第四方面，本申请实施例提供一种通信方法，包括：第二数据分析网元向第一数据分析网元发送第一请求，所述第一请求携带分析类型标识和第二模型需求信息，所述第一请求用于请求所述分析类型标识对应的、且满足所述第二模型需求信息的模型的模型索引信息；所述第二数据分析网元接收来自所述第一数据分析网元的所述模型索引信息。

基于上述方案，通过第二模型需求信息，可以辅助第二数据分析网元快速从第一数据分析网元获取到尽量满足第二模型需求信息的模型，从而可以实现为第二数据分析网元精准提供模型。

在一种可能的实现方法中，所述第二数据分析网元向网络存储网元发送网络功能发现请求，所述网络功能发现请求包括所述分析类型标识和第一模型需求信息，所述网络功能发现请求用于请求获取能够提供与所述分析类型标识对应的、且满足所述第一模型需求信息的模型的数据分析网元；所述第二数据分析网元接收来自所述网络存储网元的所述第一数据分析网元的地址信息。

在一种可能的实现方法中，所述第一模型需求信息包括以下信息中的一项或多项：分析过滤信息、分析报告目标信息模型性能信息或模型部署环境信息；

其中，所述分析过滤信息用于表示所述第二数据分析网元需要请求的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

所述分析报告目标信息用于表示所述第二数据分析网元需要请求的模型对应的终端，

所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述第二数据分析网元需要请求的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；

所述模型部署环境信息用于表示所述第二数据分析网元需要请求的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

10 在一种可能的实现方法中，所述第二模型需求信息包括所述第一模型需求信息中的部分信息或全部信息。

在一种可能的实现方法中，所述第二模型需求信息包括时间信息，所述时间信息用于指示期望收到来自所述第一数据分析网元的模型索引信息的时间。

15 在一种可能的实现方法中，所述第二数据分析网元接收来自所述第一数据分析网元的第一指示信息，所述第一指示信息指示无法在所述时间信息指示的时间内发送所述模型索引信息。

第五方面，本申请实施例提供一种通信方法，包括：第一数据分析网元接收来自第二数据分析网元的第一请求，所述第一请求携带分析类型标识和第二模型需求信息，所述第一请求用于请求所述分析类型标识对应的、且满足所述第二模型需求信息的模型的模型索引信息；所述第一数据分析网元根据所述第二模型需求信息和所述分析类型标识，获取所述模型索引信息；所述第一数据分析网元向所述第二数据分析网元发送所述模型索引信息。

基于上述方案，通过第二模型需求信息，可以辅助第二数据分析网元快速从第一数据分析网元获取到尽量满足第二模型需求信息的模型，从而可以实现为第二数据分析网元精准提供模型。

25 在一种可能的实现方法中，所述第二模型需求信息包括时间信息，所述时间信息用于指示期望收到来自所述第一数据分析网元的模型索引信息的时间。

在一种可能的实现方法中，所述第一数据分析网元向所述第二数据分析网元发送第一指示信息，所述第一指示信息指示无法在所述时间信息指示的时间内发送所述模型索引信息。

30 在一种可能的实现方法中，所述第一数据分析网元向网络存储网元发送网络功能注册请求，所述网络功能注册请求携带所述分析类型标识和模型信息，所述模型信息包括第二指示信息，所述第二指示信息指示所述分析类型标识对应的模型是否已经训练完成或者是否已经准备完成。

35 在一种可能的实现方法中，在所述第二指示信息指示所述分析类型标识对应的模型已经训练完成或者已经准备完成的情况下，所述模型信息还包括模型描述信息，所述模型描述信息包括以下信息中的一项或多项：分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息；

40 其中，所述分析过滤信息用于表示所述分析类型标识对应的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

所述分析报告目标信息用于表示所述分析类型标识对应的模型对应的终端，所述分析报告信息目标包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述分析类型标识对应的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；

所述模型部署环境信息用于表示所述分析类型标识对应的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

10 第六方面，本申请实施例提供一种通信方法，包括：网络存储网元接收来自第二数据分析网元的网络功能发现请求，所述网络功能发现请求包括分析类型标识和第一模型需求信息，所述网络功能发现请求用于请求获取能够提供与所述分析类型标识对应的、且满足所述第一模型需求信息的模型的数据分析网元；所述网络存储网元根据所述第一模型需求信息和所述分析类型标识，获取第一数据分析网元的地址信息；所述网络存储网元向所述
15 第二数据分析网元发送所述第一数据分析网元的地址信息。

基于上述方案，通过第一模型需求信息，可以辅助第二数据分析网元快速从网络存储网元获取到一个尽量满足第一模型需求信息的第一数据分析网元，从而可以实现为第二数据分析网元精准提供第一数据分析网元。

在一种可能的实现方法中，所述网络存储网元接收来自所述第一数据分析网元的网络
20 功能注册请求，所述网络功能注册请求携带所述分析类型标识和模型信息，所述模型信息包括第二指示信息，所述第二指示信息指示所述分析类型标识对应的模型是否已经训练完成或者是否已经准备完成。

在一种可能的实现方法中，在所述第二指示信息指示所述分析类型标识对应的模型已经训练完成或者已经准备完成的情况下，所述模型信息还包括模型描述信息，所述模型描述信息包括以下信息中的一项或多项：分析过滤信息、分析报告目标信息、模型性能信息
25 或模型部署环境信息；

其中，所述分析过滤信息用于表示所述分析类型标识对应的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

30 所述分析报告目标信息用于表示所述分析类型标识对应的模型对应的终端，所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述分析类型标识对应的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型
35 可解释性；

所述模型部署环境信息用于表示所述分析类型标识对应的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

在一种可能的实现方法中，所述第一模型需求信息包括以下信息中的一项或多项：分析
40 过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息；

其中，所述分析过滤信息用于表示所述第二数据分析网元需要请求的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

所述分析报告目标信息用于表示所述第二数据分析网元需要请求的模型对应的终端，
5 所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述第二数据分析网元需要请求的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；
10

所述模型部署环境信息用于表示所述第二数据分析网元需要请求的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

第七方面，本申请实施例提供一种通信装置，该装置可以是数据分析网元，还可以是用于数据分析网元的芯片。该装置具有实现上述第一方面至第六方面、或第一方面至第六方面的各可能的实现方法的功能。该功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块。
15

第八方面，本申请实施例提供一种通信装置，包括处理器和存储器；该存储器用于存储计算机执行指令，当该装置运行时，该处理器执行该存储器存储的该计算机执行指令，以使该装置执行如上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法。
20

第九方面，本申请实施例提供一种通信装置，包括用于执行上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法的各个步骤的单元或手段（means）。

第十方面，本申请实施例提供一种通信装置，包括处理器和接口电路，所述处理器用于通过接口电路与其它装置通信，并执行上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法。该处理器包括一个或多个。
25

第十一方面，本申请实施例提供一种通信装置，包括处理器，用于与存储器相连，用于调用所述存储器中存储的程序，以执行上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法。该存储器可以位于该装置之内，也可以位于该装置之外。且该处理器包括一个或多个。
30

第十二方面，本申请实施例还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得处理器执行上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法。

第十三方面，本申请实施例还提供一种计算机程序产品，该计算机产品包括计算机程序，当计算机程序运行时，使得上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法。
35

第十四方面，本申请实施例还提供一种芯片系统，包括：处理器，用于执行上述第一方面至第六方面的方法及第一方面至第六方面的各可能的实现方法中的任意方法。

第十五方面，本申请实施例还提供一种通信系统，包括：用于执行上述第一方面或第
40

一方面的任意实现方法的第一数据分析网元, 和用于执行上述第二方面或第二方面的任意实现方法的第二数据分析网元。

- 第十六方面, 本申请实施例还提供一种通信系统, 包括: 用于执行上述第五方面或第五方面的任意实现方法的第一数据分析网元, 和用于执行上述第四方面或第四方面的任意实现方法的第二数据分析网元。可选的, 该系统还包括用于执行上述第六方面或第六方面的任意实现方法的网络存储网元。

附图说明

- 图 1 为 5G 网络架构示意图;
- 图 2 为 5G 网络中的 NF 注册/发现/更新流程示意图;
- 图 3 为训练和推断分离架构下的训练 NWDAF 和推断 NWDAF 的工作流程示意图;
- 图 4 为本申请实施例适用的网络架构示意图;
- 图 5 至图 12 为本申请实施例提供的训练-推断分离场景下保证模型有效性的八种方法示意图;
- 图 13 为本申请实施例提供的一种通信方法示意图;
- 图 14(a)为横向联邦学习的训练过程;
- 图 14(b)为本申请实施例提供的又一种通信方法示意图;
- 图 14(c)为本申请实施例提供的又一种通信方法示意图;
- 图 15 为本申请实施例提供的一种通信装置示意图;
- 图 16 为本申请实施例提供的又一种通信装置示意图。

具体实施方式

- 为了使本申请的目的、技术方案和优点更加清楚, 下面将结合附图对本申请作进一步地详细描述。方法实施例中的具体操作方法也可以应用于装置实施例或系统实施例中。其中, 在本申请的描述中, 除非另有说明, “多个”的含义是两个或两个以上。

- 无线机器学习模型驱动网络架构 (wireless Machine Learning-based Network, wMLN) 主要解决机器学习模型在无线网络中的生命周期管理问题。该网络架构中的模型训练功能和模型推断功能是与机器学习模型紧密相关的两个核心功能模块。模型训练功能对计算能力要求较高, 且需要较大的数据量, 通常需要部署在算力和数据强大的集中网元。考虑到推断实时性等要求, 模型推断功能通常部署在靠近业务功能的本地网元中, 以减少传输和处理时延。因此, 模型训练功能和推断功能分离是一种典型的部署场景。

- 使能网络自动化 (enabler of Network Automation, eNA) 架构是一个基于网络数据分析功能 (Network Data Analytics Function, NWDAF) 的智能网络架构。如图 1 所示, NWDAF 是第三代合作伙伴计划 (3rd generation partnership project, 3GPP) 引入的标准化网元, 主要可以用于收集网络数据 (包括终端数据、基站数据、传输网数据、核心网数据、网管数据以及第三方应用数据中的一种或者多种), 并提供网络数据分析服务, 可以输出数据分析结果, 供网络、网管及应用执行策略决策使用。NWDAF 可以利用机器学习模型进行数据分析。3GPP Release 17 中 NWDAF 的功能被分解, 包括数据收集功能、模型训练功能以及模型推断功能。在训练功能和推断功能分离的场景下, 同一模型的训练功能和推断功能分开部署在不同 NWDAF 实例中, 部署训练功能的 NWDAF (简称为训练 NWDAF) 可以

提供训练后的模型,部署推断功能的NWDAF(简称为推断NWDAF)通过获取训练NWDAF提供的模型进行模型推断,提供数据分析服务。

机器学习模型的训练通常是通过学习某一组输入特征与输出目标之间的映射来进行的,通过优化某些损失函数来使得机器学习模型的输出结果(即预测值)与实际结果(即标签值/真实值)之间的误差最小化。当训练出最优模型后,利用该模型的输出对未来的情况进行预测。在理想情况下,假定未来将使用到的数据与模型训练期间所使用的数据类似,具体来说,可能假设训练时的输入特征和预测时的输入特征的分布保持恒定。但是实际中这种假设通常不成立,数据的特征会随着时间的推移由于网络部署变化、应用层业务要求变化、网络实际用户分布变化等而变化,因此,模型的性能(即泛化能力)会随着时间推移逐渐下降。具体的表现可能是,模型的准确率降低,即模型的预测值与真实值之间的误差变大。

图1所示的5G网络架构中可包括三部分,分别是终端设备部分、数据网络(data network, DN)和运营商网络部分。下面对其中的部分网元的功能进行简单介绍说明。

其中,运营商网络可包括以下网元中的一个或多个:鉴权服务器功能(Authentication Server Function, AUSF)网元、网络开放功能(network exposure function, NEF)网元、策略控制功能(Policy Control Function, PCF)网元、统一数据管理(unified data management, UDM)、统一数据库(Unified Data Repository, UDR)、网络存储功能(Network Repository Function, NRF)网元、应用功能(Application Function, AF)网元、接入与移动性管理功能(Access and Mobility Management Function, AMF)网元、会话管理功能(session management function, SMF)网元、RAN以及用户面功能(user plane function, UPF)网元、NWDAF网元等。上述运营商网络中,除无线接入网部分之外的部分可以称为核心网络部分。

在具体实现中,本申请实施例中的终端设备,可以是用于实现无线通信功能的设备。其中,终端设备可以是5G网络或者未来演进的公共陆地移动网络(public land mobile network, PLMN)中的用户设备(user equipment, UE)、接入终端、终端单元、终端站、移动站、移动台、远方站、远程终端、移动设备、无线通信设备、终端代理或终端装置等。接入终端可以是蜂窝电话、无绳电话、会话启动协议(session initiation protocol, SIP)电话、无线本地环路(wireless local loop, WLL)站、个人数字助理(personal digital assistant, PDA)、具有无线通信功能的手持设备、计算设备或连接到无线调制解调器的其它处理设备、车载设备或可穿戴设备,虚拟现实(virtual reality, VR)终端设备、增强现实(augmented reality, AR)终端设备、工业控制(industrial control)中的无线终端、无人驾驶(self driving)中的无线终端、远程医疗(remote medical)中的无线终端、智能电网(smart grid)中的无线终端、运输安全(transportation safety)中的无线终端、智慧城市(smart city)中的无线终端、智慧家庭(smart home)中的无线终端等。终端可以是移动的,也可以是固定的。

上述终端设备可通过运营商网络提供的接口(例如N1等)与运营商网络建立连接,使用运营商网络提供的数据和/或语音等服务。终端设备还可通过运营商网络访问DN,使用DN上部署的运营商业务,和/或第三方提供的业务。其中,上述第三方可为运营商网络和终端设备之外的服务方,可为终端设备提供其他数据和/或语音等服务。其中,上述第三方的具体表现形式,具体可根据实际应用场景确定,在此不做限制。

RAN 作为接入网网元是运营商网络的子网络,是运营商网络中业务节点与终端设备之间的实施系统。终端设备要接入运营商网络,首先是经过 RAN,进而可通过 RAN 与运营商网络的业务节点连接。本申请中的 RAN 设备,是一种为终端设备提供无线通信功能的设备,RAN 设备也称为接入网设备。本申请中的 RAN 设备包括但不限于:5G 中的下一代

5 基站(g nodeB, gNB)、演进型节点 B (evolved node B, eNB)、无线网络控制器 (radio network controller, RNC)、节点 B (node B, NB)、基站控制器 (base station controller, BSC)、基站收发台 (base transceiver station, BTS)、家庭基站 (例如, home evolved nodeB, 或 home node B, HNB)、基带单元 (baseBand unit, BBU)、传输点(transmitting and receiving point, TRP)、发射点(transmitting point, TP)、移动交换中心等。

10 AMF 网元,主要进行移动性管理、接入鉴权/授权等功能。此外,还负责在 UE 与 PCF 间传递用户策略。

SMF 网元,主要进行会话管理、PCF 下发控制策略的执行、UPF 的选择、UE 互联网协议 (internet protocol, IP) 地址分配等功能。

UPF 网元,作为和数据网络的接口 UPF,完成用户面数据转发、基于会话/流级的计费

15 统计,带宽限制等功能。

UDM 网元,主要负责管理签约数据、用户接入授权等功能。

UDR,主要负责签约数据、策略数据、应用数据等类型数据的存取功能。

NEF 网元,主要用于支持能力和事件的开放。

AF 网元,主要传递应用侧对网络侧的需求,例如,服务质量 (Quality of Service, QoS) 需求或用户状态事件订阅等。AF 可以是第三方功能实体,也可以是运营商部署的应用服

20 务,如 IP 多媒体子系统 (IP Multimedia Subsystem, IMS) 语音呼叫业务。

PCF 网元,主要负责针对会话、业务流级别进行计费、QoS 带宽保障及移动性管理、UE 策略决策等策略控制功能。

NRF 网元,可用于提供网元发现功能,基于其他网元的请求,提供网元类型对应的网

25 元信息。NRF 还提供网元管理服务,如网元注册、更新、去注册以及网元状态订阅和推送等。

AUSF 网元:主要负责对用户进行鉴权,以确定是否允许用户或设备接入网络。

DN,是位于运营商网络之外的网络,运营商网络可以接入多个 DN, DN 上可部署多种业务,可为终端设备提供数据和/或语音等服务。例如, DN 是某智能工厂的私有网络,

30 智能工厂安装在车间的传感器可为终端设备, DN 中部署了传感器的控制服务器,控制服务器可为传感器提供服务。传感器可与控制服务器通信,获取控制服务器的指令,根据指令将采集的传感器数据传送给控制服务器等。又例如, DN 是某公司的内部办公网络,该公司员工的手机或者电脑可为终端设备,员工的手机或者电脑可以访问公司内部办公网络上的信息、数据资源等。

35 图 1 中 Nnwdaf、Nausf、Nnef、Npcf、Nudm、Naf、Namf、Nsmf、N1、N2、N3、N4,以及 N6 为接口序列号。这些接口序列号的含义可参见 3GPP 标准协议中定义的含义,在此不做限制。

需要说明的是,本申请实施例中,数据分析网元可以是图 1 所示的 NWDAF 网元,也可以是未来通信系统中具有上述 NWDAF 网元的功能的其它网元。网络存储网元可以是图

40 1 所示的 NRF 网元,也可以是未来通信系统中具有上述 NRF 网元的功能的其它网元。为

便于说明，本申请实施例中，以数据分析网元为 NWDAF 网元，网络存储网元为 NRF 网元为例进行说明。并且，将 NWDAF 网元进一步划分为训练 NWDAF 网元和推断 NWDAF 网元。

5 如图 2 所示，为 5G 网络中的 NF 注册/发现/更新流程示意图。5G 网络中的 NRF 主要用于网络功能 (Network Function, NF) 的管理，这里的网络功能比如可以是 SMF、AMF、NEF、AUSF、NWDAF、PCF 等等。NRF 支持的功能包括：

1) NF 注册/更新/去注册：可用的 NF 实例 (NF instance) 将自身可提供的服务注册在 NRF 中，注册信息通过 NF 配置文件 (NF profile) 描述，NF 配置文件包括 NF 类型，NF 服务名称，NF 地址等信息。NRF 负责维护这些 NF 配置文件。当 NF 需要更新或删除时，
10 NRF 对 NF 配置文件进行相应的修改和删除。

2) NF 发现：NRF 接收来自 NF 实例的 NF 发现请求，并将发现的 NF 实例信息提供给请求的 NF 实例。例如，AMF 向 NRF 请求发现 SMF 实例。再比如，某个 AMF 向 NRF 请求发现另一个 AMF 实例。

15 3) NF 状态通知：NRF 向订阅的 NF 服务消费者通知新注册/更新/注销的 NF 实例和其提供的 NF 服务。

在图 2 中，NF 注册过程包括步骤 201 至步骤 203。

步骤 201，NF1 向 NRF 发送 NF 注册请求，携带 NF 配置文件。

该 NF 配置文件包括 NF 类型，NF 服务名称，NF 地址等信息。

20 步骤 202，NRF 存储该 NF 配置文件。

步骤 203，NRF 向 NF1 发送 NF 注册响应。

该 NF 注册响应用于通知 NF 注册成功。

NF 发现过程包括步骤 204 至步骤 205。

25 步骤 204，NF2 向 NRF 发送 NF 发现请求消息，携带需要查找的 NF 的条件信息，如 NF 类型 (NF type)。

步骤 205，NRF 向 NF2 发送 NF 发现响应，携带符合条件的 NF 实例信息，如 NF 标识 (NF ID) 或 NF IP 地址。

NF 更新过程包括步骤 206a 至步骤 210。

30 步骤 206a，NF2 向 NRF 发送 NF 状态订阅请求，携带 NF 实例，用于请求订阅该 NF 实例的状态信息。

当 NF2 向 NRF 订阅了某个 NF 实例的状态信息 (以下以订阅 NF1 的状态信息为例) 之后，后续 NRF 发现该 NF 实例的状态信息发生改变，则 NRF 会向 NF2 发送该 NF 实例的更新的状态信息。

步骤 206b，NRF 向 NF2 发送 NF 状态订阅响应。

35 该 NF 状态订阅响应用于通知 NF 状态订阅成功。

步骤 207，NF1 向 NRF 发送 NF 更新请求，携带更新的 NF 配置文件。

步骤 208，NRF 更新 NF 配置文件。

也即，NRF 根据接收到的更新的 NF 配置文件，对存储的 NF 配置文件进行更新。

步骤 209，NRF 向 NF1 发送 NF 更新响应。

40 NF 更新响应用于指示 NF 配置文件更新成功。

步骤 210, NRF 向 NF2 发送 NF 状态改变通知, 携带更新的 NF 配置文件。

也即, NRF 向之前订阅过 NF1 的状态信息的 NF2 发送给 NF 状态改变通知。

基于上述过程, 可以结合 NRF, 实现 NF 的注册, 发现以及更新的功能。

需要说明的是, 上述 NF 注册、NF 发现、NF 更新的过程并不一定是连续发生的, 这里只是给出一个流程示例, 说明通常发生的先后顺序。

如图 3 所示, 为训练和推断分离架构下的训练 NWDAF 和推断 NWDAF 的工作流程示意图。其中, 各个网元功能介绍如下:

NRF: 负责 NF 管理, 提供的接口服务包括 NF 注册/去注册/更新, NF 状态订阅/通知等。

训练 NWDAF: 负责模型训练, 训练好的模型可被其它 NWDAF (如推断 NWDAF) 使用。

推断 NWDAF: 负责模型推断, 利用推断结果进行数据分析, 并输出数据分析结果。

NF: 负责某个特定业务功能, 可以调用推断 NWDAF 的服务获取数据分析结果。

该图 3 所示的流程包括以下步骤:

步骤 301, 训练 NWDAF 向 NRF 发送 NF 注册请求, 携带 NF 配置文件。

该 NF 配置文件包括 NF 类型, NF 服务名称 (NF Service), 分析类型标识 (Analytics ID) 等信息。

其中, NF 类型可以是 NWDAF。

NF 服务名称可以是提供模型服务 (ModelProvision)。

分析类型标识用于指示训练 NWDAF 提供的某种特定分析类型, 比如可以是 Service Experience, Network Performance, UE Mobility 等。

步骤 302, NRF 存储该 NF 配置文件。

步骤 303, NRF 向训练 NWDAF 发送 NF 注册响应。

该 NF 注册响应用于通知训练 NWDAF 注册成功。

步骤 304, 推断 NWDAF 向 NRF 发送 NF 发现请求, 携带 NF 配置文件。

比如携带的 NF 配置文件包含 NF 类型 (如 NWDAF), NF 服务名称 (如 ModelProvision), 以及携带 Analytics ID, 则该 NF 发现请求用于请求从 NRF 获取一个对应该 Analytics ID 的训练 NWDAF。

步骤 305, NRF 向推断 NWDAF 发送 NF 发现响应, 其中携带 NWDAF 实例。

其中, 携带的 NWDAF 实例是训练 NWDAF 的一个实例, 可以用训练 NWDAF 的 ID 或 IP 地址来表示。

其中, 上述步骤 301-步骤 305 可选步骤。例如, 如果推断 NWDAF 上配置了训练 NWDAF 的 NF 配置信息, 步骤 301-步骤 305 可以不执行。

步骤 306, 推断 NWDAF 向训练 NWDAF 发送模型请求, 其中携带 Analytics ID。

其中, 推断 NWDAF 可以基于从 NRF 获取到的训练 NWDAF 的 ID 或 IP 地址, 向训练 NWDAF 发送模型请求, 携带的 Analytics ID 用于指示请求获取与该 Analytics ID 对应的模型。

步骤 307, 训练 NWDAF 向推断 NWDAF 发送模型响应, 携带模型信息。

其中, 模型 (也称为机器学习模型, Machine Learning Model, ML Model) 信息用于

描述根据样本输入数据确定样本输出数据的方法，模型信息中可以包括但不限于以下信息中的一个或者多个：输入数据对应的特征类型、输入数据对应的特征类型的特征提取方法（函数关系）、输出数据对应的类型（类别标签、连续数值等）、模型使用的算法类型、模型的类别（分类、回归、聚类等）、模型的参数。以猫-狗分类模型为例，该模型可以根据未知动物的形体样本输入数据确定该样本是猫还是狗，这其中，输入数据的特征类型可以是动物体重、毛长、叫声，输入数据对应的特征类型动物体重的提取方法可以是最大最小归一化，输出数据对应的类型为猫或者狗，模型使用的算法类型可以为深度神经网络（deep neural network, DNN），模型的类别为分类，模型的参数包括但不限于：神经网络的层数、每一层使用的激活函数、每一层激活函数对应的一个或者多个函数参数值。值得说明的是，本发明中所有涉及模型信息（如第一模型信息、第二模型信息等）、模型的信息（如模型的第一信息、模型的第二信息等）都可以参考关于模型信息的描述，其他地方不再赘述。

上述步骤 301 至步骤 307 是训练 NWDAF 提供模型服务的流程。基于该流程，训练 NWDAF 将 NF 配置文件注册至 NRF，后续推断 NWDAF 可以从 NRF 获取到训练 NWDAF 实例，进而推断 NWDAF 可以向训练 NWDAF 请求获取特定类型的模型信息。也即，训练 NWDAF 可以向推断 NWDAF 提供模型服务。

步骤 308，推断 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件。

该 NF 配置文件包括 NF 类型，NF 服务名称（NF Service），分析类型标识（Analytics ID）等信息。

其中，NF 类型可以是 NWDAF。

NF 服务名称可以是提供分析服务（Analytics）。

分析类型标识用于指示训练 NWDAF 提供的某种特定分析类型，比如可以是 Service Experience，Network Performance，UE Mobility 等。

步骤 309，NRF 存储该 NF 配置文件。

步骤 310，NRF 向推断 NWDAF 发送 NF 注册响应。

该 NF 注册响应用于通知推断 NWDAF 注册成功。

步骤 311，NF 向 NRF 发送 NF 发现请求，携带 NF 配置文件。

该 NF 指的是一个 NF 消费者（NF consumer），比如可以是 SMF、AMF 或 UPF 等等。

比如携带的 NF 配置文件包含 NF 类型（如 NWDAF），NF 服务名称（如 ModelProvision），以及携带 Analytics ID，则该 NF 发现请求用于请求从 NRF 获取对应该 Analytics ID 的推断 NWDAF。

步骤 312，NRF 向 NF 发送 NF 发现响应，其中携带 NWDAF 实例。

其中，携带的 NWDAF 实例是推断 NWDAF 的一个实例，可以用推断 NWDAF 的 ID 或 IP 地址来表示。

值得说明的是，上述步骤 308-步骤 312 可选执行。例如，如果 NF 上配置了推断 NWDAF 的 NF 配置信息，步骤 308-步骤 312 可以不执行。

步骤 313，NF 向推断 NWDAF 发送分析订阅，其中携带 Analytics ID。

其中，NF 可以基于从 NRF 获取到的推断 NWDAF 的 ID 或 IP 地址，向推断 NWDAF 发送分析订阅，携带的 Analytics ID 用于指示订阅获取与该 Analytics ID 对应的数据分析结果。

步骤 314, 推断 NWDAF 向 NF 发送分析结果通知, 携带数据分析结果。

上述步骤 308 至步骤 314 是推断 NWDAF 提供分析服务的流程。基于该流程, 推断 NWDAF 将 NF 配置文件注册至 NRF, 后续 NF 可以从 NRF 获取到推断 NWDAF 实例, 进而 NF 可以向推断 NWDAF 请求获取特定类型的数据分析结果。也即, 推断 NWDAF 可以向 NF 提供数据分析服务。

作为一种可替代的实现方法, 上述步骤 313 至步骤 314 可以使用以下步骤 313' 至步骤 314' 替换:

步骤 313', NF 向推断 NWDAF 发送分析请求, 其中携带 Analytics ID。

其中, NF 可以基于从 NRF 获取到的推断 NWDAF 的 ID 或 IP 地址, 向推断 NWDAF 发送分析订阅, 携带的 Analytics ID 用于指示请求获取与该 Analytics ID 对应的数据分析结果。

步骤 314', 推断 NWDAF 向 NF 发送分析结果响应, 携带数据分析结果。

该步骤 313' 至步骤 314' 是每次需要主动发送分析请求, 推断 NWDAF 才向 NF 发送数据分析结果, 而上述步骤 313 至步骤 314 是只需要订阅一次, 后续推断 NWDAF 在产生新的数据分析结果时主动向 NF 发送数据分析结果。

上述图 3 所示的模型训练与模型使用过程中存在的问题是: 随着时间推移, 推断 NWDAF 本地可以根据推断数据确定推断结果, 然后根据推断数据的真实结果以及推断结果确定模型的使用效果 (也就是模型性能评估结果), 推断 NWDAF 根据使用效果确定机器学习模型性能下降, 但是训练-推断分离的场景下训练 NWDAF 无法感知推断 NWDAF 中的模型使用效果, 且推断 NWDAF 没有能力进行模型训练, 因此, 现有技术中无法实现在模型性能下降时进行再训练和模型更新, 也就无法保证模型在运行过程中性能一直良好。如果推断 NWDAF 继续使用性能下降的模型进行数据分析, 可能导致数据分析结果不准确, 影响模型性能。

为解决上述问题, 本申请实施例提出建立一种模型性能监控和反馈机制, 对推断 NWDAF 中运行的模型性能进行评估, 当模型性能下降至一定程度时训练 NWDAF 可以感知并及时进行再训练, 推断 NWDAF 可利用再训练得到的性能良好的新模型进行模型更新 (或者替换), 保证模型的使用效果。其中, 监控、反馈、再训练和更新机制可以通过 NRF 实现, 也可以由训练 NWDAF 与推断 NWDAF 直接交互实现。

本申请实施例应用的系统架构是 eNA 架构, 具体地, 本申请实施例针对的是模型训练和推断功能分离部署的场景, 即训练功能和推断功能部署在不同的 NWDAF 实例中。如图 4 所示, 为本申请实施例适用的一种网络架构示意图。训练 NWDAF, 推断 NWDAF 和 NF, 都需要通过 Nnrf 接口服务在 NRF 中进行注册。推断 NWDAF 通过 Nnwdaf 接口服务向训练 NWDAF 请求模型, NF 通过 Nnwdaf 接口服务向推断训练 NWDAF 请求数据分析结果。

下面对本申请实施例提供的方案进行说明。

实施例一

如图 5 所示, 为本申请实施例提供的一种训练-推断分离场景下保证模型有效性的方法流程图示意图。

该实施例一考虑通过 NRF 更新注册信息, 实现性能监控和模型更新。主要涉及:

1、模型性能监控与反馈: 推断 NWDAF 在 NRF 处的注册信息增加模型状态信息, 推

断 NWDAF 进行模型性能监控，当判断模型需要再训练时通过 NRF 更新模型状态信息，NRF 通知训练 NWDAF 模型状态更新，触发训练 NWDAF 对模型进行再训练。

- 2、模型更新：训练 NWDAF 在 NRF 处的注册信息增加模型索引信息，训练 NWDAF 再训练得到新模型后通过 NRF 更新模型索引信息，NRF 通知推断 NWDAF 有新模型可以使用，推断 NWDAF 主动向训练 NWDAF 请求新模型并完成模型更新。

该实施例包括以下步骤：

步骤 501，训练 NWDAF 向 NRF 注册。

- 训练 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件，该 NF 配置文件包括 NF 类型，NF 服务名称（例如，NF Service），分析类型标识（例如，Analytics ID）等信息，以及还包括模型索引信息。该模型索引信息可以是模型版本号（例如，version）、位置信息（例如，location）或统一资源定位符（Uniform Resource Locator，URL）等。其中，version 表示模型版本，location 或 URL 表示模型的存储位置，三者中使用任意一个都可以。可选的，当模型索引信息是位置信息或 URL，该位置信息或 URL 中也可以包含模型版本。可选的，位置信息可以是 IP 地址。

- 相应地，NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

步骤 502，推断 NWDAF 向 NRF 注册。

推断 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件，该 NF 配置文件包括 NF 类型，NF 服务名称（例如，NF Service），分析类型标识（例如，Analytics ID）等信息，以及还包括模型状态信息。该模型状态信息用于指示模型使用状态。

- 该模型状态信息的可选取值包括但不限于：

- 1）空 ‘null’：表示没有模型可以使用。
- 2）允许 ‘ok’：表示模型性能良好，可以对外提供分析服务。
- 3）限定 ‘limited’：表示模型性能有所下降，但仍可以提供服务，需要进行再训练。
- 4）停止 ‘stopped’：表示模型关闭，已停止提供服务。

- 其中，在注册过程中，NF 注册请求携带的模型状态信息为 ‘null’，即推断 NWDAF 上当前没有模型可以使用。

相应地，NRF 存储 NF 配置文件以及向推断 NWDAF 发送 NF 注册响应。

步骤 503，推断 NWDAF 发现训练 NWDAF，并向训练 NWDAF 请求获取模型信息。

- 该过程可以参考图 3 实施例中的步骤 304 至步骤 307，不再赘述。基于该过程，推断 NWDAF 可以从训练 NWDAF 中获取到模型信息。

步骤 504，推断 NWDAF 向 NRF 订阅训练 NWDAF 的状态。

后续训练 NWDAF 在 NRF 中注册的 NF 配置文件发生更新时，NRF 通知推断 NWDAF。

步骤 505，训练 NWDAF 向 NRF 订阅推断 NWDAF 的状态。

后续推断 NWDAF 在 NRF 中注册的 NF 配置文件发生更新时，NRF 通知训练 NWDAF。

- 上述步骤 504 与步骤 505 相互之间没有固定的先后顺序。

步骤 506，推断 NWDAF 向 NRF 发送 NF 更新请求，携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型状态信息，该更新的模型状态信息比如可以是 “ok”。可选的，该更新的 NF 配置文件还携带 Analytics ID，用于标识要更新的模型。可选的，该更新的 NF 配置文件还携带 NF 类型，NF 服务名称（NF Service）等。

- 步骤 507，NRF 更新 NF 配置文件。

也即, NRF 根据接收到的更新的 NF 配置文件, 对存储的 NF 配置文件进行更新。

步骤 508, NRF 向推断 NWDAF 发送 NF 更新响应。

该 NF 更新响应用于通知推断 NWDAF 的 NF 配置文件更新成功。

5 步骤 509, NRF 向训练 NWDAF 发送 NF 状态更新通知, 其中携带更新的模型状态信息。

该更新的模型状态信息比如可以是 'ok'。

可选的, 该 NF 状态更新通知还携带指示信息, 用于指示更新的类型是模型状态信息更新。

10 由于上述步骤 505 中, 训练 NWDAF 向 NRF 订阅了推断 NWDAF 的状态, 因此当 NRF 中存储的推断 NWDAF 的 NF 配置文件发生更新后, NRF 通知训练 NWDAF。

步骤 510, 推断 NWDAF 确定模型需要再训练。

15 判断依据可以是对模型性能的评估结果不满足模型性能要求(比如模型精度下降到 80% 以下, 其中 80% 为模型精度要求), 也可以是 NF 上报的业务关键性能指标(Key Performance Indicator, KPI)不满足 KPI 要求(比如 KPI 下降到 KPI 要求以下)。本发明其他实施例中涉及判定模型是否需要再训练或者需要更新的方法都可以参考此处描述, 不再赘述。

需要说明的是, 该步骤 510 发生在推断 NWDAF 中的模型运行过程中, 出现的时间不固定。

步骤 511, 推断 NWDAF 向 NRF 发送 NF 更新请求, 携带更新的 NF 配置文件。

20 该更新的 NF 配置文件至少携带更新的模型状态信息, 该更新的模型状态信息比如可以是 'limited'。可选的, 该更新的 NF 配置文件还携带 Analytics ID, 用于标识要更新的模型。可选的, 该更新的 NF 配置文件还携带 NF 类型, NF 服务名称(NF Service)等。

NRF 更新存储的 NF 配置文件, 然后向推断 NWDAF 发送 NF 更新响应。

步骤 512, NRF 向训练 NWDAF 发送 NF 状态更新通知, 其中携带更新的模型状态信息。

25 该更新的模型状态信息可以是 'limited'。

可选的, 该 NF 状态更新通知还携带指示信息, 用于指示更新的类型是模型状态信息更新。

由于上述步骤 505 中, 训练 NWDAF 向 NRF 订阅了推断 NWDAF 的状态, 因此当 NRF 中存储的推断 NWDAF 的 NF 配置文件发生更新后, NRF 通知训练 NWDAF。

30 步骤 513, 训练 NWDAF 对模型启动再训练。

训练 NWDAF 对模型启动再训练, 得到训练后的模型以及相应的模型索引信息, 如模型版本号、位置信息或 URL 等。

步骤 514, 训练 NWDAF 向 NRF 发送 NF 更新请求, 携带更新的 NF 配置文件。

35 该更新的 NF 配置文件至少携带更新的模型索引信息, 该更新的模型索引信息比如可以是更新的模型版本信息、更新的模型位置信息或更新的模型 URL 等。可选的, 该更新的 NF 配置文件还携带 Analytics ID, 用于标识要更新的模型。可选的, 该更新的 NF 配置文件还携带 NF 类型, NF 服务名称(NF Service)等。

NRF 更新存储的 NF 配置文件, 然后向训练 NWDAF 发送 NF 更新响应。

40 步骤 515, NRF 向推断 NWDAF 发送 NF 状态更新通知, 其中携带更新的模型索引信息。

该更新的模型索引信息比如可以是更新的模型版本信息、更新的模型位置信息或更新的模型 URL 等。

可选的，该 NF 状态更新通知还携带指示信息，用于指示更新的类型是模型索引信息更新。

- 5 由于上述步骤 504 中，推断 NWDAF 向 NRF 订阅了训练 NWDAF 的状态，因此当 NRF 中存储的训练 NWDAF 的 NF 配置文件发生更新后，NRF 通知推断 NWDAF。

步骤 516，推断 NWDAF 向训练 NWDAF 发送模型请求，携带 Analytics ID 和更新的模型索引信息。

Analytics ID 用于指示与该 Analytics ID 对应的模型。

- 10 步骤 517，训练 NWDAF 向推断 NWDAF 发送模型响应，携带模型信息。

该模型信息包括与更新的模型索引信息对应的模型信息，也即获取到新的模型对应的模型信息。

可选的，模型响应中携带的模型信息可以是新模型的参数项取值、或新的模型（如模型文件或包含模型的镜像文件）、或新的模型地址（如 URL 或 IP 地址）。

- 15 其中，模型文件是使用第三方框架保存下来的模型持久化文件，如使用人工智能框架 TensorFlow 保存的.pb 格式的模型文件。模型镜像文件是一个包含模型的镜像软件包，其中可以包含模型文件以及与模型使用相关的其他多个文件。

- 需要说明的是，如果步骤 516 中携带的模型索引信息是新的模型地址，则推断 NWDAF 可以直接根据该地址信息进一步获取新的模型信息，无需执行步骤 516-517。比如，推断
20 NWDAF 根据 URL 通过文件传输协议（File Transfer Protocol, FTP）获取包含新的模型信息的文件（如包含新模型的参数项取值的文件、或新的模型文件、或包含新模型的镜像文件）。

- 如果步骤 516 中携带的模型索引信息是模型版本号，则模型响应中携带的模型信息可以是新模型的参数项取值、或新的模型（模型文件或包含模型的镜像），也可以是新模型
25 的地址（如 URL 或 IP 地址）。如果模型响应中携带的模型信息是新模型的地址（如 URL 或 IP 地址），推断 NWDAF 可以根据该地址信息进一步获取新的模型信息。

步骤 518，推断 NWDAF 进行模型更新。

也即，推断 NWDAF 根据接收到的新的模型信息，对正在使用的旧模型进行更新或者替换。

- 30 可选的，推断 NWDAF 在模型更新前，对新的模型信息进行本地测试，测试通过后再进行更新或者替换。

步骤 519，推断 NWDAF 向 NRF 发送 NF 更新请求，携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型状态信息，该更新的模型状态信息比如可以是 'ok'。可选的，该更新的 NF 配置文件还携带 Analytics ID，用于标识要更新的模型。

- 35 可选的，该更新的 NF 配置文件还携带 NF 类型，NF 服务名称（NF Service）等。

NRF 更新存储的 NF 配置文件，然后向推断 NWDAF 发送 NF 更新响应。

步骤 520，NRF 向训练 NWDAF 发送 NF 状态更新通知，其中携带更新的模型状态信息。

该更新的模型状态信息可以是 'ok'。

- 40 可选的，该 NF 状态更新通知还携带指示信息，用于指示更新的类型是模型状态信息

更新。

其中，上述步骤 511-步骤 520 可选步骤。例如，如果步骤 510 中推断 NWDAF 确定模型不需要再训练或者推断 NWDAF 可以容忍模型性能下降到模型性能要求以下，步骤 511-步骤 520 可以不执行。

5 由于上述步骤 505 中，训练 NWDAF 向 NRF 订阅了推断 NWDAF 的状态，因此当 NRF 中存储的推断 NWDAF 的 NF 配置文件发生更新后，NRF 通知训练 NWDAF。

基于上述实施例，一方面，当推断 NWDAF 中使用的模型的性能下降时，可以通过 NRF 通知训练 NWDAF 进行模型再训练，训练结束后推断 NWDAF 可利用新模型更新或者替换旧模型，保证模型的使用效果。

10 另一方面，当推断 NWDAF 为多个时，如果只是其中一个推断 NWDAF X 的性能下降，并且通知训练 NWDAF 进行再训练，训练结束后，除了推断 NWDAF X 可以利用新模型更新或者替换旧模型，多个推断 NWDAF 中的其他任何一个推断 NWDAF 也可以利用上述机制获取新模型，并且利用新模型更新或者替换旧模型，这样可以保证多个推断 NWDAF 上模型的使用效果。针对多推断 NWDAF 的场景，详细过程的示例参见实施例二。

15

实施例二

如图 6 所示，为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程图示意图。

20 该实施例二，是在上述实施例一的基础上，考虑有多个推断 NWDAF 存在的场景，以下以有两个推断 NWDAF（分别用推断 NWDAF1 和推断 NWDAF2 来表示）为例，推断 NWDAF1 和推断 NWDAF2 中使用相同的模型进行数据分析，该模型都来自同一训练 NWDAF。某一时刻，NWDAF1 中模型性能下降需要再训练，推断 NWDAF2 中模型不需要再训练。这里仅以推断 NWDAF2 为例表示存在除了请求再训练的推断 NWDAF1 外，还存在其他使用相同模型的推断 NWDAF。

25 由于推断 NWDAF2 也订阅了训练 NWDAF 状态，因此，当训练 NWDAF 再训练得到新模型之后，推断 NWDAF2 也会收到来自 NRF 的通知。一方面，如果新模型比推断 NWDAF2 中模型效果更好，推断 NWDAF2 可以利用新模型进一步改善数据分析效果，另一方面，由于推断 NWDAF2 中模型此时并不是必须要更新，推断 NWDAF2 需要先获取新模型并进行本地评估后才可以确定是否需要更新。如果获取新模型后推断 NWDAF2 最终
30 决定不更新，则浪费了一部分传输资源。本实施例考虑在训练 NWDAF 的注册信息进一步增加模型性能信息，包括精度、所需计算量等，帮助其它暂不需要更新的推断 NWDAF 判断是否需要请求新模型。

该实施例包括以下步骤：

步骤 601，训练 NWDAF 向 NRF 注册。

35 训练 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件，该 NF 配置文件包括 NF 类型，NF 服务名称（NF Service），分析类型标识（Analytics ID）等信息，以及还包括模型索引信息和模型性能信息。该模型索引信息可以是模型版本号（version）、位置信息（location）或 URL 等。其中，version 表示模型版本，location 或 URL 表示模型的存储位置，三者中使用任意一个都可以。可选的，当模型索引信息是 location 或 URL，该 location
40 或 URL 中也可以包含 version。该模型性能信息用于指示模型的性能，比如可以包括模型

准确度、达到该准确度所需的硬件能力信息、模型推理所需计算量、模型的推理时长、模型的大小等。

可选的，NF 配置文件还可以包含模型使用的算法、人工智能框架、模型的输入特征等信息。

5 相应地，NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

步骤 602 至步骤 613，类似于上述实施例一中的步骤 502 至步骤 513。

需要说明的是，该步骤 602 至步骤 613 中，涉及推断 NWDAF1 的相关操作以及推断 NWDAF2 的相关操作，均可以分别参考上述步骤 502 至步骤 513 中关于推断 NWDAF 的相关操作。并且，在步骤 610 至步骤 611（参考步骤 510 至步骤 611）中，是推断 NWDAF1
10 确定模型需要再训练，然后向 NRF 发送 NF 更新请求，从而触发训练 NWDAF 对模型启动再训练。

步骤 614，训练 NWDAF 向 NRF 发送 NF 更新请求，携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型索引信息以及更新的模型性能信息，该更新的模型索引信息比如可以是更新的模型版本信息、更新的模型位置信息或更新的模型
15 URL 等。可选的，该更新的 NF 配置文件还携带 Analytics ID，用于标识要更新的模型。该更新的模型性能信息比如可以包括模型准确度、达到该准确度所需的硬件能力信息、模型推理所需计算量、模型的推理时长、模型的大小等。

可选的，该更新的 NF 配置文件还携带 NF 类型，NF 服务名称（NF Service）等。

NRF 更新存储的 NF 配置文件，然后向训练 NWDAF 发送 NF 更新响应。

20 步骤 615，NRF 分别向推断 NWDAF1 和推断 NWDAF2 发送 NF 状态更新通知，其中携带更新的模型索引信息和更新的模型性能信息。

该更新的模型索引信息比如可以是更新的模型版本信息、更新的模型位置信息或更新的模型 URL 等。

该更新的模型性能信息比如可以包括模型准确度、达到该准确度所需的硬件能力信息、
25 模型推理所需计算量、模型的推理时长、模型的大小等。

可选的，该 NF 状态更新通知还携带指示信息，用于指示更新的类型是模型索引信息更新以及模型性能信息更新。

由于上述步骤中，推断 NWDAF1 和推断 NWDAF2 分别向 NRF 订阅了训练 NWDAF 的状态，因此当 NRF 中存储的训练 NWDAF 的 NF 配置文件发生更新后，NRF 通知推断
30 NWDAF1 和推断 NWDAF2。

步骤 616，推断 NWDAF2 判断是否需要更新模型。

由于推断 NWDAF2 不是模型训练的触发者，因此当推断 NWDAF2 收到更新的模型索引信息后，需要判断是否需要更新模型。

可选的，推断 NWDAF2 可以根据自身算力、模型性能要求以及接收到的更新的模型性能信息，判断是否需要更新模型。或者，推断 NWDAF2 还可以根据正在使用的模型性能状况以及接收到的更新的模型性能信息，判断是否需要更新模型。
35

步骤 617，推断 NWDAF1 向训练 NWDAF 发送模型请求，携带 Analytics ID 和更新的模型索引信息。

Analytics ID 用于指示与该 Analytics ID 对应的模型。

40 由于推断 NWDAF1 是模型训练的触发者，因此当推断 NWDAF1 收到更新的模型索引

信息后，需要更新模型。

步骤 618，训练 NWDAF 向推断 NWDAF1 发送模型响应，携带模型信息。

该模型信息包括与更新后的模型标识对应的模型信息，也即获取到新的模型对应的模型信息。

5 其中，模型信息的具体实现，可以参考前述实施例的描述。

步骤 619，推断 NWDAF1 进行模型更新。

也即，推断 NWDAF 根据接收到的新的模型信息，对正在使用的旧模型进行更新或者替换。

10 可选的，推断 NWDAF 在模型更新前，对新的模型信息进行本地测试，测试通过后再进行更新或者替换。

需要说明的是，如果上述步骤 616 中，推断 NWDAF2 确定需要更新模型，则推断 NWDAF2 还需要执行类似于上述步骤 617 至步骤 619 的操作过程，以向训练 NWDAF 请求获取到更新的模型信息，然后根据接收到的新的模型信息，对正在使用的旧模型进行更新。如果上述步骤 616 中，推断 NWDAF2 确定不需要更新模型，则无需执行该模型更新流程。

15 基于上述实施例，订阅相同模型的推断 NWDAF 可以根据模型性能信息，确定是否需要请求新模型，以避免不必要的模型传输和本地评估过程，从而可以提升模型更新流程的效率以及节约资源。

20 实施例三

如图 7 所示，为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程图示意图。

25 该实施例三是在上述实施例一的基础上，考虑针对同一个 Analytics ID，需要多个子模型共同协作完成分析的场景。该场景下任意一个子模型的性能下降都会导致 Analytics ID 对应的模型性能下降，如果仅根据 Analytics ID 进行模型监控，则无法精确定位到子模型的性能，会将 Analytics ID 对应的所有子模型都进行再训练和更新，而实际可能某些子模型性能良好并不需要更新，这就导致了不必要的训练和更新。

本实施例考虑进一步地增加模型标识（model ID）来表示各个子模型。

该实施例包括以下步骤：

30 步骤 701，训练 NWDAF 向 NRF 注册。

训练 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件，该 NF 配置文件包括 NF 类型，NF 服务名称（NF Service），分析类型标识（Analytics ID）等信息，以及还包括模型标识（model ID）和模型索引信息。该模型索引信息可以是模型版本号（version）、位置信息（location）或 URL 等。其中，version 表示模型版本，location 或 URL 表示模型的存储位置，三者中使用任意一个都可以。可选的，当模型索引信息是 location 或 URL，该 location 或 URL 中也可以包含 version。该模型标识用于唯一标识一个模型，例如，该模型标识可以由 NWDAF 地址、PLMN ID、在某个 NWDAF 范围内唯一的 modelID 组成。

需要说明的是，NF 配置文件可以携带多个模型索引信息，每个模型标识对应一个模型索引信息。

40 相应地，NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

可选的, 该更新的 NF 配置文件还可以携带多个模型标识, 每个模型标识用于标识多个更新的模型中的一个模型。

步骤 702, 推断 NWDAF 向 NRF 注册。

推断 NWDAF 向 NRF 发送 NF 注册请求, 携带 NF 配置文件, 该 NF 配置文件包括 NF 类型, NF 服务名称 (NF Service), 分析类型标识 (Analytics ID) 等信息, 以及还包括模型状态信息和模型标识。其中, 每个模型标识对应一个模型状态信息。

模型状态信息用于指示模型标识对应的模型的使用状态。该模型状态信息的可选取值包括但不限于:

- 1) 'null': 表示没有模型可以使用。
- 2) 'ok': 表示模型性能良好, 可以对外提供分析服务。
- 3) 'limited': 表示模型性能有所下降, 但仍可以提供服务, 需要进行再训练。
- 4) 'stopped': 表示模型关闭, 已停止提供服务。

其中, 在注册过程中, NF 注册请求携带的模型状态信息为 'null', 即推断 NWDAF 上当前没有模型可以使用。

示例性地, NF 配置文件携带的模型状态信息和模型标识如下:

{ (model ID 1, null), (model ID 2, null), (model ID 3, null) }。

或者, NF 配置文件携带的模型状态信息和模型标识如下:

{ (model ID 1, model ID 2, model ID 3), (null, null, null) }。

可选的, 在实际应用中 NF 配置文件携带的模型状态信息和模型标识可以是一个列表, 该列表包含多个项信息, 每个项信息包含一个模型状态信息和一个模型标识。

相应地, NRF 存储 NF 配置文件以及向推断 NWDAF 发送 NF 注册响应。

步骤 703, 推断 NWDAF 发现训练 NWDAF, 并向训练 NWDAF 请求获取模型信息。

该过程可以参考图 5 实施例中的步骤 503, 不再赘述。基于该过程, 推断 NWDAF 可以从训练 NWDAF 中获取到模型信息。

步骤 704, 推断 NWDAF 向 NRF 订阅训练 NWDAF 的状态。

后续训练 NWDAF 在 NRF 中注册的 NF 配置文件发生更新时, NRF 通知推断 NWDAF。

步骤 705, 训练 NWDAF 向 NRF 订阅推断 NWDAF 的状态。

后续推断 NWDAF 在 NRF 中注册的 NF 配置文件发生更新时, NRF 通知训练 NWDAF。

上述步骤 704 与步骤 705 相互之间没有固定的先后顺序。

步骤 706, 推断 NWDAF 向 NRF 发送 NF 更新请求, 携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型状态信息, 该更新的模型状态信息比如可以是 'ok'。其中, 每个模型标识对应一个更新的模型状态信息。

可选的, 该更新的 NF 配置文件还携带 Analytics ID, 用于标识要更新的模型。可选的, 该更新的 NF 配置文件还携带 NF 类型, NF 服务名称 (NF Service) 等。

可选的, 该更新的 NF 配置文件还可以携带模型标识, 用于标识更新的模型。

步骤 707, NRF 更新存储的 NF 配置文件。

步骤 708, NRF 向推断 NWDAF 发送 NF 更新响应。

该 NF 更新响应用于通知 NF 配置文件更新成功。

步骤 709, NRF 向训练 NWDAF 发送 NF 状态更新通知, 其中携带更新的模型状态信息。

其中，每个模型标识对应一个更新的模型状态信息。更新的模型状态信息比如可以是‘ok’。

可选的，该 NF 状态更新通知还携带指示信息，用于指示更新的类型是模型状态信息更新。

5 可选的，该 NF 状态更新通知还可以携带模型标识，用于标识更新的模型。

由于上述步骤 705 中，训练 NWDAF 向 NRF 订阅了推断 NWDAF 的状态，因此当 NRF 中存储的推断 NWDAF 的 NF 配置文件发生更新后，NRF 通知训练 NWDAF。

步骤 710，推断 NWDAF 确定模型需要再训练。

判断依据可以是对模型性能的评估结果（比如模型精度下降），也可以是 NF 上报的业务 KPI（比如 KPI 下降）。10

需要说明的是，该步骤 710 发生在推断 NWDAF 中的模型运行过程中，出现的时间不固定。

需要说明的是，该步骤中，判断的结果可以是：某一个或某几个子模型需要再训练。比如针对某个 Analytics ID 对应的模型，一共有 10 个子模型，分别用 model ID 1 至 model ID 15 10 表示。该步骤 710 的判断结果比如是：model ID 1 至 model ID 3 对应的子模型需要再训练，model ID 4 至 model ID 10 对应的子模型不需要再训练。

步骤 711，推断 NWDAF 向 NRF 发送 NF 更新请求，携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型状态信息，其中，每个模型标识对应一个更新的模型状态信息。该更新的模型状态信息比如可以是‘limited’。可选的，该更新的 NF 配置文件还携带 Analytics ID，用于标识要更新的模型。可选的，该更新的 NF 配置文件还携带 NF 类型，NF 服务名称（NF Service）等。20

可选的，该更新的 NF 配置文件还可以携带模型标识，用于标识更新的模型。

NRF 更新存储的 NF 配置文件，然后向推断 NWDAF 发送 NF 更新响应。

需要说明的是，该步骤 711 中的更新的 NF 配置文件中携带的模型标识即为上述步骤 25 710 中确定的需要再训练的子模型的标识信息，更新的模型状态信息即为需要再训练的子模型的标识信息对应的更新的模型状态信息。

步骤 712，NRF 向训练 NWDAF 发送 NF 状态更新通知，其中携带更新的模型状态信息。

其中，每个模型标识对应一个更新的模型状态信息。该更新的模型状态信息可以是‘limited’。30

可选的，该 NF 状态更新通知还携带指示信息，用于指示更新的类型是模型状态信息更新。

可选的，该 NF 状态更新通知还可以携带模型标识，用于标识更新的模型。

由于上述步骤 705 中，训练 NWDAF 向 NRF 订阅了推断 NWDAF 的状态，因此当 NRF 中存储的推断 NWDAF 的 NF 配置文件发生更新后，NRF 通知训练 NWDAF。35

步骤 713，训练 NWDAF 对模型启动再训练。

训练 NWDAF 对模型启动再训练，得到训练后的模型以及相应的模型索引信息，如模型版本号、位置信息或 URL 等。

需要说明的是，该步骤中，仅对接收到的需要进行训练的子模型进行启动再训练。比如接收到的模型标识是 model ID 1 至 model ID 3，则对 model ID 1 至 model ID 3 对应的子40

模型进行再训练。

步骤 714, 训练 NWDAF 向 NRF 发送 NF 更新请求, 携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型索引信息, 每个模型标识对应一个更新的模型状态信息。该更新的模型索引信息比如可以是更新的模型版本信息、更新的模型位置信息或更新的模型 URL 等。可选的, 该更新的 NF 配置文件还携带 Analytics ID, 用于标识要更新的模型。可选的, 该更新的 NF 配置文件还携带 NF 类型, NF 服务名称 (NF Service) 等。

可选的, 该更新的 NF 配置文件还可以携带模型标识, 用于标识更新的模型。

NRF 更新存储的 NF 配置文件, 然后向训练 NWDAF 发送 NF 更新响应。

10 步骤 715, NRF 向推断 NWDAF 发送 NF 状态更新通知, 其中携带更新的模型索引信息。

该更新的模型索引信息比如可以是更新的模型版本信息、更新的模型位置信息或更新的模型 URL 等。

15 可选的, 该 NF 状态更新通知还携带指示信息, 用于指示更新的类型是模型索引信息更新。

可选的, 该 NF 状态更新通知还携带模型标识, 用于标识更新的模型。

由于上述步骤 704 中, 推断 NWDAF 向 NRF 订阅了训练 NWDAF 的状态, 因此当 NRF 中存储的训练 NWDAF 的 NF 配置文件发生更新后, NRF 通知推断 NWDAF。

20 步骤 716, 推断 NWDAF 向训练 NWDAF 发送模型请求, 携带 Analytics ID 和更新的模型索引信息。

Analytics ID 用于指示与该 Analytics ID 对应的模型。

模型标识用于指示该 Analytics ID 对应的模型中的子模型。

步骤 717, 训练 NWDAF 向推断 NWDAF 发送模型响应, 携带模型信息。

25 该模型信息包括与更新后的模型标识对应的模型信息, 也即获取到新的模型对应的模型信息。

其中, 模型信息的具体实现, 可以参考前述实施例的描述。

步骤 718, 推断 NWDAF 进行模型更新。

也即, 推断 NWDAF 根据接收到的新的模型信息, 对正在使用的旧模型 (具体是相应需要更新的子模型) 进行更新。

30 可选的, 推断 NWDAF 在模型更新前, 对新的模型信息进行本地测试, 测试通过后再进行更新或者替换。

步骤 719, 推断 NWDAF 向 NRF 发送 NF 更新请求, 携带更新的 NF 配置文件。

35 该更新的 NF 配置文件至少携带更新的模型状态信息, 每个模型标识对应一个更新的模型状态信息。该更新的模型状态信息比如可以是 'ok'。可选的, 该更新的 NF 配置文件还携带 Analytics ID, 用于标识要更新的模型。可选的, 该更新的 NF 配置文件还携带 NF 类型, NF 服务名称 (NF Service) 等。

NRF 更新存储的 NF 配置文件, 然后向推断 NWDAF 发送 NF 更新响应。

可选的, 该更新的 NF 配置文件还可以携带模型标识, 用于标识更新的模型。

40 步骤 720, NRF 向训练 NWDAF 发送 NF 状态更新通知, 其中携带更新的模型状态信息。

其中,每个模型标识对应一个更新的模型状态信息。更新的模型状态信息可以是‘ok’。

可选的,该 NF 状态更新通知还携带指示信息,用于指示更新的类型是模型状态信息更新。

可选的,该 NF 状态更新通知还可以携带模型标识,用于标识更新的模型。

- 5 其中,上述步骤 714-步骤 720 可选步骤。例如,如果步骤 713 中训练 NWDAF 确定模型不需要再训练或者训练 NWDAF 可以容忍模型性能下降到模型性能要求以下或者训练 NWDAF 上当前不具备模型再训练的能力(如硬件资源有限),步骤 714-步骤 720 可以不执行。

- 10 由于上述步骤 705 中,训练 NWDAF 向 NRF 订阅了推断 NWDAF 的状态,因此当 NRF 中存储的推断 NWDAF 的 NF 配置文件发生更新后,NRF 通知训练 NWDAF。

基于上述实施例,增加了模型标识(也称为子模型标识),按照子模型粒度进行性能监控,在一个 Analytics ID 对应多个子模型的场景下可以实现精准的执行模型再训练与更新,避免浪费训练和传输资源。

15 实施例四

如图 8 所示,为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程示意图。

- 20 上述实施例一至实施例三是考虑通过 NRF 实现训练 NWDAF 和推断 NWDAF 的信息交互,该实施例四考虑在训练 NWDAF 和推断 NWDAF 之间的接口上新增操作,直接交互信息。

该实施例包括以下步骤:

步骤 801,训练 NWDAF 向 NRF 注册。

训练 NWDAF 向 NRF 发送 NF 注册请求,携带 NF 配置文件,该 NF 配置文件包括 NF 类型,NF 服务名称(NF Service),分析类型标识(Analytics ID)等信息。

- 25 相应地,NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

步骤 802a,推断 NWDAF 向 NRF 发送 NF 发现请求,携带 NF 配置文件。

比如携带的 NF 配置文件包含 NF 类型(如 NWDAF),NF 服务名称(如 ModelProvision),以及携带 Analytics ID,则该 NF 发现请求用于请求从 NRF 获取对应该 Analytics ID 的训练 NWDAF。

- 30 步骤 802b,NRF 向推断 NWDAF 发送 NF 发现响应,其中携带 NWDAF 实例。

其中,携带的 NWDAF 实例是训练 NWDAF 的一个实例,可以用训练 NWDAF 的 ID 或 IP 地址来表示。

步骤 803a,推断 NWDAF 向训练 NWDAF 发送模型请求,其中携带 Analytics ID。

- 35 其中,推断 NWDAF 可以基于从 NRF 获取到的训练 NWDAF 的 ID 或 IP 地址,向训练 NWDAF 发送模型请求,携带的 Analytics ID 用于指示请求获取与该 Analytics ID 对应的模型。

步骤 803b,训练 NWDAF 向推断 NWDAF 发送模型响应,携带模型信息。

其中,模型信息的具体实现,可以参考前述实施例的描述。

- 40 例如,上述步骤 801-步骤 803b 可选步骤。例如,如果推断 NWDAF 中配置了训练 NWDAF 的 NF 配置文件,则步骤 801-步骤 803b 可以不执行。

步骤 804a, 训练 NWDAF 向推断 NWDAF 发送模型性能信息订阅请求, 其中携带 Analytics ID, 模型性能指标(如精确率(Precision)、准确率(Accuracy)、错误率(Error Rate)、召回率(Recall)、F1 分数(F-Score)、均方误差(Mean Square Error, MSE)、均方根误差(Root Mean Squared Error, RMSE)、均方根对数误差(Root Mean Squared Logarithmic Error, RMSLE)、平均绝对误差(Mean Absolute Error, MAE)、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性)以及上报周期。

其中, 精确率、准确率、错误率、召回率、F1 分数用于指示分类类型或者标注类型的模型的性能。均方误差、均方根误差、均方根对数误差、平均绝对误差用于指示回归类模型的性能。模型推理时长用于指示模型预测需要的时间。模型鲁棒性用于指示模型处理缺失值和异常值的能力。模型可拓展性用于指示处理大数据集的能力。模型可解释性用于指示模型预测标准的可理解性, 比如, 决策树模型由于产生的规则或者树结构导致模型可解释性高, 神经网络模型由于存在大量模型参数导致模型可解释性低。

步骤 804b, 推断 NWDAF 向训练 NWDAF 发送模型性能信息通知, 其中携带 Analytics ID, 模型性能指标以及模型性能指标对应的取值。

其中, 推断 NWDAF 是基于上报周期, 周期性地向训练 NWDAF 发送模型性能信息通知。

基于上述步骤 804a 至步骤 804b, 推断 NWDAF 可以周期性地向训练 NWDAF 上报模型性能信息。

可选的, 模型性能信息通知中还可以携带推断 NWDAF 对模型的模型性能要求, 和/或, 推断 NWDAF 进行模型评估使用的数据等。其中, 模型性能要求可以辅助训练 NWDAF 判断是否需要再进行再训练以及判断再训练得到的模型性能是否满足推断 NWDAF 要求, 推断 NWDAF 进行模型评估使用的数据包括模型的输入数据、模型的输出数据(推断结果)、与推断结果对应的网络实际测量值(网络数据), 可供训练 NWDAF 对模型进行再训练时使用。

可选的, 上述步骤 804a 至步骤 804b 也可以由以下步骤 804a' 至步骤 804b' 替换。

步骤 804a', 训练 NWDAF 向推断 NWDAF 发送模型性能信息订阅请求, 其中携带 Analytics ID, 模型性能指标(如精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性)以及性能门限值。

步骤 804b', 推断 NWDAF 向训练 NWDAF 发送模型性能再训练通知, 其中携带 Analytics ID。

基于上述步骤 804a' 至步骤 804b', 推断 NWDAF 确定模型性能指标对应的取值达到性能门限值, 则向训练 NWDAF 上报模型性能信息通知, 该模型性能再训练通知用于触发训练 NWDAF 对模型再训练。可选的, 上述步骤 804b' 可以不携带性能门限值, 则可以由推断 NWDAF 自行确定性能门限值。该步骤 804b' 的模型性能再训练通知也可以称为模型性能达到门限值通知或模型性能信息通知。

可选的, 模型性能信息通知中还可以携带推断 NWDAF 对模型的模型性能要求, 和/或, 推断 NWDAF 进行模型评估使用的数据等。其中, 模型性能要求可以是推断 NWDAF 自行确定的门限值, 用于辅助训练 NWDAF 判断是否需要再进行再训练以及判断再训练得到的模型性能是否满足推断 NWDAF 要求, 推断 NWDAF 进行模型评估使用的数据包括模型

的输入数据、模型的输出数据（推断结果）、与推断结果对应的网络实际测量值（网络数据），可供训练 NWDAF 对模型进行再训练时使用。

可选的，上述步骤 804a 至步骤 804b 也可以由以下步骤 804a' 至步骤 804b' 替换。

- 步骤 804a'，训练 NWDAF 向推断 NWDAF 发送模型性能信息请求，其中携带 Analytics ID，模型性能指标（如精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性）。

步骤 804b'，推断 NWDAF 向训练 NWDAF 发送模型性能信息响应，其中携带 Analytics ID，模型性能指标以及模型性能指标对应的取值。

- 10 基于上述步骤 804a' 至步骤 804b'，训练 NWDAF 可以周期性地推断 NWDAF 发送模型性能信息请求，推断 NWDAF 每次收到模型性能信息请求，则基于模型性能指标进行模型性能评估，并向训练 NWDAF 发送模型性能信息响应。

- 可选的，模型性能信息响应中还可以携带推断 NWDAF 对模型的模型性能要求，和/或，推断 NWDAF 进行模型评估使用的数据等。其中，模型性能要求可以辅助训练 NWDAF 判断是否需要再进行训练以及判断再训练得到的模型性能是否满足推断 NWDAF 要求，推断 NWDAF 进行模型评估使用的数据包括模型的输入数据、模型的输出数据（推断结果）、与推断结果对应的网络实际测量值（网络数据），可供训练 NWDAF 对模型进行再训练时使用。

可选的，上述步骤 804a 至步骤 804b 也可以由以下步骤 804a'' 至步骤 804b'' 替换。

- 20 步骤 804a''，训练 NWDAF 向推断 NWDAF 发送模型性能数据订阅请求，其中携带 Analytics ID，以及上报周期。

步骤 804b''，推断 NWDAF 向训练 NWDAF 发送模型性能数据通知，其中携带 Analytics ID 和模型性能评估参考信息。

- 25 其中，模型性能评估参考信息包括模型的输入数据、模型的输出数据（推断结果）或与推断结果对应的网络实际测量值中的至少一项。

基于上述步骤 804a'' 至步骤 804b''，推断 NWDAF 基于上报周期，周期性地向训练 NWDAF 发送模型性能数据通知，即推断 NWDAF 可以周期性地向训练 NWDAF 上报模型性能评估参考信息。

- 30 其中，与推断结果对应的网络实际测量值（网络数据）可以由推断 NWDAF 从现网中采集后上报给训练 NWDAF，也可以由训练 NWDAF 自行从现网中采集。

可选的，模型性能数据通知中还可以携带推断 NWDAF 对模型的模型性能要求。

训练 NWDAF 可以根据推断 NWDAF 周期性上报的模型性能评估参考信息，构建测试集并进行模型性能评估。

- 35 可选的，上述步骤 804a 至步骤 804b 也可以由以下步骤 804a''' 至步骤 804b''' 替换。

步骤 804a'''，训练 NWDAF 向推断 NWDAF 发送模型性能数据请求，其中携带 Analytics ID。

可选的，模型性能数据请求中还包含时间范围，用于指示请求该时间范围内的性能数据。

- 40 步骤 804b'''，推断 NWDAF 向训练 NWDAF 发送模型性能数据响应，其中携带

Analytics ID 和模型性能评估参考信息。

其中，模型性能评估参考信息包括模型的输入数据、模型的输出数据（推断结果）、或与推断结果对应的网络实际测量（网络数据）值中的至少一项。

5 基于上述步骤 804a''' 至步骤 804b'''，训练 NWDAF 可以向推断 NWDAF 发送模型性能数据请求，推断 NWDAF 向训练 NWDAF 发送模型性能数据响应，即推断 NWDAF 向训练 NWDAF 发送模型性能评估参考信息，该模型性能评估参考信息可以是一定时间范围内的。

其中，与推断结果对应的网络实际测量值可以由推断 NWDAF 从现网中采集后上报给训练 NWDAF，也可以由训练 NWDAF 自行从现网中采集。

10 可选的，模型性能数据响应中还可以携带推断 NWDAF 对模型的模型性能要求。

训练 NWDAF 可以根据推断 NWDAF 发送的模型性能评估参考信息构建测试集并进行模型性能评估。

步骤 805，训练 NWDAF 确定启动模型再训练。

15 比如，若执行上述步骤 804a 至步骤 804b，则训练 NWDAF 确定模型性能指标对应的取值达到训练 NWDAF 预设的性能门限值或者不满足推断 NWDAF 的模型性能要求，则确定启动模型再训练。

再比如，若执行上述步骤 804a' 至步骤 804b'，则训练 NWDAF 收到模型性能信息通知，则确定启动模型再训练。

20 比如，若执行上述步骤 804a'' 至步骤 804b''，则训练 NWDAF 确定模型性能指标对应的取值达到训练 NWDAF 预设的性能门限值或者不满足推断 NWDAF 的模型性能要求，则确定启动模型再训练。

再比如，若执行上述步骤 804a''' 至步骤 804b'''，或执行上述步骤 804a'''' 至步骤 804b''''，则训练 NWDAF 根据模型性能评估参考信息，确定模型性能达到训练 NWDAF 预设的性能门限值，或者不满足推断 NWDAF 的模型性能要求，则确定启动模型再训练。

25 步骤 806，训练 NWDAF 向推断 NWDAF 发送模型更新请求，携带 Analytics ID 以及新的模型信息。

可选的，模型更新请求中的新模型信息可以是新模型的参数项取值、或新模型文件或包含新模型的镜像文件、或新模型的地址（如 URL 或 IP 地址）。

30 需要说明的是，如果步骤 806 中携带的是新模型的地址，则推断 NWDAF 可以根据该地址获取包含新模型信息的文件，该文件可以是包含新模型的参数项取值的文件、或是模型文件、或是包含新模型的镜像文件。

步骤 807，推断 NWDAF 向训练 NWDAF 发送模型更新响应。

步骤 808，推断 NWDAF 进行模型更新。

35 也即，推断 NWDAF 根据接收到的新的模型信息，对正在使用的旧模型进行更新或者替换。

可选的，推断 NWDAF 在模型更新前，对新的模型信息进行本地测试，测试通过后再进行更新或者替换。

40 其中，上述步骤 806-步骤 808 可选步骤。例如，如果步骤 805 中训练 NWDAF 确定模型不需要再训练或者训练 NWDAF 可以容忍模型性能下降到模型性能要求以下或者训练 NWDAF 上当前不具备模型再训练的能力（如硬件资源有限），步骤 806-步骤 808 可以不

执行。

基于上述实施例，训练 NWDAF 向推断 NWDAF 发送模型性能订阅或模型性能请求，实现对推断 NWDAF 中模型性能的监控，当性能下降满足再训练条件时，训练 NWDAF 及时

5 模型性能。并进行再训练，并将新模型发送给推断 NWDAF 用于更新，保证了推断 NWDAF 中模型的

实施例五

如图 9 所示，为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程图示意图。

- 10 该实施例五是在实施例四的基础上，考虑有多个推断 NWDAF 存在的场景，具体场景同实施例二，可参考实施例二的场景描述。类似实施例二，本实施例考虑在模型更新请求中增加指示模型性能信息的参数，包括精度、所需计算量等，帮助其它暂不需要更新的推断 NWDAF 判断是否需要请求新模型。

该实施例包括以下步骤：

- 15 步骤 901 至步骤 905，类似于上述实施例四中的步骤 801 至步骤 805。

需要说明的是，该步骤 901 至步骤 905 中，涉及推断 NWDAF1 的相关操作以及推断 NWDAF2 的相关操作，均可以分别参考上述步骤 802a 至步骤 804b 中关于推断 NWDAF 的相关操作。并且，步骤 905 中，训练 NWDAF 是根据推断 NWDAF1 发送的模型性能信息通知或模型性能信息响应，触发启动模型再训练。

- 20 接下来，训练 NWDAF 需要通知推断 NWDAF 执行模型更新。

第一种方案是：不区分推断 NWDAF，即训练 NWDAF 始终将训练得到的新的模型信息发送给所有推断 NWDAF。该方案参考以下步骤 906a 至步骤 906b。

第二种方案是：区分不同的推断 NWDAF，仅向触发训练 NWDAF 执行模型训练的推断 NWDAF 发送新的模型信息。该方案参考以下步骤 907a 至步骤 907c。

- 25 需要说明的是，上述第一种方案与第二种方案，二者选择一种进行执行。

第一种方案：

步骤 906a，训练 NWDAF 向推断 NWDAF1 发送模型更新请求，携带 Analytics ID，新的模型信息以及模型性能信息。

- 30 步骤 906b，推断 NWDAF1 判断是否需要更新模型。

推断 NWDAF1 收到模型更新请求后，可以基于模型性能信息，和/或，对新模型信息的本地测试结果判断是否需要更新模型。如果确定需要更新，则使用新的模型信息更新或者替换旧模型。

- 35 步骤 906c，训练 NWDAF 向推断 NWDAF2 发送模型更新请求，携带 Analytics ID，新的模型信息以及模型性能信息。

步骤 906d，推断 NWDAF2 判断是否需要更新模型。

推断 NWDAF2 收到模型更新请求后，可以基于模型性能信息，和/或，对新模型的本地测试结果判断是否需要更新模型。如果确定需要更新，则使用新的模型信息更新或者替换旧模型。

40

第二种方案:

步骤 907a, 训练 NWDAF 向推断 NWDAF1 发送模型更新请求, 携带 Analytics ID 以及新的模型信息。

步骤 907b, 推断 NWDAF1 更新模型。

5 推断 NWDAF1 收到模型更新请求后, 使用新的模型信息更新或者替换旧模型。

可选的, 推断 NWDAF 在模型更新前, 对新的模型信息进行本地测试, 测试通过后再进行更新或者替换。

步骤 907c, 训练 NWDAF 向推断 NWDAF2 发送模型训练完成通知, 携带 Analytics ID 以及模型性能信息。

10 步骤 907d, 推断 NWDAF2 判断是否需要更新模型。

比如, 推断 NWDAF2 可以根据自身算力、模型性能要求, 以及接收到的模型性能信息, 决定是否需要更新模型。

当推断 NWDAF2 确定需要更新模型, 则执行以下步骤 907e 至步骤 907g, 否则不执行以下步骤 907e 至步骤 907g。

15 步骤 907e, 可选的, 推断 NWDAF2 向训练 NWDAF 发送模型请求, 携带 Analytics ID。

Analytics ID 用于指示与该 Analytics ID 对应的模型。

步骤 907f, 可选的 (取决于步骤 907e 是否执行), 训练 NWDAF 向推断 NWDAF2 发送模型响应, 携带新的模型信息。

步骤 907g, 推断 NWDAF2 更新模型。

20 也即, 推断 NWDAF2 根据接收到的新的模型信息, 对正在使用的旧模型进行更新。

可选的, 推断 NWDAF2 在模型更新前, 对新的模型信息进行本地测试, 测试通过后再进行更新或者替换。

基于上述实施例, 使用相同模型的推断 NWDAF 可以获取新模型的信息, 并根据模型性能信息确定是否需要请求新模型, 避免不必要的模型传输和本地评估过程。

25

实施例六

如图 10 所示, 为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程示意图。

30 该实施例是在实施例四的基础上, 考虑针对同一个 Analytics ID, 需要多个子模型共同协作完成分析的场景。类似实施例三的解决方案, 本实施例考虑进一步地增加 model ID 来标识各个子模型, 训练 NWDAF 给子模型分配不同的 model ID, 通过 model ID 精确监控每个子模型的性能。

该实施例包括以下步骤:

步骤 1001, 训练 NWDAF 向 NRF 注册。

35 训练 NWDAF 向 NRF 发送 NF 注册请求, 携带 NF 配置文件, 该 NF 配置文件包括 NF 类型, NF 服务名称 (NF Service), 分析类型标识 (Analytics ID) 等信息。

相应地, NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

步骤 1002a, 推断 NWDAF 向 NRF 发送 NF 发现请求, 携带 NF 配置文件。

40 比如携带的 NF 配置文件包含 NF 类型 (如 NWDAF), NF 服务名称 (如 ModelProvision), 以及携带 Analytics ID, 则该 NF 发现请求用于请求从 NRF 获取对应该 Analytics ID 的训练

NWDAF。

步骤 1002b, NRF 向推断 NWDAF 发送 NF 发现响应, 其中携带 NWDAF 实例。

其中, 携带的 NWDAF 实例是训练 NWDAF 的一个实例, 可以用训练 NWDAF 的 ID 或 IP 地址来表示。

5 步骤 1003a, 推断 NWDAF 向训练 NWDAF 发送模型请求, 其中携带 Analytics ID。

其中, 推断 NWDAF 可以基于从 NRF 获取到的训练 NWDAF 的 ID 或 IP 地址, 向训练 NWDAF 发送模型请求, 携带的 Analytics ID 用于指示请求获取与该 Analytics ID 对应的模型。

步骤 1003b, 训练 NWDAF 向推断 NWDAF 发送模型响应, 携带模型信息和模型标识。

10 其中, 模型信息的具体实现, 可以参考前述实施例的描述。

其中, 每个模型标识对应一个模型信息。

可选的, 模型信息和模型标识可以以模型列表形式实现, 也即模型响应携带模型列表, 模型列表包含模型信息与模型标识以及模型信息与模型标识之间的对应关系。示例性地, 模型列表包括: <模型信息 1, 模型标识 1>, <模型信息 2, 模型标识 2>, ...。

15 其中, 上述步骤 1001-步骤 1003b 可选步骤行。例如, 如果推断 NWDAF 中配置了训练 NWDAF 的 NF 配置文件, 则步骤 1001-步骤 1003b 可以不执行。

步骤 1004a, 训练 NWDAF 向推断 NWDAF 发送模型性能信息订阅请求, 其中携带 Analytics ID, 模型性能指标 (如精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性)、上报周期以及模型标识。

20 需要说明的是, 模型性能信息订阅请求可以携带多个模型标识以及每个模型标识对应的模型性能指标、上报周期。特别地, 当每个模型标识对应的上报周期相同时, 则可以只携带一个上报周期。

步骤 1004b, 推断 NWDAF 向训练 NWDAF 发送模型性能信息通知, 其中携带 Analytics ID, 模型性能指标以及模型性能指标对应的取值。

其中, 推断 NWDAF 是基于上报周期, 周期性地向训练 NWDAF 发送各个子模型对应的模型性能信息通知。

基于上述步骤 1004a 至步骤 1004b, 推断 NWDAF 可以周期性地向训练 NWDAF 上报各个子模型对应的模型性能信息。

30 可选的, 模型性能信息通知中还可以携带推断 NWDAF 对各子模型的模型性能要求, 和/或, 推断 NWDAF 进行各子模型评估使用的数据等。其中, 模型性能要求可以辅助训练 NWDAF 判断是否需要进行再训练以及判断再训练得到的模型性能是否满足推断 NWDAF 要求, 推断 NWDAF 进行模型评估使用的数据包括模型的输入数据、模型的输出数据 (推断结果)、与推断结果对应的网络实际测量值, 可供训练 NWDAF 对模型进行再训练时使用。

35 可选的, 上述步骤 1004a 至步骤 1004b 也可以由以下步骤 1004a' 至步骤 1004b' 替换。

步骤 1004a', 训练 NWDAF 向推断 NWDAF 发送模型性能信息订阅请求, 其中携带 Analytics ID, 模型性能指标 (如精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性)、性能门限值以及模型标识。

需要说明的是，模型性能信息订阅请求可以携带多个模型标识以及每个模型标识对应的模型性能指标、性能门限值。特别地，当每个模型标识对应的性能门限值相同时，则可以只携带一个性能门限值。

步骤 1004b'，推断 NWDAF 向训练 NWDAF 发送模型性能再训练通知，其中携带 Analytics ID。

基于上述步骤 1004a' 至步骤 1004b'，推断 NWDAF 确定子模型的模型性能指标对应的取值达到性能门限值，则向训练 NWDAF 上报子模型对应的模型性能信息通知，该模型性能再训练通知用于触发训练 NWDAF 对子模型再训练。可选的，上述步骤 1004b' 可以不携带性能门限值，则可以由推断 NWDAF 自行确定性能门限值。该步骤 1004b' 的模型性能再训练通知也可以称为模型性能达到门限值通知或模型性能信息通知。

可选的，模型性能信息通知中还可以携带推断 NWDAF 对子模型的模型性能要求，和/或，推断 NWDAF 进行模型评估使用的数据等。其中，模型性能要求可以是推断 NWDAF 自行确定的门限值，用于辅助训练 NWDAF 判断是否需要再进行再训练以及判断再训练得到的模型性能是否满足推断 NWDAF 要求，推断 NWDAF 进行模型评估使用的数据包括模型的输入数据、模型的输出数据（推断结果）、与推断结果对应的网络实际测量值，可供训练 NWDA 对模型进行再训练时使用。

可选的，上述步骤 1004a 至步骤 1004b 也可以由以下步骤 1004a'' 至步骤 1004b'' 替换。

步骤 1004a''，训练 NWDAF 向推断 NWDAF 发送模型性能信息请求，其中携带 Analytics ID，模型性能指标（如精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性）以及模型标识。

需要说明的是，模型性能信息请求可以携带多个模型标识以及每个模型标识对应的模型性能指标。

步骤 1004b''，推断 NWDAF 向训练 NWDAF 发送模型性能信息响应，其中携带 Analytics ID，模型性能指标以及模型性能指标对应的取值。

基于上述步骤 1004a'' 至步骤 1004b''，训练 NWDAF 可以周期性地推断 NWDAF 发送模型性能信息请求，推断 NWDAF 每次收到模型性能信息请求，则基于模型性能指标进行模型性能评估，并向训练 NWDAF 发送子模型对应的模型性能信息响应。

可选的，模型性能信息响应中还可以携带推断 NWDAF 对各子模型的模型性能要求，和/或，推断 NWDAF 进行模型评估使用的数据等。其中，模型性能要求可以辅助训练 NWDAF 判断是否需要再进行再训练以及判断再训练得到的模型性能是否满足推断 NWDAF 要求，推断 NWDAF 进行模型评估使用的数据包括模型的输入数据、模型的输出数据（推断结果）、与推断结果对应的网络实际测量值（网络数据），可供训练 NWDAF 对模型进行再训练时使用。

可选的，上述步骤 1004a 至步骤 1004b 也可以由以下步骤 1004a''' 至步骤 1004b''' 替换。

步骤 1004a'''，训练 NWDAF 向推断 NWDAF 发送模型性能数据订阅请求，其中携带 Analytics ID，上报周期以及模型标识。

需要说明的是，模型性能数据订阅请求可以携带多个模型标识以及每个模型标识对应

的上报周期。

步骤 1004b''', 推断 NWDAF 向训练 NWDAF 发送模型性能数据通知, 其中携带 Analytics ID 和模型性能评估参考信息。

其中, 模型性能评估参考信息包括模型的输入数据、模型的输出数据(推断结果)或
5 与推断结果对应的网络实际测量值(网络数据)中的至少一项。

需要说明的是, 这里的模型性能评估参考信息可以是多个模型性能评估参考信息, 具体的, 每个模型标识对应一个模型性能评估参考信息。

基于上述步骤 1004a''' 至步骤 1004b''', 推断 NWDAF 基于上报周期, 周期性地向训练 NWDAF 发送模型性能数据通知, 即推断 NWDAF 可以周期性地向训练 NWDAF 上
10 报各个子模型分别对应的模型性能评估参考信息。

其中, 与推断结果对应的网络实际测量值(网络数据)可以由推断 NWDAF 从现网中采集后上报给训练 NWDAF, 也可以由训练 NWDAF 自行从现网中采集。

可选的, 模型性能数据通知中还可以携带推断 NWDAF 对各个子模型的模型性能要求。

训练 NWDAF 可以根据推断 NWDAF 周期性上报的模型性能评估参考信息, 构建测试
15 集并进行模型性能评估。

可选的, 上述步骤 1004a 至步骤 1004b 也可以由以下步骤 1004a'''' 至步骤 1004b'''' 替换。

步骤 1004a''''', 训练 NWDAF 向推断 NWDAF 发送模型性能数据请求, 其中携带 Analytics ID 以及模型标识。

20 需要说明的是, 模型性能数据订阅请求可以携带多个模型标识。

可选的, 模型性能数据请求中还包含时间范围, 用于指示请求该时间范围内的性能数据。具体的, 每个模型标识可以对应一个时间范围。

步骤 1004b''''', 推断 NWDAF 向训练 NWDAF 发送模型性能数据响应, 其中携带 Analytics ID 和模型性能评估参考信息。

25 其中, 模型性能评估参考信息包括模型的输入数据、模型的输出数据(推断结果), 和/或, 或与推断结果对应的网络实际测量值(网络数据)中的至少一项。

需要说明的是, 这里的模型性能评估参考信息可以是多个模型性能评估参考信息, 具体的, 每个模型标识对应一个模型性能评估参考信息。

基于上述步骤 1004a'''' 至步骤 1004b''''', 训练 NWDAF 可以向推断 NWDAF 发送模型性能数据请求, 推断 NWDAF 向训练 NWDAF 发送模型性能数据响应, 即推断 NWDAF 向训练 NWDAF 发送模型性能评估参考信息, 该模型性能评估参考信息可以是一定时间范围内的。
30

其中, 与推断结果对应的网络实际测量值(网络数据)可以由推断 NWDAF 从现网中采集后上报给训练 NWDAF, 也可以由训练 NWDAF 自行从现网中采集。

35 可选的, 模型性能数据响应中还可以携带推断 NWDAF 对模型的模型性能要求。

训练 NWDAF 可以根据推断 NWDAF 发送的模型性能评估参考信息构建测试集并进行模型性能评估。

步骤 1005, 训练 NWDAF 确定启动模型再训练。

比如, 若执行上述步骤 1004a 至步骤 1004b, 则训练 NWDAF 确定模型性能指标对应的取值达到训练 NWDAF 预设的性能门限值或者不满足推断 NWDAF 的模型性能要求, 则
40

确定启动模型再训练。

再比如，若执行上述步骤 1004a' 至步骤 1004b'，则训练 NWDAF 收到模型性能信息通知，则确定启动模型再训练。

比如，若执行上述步骤 1004a'' 至步骤 1004b''，则训练 NWDAF 确定模型性能指标对应的取值达到训练 NWDAF 预设的性能门限值或者不满足推断 NWDAF 的模型性能要求，则确定启动模型再训练。

再比如，若执行上述步骤 1004a''' 至步骤 1004b'''，或执行上述步骤 1004a'''' 至步骤 1004b''''，则训练 NWDAF 根据模型性能评估参考信息，确定模型性能达到训练 NWDAF 预设的性能门限值，或者不满足推断 NWDAF 的模型性能要求，则确定启动模型再训练。

步骤 1006，训练 NWDAF 向推断 NWDAF 发送模型更新请求，携带 Analytics ID、新的模型信息以及模型标识。

需要说明的是，模型更新请求可以携带多个模型标识以及每个模型标识对应的新的模型信息。

步骤 1007，推断 NWDAF 向训练 NWDAF 发送模型更新响应。

步骤 1008，推断 NWDAF 进行模型更新。

也即，推断 NWDAF 根据接收到的新的模型信息，对正在使用的旧模型（具体是旧的子模型）进行更新或者替换。

可选的，推断 NWDAF 在模型更新前，对新的模型信息进行本地测试，测试通过后再进行更新或者替换。

其中，上述步骤 1006-步骤 1008 可选步骤。例如，如果步骤 805 中训练 NWDAF 确定模型不需要再训练或者训练 NWDAF 可以容忍模型性能下降到模型性能要求以下或者训练 NWDAF 上当前不具备模型再训练的能力（如硬件资源有限），步骤 1006-步骤 1008 可以不执行。

基于上述实施例，通过增加子模型的标识，按照模型粒度进行性能监控，在一个 Analytics ID 对应多个子模型的场景下可以实现精准的模型再训练与更新，避免浪费训练和传输资源。

实施例七

如图 11 所示，为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程示意图。

该实施例考虑由训练 NWDAF 进行周期性再训练，并通知推断 NWDAF 有可用的新模型。该实施例适用于推断 NWDAF 不具备评估功能的场景，即无法获取推断 NWDAF 关于模型性能的实时反馈。为了维持模型的性能，训练 NWDAF 可以周期性的进行再训练。

该实施例包括以下步骤：

步骤 1101，训练 NWDAF 向 NRF 注册。

训练 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件，该 NF 配置文件包括 NF 类型，NF 服务名称（NF Service），分析类型标识（Analytics ID）等信息，以及还包括模型索引信息。该模型索引信息可以是模型版本号（version）、位置信息（location）或统一资源定位符（Uniform Resource Locator，URL）等。其中，version 表示模型版本，location

或 URL 表示模型的存储位置，三者中使用任意一个都可以。可选的，当模型索引信息是 location 或 URL，该 location 或 URL 中也可以包含 version。

相应地，NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

步骤 1102，推断 NWDAF 发现训练 NWDAF，并向训练 NWDAF 请求获取模型信息。

5 该过程可以参考图 3 实施例中的步骤 304 至步骤 307，不再赘述。基于该过程，推断 NWDAF 可以从训练 NWDAF 中获取到模型信息。

步骤 1103，推断 NWDAF 向 NRF 订阅训练 NWDAF 的状态。

后续训练 NWDAF 在 NRF 中注册的 NF 配置文件发生更新时，NRF 通知推断 NWDAF。

步骤 1104，训练 NWDAF 周期性地启动模型再训练。

10 比如，训练 NWDAF 可以设置定时器，每隔固定的时间再训练一次。

步骤 1105，训练 NWDAF 向 NRF 发送 NF 更新请求，携带更新的 NF 配置文件。

该更新的 NF 配置文件至少携带更新的模型索引信息。可选的，该更新的 NF 配置文件还携带 NF 类型，NF 服务名称（NF Service）等。

15 可选的，该更新的 NF 配置文件中还可以携带更新的模型性能信息，如模型准确度、达到该准确度所需的硬件能力信息、模型推理所需计算量、模型的推理时长、模型的大小等。

步骤 1106，NRF 更新存储的 NF 配置文件。

步骤 1107，NRF 向训练 NWDAF 发送 NF 更新响应。

该 NF 更新响应用于通知 NF 配置文件更新成功。

20 步骤 1108，NRF 向推断 NWDAF 发送 NF 状态更新通知，其中携带更新的模型索引信息。

可选的，该 NF 状态更新通知还携带指示信息，用于指示更新的类型是模型索引信息更新。

25 步骤 1109，推断 NWDAF 向训练 NWDAF 发送模型请求，携带 Analytics ID 和更新的模型索引信息。

Analytics ID 用于指示与该 Analytics ID 对应的模型。

步骤 1110，训练 NWDAF 向推断 NWDAF 发送模型响应，携带模型信息。

该模型信息包括与更新的模型索引信息对应的模型信息，也即获取到新的模型对应的模型信息。

30 步骤 1111，推断 NWDAF 进行模型更新。

也即，推断 NWDAF 根据接收到的新的模型信息，对正在使用的旧模型进行更新。

可选的，推断 NWDAF 在模型更新前，对新的模型信息进行本地测试，测试通过后再进行更新或者替换。

35 其中，上述步骤 1109-步骤 1111 可选步骤。例如，如果步骤 1108 后中推理 NWDAF 可以容忍模型性能下降到模型性能要求，步骤 1109-步骤 1111 可以不执行。

需要说明的是，上述步骤 1104-步骤 1108 是周期性进行的，因此步骤 1109-步骤 1111 是可选的，因为训练 NWDAF 只负责进行周期性再训练，是否请求新的模型进行更新由推断 NWDAF 自行确定。

40 基于该实施例，在推断 NWDAF 不具备评估功能的场景下，即训练 NWDAF 无法获取推断 NWDAF 关于模型性能的实时反馈时可以进行周期性再训练，以保证模型的性能。

需要说明的是，基于上述实施例七，在当有多个推断 NWDAF 时，上述步骤 1107 之后，NRF 可以向多个推断 NWDAF 发送 NF 状态更新通知，从而使得多个推断 NWDAF 均可以向训练 NWDAF 发送模型请求，从而实现多个推断 NWDAF 的模型更新。

需要说明的是，基于上述实施例七，当一个分析类型标识对应多个子模型，每个子模型用一个模型标识进行标识，则上述步骤 1105 中还可以携带一个或多个模型标识，以及上述步骤 1108 中，还可以携带该一个或多个模型标识，进而上述步骤 1109 中可以携带该一个或多个模型标识，从而实现推断 NWDAF 中的一个或多个子模型的更新。

实施例八

如图 12 所示，为本申请实施例提供的又一种训练-推断分离场景下保证模型有效性的方法流程示意图。

该实施例八与实施例七场景相同，即推断 NWDAF 不具备评估功能时，由训练 NWDAF 进行周期性再训练，并给推断 NWDAF 发送模型更新消息。

该实施例包括以下步骤：

步骤 1201，训练 NWDAF 向 NRF 注册。

训练 NWDAF 向 NRF 发送 NF 注册请求，携带 NF 配置文件，该 NF 配置文件包括 NF 类型，NF 服务名称（NF Service），分析类型标识（Analytics ID）等信息。

相应地，NRF 存储 NF 配置文件以及向训练 NWDAF 发送 NF 注册响应。

步骤 1202，推断 NWDAF 发现训练 NWDAF，并向训练 NWDAF 请求获取模型。

该过程可以参考图 3 实施例中的步骤 304 至步骤 307，不再赘述。基于该过程，推断 NWDAF 可以从训练 NWDAF 中获取到模型信息。

步骤 1203，训练 NWDAF 周期性地启动模型再训练。

比如，训练 NWDAF 可以设置定时器，每隔固定的时间再训练一次。

步骤 1204，训练 NWDAF 向推断 NWDAF 发送模型更新请求，携带 Analytics ID 以及新的模型信息。

可选的，模型更新请求中还可以携带新模型的性能信息，如模型准确度、达到该准确度所需的硬件能力信息、模型推理所需计算量、模型的推理时长、模型的大小等。

步骤 1205，推断 NWDAF 向训练 NWDAF 发送模型更新响应。

步骤 1206，推断 NWDAF 进行模型更新。

也即，推断 NWDAF 根据接收到的新的模型信息，对正在使用的旧模型进行更新。

可选的，推断 NWDAF 在模型更新前，对新的模型信息进行本地测试，测试通过后再进行更新或者替换。

其中，上述步骤 1204-步骤 1206 可选步骤。例如，如果步骤 1203 中训练 NWDAF 确定训练更新后的模型的模型性能评估结果小于或者等于步骤 1202 中训练 NWDAF 向推断 NWDAF 提供的模型的性能评估结果，步骤 1204-步骤 1206 可以不执行。

基于上述实施例，针对推断 NWDAF 不具备评估功能的场景，当训练 NWDAF 无法获取推断 NWDAF 关于模型性能的实时反馈时可以进行周期性再训练，以保证模型的性能。

需要说明的是，基于上述实施例八，在当有多个推断 NWDAF 时，上述步骤 1204 中，训练 NWDAF 可以向多个推断 NWDAF 发送模型更新请求，从而实现多个推断 NWDAF 的模型更新。

需要说明的是，基于上述实施例八，当一个分析类型标识对应多个子模型，每个子模型用一个模型标识进行标识，则上述步骤 1204 中还可以携带一个或多个模型标识，从而实现推断 NWDAF 中的一个或多个子模型的更新。

5 实施例九

该实施例九与上述实施例一至实施例八之间的关系是：上述实施例一至实施例八是该实施例九的各种不同的具体实现方式。如图 13 所示，为本申请实施例提供的一种通信方法流程图。需要说明的是，该实施例九中的第一 NWDAF 可以是上述实施例一至实施例八中的训练 NWDAF，第二 NWDAF 可以是上述实施例一至实施例八中的推断 NWDAF1，第三 NWDAF 可以是上述实施例一至实施例八中的推断 NWDAF2。

该方法包括以下步骤：

步骤 1301，第一 NWDAF 向第二 NWDAF 发送第三信息。相应地，第二 NWDAF 收到第三信息。

该第三信息包括模型的性能指标，该模型的性能指标用于获取模型的性能的评估结果。可选的，模型性能指标包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性。也即，第二 NWDAF 根据接收到的模型的性能指标，对正在使用的模型进行性能评估，进而得到性能的评估结果，并生成模型的性能报告。

可选的，该第三信息还包括以下一项或多项：分析类型标识、模型的标识、子模型的标识。其中，分析类型标识（Analytics ID）用于指示模型的分析类型，比如可以是 Service Experience，Network Performance，UE Mobility 等。模型的标识用于标识模型。子模型的标识用于标识该模型的子模型。需要说明的是，当该模型没有子模型，则第三信息可以携带模型的标识，且不需要携带子模型的标识，或者第三信息既不携带模型的标识，也不携带子模型的标识。当该模型有子模型，则需要同时携带模型的标识，以及携带一个或多个子模型的标识。需要说明的是，当第三信息携带子模型的标识，则模型的性能指标用于获取该模型的子模型的性能的评估结果。

其中，关于子模型的具体示例，可以参考上述实施例三以及实施例六的描述。

可选的，该第三信息还包括以下一项或多项：上报周期、门限信息。其中，该上报周期用于指示上报模型的性能报告的时间，也即用于指示第二 NWDAF 基于该上报周期向第一 NWDAF 上报模型的性能报告。该门限信息用于指示上报模型的性能报告的条件，也即当第二 NWDAF 得到的模型的评估结果达到了该门限信息对应的门限值，则第二 NWDAF 向第一 NWDAF 上报模型的性能报告。

需要说明的是，该步骤 1301 为可选步骤。当不执行该步骤 1301，则可以预先在第二 NWDAF 上预先配置上述第三信息，或者是由其它网元向第二 NWDAF 配置上述第三信息。

步骤 1302，第二 NWDAF 向第一 NWDAF 发送第一信息。相应地，第一 NWDAF 收到第一信息。

该第一信息包括模型的性能报告，该模型的性能报告用于指示模型的性能的评估结果，或者，该模型的性能报告用于指示模型的性能的评估结果不满足模型的性能指标的要求。

可选的，第一信息还包括模型的性能报告对应的以下一项或多项信息：时间、区域、切片。该时间指的是生成该模型的性能报告的时间范围，该区域指的是该模型的性能报告

对应的区域范围，该切片指的是该模型的性能报告对应的切片信息。

步骤 1303，第一 NWDAF 根据模型的性能报告更新模型的第一模型信息，获得模型的第二模型信息。

5 步骤 1304，第一 NWDAF 向第二 NWDAF 发送第二信息。相应地，第二 NWDAF 收到第二信息。

该第二信息包括第二模型信息。

可选的，第二信息还包括以下一项或多项：模型的标识、子模型的标识、模型的性能评估结果、模型的性能评估结果对应的硬件能力信息、模型的大小、模型的推理时长。其中，模型的性能评估结果对应的硬件能力信息指的是运行该模型所需要的硬件能力要求，
10 如要求具备图形处理器（Graphic Processing Unit，GPU）加速能力，模型的推理时长指的是该模型接收输入到产生输出之间的时延。可选的，每种硬件能力信息对应一种推理时长，硬件能力越强，则推断时长越短。

步骤 1305，第二 NWDAF 根据第二信息，更新模型。

15 比如，第二 NWDAF 基于第二信息，使用第二模型信息替换掉第一模型信息，实现模型更新。

可选的，第一 NWDAF 还可以向除第二 NWDAF 之外的其它 NWDAF（如第三 NWDAF）发送第二信息。也即，由第二 NWDAF 触发第一 NWDAF 更新模型，得到第二模型信息，但第一 NWDAF 不仅将第二信息发送给第二 NWDAF，还将第二信息发送给第三 NWDAF，以实现第三 NWDAF 对模型进行更新，从而避免第三 NWDAF 需要单独向第一 NWDAF
20 请求模型更新，可以节约信令开销。

其中，关于第一 NWDAF 向第三 NWDAF 发送第二信息的具体示例，可以参考上述实施例二以及实施例五的描述。

作为一种实现方法，上述步骤 1301 具体可以是：第一 NWDAF 通过 NRF 向第二 NWDAF 发送第三信息，上述步骤 1302 具体可以是：第一 NWDAF 通过 NRF 接收来自第二 NWDAF 的第一信息，上述步骤 1304 具体可以是：第一 NWDAF 通过 NRF 向第二 NWDAF 发送第二信息。也即，当第一 NWDAF 与第二 NWDAF 之间没有接口时，则可以通过 NRF 作为中间网元，实现第一 NWDAF 与第二 NWDAF 之间的交互。
25

其中，关于通过 NRF 作为中间网元的具体示例，可以参考上述实施例一至实施例三的描述。

30 基于上述方案，当第二 NWDAF 无法完成模型训练时，则第二 NWDAF 可以向第一 NWDAF 发送模型的性能报告，从而第一 NWDAF 可以根据该模型的性能报告对模型进行更新，得到模型的第二模型信息，并将第二模型信息发送给第二 NWDAF，使得第二 NWDAF 可以基于第二模型信息更新模型，从而可以实现在模型性能下降时及时对模型进行训练，进而可以保证模型性能。

35 示例性的，假设模型为业务体验模型，该模型可以用于基于业务流对应的网络数据（比如业务流对应的终端在基站侧的空口质量、该业务流对应的终端的会话的服务质量流在用户面管理网元上的带宽、时延、抖动等）评估该业务流的业务体验，网络侧策略控制网元（Policy Charging Function，PCF）可以根据该模型的业务体验输出结果确定该业务流的体验要求是否满足，如果不满足，则可以调整该业务的 QoS 参数。这里，PCF 进行 QoS 参数调整的前提是业务体验模型性能足够好，否则会影响业务体验。比如，以语音业务的业
40

务体验,也就是 MOS 分 (Mean Opinion Score, 平均意见分) 为例, MOS 分要求为 3.0 分, 如果业务流的实际 MOS 分为 2.5 分, 但是模型的输出 MOS 分为 3.5 分, 那么 PCF 不会对业务的 QoS 参数进行调整, 这样就导致业务体验很差, 如果模型的性能足够好, 那么模型的输出 MOS 分应该为 2.5 分, 这样 PCF 就会对业务 QoS 参数进行调整, 使得 MOS 分达到 3.0 分以上。针对该示例, 模型的性能影响业务体验。并且如果模型性能持续下降, 最终可能会恶化到模型完全不可用的程度, 造成极差的业务体验或者业务中断。

联邦学习作为一种新型人工智能技术, 可以在原始数据不出本域的情况下实现模型跨域联合训练, 既可以提高训练的效率, 最重要的, 可以通过联邦学习技术, 避免数据汇聚到数据分析中心时带来的安全问题 (比如, 原始数据在传输过程中被劫持, 原始数据被数据中心错误使用等)。横向联邦学习, 作为一种联邦学习技术, 适合“特征重复度非常高, 但是数据样本之间差异较大”的训练数据场景。

如图 14(a)所示, 为横向联邦学习的训练过程 (以线性回归为例)。可以看到, 横向联邦包括一个中心服务器 (server) 节点以及多个边缘客户端 (client) 节点 (例如, client 节点 A、client 节点 B 以及 client 节点 K), 这其中, 原始数据都分布在各个 client 节点, server 节点不具有原始数据, 并且 client 节点不允许将原始数据发送给 server 节点。

首先, 各个 client 节点上的数据集 (假设共 K 个 client 节点, 也就是存在 K 个数据集) 分别是:

$$\{x_i^A, y_i^A\}_i \in D_A, \{x_j^B, y_j^B\}_j \in D_B, \dots, \{x_k^K, y_k^K\}_k \in D_K,$$

其中, x 为样本数据, y 为样本数据对应的标签数据。横向联邦学习中每个样本数据都包括标签, 即标签和数据存放在一起。

然后, 每个 client 节点上的数据分析模块可以根据线性回归算法各自训练自己的模型, 称之为子模型, 即:

$$h(x_i) = \Theta_A x_i^A, h(x_j) = \Theta_B x_j^B, \dots, h(x_k) = \Theta_K x_k^K.$$

假设线性回归所使用的损失函数是均方误差 (Mean Squared Error, MSE), 那么每个子模型训练的目标函数 (整个训练的过程就是使得上述损失函数的值最小) 为:

$$\min L_I = \sum_i \|\Theta_I x_i^I - y_i^I\|^2 + \frac{\lambda}{2} \|\Theta_I\|^2, I = A, B, \dots, K$$

下面才真正开始训练过程, 针对每一次迭代过程,

(1) 每个 client 节点生成的子模型梯度如下:

$$\frac{\partial L_I}{\partial \Theta_I} = \sum_i (\Theta_I x_i^I - y_i^I) x_i^I + \lambda \Theta_I, I = A, B, \dots, K$$

(2) 每个 client 上报样本个数以及本地梯度值, 即:

$$N_I \text{ 以及 } \frac{\partial L_I}{\partial \Theta_I}, \text{ 其中, } N_I \text{ 表示样本个数, } \frac{\partial L_I}{\partial \Theta_I} \text{ 表示本地梯度值。}$$

(3) server 节点收到上述信息后, 对梯度进行聚合, 如下:

$$\frac{1}{\|K\|} \sum_I \frac{\partial L_I}{\partial \Theta_I} * P_I,$$

$$\text{其中, } \|K\| \text{ 为 client 节点的个数, } P_I = N_I / \sum_I N_I.$$

(4) server 节点将聚合后的梯度下发给每一个参与训练的 client 节点, 然后 client 节点本地更新模型参数, 如下:

$$\Theta_I := \Theta_I + \alpha \frac{1}{\|K\|} \sum_I \frac{\partial L_I}{\partial \Theta_I} * P_I, \quad I = A, B, \dots, K$$

(5) client节点进行模型参数更新后, 计算损失函数值 L_I , 转至步骤(1)。

上述训练过程, server节点可以通过迭代次数控制训练结束, 比如训练10000次终止训练, 或者通过设置损失函数的阈值控制训练结束, 比如 $L_I \leq 0.0001$ 时, 训练结束。

- 5 训练结束后, 每个client节点都会保留着同一份模型(可以来自server节点, 也可以是本地进一步根据来自server节点本地个性化所得), 用于本地推理。

本申请实施例可以将横向联邦学习与 NWDAF 相结合, 实现模型训练与更新过程。其中, 第一 NWDAF (也称为 Server NWDAF) 可以训练模型或者聚合模型, 第二 NWDAF (也称为 Client NWDAF) 可以训练模型、更新模型以及使用模型进行推断。

如图 14(b)所示, 为本申请实施例提供的又一种通信方法流程示意图。该方法包括以下步骤:

步骤 1401b, 第一 NWDAF 向 NRF 注册。

- 15 第一 NWDAF 向 NRF 发送 NF 注册请求, 携带 NF 配置文件, 该 NF 配置文件包括 NF 类型, NF 服务名称 (NF Service, 如 ModelProvision), 分析类型标识 (Analytics ID) 等信息。

相应地, NRF 存储 NF 配置文件以及向第一 NWDAF 发送 NF 注册响应。

步骤 1402b, 第二 NWDAF 向 NRF 注册。

- 20 第二 NWDAF 向 NRF 发送 NF 注册请求, 携带 NF 配置文件, 该 NF 配置文件包括 NF 类型, NF 服务名称 (NF Service, 如 ModelUpdate), 分析类型标识 (Analytics ID) 等信息。

相应地, NRF 存储 NF 配置文件以及向第二 NWDAF 发送 NF 注册响应。

步骤 1403b, 第二 NWDAF 向 NRF 发送 NF 发现请求, 携带 NF 配置文件。

- 25 比如携带的 NF 配置文件包含 NF 类型 (如 NWDAF), NF 服务名称 (NF Service, 如 ModelProvision, 以及携带 Analytics ID, 则该 NF 发现请求用于请求从 NRF 获取对应该 Analytics ID 的 Server NWDAF。

步骤 1404b, NRF 向第二 NWDAF 发送 NF 发现响应, 其中携带 NWDAF 实例。

其中, 携带的 NWDAF 实例是 Server NWDAF 的一个实例, 可以用 Server NWDAF 的 ID 或 IP 地址来表示。

步骤 1405b, 第一 NWDAF 向 NRF 发送 NF 发现请求, 携带 NF 配置文件。

- 30 比如携带的 NF 配置文件包含 NF 类型 (如 NWDAF), NF 服务名称 (NF Service, 如 ModelUpdate), 以及携带 Analytics ID, 则该 NF 发现请求用于请求从 NRF 获取对应该 Analytics ID 的 Client NWDAF。

步骤 1406b, NRF 向第一 NWDAF 发送 NF 发现响应, 其中携带 NWDAF 实例。

- 35 其中, 携带的 NWDAF 实例是 Client NWDAF 的一个实例, 可以用 Client NWDAF 的 ID 或 IP 地址来表示。

需要说明的是, NF 发现响应可以包含一个或多个 Client NWDAF 实例。

需要说明的是, “步骤 1403b-步骤 1404b” 以及 “步骤 1405b-步骤 1406b” 可以只执行其中一个, 这样, 在联邦学习中, 可以是 Client NWDAF 主动触发到 Server NWDAF 的横向联邦训练, 也可以是 Server NWDAF 主动触发到 Client NWDAF 横向联邦训练。

其中,上述步骤 1401b-步骤 1406b 可选步骤。例如,第一 NWDAF 中配置了第二 NWDAF 的 NF 配置文件和/或第二 NWDAF 中配置了第一 NWDAF 的 NF 配置文件,步骤 1401b-步骤 1406b 可以不执行。

步骤 1407b,第二 NWDAF 向第一 NWDAF 发送模型订阅请求,携带 Analytics ID。

5 该模型订阅请求用于向第一 NWDAF 订阅该 Analytics ID 对应的模型索引信息。

步骤 1408b,第一 NWDAF 向第二 NWDAF 发送模型通知 1,携带模型索引信息 1。

该模型索引信息 1 即为与 Analytics ID 对应的模型的索引信息。

该模型通知为与步骤 1707 的模型订阅请求对应的模型通知。

进一步的,第二 NWDAF 可以根据该模型索引信息 1,获取到相应的模型的第一信息。

10 步骤 1409b,第一 NWDAF 向第二 NWDAF 发送模型订阅请求,携带模型索引信息 1。

该模型订阅请求用于向第二 NWDAF 请求更新模型索引信息 1 对应的模型的第一信息,并订阅更新的模型的信息。

步骤 1410b,模型更新。

具体的,第二 NWDAF 利用模型索引信息 1 对应的模型信息进行本地训练,获得的模型的第二信息,并确定的模型的第二信息对应的模型索引信息 2。

15 步骤 1411b,第二 NWDAF 向第一 NWDAF 发送模型通知,携带模型索引信息 2。

该模型通知为与步骤 1409b 的模型订阅请求对应的模型通知。

步骤 1412b,模型更新。

具体的,第一 NWDAF 利用模型索引信息 2 对应的模型的第二信息进行本地训练,获得的模型的第三信息,并确定的模型的第三信息对应的模型索引信息 3。

可选的,步骤 1407b-1410b 中的第二 NWDAF 可以是多个 Client NWDAF 的实例,则第一 NWDAF 在步骤 1411b 可以接收到来自多个第二 NWDAF 实例的模型索引信息,第一 NWDAF 根据多个模型索引信息获取对应的多个模型信息,并将多个模型信息进行聚合训练得到更新的模型信息。

25 步骤 1413b,第一 NWDAF 向第二 NWDAF 发送模型通知,携带模型索引信息 3。

该模型通知为与步骤 1407b 的模型订阅请求 1 对应的模型通知。

后续,可以重复上述步骤 1410b 至步骤 1413b,并且模型索引信息一直在发生变化,直到第一 NWDAF 确定停止迭代。可选的,第一 NWDAF 可以向第二 NWDAF 发送模型取消订阅消息,也即取消步骤 1409b 对应的模型订阅请求,实现停止迭代。

30 需要说明的是,该实施例中,模型索引信息可以包括标识信息,该标识信息用于指示该模型索引信息对应的模型的信息。可选的,模型索引信息还包括以下一项或多项:模型对应的分析类型标识、模型的标识、模型的信息的版本信息。

基于上述方案,第一 NWDAF 与第二 NWDAF 均可以对模型进行更新得到新的模型信息,并将新的模型信息对应的模型索引信息发送给对方,如此可以实现模型反复迭代,从而可以实现模型性能的逐步提升,最终得到一个模型性能最优的模型,进而可以保证模型性能。

如图 14(c)所示,为本申请实施例提供的又一种通信方法流程示意图。该实施例中,第一 NWDAF 也可以称为训练 NWDAF,第二 NWDAF 也可以称为推断 NWDAF 或者不同于第一 NWDAF 的其他训练 NWDAF。

该方法包括以下步骤:

步骤 1401c, 第一 NWDAF 向 NRF 注册。

第一 NWDAF 向 NRF 发送 NF 注册请求, 携带 NF 配置文件, 该 NF 配置文件包括 NF 类型, NF 服务名称 (NF Service Name), 分析类型标识 (Analytics ID) 等信息。

5 其中, NF 类型可以是 NWDAF。

NF 服务名称可以是提供模型服务 (ModelProvision)。

分析类型标识用于指示训练 NWDAF 提供的某种特定分析类型, 比如可以是业务体验 (Service Experience), 网络性能 (Network Performance), 终端移动性 (UE Mobility) 等。

相应地, NRF 存储 NF 配置文件以及向第一 NWDAF 发送 NF 注册响应。

10 作为一种实现方法, 针对某个 Analytics ID, 第一 NWDAF 可以是在具备了该 Analytics ID 对应的模型的模型训练能力时, 向 NRF 注册。作为另一种实现方法, 针对某个 Analytics ID, 第一 NWDAF 也可以是在已经训练得到该 Analytics ID 对应的模型时, 向 NRF 注册。

作为一种实现方法, 第一 NWDAF 还可以在上述 NF 配置文件中携带第二指示信息, 该第二指示信息指示每个 Analytics ID 对应的模型是否已经训练完成或是否已经准备完成。
15 或者在 NF 配置文件中的模型信息中携带该第二指示信息。可选的, 当第二指示信息指示某个 Analytics ID 对应的模型已经训练完成或已经准备完成, 则该 NF 配置文件还可以携带该 Analytics ID 对应的模型描述信息。

作为另一种实现方法, NF 配置文件携带该 Analytics ID 对应的模型描述信息。所述模型描述信息用于指示该 Analytics ID 对应的模型已经训练完成或已经准备完成。

20 其中, 模型描述信息包括以下信息中的一项或多项: 分析过滤信息 (analytics filter)、分析报告目标信息 (target of analytics reporting)、模型性能信息 (model performance) 或模型部署环境信息 (model deployment environment)。其中, 分析过滤信息用于表示 Analytics ID 对应的模型适用的范围, 分析过滤信息包括以下一项或多项: 区域、时间段、单网络切片选择辅助信息 (single network slice selection assistance information, S-NSSAI) 或数据网
25 名称 (data network name, DNN)。

分析报告目标信息用于表示 Analytics ID 对应的模型对应的终端, 分析报告目标信息包括以下一项或多项: 终端的标识、终端组的标识或指示任意终端的信息。其中, 该指示任意终端的信息用于表示 Analytics ID 对应的模型对应的终端可以为任意终端。

模型性能信息用于表示 Analytics ID 对应的模型的性能, 模型性能信息包括以下一项或多项: 准确率 (Accuracy)、精确率 (Precision) 或召回率 (Recall)、错误率 (Error Rate)、F1 分数 (F1 Score)、均方误差 (Mean Square error)、均方根误差 (Standard Deviation)、均方根对数误差 (Logarithm Standard Deviation)、平均绝对误差 (Mean Absolute Deviation)、模型推理时长 (Model Inference Delay)、模型鲁棒性 (Model Robustness)、模型可扩展性 (Model Expandability)、模型可解释性 (Model Interpretability)。
30

模型部署环境信息用于表示 Analytics ID 对应的模型部署的硬件环境, 模型部署环境信息包括以下一项或多项: 中央处理器 (central processing unit, CPU) 个数、图形处理器 (Graphic Processing Unit, GPU) 个数、内存大小或硬盘大小。

步骤 1402c, 第二 NWDAF 向 NRF 发送 NF 发现请求, 携带 NF 需求信息。

例如, NF 发现请求用于根据 NF 需求信息从 NRF 获取训练 NWDAF。比如携带的 NF
40 需求信息包含 NF 类型 (如 NWDAF), NF 服务名称 (如 ModelProvision), Analytics ID 以

及针对该 Analytics ID 对应的第一模型需求信息。该 NF 发现请求用于请求从 NRF 获取能够提供与该 Analytics ID 对应的训练 NWDAF, 且该训练 NWDAF 根据该 Analytics ID 训练的模型满足第一模型需求信息。其中, 模型满足第一模型需求信息的模型可以理解为: 训练 NWDAF 针对该 Analytics ID 训练的模型完全匹配或者部分匹配第一模型需求信息。

- 5 其中, 第一模型需求信息包括以下信息中的一项或多项: 分析过滤信息 (analytics filter) 需求、分析报告目标信息 (target of analytics reporting)、模型性能信息 (model performance) 或模型部署环境信息 (model deployment environment)。

例如, Analytics ID 训练的模型完全匹配第一模型需求信息可以理解为: 该模型对应的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息完全满足第一模型需求信息中的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息。
10 Analytics ID 训练的模型部分匹配第一模型需求信息可以理解为: 该模型对应的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息完全满足第一模型需求信息中的分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息中的一项或多项。

- 15 其中, 分析过滤信息用于表示第二 NWDAF 需要请求的模型适用的范围, analytics filter 包括以下一项或多项: 区域、时间段、S-NSSAI、DNN。

分析报告目标信息用于表示第二 NWDAF 需要请求的模型对应的终端, 分析报告目标信息包括以下一项或多项: 终端的标识、终端组的标识或指示任意终端的信息。其中, 该指示任意终端的信息用于表示 Analytics ID 对应的模型对应的终端可以为任意终端。

- 20 模型性能信息用于表示第二 NWDAF 需要请求的模型的性能, 模型性能信息包括以下一项或多项: 准确率 (Accuracy)、精确率 (Precision) 或召回率 (Recall)、错误率 (Error Rate)、F1 分数 (F1 Score)、均方误差 (Mean Square error)、均方根误差 (Standard Deviation)、均方根对数误差 (Logarithm Standard Deviation)、平均绝对误差 (Mean Absolute Deviation)、模型推理时长 (Model Inference Delay)、模型鲁棒性 (Model Robustness)、模型可扩展性 (Model Expandability)、模型可解释性 (Model Interpretability)。

- 25 模型部署环境信息用于表示第二 NWDAF 需要请求模型部署的硬件环境, 模型部署环境信息包括以下一项或多项: CPU 个数、GPU 个数、内存大小或硬盘大小。

步骤 1403c, NRF 向第二 NWDAF 发送 NF 发现响应, 其中携带第一 NWDAF 的地址信息。

- 30 例如, 该地址信息可以是 ID、IP 地址或者完全限定域名 (Fully Qualified Domain Name, FQDN)。

可选的, 在 NRF 向第二 NWDAF 返回第一 NWDAF 的地址信息之前, NRF 确定第一 NWDAF 为与步骤 1402c 中 Analytics ID 对应的训练 NWDAF, 且第一 NWDAF 根据该 Analytics ID 训练的模型满足步骤 1402c 中第一模型需求信息。

- 35 其中, 携带的第一 NWDAF 是训练 NWDAF 的一个实例, 可以用训练 NWDAF 实例的 ID 或 IP 地址来表示。

具体的, NRF 根据第二 NWDAF 发送的上述第一模型需求信息, 获取第一 NWDAF。其中, 获取的第一 NWDAF 可以满足该第一模型需求信息。其中, 获取的第一 NWDAF 可以满足该第一模型需求信息可以理解为: 第一 NWDAF 针对该 Analytics ID 训练的模型完全匹配或者部分匹配第一模型需求信息。

- 40 例如, Analytics ID 训练的模型完全匹配第一模型需求信息可以理解为: 该模型对应的

分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息分别与第一模型需求信息中的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息相同。Analytics ID 训练的模型部分匹配第一模型需求信息可以理解为以下情况中的任一种或者多种满足：该模型对应的分析过滤信息与第一模型需求信息中的分析过滤信息相同，该模型对应的分析报告目标信息与第一模型需求信息中的分析报告目标信息相同，该模型对应的模型性能信息与第一模型需求信息中的模型性能信息相同，或者该模型对应的模型部署环境信息与第一模型需求信息中的分析过滤信息相同。

步骤 1404c，第二 NWDAF 向第一 NWDAF 发送第一请求。

可选的，该第一请求是模型订阅请求，该模型订阅请求用于向第一 NWDAF 订阅该 Analytics ID 对应的、且满足第二模型需求信息的模型的模型索引信息。

该第一请求携带 Analytics ID 和第二模型需求信息，该第一请求用于向第一 NWDAF 请求该 Analytics ID 对应的、且满足第二模型需求信息的模型的模型索引信息。其中，满足第二模型需求信息的模型可以理解为：第一 NWDAF 针对该 Analytics ID 训练的模型完全匹配或者部分匹配第二模型需求信息。

例如，Analytics ID 训练的模型完全匹配第二模型需求信息可以理解为：该模型对应的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息完全满足第二模型需求信息中的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息。Analytics ID 训练的模型部分匹配第二模型需求信息可以理解为：该模型对应的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息完全满足第二模型需求信息中的分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息中的一项或多项。

该模型索引信息可以是以下一种或者多种信息：模型标识 (model ID)、模型版本号 (例如，version)、位置信息 (例如，location) 或地址信息等。其中，模型标识用于标识该模型，模型版本号表示模型版本，位置信息或地址信息表示模型的存储位置。可选的，当模型索引信息是位置信息或地址信息，该位置信息或地址信息中也可以包含模型版本。可选的，地址信息可以包括以下信息中的一种或多种：IP 地址、FQDN 或者 URL。

其中，这里的第二模型需求信息包括时间信息，以及包括上述第一模型需求信息中的部分或全部信息。该时间信息用于指示期望收到来自第一 NWDAF 的模型索引信息的时间。该时间信息可以是绝对时间或相对时间。

作为一种实现方法，第二 NWDAF 可以通过触发到第一 NWDAF 的 Nnwdaf_MLModelProvision_Subscribe 服务操作或 Nnwdaf_MLModelProvision_Unsubscribe 服务操作，以实现向第一 NWDAF 发送模型订阅请求。

步骤 1405c，第一 NWDAF 根据该第二模型需求信息和该 Analytics ID，获取模型索引信息。

其中，第一 NWDAF 获取的是与 Analytics ID 对应的、且满足第二模型需求信息的模型的模型索引信息。

这里的 Analytics ID 即为模型订阅请求中的 Analytics ID。

第一 NWDAF 获取模型索引信息的方式可以是：

1) 如果第一 NWDAF 确定存在与 Analytics ID 对应的本地模型，且该本地模型满足第二模型需求信息，则第一 NWDAF 无需进一步向其它网元收集数据，而是可以直接获取该本地模型的模型索引信息；

2) 如果第一 NWDAF 确定存在与 Analytics ID 对应的本地模型, 但该本地模型不满足第二模型需求信息, 则第一 NWDAF 进一步向其它网元收集数据, 并根据收集的数据, 对该本地模型进行再训练, 得到新的模型, 并获取该新的模型的模型索引信息; 其中, 本地模型不满足第二模型需求信息可以理解为该本地模型对应的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息不满足第一模型需求信息中的分析过滤信息、分析报告目标信息、模型性能信息和模型部署环境信息中的任一项。

3) 如果第一 NWDAF 确定不存在与 Analytics ID 对应的本地模型, 则第一 NWDAF 进一步向其它网元收集数据, 并根据收集的数据训练得到初始模型, 并获取该初始模型的模型索引信息。

其中, 第二 NWDAF 针对以上三种操作获取模型索引信息的时间依次延长。这是因为, 模型再次训练所需时间要比模型初始训练所需时间要短很多。

步骤 1406c, 第一 NWDAF 向第二 NWDAF 发送模型索引信息。

该模型索引信息对应的模型, 即为第一 NWDAF 获取到与 Analytics ID 对应、且满足第二模型需求信息的模型。

第二 NWDAF 可以根据模型索引信息获取到 Analytics ID 对应的模型, 用于推理得到 Analytics ID 对应的数据分析结果。

可选的, 如果第一 NWDAF 判定在第二模型需求信息中的时间信息指示的时间内, 无法向第二 NWDAF 发送模型索引信息, 则第一 NWDAF 可以向第二 NWDAF 发送第一指示信息, 该第一指示信息指示无法在该时间信息指示的时间内发送模型索引信息。该第一指示信息可以称为 Error Response 或者 Error Notification。

基于上述方案, 通过第一模型需求信息, 可以辅助第二 NWDAF 快速从 NRF 获取到一个尽量满足第一模型需求信息的第一 NWDAF, 从而可以实现为第二 NWDAF 精准提供第一 NWDAF。通过第二模型需求信息, 可以辅助第二 NWDAF 快速从第一 NWDAF 获取到尽量满足第二模型需求信息的模型, 从而可以实现为第二 NWDAF 精准提供模型。

参考图 15, 为本申请实施例提供的一种通信装置示意图, 该通信装置 1500 包括收发单元 1510 和处理单元 1520。

在第一个实施例中, 该通信装置用于实现上述各实施例中对应第一数据分析网元的各个步骤:

收发单元 1510, 用于接收来自第二数据分析网元的第一信息, 所述第一信息包括模型的性能报告, 所述模型的性能报告用于指示所述模型的性能的评估结果, 或者, 所述模型的性能报告用于指示所述模型的性能的评估结果不满足所述模型的性能指标的要求; 以及, 用于向所述第二数据分析网元发送第二信息, 所述第二信息包括所述模型的第二模型信息。处理单元 1520, 用于根据所述模型的性能报告更新所述模型的第一模型信息, 获得所述第二模型信息。

在一种可能的实现方法中, 所述收发单元 1510, 还用于向所述第二数据分析网元发送第三信息, 所述第三信息包括所述模型的性能指标, 所述模型的性能指标用于获取所述模型的性能的评估结果。

在一种可能的实现方法中, 所述收发单元 1510, 还用于向第三数据分析网元发送所述第二信息。

在一种可能的实现方法中,所述收发单元 1510,用于接收来自第二数据分析网元的第一信息,具体包括:用于通过网络存储网元接收来自所述第二数据分析网元的所述第一信息。所述收发单元 1510,用于向所述第二数据分析网元发送第二信息,具体包括:用于通过网络存储网元向所述第二数据分析网元发送所述第二信息。

5 在一种可能的实现方法中,所述模型性能指标包括以下一项或多项:精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性。

在一种可能的实现方法中,所述第三信息还包括以下一项或多项:分析类型标识、所述模型的标识、子模型的标识,所述分析类型标识用于指示所述模型的分析类型。

10 在一种可能的实现方法中,所述第三信息还包括以下一项或多项:上报周期、门限信息,所述上报周期用于指示上报所述模型的性能报告的时间,所述门限信息用于指示上报所述模型的性能报告的条件。

在一种可能的实现方法中,所述第一信息还包括所述模型的性能报告对应的以下一项或多项信息:时间、区域、切片。

15 在一种可能的实现方法中,所述第二信息还包括以下一项或多项:所述模型的标识、子模型的标识、所述模型的性能评估结果、所述模型的性能评估结果对应的硬件能力信息、所述模型的大小、所述模型的推理时长。

20 在第二个实施例中,该通信装置用于实现上述各实施例中对应第二数据分析网元的各个步骤:

收发单元 1510,用于向第一数据分析网元发送第一信息,所述第一信息包括模型的性能报告,所述模型的性能报告用于指示所述模型的性能的评估结果,或者,所述模型的性能报告用于指示所述模型的性能的评估结果不满足所述模型的性能指标的要求;以及,用于接收来自所述第一数据分析网元的第二信息,所述第二信息包括所述模型的第二模型信息,所述模型的第二信息是根据所述模型的性能报告更新所述模型的第一模型信息得到的。
25 处理单元 1520,用于根据所述第二模型信息,更新所述模型。

在一种可能的实现方法中,所述收发单元 1510,还用于接收来自所述第一数据分析网元的第三信息,所述第三信息包括所述模型的性能指标,所述模型的性能指标用于获取所述模型的性能的评估结果。

30 在一种可能的实现方法中,所述收发单元 1510,用于向第一数据分析网元发送第一信息,具体包括:用于通过网络存储网元向所述第一数据分析网元发送所述第一信息。所述收发单元 1510,用于接收来自所述第一数据分析网元的第二信息,具体包括:用于通过网络存储网元接收来自所述第一数据分析网元的第二信息。

35 在一种可能的实现方法中,所述模型性能指标包括以下一项或多项:精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性。

在一种可能的实现方法中,所述第三信息还包括以下一项或多项:分析类型标识、所述模型的标识、子模型的标识,所述分析类型标识用于指示所述模型的分析类型。

40 在一种可能的实现方法中,所述第三信息还包括以下一项或多项:上报周期、门限信息,所述上报周期用于指示上报所述模型的性能报告的时间,所述门限信息用于指示上报

所述模型的性能报告的条件。

在一种可能的实现方法中，所述第一信息还包括所述模型的性能报告对应的以下一项或多项信息：时间、区域、切片。

- 5 在一种可能的实现方法中，所述第二信息还包括以下一项或多项：所述模型的标识、子模型的标识、所述模型的性能评估结果、所述模型的性能评估结果对应的硬件能力信息、所述模型的大小、所述模型的推理时长。

在第三个实施例中，该通信装置用于实现上述各实施例中对应第一数据分析网元的各个步骤：

- 10 处理单元 1520，用于将模型的第一信息更新为所述模型的第二信息；所述第一数据分析网元确定所述模型的第二信息的索引信息，所述第二信息的索引信息包括第一标识信息，所述第一标识信息用于指示所述模型的第二信息；收发单元 1510，用于向所述第二数据分析网元发送所述第二信息的索引信息，所述第二信息的索引信息用于所述模型的第二信息的获取。其中，模型的第二信息的索引信息，也可以称为第二信息对应的模型索引信息。

- 15 在一种可能的实现方法中，所述第二信息的索引信息还包括以下一项或多项：所述模型对应的分析类型标识、所述模型的标识、所述模型的第二信息的版本信息。

- 在一种可能的实现方法中，收发单元 1510，用于从所述第二数据分析网元接收所述模型的第一信息的索引信息，所述第一信息的索引信息包括第二标识信息，所述第二标识信息用于指示所述模型的第一信息；处理单元 1520，用于根据所述第一信息的索引信息获取
20 所述模型的第一信息。

在一种可能的实现方法中，所述第一信息的索引信息还包括以下一项或多项：所述模型对应的分析类型标识、所述模型的标识、所述模型的第一信息的版本信息。

- 在一种可能的实现方法中，处理单元 1520，用于从所述第二数据分析网元获取第一请求，所述第一请求用于更新所述模型的第一信息，所述第一请求包括所述模型的第一信息的索引信息；根据所述第一信息的索引信息获取所述模型的第一信息；所述第一数据分析网元更新所述模型的第一信息，得到所述模型的第二信息。
25

- 在一种可能的实现方法中，收发单元 1510，用于向所述第二数据分析网元发送第二请求，所述第二请求用于请求所述模型的第一信息的索引信息，所述第二请求包括所述模型对应的分析类型标识；从所述第二数据分析网元接收第二响应，所述第二响应包括所述模型的第一信息的索引信息。
30

在一种可能的实现方法中，收发单元 1510，用于通过网络存储网元从第二数据分析网元接收所述模型的第一信息的索引信息。

在一种可能的实现方法中，收发单元 1510，用于通过网络存储网元向第二数据分析网元发送所述模型的第二信息的索引信息。

- 35 在一种可能的实现方法中，所述第一数据分析网元为分布式学习中的客户端数据分析网元，所述第二数据分析网元为分布式学习中的服务端数据分析网元。

在一种可能的实现方法中，所述分布式学习为联邦学习。

在一种可能的实现方法中，所述第一数据分析网元为支持推理功能的数据分析网元，所述第二数据分析网元为支持训练功能的数据分析网元。

在第四个实施例中，该通信装置用于实现上述各实施例中对应第二数据分析网元的各个步骤：

收发单元 1510，用于向第一数据分析网元发送第一请求，所述第一请求携带分析类型标识和第二模型需求信息，所述第一请求用于请求所述分析类型标识对应的、且满足所述第二模型需求信息的模型的模型索引信息；接收来自所述第一数据分析网元的所述模型索引信息。

在一种可能的实现方法中，收发单元 1510，用于向网络存储网元发送网络功能发现请求，所述网络功能发现请求包括所述分析类型标识和第一模型需求信息，所述网络功能发现请求用于请求获取能够提供与所述分析类型标识对应的、且满足所述第一模型需求信息的模型的数据分析网元；接收来自所述网络存储网元的所述第一数据分析网元的地址信息。

在一种可能的实现方法中，所述第一模型需求信息包括以下信息中的一项或多项：分析过滤信息、分析报告目标信息模型性能信息或模型部署环境信息；

其中，所述分析过滤信息用于表示所述第二数据分析网元需要请求的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

所述分析报告目标信息用于表示所述第二数据分析网元需要请求的模型对应的终端，所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述第二数据分析网元需要请求的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；

所述模型部署环境信息用于表示所述第二数据分析网元需要请求的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

在一种可能的实现方法中，所述第二模型需求信息包括所述第一模型需求信息中的部分信息或全部信息。

在一种可能的实现方法中，所述第二模型需求信息包括时间信息，所述时间信息用于指示期望收到来自所述第一数据分析网元的模型索引信息的时间。

在一种可能的实现方法中，收发单元 1510，用于接收来自所述第一数据分析网元的第一指示信息，所述第一指示信息指示无法在所述时间信息指示的时间内发送所述模型索引信息。

在第五个实施例中，该通信装置用于实现上述各实施例中对应第一数据分析网元的各个步骤：

收发单元 1510，用于接收来自第二数据分析网元的第一请求，所述第一请求携带分析类型标识和第二模型需求信息，所述第一请求用于请求所述分析类型标识对应的、且满足所述第二模型需求信息的模型的模型索引信息；向所述第二数据分析网元发送模型索引信息；处理单元 1520，用于根据所述第二模型需求信息和所述分析类型标识，获取所述模型索引信息。

在一种可能的实现方法中,所述第二模型需求信息包括时间信息,所述时间信息用于指示期望收到来自所述第一数据分析网元的模型索引信息的时间。

在一种可能的实现方法中,收发单元 1510,用于向所述第二数据分析网元发送第一指示信息,所述第一指示信息指示无法在所述时间信息指示的时间内发送所述模型索引信息。

5 在一种可能的实现方法中,收发单元 1510,用于向网络存储网元发送网络功能注册请求,所述网络功能注册请求携带所述分析类型标识和模型信息,所述模型信息包括第二指示信息,所述第二指示信息指示所述分析类型标识对应的模型是否已经训练完成或者是否已经准备完成。

10 在一种可能的实现方法中,在所述第二指示信息指示所述分析类型标识对应的模型已经训练完成或者已经准备完成的情况下,所述模型信息还包括模型描述信息,所述模型描述信息包括以下信息中的一项或多项:分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息;

15 其中,所述分析过滤信息用于表示所述分析类型标识对应的模型适用的范围,所述分析过滤信息包括以下一项或多项:区域、时间段、单网络切片选择辅助信息或数据网络名称;

所述分析报告目标信息用于表示所述分析类型标识对应的模型对应的终端,所述分析报告信息目标包括以下一项或多项:终端的标识、终端组的标识或指示任意终端的信息;

20 所述模型性能信息用于表示所述分析类型标识对应的模型的性能,所述模型性能信息包括以下一项或多项:精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性;

所述模型部署环境信息用于表示所述分析类型标识对应的模型部署的硬件环境,所述模型部署环境信息包括以下一项或多项:中央处理器个数、图形处理器个数、内存大小或硬盘大小。

25

在第六个实施例中,该通信装置用于实现上述各实施例中对应网络存储网元的各个步骤:

收发单元 1510,用于接收来自第二数据分析网元的网络功能发现请求,所述网络功能发现请求包括分析类型标识和第一模型需求信息,所述网络功能发现请求用于请求获取能够
30 提供与所述分析类型标识对应的、且满足所述第一模型需求信息的模型的数据分析网元;向所述第二数据分析网元发送所述第一数据分析网元的地址信息;处理单元 1520,用于根据所述第一模型需求信息和所述分析类型标识,获取第一数据分析网元的地址信息;

35 在一种可能的实现方法中,收发单元 1510,用于接收来自所述第一数据分析网元的网络功能注册请求,所述网络功能注册请求携带所述分析类型标识和模型信息,所述模型信息包括第二指示信息,所述第二指示信息指示所述分析类型标识对应的模型是否已经训练完成或者是否已经准备完成。

40 在一种可能的实现方法中,在所述第二指示信息指示所述分析类型标识对应的模型已经训练完成或者已经准备完成的情况下,所述模型信息还包括模型描述信息,所述模型描述信息包括以下信息中的一项或多项:分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息;

其中，所述分析过滤信息用于表示所述分析类型标识对应的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

所述分析报告目标信息用于表示所述分析类型标识对应的模型对应的终端，所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述分析类型标识对应的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；

所述模型部署环境信息用于表示所述分析类型标识对应的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

在一种可能的实现方法中，所述第一模型需求信息包括以下信息中的一项或多项：分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息；

其中，所述分析过滤信息用于表示所述第二数据分析网元需要请求的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

所述分析报告目标信息用于表示所述第二数据分析网元需要请求的模型对应的终端，所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

所述模型性能信息用于表示所述第二数据分析网元需要请求的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；

所述模型部署环境信息用于表示所述第二数据分析网元需要请求的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

可选地，上述通信装置还可以包括存储单元，该存储单元用于存储数据或者指令（也可以称为代码或者程序），上述各个单元可以和存储单元交互或者耦合，以实现对应的方法或者功能。例如，处理单元 1520 可以读取存储单元中的数据或者指令，使得通信装置实现上述实施例中的方法。

应理解以上通信装置中单元的划分仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。且通信装置中的单元可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分单元以软件通过处理元件调用的形式实现，部分单元以硬件的形式实现。例如，各个单元可以为单独设立的处理元件，也可以集成在通信装置的某一个芯片中实现，此外，也可以以程序的形式存储于存储器中，由通信装置的某一个处理元件调用并执行该单元的功能。此外这些单元全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件又可以成为处理器，可以是一种具有信号的处理能力的集成电路。在实现过程中，上述方法的各步骤或以上各个单元可以通过处理器元件中的硬件的集成逻辑电路实现或者以软件通过处理元件调用

的形式实现。

在一个例子中，以上任一通信装置中的单元可以是配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路(application specific integrated circuit, ASIC)，或，一个或多个微处理器(digital signal processor, DSP)，或，一个或者多个现场可编程门阵列(field programmable gate array, FPGA)，或这些集成电路形式中至少两种的组合。再如，当通信装置中的单元可以通过处理元件调度程序的形式实现时，该处理元件可以是通用处理器，例如中央处理器(central processing unit, CPU)或其它可以调用程序的处理器。再如，这些单元可以集成在一起，以片上系统(system-on-a-chip, SOC)的形式实现。

10 参考图 16，为本申请实施例提供的一种通信装置示意图，用于实现以上实施例中第一数据分析网元或第二数据分析网元的操作。如图 16 所示，该通信装置包括：处理器 1610 和接口 1630，可选地，该通信装置还包括存储器 1620。接口 1630 用于实现与其他设备进行通信。

15 以上实施例中第一数据分析网元或第二数据分析网元执行的方法可以通过处理器 1610 调用存储器(可以是第一数据分析网元或第二数据分析网元中的存储器 1620，也可以是外部存储器)中存储的程序来实现。即，第一数据分析网元或第二数据分析网元可以包括处理器 1610，该处理器 1610 通过调用存储器中的程序，以执行以上方法实施例中第一数据分析网元或第二数据分析网元执行的方法。这里的处理器可以是一种具有信号的处理能力的集成电路，例如 CPU。第一数据分析网元或第二数据分析网元可以通过配置成实施以上方法的一个或多个集成电路来实现。例如：一个或多个 ASIC，或，一个或多个微处理器 DSP，或，一个或者多个 FPGA 等，或这些集成电路形式中至少两种的组合。或者，可以结合以上实现方式。

20 具体的，图 15 中的收发单元 1510 和处理单元 1520 的功能/实现过程可以通过图 16 所示的通信装置 1600 中的处理器 1610 调用存储器 1620 中存储的计算机可执行指令来实现。或者，图 15 中的处理单元 1520 的功能/实现过程可以通过图 16 所示的通信装置 1600 中的处理器 1610 调用存储器 1620 中存储的计算机执行指令来实现，图 15 中的收发单元 1510 的功能/实现过程可以通过图 16 中所示的通信装置 1600 中的接口 1630 来实现，示例性的，收发单元 1510 的功能/实现过程可以通过处理器调用存储器中的程序指令以驱动接口 1630 来实现。

30 本领域普通技术人员可以理解：本申请中涉及的第一、第二等各种数字编号仅为描述方便进行的区分，并不用来限制本申请实施例的范围，也表示先后顺序。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。“至少一个”是指一个或者多个。至少两个是指两个或者多个。“至少一个”、“任意一个”或其类似表达，是指的这些项中的任意组合，包括单项(个)或复数项(个)的任意组合。例如，a，b，或 c 中的至少一项(个、种)，可以表示：a，b，c，a-b，a-c，b-c，或 a-b-c，其中 a，b，c 可以是单个，也可以是多个。“多个”是指两个或两个以上，其它量词与之类似。

40 应理解，在本申请的各种实施例中，上述各过程的序号的大小并不意味着执行顺序的

先后, 各过程的执行顺序应以其功能和内在逻辑确定, 而不应对本发明实施例的实施过程构成任何限定。

所属领域的技术人员可以清楚地了解到, 为描述的方便和简洁, 上述描述的系统、装置和单元的具体工作过程, 可以参考前述方法实施例中的对应过程, 在此不再赘述。

- 5 在上述实施例中, 可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时, 可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时, 全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中, 或者从一个计算机可读存储介质向另一个计算机可读存储介质传输, 例如, 所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包括一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质, (例如, 软盘、硬盘、磁带)、光介质(例如, DVD)、或者半导体介质(例如固态硬盘(solid state disk, SSD))等。
- 10
- 15

- 本申请实施例中所描述的各种说明性的逻辑单元和电路可以通过通用处理器, 数字信号处理器, 专用集成电路(ASIC), 现场可编程门阵列(FPGA)或其它可编程逻辑装置, 离散门或晶体管逻辑, 离散硬件部件, 或上述任何组合的设计来实现或操作所描述的功能。
- 20 通用处理器可以为微处理器, 可选地, 该通用处理器也可以为任何传统的处理器、控制器、微控制器或状态机。处理器也可以通过计算装置的组合来实现, 例如数字信号处理器和微处理器, 多个微处理器, 一个或多个微处理器联合一个数字信号处理器核, 或任何其它类似的配置来实现。

- 本申请实施例中所描述的方法或算法的步骤可以直接嵌入硬件、处理器执行的软件单元、或者这两者的结合。软件单元可以存储于随机存取存储器(Random Access Memory, RAM)、闪存、只读存储器(Read-Only Memory, ROM)、EPROM 存储器、EEPROM 存储器、寄存器、硬盘、可移动磁盘、CD-ROM 或本领域中其它任意形式的存储媒介中。示例性地, 存储媒介可以与处理器连接, 以使得处理器可以从存储媒介中读取信息, 并可以向存储媒介存写信息。可选地, 存储媒介还可以集成到处理器中。处理器和存储媒介可以
- 25
- 30 设置于 ASIC 中。

- 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设备上, 使得在计算机或其他可编程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。

- 35 在一个或多个示例性的设计中, 本申请所描述的上述功能可以在硬件、软件、固件或这三者的任意组合来实现。如果在软件中实现, 这些功能可以存储与电脑可读的媒介上, 或以一个或多个指令或代码形式传输于电脑可读的媒介上。电脑可读媒介包括电脑存储媒介和便于使得让电脑程序从一个地方转移到其它地方的通信媒介。存储媒介可以是任何通用或特殊电脑可以接入访问的可用媒体。例如, 这样的电脑可读媒体可以包括但不限于
- 40 RAM、ROM、EEPROM、CD-ROM 或其它光盘存储、磁盘存储或其它磁性存储装置, 或

其它任何可以用于承载或存储以指令或数据结构和其它可被通用或特殊电脑、或通用或特殊处理器读取形式的程序代码的媒介。此外，任何连接都可以被适当地定义为电脑可读媒介，例如，如果软件是从一个网站站点、服务器或其它远程资源通过一个同轴电缆、光纤电脑、双绞线、数字用户线（DSL）或以例如红外、无线和微波等无线方式传输的也被包含在所定义的电脑可读媒介中。所述的碟片（disk）和磁盘（disc）包括压缩磁盘、镭射盘、光盘、数字通用光盘（英文：Digital Versatile Disc，简称：DVD）、软盘和蓝光光盘，磁盘通常以磁性复制数据，而碟片通常以激光进行光学复制数据。上述的组合也可以包含在电脑可读媒介中。

尽管结合具体特征及其实施例对本申请进行了描述，显而易见的，在不脱离本申请的精神和范围的情况下，可对其进行各种修改和组合。相应地，本说明书和附图仅仅是所附权利要求所界定的本申请的示例性说明，且视为已覆盖本申请范围内的任意和所有修改、变化、组合或等同物。显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包括这些改动和变型在内。

15

权利要求

1.一种通信方法，其特征在于，包括：

第一数据分析网元接收来自第二数据分析网元的第一信息，所述第一信息包括模型的性能报告，所述模型的性能报告用于指示所述模型的性能的评估结果，或者，所述模型的性能报告用于指示所述模型的性能的评估结果不满足所述模型的性能指标的要求；

所述第一数据分析网元根据所述模型的性能报告更新所述模型的第一模型信息，获得所述模型的第二模型信息；

所述第一数据分析网元向所述第二数据分析网元发送第二信息，所述第二信息包括所述第二模型信息。

2.如权利要求1所述的方法，其特征在于，还包括：

所述第一数据分析网元向所述第二数据分析网元发送第三信息，所述第三信息包括所述模型的性能指标，所述模型的性能指标用于获取所述模型的性能的评估结果。

3.如权利要求2所述的方法，其特征在于，所述模型性能指标包括以下一项或多项：精确率、准确率、错误率、召回率、F1分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性。

4.如权利要求2或3所述的方法，其特征在于，所述第三信息还包括以下一项或多项：分析类型标识、所述模型的标识、子模型的标识，所述分析类型标识用于指示所述模型的分析类型。

5.如权利要求2-4任一所述的方法，其特征在于，所述第三信息还包括以下一项或多项：上报周期、门限信息，所述上报周期用于指示上报所述模型的性能报告的时间，所述门限信息用于指示上报所述模型的性能报告的条件。

6.如权利要求1-5任一所述的方法，其特征在于，所述第一信息还包括所述模型的性能报告对应的以下一项或多项信息：时间、区域、切片。

7.如权利要求1-6任一所述的方法，其特征在于，所述第二信息还包括以下一项或多项：所述模型的标识、子模型的标识、所述模型的性能评估结果、所述模型的性能评估结果对应的硬件能力信息、所述模型的大小、所述模型的推理时长。

8.如权利要求1-7任一所述的方法，其特征在于，还包括：

所述第一数据分析网元向第三数据分析网元发送所述第二信息。

9.如权利要求1-8任一所述的方法，其特征在于，所述第一数据分析网元接收来自第二数据分析网元的第一信息，包括：

所述第一数据分析网元通过网络存储网元接收来自所述第二数据分析网元的所述第一信息；

所述第一数据分析网元向所述第二数据分析网元发送第二信息，包括：

所述第一数据分析网元通过网络存储网元向所述第二数据分析网元发送所述第二信息。

10.一种通信方法，其特征在于，包括：

第二数据分析网元向第一数据分析网元发送第一信息，所述第一信息包括模型的性能报告，所述模型的性能报告用于指示所述模型的性能的评估结果，或者，所述模型的性能报告用于指示所述模型的性能的评估结果不满足所述模型的性能指标的要求；

所述第二数据分析网元接收来自所述第一数据分析网元的第二信息，所述第二信息包括所述模型的第二模型信息，所述第二信息是根据所述模型的性能报告更新所述模型的第一模型信息得到的；

所述第二数据分析网元根据所述第二模型信息，更新所述模型。

5 11.如权利要求 10 所述的方法，其特征在于，还包括：

所述第二数据分析网元接收来自所述第一数据分析网元的第三信息，所述第三信息包括所述模型的性能指标，所述模型的性能指标用于获取所述模型的性能的评估结果。

12.如权利要求 11 所述的方法，其特征在于，所述模型性能指标包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性。

13.如权利要求 11 或 12 所述的方法，其特征在于，所述第三信息还包括以下一项或多项：分析类型标识、所述模型的标识、子模型的标识，所述分析类型标识用于指示所述模型的分析类型。

14.如权利要求 11-13 任一所述的方法，其特征在于，所述第三信息还包括以下一项或多项：上报周期、门限信息，所述上报周期用于指示上报所述模型的性能报告的时间，所述门限信息用于指示上报所述模型的性能报告的条件。

15.如权利要求 10-14 任一所述的方法，其特征在于，所述第一信息还包括所述模型的性能报告对应的以下一项或多项信息：时间、区域、切片。

16.如权利要求 10-15 任一所述的方法，其特征在于，所述第二信息还包括以下一项或多项：所述模型的标识、子模型的标识、所述模型的性能评估结果、所述模型的性能评估结果对应的硬件能力信息、所述模型的大小、所述模型的推理时长。

17.如权利要求 10-16 任一所述的方法，其特征在于，所述第二数据分析网元向第一数据分析网元发送第一信息，包括：

25 所述第二数据分析网元通过网络存储网元向所述第一数据分析网元发送所述第一信息；

所述第二数据分析网元接收来自所述第一数据分析网元的第二信息，包括：

所述第二数据分析网元通过网络存储网元接收来自所述第一数据分析网元的第二信息。

18.一种通信方法，其特征在于，包括：

30 第二数据分析网元向第一数据分析网元发送第一请求，所述第一请求携带分析类型标识和第二模型需求信息，所述第一请求用于请求所述分析类型标识对应的、且满足所述第二模型需求信息的模型的模型索引信息；

所述第二数据分析网元接收来自所述第一数据分析网元的所述模型索引信息。

19.如权利要求 18 所述的方法，其特征在于，还包括：

35 所述第二数据分析网元向网络存储网元发送网络功能发现请求，所述网络功能发现请求包括所述分析类型标识和第一模型需求信息，所述网络功能发现请求用于请求获取能够提供与所述分析类型标识对应的、且满足所述第一模型需求信息的模型的数据分析网元；

所述第二数据分析网元接收来自所述网络存储网元的所述第一数据分析网元的地址信息。

40 20.如权利要求 19 所述的方法，其特征在于，所述第一模型需求信息包括以下信息中

的一项或多项：分析过滤信息、分析报告目标信息模型性能信息或模型部署环境信息；

其中，所述分析过滤信息用于表示所述第二数据分析网元需要请求的模型适用的范围，所述分析过滤信息包括以下一项或多项：区域、时间段、单网络切片选择辅助信息或数据网络名称；

- 5 所述分析报告目标信息用于表示所述第二数据分析网元需要请求的模型对应的终端，所述分析报告目标信息包括以下一项或多项：终端的标识、终端组的标识或指示任意终端的信息；

- 10 所述模型性能信息用于表示所述第二数据分析网元需要请求的模型的性能，所述模型性能信息包括以下一项或多项：精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性；

所述模型部署环境信息用于表示所述第二数据分析网元需要请求的模型部署的硬件环境，所述模型部署环境信息包括以下一项或多项：中央处理器个数、图形处理器个数、内存大小或硬盘大小。

- 15 21.如权利要求 20 所述的方法，其特征在于，所述第二模型需求信息包括所述第一模型需求信息中的部分信息或全部信息。

22.如权利要求 18-21 任一所述的方法，其特征在于，所述第二模型需求信息包括时间信息，所述时间信息用于指示期望收到来自所述第一数据分析网元的模型索引信息的时间。

23.如权利要求 22 所述的方法，其特征在于，还包括：

- 20 所述第二数据分析网元接收来自所述第一数据分析网元的第一指示信息，所述第一指示信息指示无法在所述时间信息指示的时间内发送所述模型索引信息。

24.一种通信方法，其特征在于，包括：

- 25 第一数据分析网元接收来自第二数据分析网元的第一请求，所述第一请求携带分析类型标识和第二模型需求信息，所述第一请求用于请求所述分析类型标识对应的、且满足所述第二模型需求信息模型的模型索引信息；

所述第一数据分析网元根据所述第二模型需求信息和所述分析类型标识，获取所述模型索引信息；

所述第一数据分析网元向所述第二数据分析网元发送所述模型索引信息。

- 30 25.如权利要求 24 所述的方法，其特征在于，所述第二模型需求信息包括时间信息，所述时间信息用于指示期望收到来自所述第一数据分析网元的模型索引信息的时间。

26.如权利要求 25 所述的方法，其特征在于，还包括：

所述第一数据分析网元向所述第二数据分析网元发送第一指示信息，所述第一指示信息指示无法在所述时间信息指示的时间内发送所述模型索引信息。

27.如权利要求 24-26 任一所述的方法，其特征在于，还包括：

- 35 所述第一数据分析网元向网络存储网元发送网络功能注册请求，所述网络功能注册请求携带所述分析类型标识和模型信息，所述模型信息包括第二指示信息，所述第二指示信息指示所述分析类型标识对应的模型是否已经训练完成或者是否已经准备完成。

- 40 28.如权利要求 27 所述的方法，其特征在于，在所述第二指示信息指示所述分析类型标识对应的模型已经训练完成或者已经准备完成的情况下，所述模型信息还包括模型描述信息，所述模型描述信息包括以下信息中的一项或多项：分析过滤信息、分析报告目标信

息、模型性能信息或模型部署环境信息;

其中,所述分析过滤信息用于表示所述分析类型标识对应的模型适用的范围,所述分析过滤信息包括以下一项或多项:区域、时间段、单网络切片选择辅助信息或数据网络名称;

- 5 所述分析报告目标信息用于表示所述分析类型标识对应的模型对应的终端,所述分析报告信息目标包括以下一项或多项:终端的标识、终端组的标识或指示任意终端的信息;

所述模型性能信息用于表示所述分析类型标识对应的模型的性能,所述模型性能信息包括以下一项或多项:精确率、准确率、错误率、召回率、F1分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性;

10 所述模型部署环境信息用于表示所述分析类型标识对应的模型部署的硬件环境,所述模型部署环境信息包括以下一项或多项:中央处理器个数、图形处理器个数、内存大小或硬盘大小。

29.一种通信方法,其特征在于,包括:

- 15 网络存储网元接收来自第二数据分析网元的网络功能发现请求,所述网络功能发现请求包括分析类型标识和第一模型需求信息,所述网络功能发现请求用于请求获取能够提供与所述分析类型标识对应的、且满足所述第一模型需求信息的模型的数据分析网元;

所述网络存储网元根据所述第一模型需求信息和所述分析类型标识,获取第一数据分析网元的地址信息;

- 20 所述网络存储网元向所述第二数据分析网元发送所述第一数据分析网元的地址信息。

30.如权利要求29所述的方法,其特征在于,还包括:

所述网络存储网元接收来自所述第一数据分析网元的网络功能注册请求,所述网络功能注册请求携带所述分析类型标识和模型信息,所述模型信息包括第二指示信息,所述第二指示信息指示所述分析类型标识对应的模型是否已经训练完成或者是否已经准备完成。

- 25 31.如权利要求30所述的方法,其特征在于,在所述第二指示信息指示所述分析类型标识对应的模型已经训练完成或者已经准备完成的情况下,所述模型信息还包括模型描述信息,所述模型描述信息包括以下信息中的一项或多项:分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息;

30 其中,所述分析过滤信息用于表示所述分析类型标识对应的模型适用的范围,所述分析过滤信息包括以下一项或多项:区域、时间段、单网络切片选择辅助信息或数据网络名称;

所述分析报告目标信息用于表示所述分析类型标识对应的模型对应的终端,所述分析报告目标信息包括以下一项或多项:终端的标识、终端组的标识或指示任意终端的信息;

- 35 所述模型性能信息用于表示所述分析类型标识对应的模型的性能,所述模型性能信息包括以下一项或多项:精确率、准确率、错误率、召回率、F1分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性;

40 所述模型部署环境信息用于表示所述分析类型标识对应的模型部署的硬件环境,所述模型部署环境信息包括以下一项或多项:中央处理器个数、图形处理器个数、内存大小或硬盘大小。

32.如权利要求 29 或 30 所述的方法,其特征在于,所述第一模型需求信息包括以下信息中的一项或多项:分析过滤信息、分析报告目标信息、模型性能信息或模型部署环境信息;

其中,所述分析过滤信息用于表示所述第二数据分析网元需要请求的模型适用的范围,
5 所述分析过滤信息包括以下一项或多项:区域、时间段、单网络切片选择辅助信息或数据网络名称;

所述分析报告目标信息用于表示所述第二数据分析网元需要请求的模型对应的终端,所述分析报告目标信息包括以下一项或多项:终端的标识、终端组的标识或指示任意终端的信息;

10 所述模型性能信息用于表示所述第二数据分析网元需要请求的模型的性能,所述模型性能信息包括以下一项或多项:精确率、准确率、错误率、召回率、F1 分数、均方误差、均方根误差、均方根对数误差、平均绝对误差、模型推理时长、模型鲁棒性、模型可扩展性、模型可解释性;

所述模型部署环境信息用于表示所述第二数据分析网元需要请求的模型部署的硬件
15 环境,所述模型部署环境信息包括以下一项或多项:中央处理器个数、图形处理器个数、内存大小或硬盘大小。

33.一种通信装置,其特征在于,包括:处理器,所述处理器与存储器耦合,所述存储器用于存储程序或指令,当所述程序或指令被所述处理器执行时,使得所述装置执行如权利要求 1-9、24-28 任一所述的方法。

20 34.一种通信装置,其特征在于,包括:处理器,所述处理器与存储器耦合,所述存储器用于存储程序或指令,当所述程序或指令被所述处理器执行时,使得所述装置执行如权利要求 10-23 任一所述的方法。

35.一种通信装置,其特征在于,包括:处理器,所述处理器与存储器耦合,所述存储器用于存储程序或指令,当所述程序或指令被所述处理器执行时,使得所述装置执行如权利要求 29-32 任一所述的方法。

36.一种芯片系统,其特征在于,包括:所述芯片系统包括至少一个处理器,和接口电路,所述接口电路和所述至少一个处理器耦合,所述处理器通过运行指令,以执行权利要求 1-9、24-28 任一所述的方法。

30 37.一种芯片系统,其特征在于,包括:所述芯片系统包括至少一个处理器,和接口电路,所述接口电路和所述至少一个处理器耦合,所述处理器通过运行指令,以执行权利要求 10-23 任一所述的方法。

38.一种芯片系统,其特征在于,包括:所述芯片系统包括至少一个处理器,和接口电路,所述接口电路和所述至少一个处理器耦合,所述处理器通过运行指令,以执行权利要求 29-32 任一所述的方法。

35 39.一种通信系统,其特征在于,包括用于执行权利要求 1-9 任一所述方法的第一数据分析网元,和用于执行权利要求 10-17 任一所述方法的第二数据分析网元。

40.一种通信系统,其特征在于,包括用于执行权利要求 24-28 任一所述方法的第一数据分析网元,和用于执行权利要求 18-23 任一所述方法的第二数据分析网元。

41.如权利要求 40 所述的系统,其特征在于,所述系统还包括用于执行权利要求 29-32
40 任一所述方法的网络存储网元。