

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 11 月 3 日 (03.11.2022)



(10) 国际公布号
WO 2022/227644 A1

(51) 国际专利分类号:
G06N 20/20 (2019.01)

(21) 国际申请号: PCT/CN2021/140955

(22) 国际申请日: 2021 年 12 月 23 日 (23.12.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110454684.2 2021 年 4 月 26 日 (26.04.2021) CN

(71) 申请人: 深圳前海微众银行股份有限公司
(WEBANK CO., LTD.) [CN/CN]; 中国广东省深圳市前海深港合作区前湾一路 1 号 A 栋 201 室, Guangdong 518052 (CN)。

(72) 发明人: 魏文斌(WEI, Wenbin); 中国广东省深圳市前海深港合作区前湾一路 1 号 A 栋 201 室, Guangdong 518052 (CN)。 范涛(FAN, Tao); 中国广东省深圳市前海深港合作区前湾一路 1 号 A 栋 201 室, Guangdong 518052 (CN)。 陈天健(CHEN, Tianjian); 中国广东省深圳市前海深港合作区前湾一路 1 号 A 栋 201 室, Guangdong 518052 (CN)。

(74) 代理人: 北京派特恩知识产权代理有限公司(CHINA PAT INTELLECTUAL PROPERTY OFFICE); 中国北京市海淀区海淀南路 21 号中关村知识产权大厦 B 座 2 层, Beijing 100080 (CN)。

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

(54) Title: DATA PROCESSING METHOD AND APPARATUS, AND DEVICE, STORAGE MEDIUM AND PROGRAM PRODUCT

(54) 发明名称: 数据处理方法、装置、设备、存储介质及程序产品

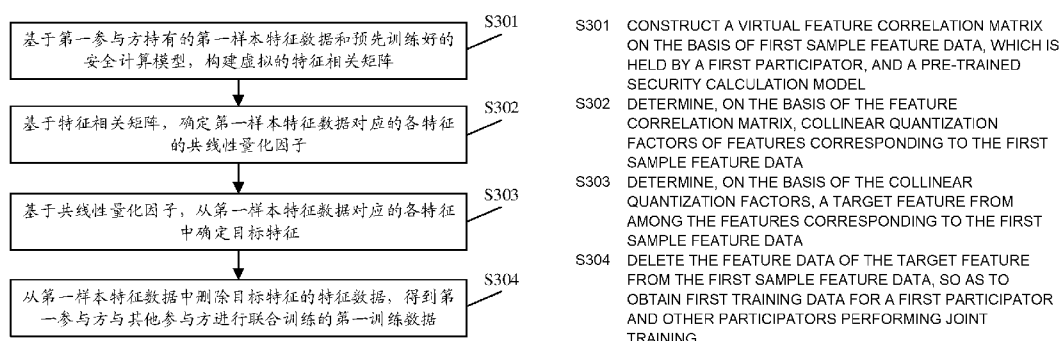


图 3

(57) Abstract: Provided are a data processing method and apparatus, and a device, a storage medium and a program product. The data processing method comprises: constructing a virtual feature correlation matrix on the basis of first sample feature data and a pre-trained security calculation model (S301); determining, on the basis of the feature correlation matrix, collinear quantization factors of features corresponding to the first sample feature data (S302); determining, on the basis of the collinear quantization factors, a target feature from among the features corresponding to the first sample feature data (S303); deleting feature data of the target feature from the first sample feature data, so as to obtain first training data for performing joint training (S304), wherein there is a linear relationship between the feature data of the target feature and the feature data of at least one of other features.

(57) 摘要: 提供一种数据处理方法、装置、设备、存储介质及程序产品, 数据处理方法包括: 基于第一样本特征数据和预先训练好的安全计算模型, 构建虚拟的特征相关矩阵 (S301); 基于特征相关矩阵, 确定第一样本特征数据对应的各特征的共线性量化因子 (S302); 基于共线性量化因子, 从第一样本特征数据对应的各特征中确定目标特征 (S303); 从第一样本特征数据中删除目标特征的特征数据, 得到用于进行联合训练的第一训练数据 (S304); 其中, 目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系。

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

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

本国际公布：

— 包括国际检索报告 (条约第21条(3))。

数据处理方法、装置、设备、存储介质及程序产品

相关申请的交叉引用

本申请基于申请号为 202110454684.2、申请日为 2021 年 04 月 26 日、申请名称为“数据处理方法、装置、设备、存储介质及程序产品”的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请作为参考。

技术领域

本申请涉及人工智能技术领域，涉及但不限于一种数据处理方法、装置、设备、存储介质及程序产品。

背景技术

机器学习是研究怎样使用计算机模拟或实现人类学习活动的科学，是人工智能中最具智能特征，最前沿的研究领域之一。机器学习的研究主要分为两类研究方向：第一类是传统机器学习的研究，该类研究主要是研究学习机制，注重探索模拟人的学习机制；第二类是大数据环境下机器学习的研究，该类研究主要是研究如何有效利用信息，注重从巨量数据中获取隐藏的、有效的、可理解的知识。

联邦学习技术是新兴的一种隐私保护技术，能够保证数据在不出本地的前提下，有效联合各方数据进行模型训练。对于大数据领域的诸多业务问题，都可通过相应的机器学习模型来解决。而剔除共线数据是训练出好的模型的关键。在保护数据隐私的情况下，相关技术中无法量化联邦学习中多方数据的共线性，无法对存在共线性的训练数据进行高效的筛选与剔除，导致训练得到的模型准确率低、稳定性较差。

发明内容

本申请实施例提供一种数据处理方法、装置、设备、计算机可读存储介质及计算机程序产品，能够在线性联邦建模中剔除存在共线性的数据，提高联邦模型的准确率和稳定性，提高模型的建模效果。

本申请实施例提供一种数据处理方法，应用于联邦学习的第一参与方，所述方法包括：

基于所述第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵，所述安全计算模型由所述第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到；

基于所述特征相关矩阵，确定所述第一样本特征数据对应的各特征的共线性量化因子；

基于所述共线性量化因子，从所述第一样本特征数据对应的各特征中确定目标特征；

从所述第一样本特征数据中删除所述目标特征的特征数据，得到所述第一参与方与所述其他参与方进行联合训练的第一训练数据；

其中，所述目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，所述其他特征包括所述第一参与方持有的除所述目标特征以外的特征和所述其他参与方持有的特征。

本申请实施例提供一种数据处理装置，应用于联邦学习的第一参与方，所述装置包括：

构建模块，配置为基于所述第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵，所述安全计算模型由所述第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到；

第一确定模块，配置为基于所述特征相关矩阵，确定所述第一样本特征数据对应的各特征的共线性量化因子；

第二确定模块，配置为基于所述共线性量化因子，从所述第一样本特征数据对应的各特征中确定目标特征；

删除模块，配置为从所述第一样本特征数据中删除所述目标特征的特征数据，得到所述第一参与方与所述其他参与方进行联合训练的第一训练数据；

其中，所述目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，所述其他特征包括所述第一参与方持有的除所述目标特征以外的特征和所述其他参与方持有的特征。

本申请实施例提供一种数据处理设备，所述设备包括：

存储器，配置为存储可执行指令；

处理器，配置为执行所述存储器中存储的可执行指令时，实现本申请实施例提供的方法。

本申请实施例提供一种计算机可读存储介质，所述计算机可读存储介质上存储有可执行指令，配置为引起处理器执行时，实现本申请实施例提供的方法。

本申请实施例提供一种计算机程序产品，包括计算机程序，所述计算机程序被处理器执行时实现本申请实施例提供的方法。

本申请实施例具有以下有益效果：在本申请实施例提供的数据处理方法中，在进行数据处理时，首先第一参与方与联邦学习的其他参与方基于安全多方计算预先训练得到安全计算模型；然后第一参与方获取第一参与方持有的第一样本特征数据，基于第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵；然后基于特征相关矩阵确定第一样本特征数据对应的各特征的共线性量化因子；由共线性量化因子确定第一样本特征数据对应的各特征中的目标特征，这些目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，这里的其他特征不仅包括第一参与方持有的除目标特征以外的特征，还包括其他参与方持有的所有特征；确定目标特征后，从第一样本特征数据中删除目标特征的特征数据，得到第一参与方与其他参与方进行联合训练的第一训练数据。如此，能够在保护数据隐私的前提下，对各参与方持有的特征数据中存在共线性的数据进行筛选与剔除，得到不存在线性关系的训练数据，从而在进行联合训练时，各参与方利用不存在线性关系的训练数据进行联合训练，能够提高联邦模型的准确率和稳定性，提高联邦模型的建模效果。

附图说明

图 1 为本申请实施例提供的数据处理方法的网络架构示意图；

图 2 是本申请实施例提供的数据处理设备的组成结构示意图；

图 3 为本申请实施例提供的数据处理方法的一种实现流程示意图；

5 图 4 为本申请实施例提供的数据处理方法的另一种实现流程示意图；

图 5 为本申请实施例提供的纵向联邦情形下的方差膨胀因子的计算流程示意图；

图 6 为本申请实施例提供的相关矩阵的行列式的计算流程示意图。

具体实施方式

10 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，所描述的实施例不应视为对本申请的限制，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

在以下的描述中，涉及到“一些实施例”，其描述了所有可能实施例的子集，但是可以理解，“一些实施例”可以是所有可能实施例的相同子集或不同子集，并且可以在不冲突的情况下相互结合。

15 在以下的描述中，所涉及的术语“第一\第二\第三”仅是区别类似的对象，不代表针对对象的特定排序，可以理解地，“第一\第二\第三”在允许的情况下可以互换特定的顺序或先后次序，以使这里描述的本申请实施例能够以除了在这里图示或描述的以外的顺序实施。

20 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本申请实施例的目的，不是旨在限制本申请。

对本申请实施例进行进一步详细说明之前，对本申请实施例中涉及的名词和术语进行说明，本申请实施例中涉及的名词和术语适用于如下的解释。

25 1) 纵向联邦学习 (Vertical Federated Learning)，在两个数据集的用户重叠较多、用户特征重叠较少的情况下，把数据集按照纵向 (即特征维度) 进行切分，并取出双方用户相同而用户特征不完全相同的部分数据进行训练的机器学习。

2) 方差膨胀因子 (VIF, Variance Inflation Factor)，又称方差扩大因子，是表征自变量观察值之间复共线性程度的数值，用于衡量多元线性回归模型中复 (多重) 共线性严重程度。

30 3) 同态加密 (Homomorphic Encryption)，同态加密是基于数学难题的计算复杂性理论的密码学技术。对经过同态加密的数据进行处理得到一个输出，将这一输出进行解密，其结果与用同一方法处理未加密的原始数据得到的输出结果是一样的。

下面说明实现本申请实施例的装置的示例性应用，本申请实施例提供的装置可以实施为终端设备。下面，将说明装置实施为终端设备时涵盖终端设备的示例性应用。

35 图 1 为本申请实施例提供的数据处理方法的网络架构示意图，如图 1 所示，在该网络架构中至少包括第一参与方 100、第二参与方 200 和网络 300。为实现支撑一个示例性应用，第一参与方 100 和第二参与方 200 可以为纵向联邦学习中联合训练机器学习模

型的各参与方。其中，第一参与方 100 和第二参与方 200 可以为客户端，例如各银行或医院等存储有用户特征数据的参与方设备，客户端可以是笔记本电脑，平板电脑，台式计算机，专用训练设备等具有模型训练功能的设备。第一参与方 100 通过网络 300 连接第二参与方 200，网络 300 可以是广域网或者局域网，又或者是二者的组合，使用无线或有

线链路实现数据传输。

第一参与方 100 首先从自身的数据中获取第一样本特征数据，对该第一样本特征数据进行处理，得到处理后的第一样本特征数据。第二参与方 200 从自身的数据中获取第二样本特征数据，对该第二样本特征数据进行处理得到处理后的第二样本特征数据。这里的第一样本特征数据和第二样本特征数据具有相同的标识，即第一样本特征数据和第二样本特征数据分别为第一参与方 100 和第二参与方 200 各自持有的同一批样本的不同特征的数据。第一参与方 100 和第二参与方 200 基于安全多方计算，利用处理后的第一样本特征数据和处理后的第二样本特征数据确定第一矩阵 E，第一参与方 100 和第二参与方 200 均持有该第一矩阵 E。基于安全多方计算，第一参与方 100 只能得到第一矩阵 E，是无法获知第二参与方 200 所持有的处理后的第二样本特征数据的；同样的，第二参与方 200 也只能得到第一矩阵 E，是无法获知第一参与方 100 所持有的处理后的第一样本特征数据的。

第一参与方 100 得到第一矩阵 E 后，根据第一矩阵 E 和处理后的第一样本特征数据，构建一虚拟的特征相关矩阵；然后计算该特征相关矩阵的行列式，以及该特征相关矩阵对应的各余子式的行列式，基于特征相关矩阵的行列式和各余子式的行列式，确定第一样本特征数据对应的各特征的共线性量化因子，该共线性量化因子可以为方差膨胀系数，用于量化每一特征与其他所有特征的共线性，这里的其他所有特征不仅包括第一参与方 100 持有的除本次量化特征以外的其他特征，还包括第二参与方 200 持有的所有特征。第一参与方 100 根据各特征的共线性量化因子，确定哪些特征的特征数据之间存在共线性，将该些特征确定为目标特征，最后从第一样本特征数据中删除目标特征的特征数据，得到第一参与方 100 与第二参与方 200 进行联合训练的第一训练数据。基于同样的方法，第二参与方 200 得到进行联合训练的第二训练数据。从而在进行联合训练时，第一参与方 100 和第二参与方 200 利用不存在共线性的第一训练数据和第二训练数据进行联合训练，能够得到准确率高、稳定性好的联邦模型。

通过本申请实施例提供的方法，能够在保护数据隐私的前提下，对各参与方持有的特征数据中存在共线性的数据进行筛选与剔除，得到不存在线性关系的训练数据，从而在进行联合训练时，各参与方利用不存在线性关系的训练数据进行联合训练，能够提高联邦模型的准确率和稳定性，提高联邦模型的建模效果。

本申请实施例提供的装置可以实施为硬件或者软硬件结合的方式，下面说明本申请实施例提供的装置的各种示例性实施。

根据图 2 示出的数据处理设备 10 的示例性结构，这里，数据处理设备 10 以应用于联邦学习的第一参与方的设备为例示出，可以预见数据处理设备 10 的其他的示例性结构，因此这里所描述的结构不应视为限制，例如可以省略下文所描述的部分组件，或者，增设下文所未记载的组件以适应某些应用的特殊需求。

图 2 所示的数据处理设备 10 包括：至少一个处理器 110、存储器 140、至少一个网

络接口 120 和用户接口 130。数据处理设备 10 中的每个组件通过总线系统 150 耦合在一起。可理解，总线系统 150 用于实现这些组件之间的连接通信。总线系统 150 除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见，在图 2 中将各种总线都标为总线系统 150。

5 用户接口 130 可以包括显示器、键盘、鼠标、触感板和触摸屏等。

存储器 140 可以是易失性存储器或非易失性存储器，也可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（ROM，Read Only Memory）。易失性存储器可以是随机存取存储器（RAM，Random Access Memory）。本申请实施例描述的存储器 140 旨在包括任意适合类型的存储器。

10 本申请实施例中的存储器 140 能够存储数据以支持数据处理设备 10 的操作。这些数据的示例包括：用于在数据处理设备 10 上操作的任何计算机程序，如操作系统和应用程序。其中，操作系统包含各种系统程序，例如框架层、核心库层、驱动层等，用于实现各种基础业务以及处理基于硬件的任务。应用程序可以包含各种应用程序。

作为本申请实施例提供的方法采用软件实施的示例，本申请实施例所提供的方法可以
15 以直接体现为由处理器 110 执行的软件模块组合，软件模块可以位于存储介质中，存储介质位于存储器 140，处理器 110 读取存储器 140 中软件模块包括的可执行指令，结合必要的硬件（例如，包括处理器 110 以及连接到总线 150 的其他组件）完成本申请实施例提供的方法。

作为示例，处理器 110 可以是一种集成电路芯片，具有信号的处理能力，例如通用
20 处理器、数字信号处理器（DSP，Digital Signal Processor），或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等，其中，通用处理器可以是微处理器或者任何常规的处理器等。

将结合本申请实施例提供的设备的示例性应用和实施，说明本申请实施例提供的数据处理方法。

25 图 3 为本申请实施例提供的数据处理方法的一种实现流程示意图，应用于图 1 所示网络架构的第一参与方，将结合图 3 示出的步骤进行说明。

步骤 S301，基于第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵。

30 本申请实施例中，第一样本特征数据对应的各特征中，不同特征的特征数据可能存在线性关系，若采用存在共线性的数据训练模型，训练好的模型准确率低、稳定性差。为了剔除存在线性关系的数据，第一参与方根据第一样本特征数据和预先训练好的安全计算模型，构建一虚拟的特征相关矩阵，记为 H。

该安全计算模型由第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到，能够进行隐私下的矩阵加减法和矩阵乘法。示例性的，本申请实施例中，以两个参与方进行联邦学习为例对该数据处理方法进行说明。在实际应用中，该数据处理方法也可以应用于三个及三个以上参与方进行联邦学习。在各参与方持有的数据不出本地、保证数据隐私的前提下，第一参与方和第二参与方基于隐私保护技术进行训练，得到训练好的安全计算模型，训练好的安全计算模型为各参与方所持有。

本申请实施例中，由于第一参与方和第二参与方进行联合训练，进行训练的数据需

要来自相同的用户才能进行联合训练，即第一参与方用于进行联合训练的第一样本特征数据，与第二参与方用于进行联合训练的第二样本特征数据，其各自对应的用户标识（ID，Identity Document）是相同的。因此在步骤 S301 之前，该方法还包括：第一参与方获取第一样本特征数据。

5 在一种实现方式中，获取第一样本特征数据可以实现为：获取第一参与方和其他参与方持有的共同训练样本；从共同训练样本中筛选出本次训练样本；将本次训练样本的特征数据确定为第一样本特征数据。共同训练样本即各参与方持有的共同用户，各参与方获取自身持有数据的标识，基于加密计算，利用各参与方持有数据的标识，确定该些参与方的共同用户。然后从这些共同用户中筛选出参与本次训练的目标用户作为本次训练
10 样本，将这些目标用户的特征数据作为第一样本特征数据。

步骤 S302，基于特征相关矩阵，确定第一样本特征数据对应的各特征的共线性量化因子。

在实际应用中，确定多个特征的特征数据之间的共线性定量因子的方法包括方差膨胀因子法、特征根分析法和条件数法。另外，还存在一可定性的分析多个特征的特征数
15 据之间共线程度的直观判定法，直观判断法一般用于初步判断，无法进行定量的分析。

第一参与方持有的第一样本特征数据，以及第二参与方持有的第二样本特征数据，与特征相关矩阵中不同列的数据相对应，确定特征相关矩阵中各列数据与其他列数据之间是否存在共线性，即可确定第一样本特征中各特征的特征数据与其他特征的特征数据之间是否存在共线性。这里不同特征的特征数据之间存在共线性包括以下可能：存在一个特征的特征数据是另一个特征的特征数据的倍数；存在一个特征的特征数据等于另一个特征的特征数据加上一个常数项；存在一个特征的特征数据等于另两个特征的特征数据相加。
20 数据相加。

本申请实施例中，根据特征相关矩阵，确定第一样本特征数据对应的各特征的共线性量化因子，根据第一样本特征数据对应的各特征的共线性量化因子，即可确定哪些特征存在与之共线的特征数据。
25 特征存在与之共线的特征数据。

步骤 S303，基于共线性量化因子，从第一样本特征数据对应的各特征中确定目标特征。

根据共线性量化因子，确定出存在共线性的目标特征。以方差膨胀因子为例，在不存在多重共线性的情况下，方差膨胀因子接近于 1，多重共线性越强，方差膨胀因子越大。实际上数据之间总是或多或少的存在多重共线性，因此将方差膨胀因子等于 1 作为评价共线性的标准是不现实的，一般的，可根据实际应用场景，预设一边界值，本申请实施例中，预设边界值可以取值为 10。判断各特征的方差膨胀因子是否大于 10，当存在某特征的方差膨胀因子大于 10 时，认为该特征的特征数据与其他特征的特征数据之间具有较强的共线性，将该特征确定为目标特征。
30 将该特征确定为目标特征。

步骤 S304，从第一样本特征数据中删除目标特征的特征数据，得到第一参与方与其他参与方进行联合训练的第一训练数据。

其中，目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，这里的其他特征包括第一参与方持有的除目标特征以外的特征和其他参与方持有的特征。本申请实施例中，可基于特征相关矩阵计算第一参与方的第一样本特征数据中各特

征的方差膨胀因子，确定出方差膨胀因子大于预设边界值的目标特征，然后对第一样本特征数据中目标特征的特征数据进行删除，得到不存在共线性的特征数据作为第一训练数据。从而在进行联合训练时，第一参与方利用不存在共线性的第一训练数据进行联合训练，能够得到准确率高、稳定性好的联邦模型。

- 5 本申请实施例提供的数据处理方法，应用于联邦学习的第一参与方，该数据处理方法包括：基于第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵，该安全计算模型由第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到；基于特征相关矩阵，确定第一样本特征数据对应的各特征的共线性量化因子；基于共线性量化因子，从第一样本特征数据对应的各特征中确定目标特征；从第一样本特征数据中删除目标特征的特征数据，得到第一参与方与其他参与方进行联合训练的第一训练数据；其中，目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，其他特征包括第一参与方持有的除目标特征以外的特征和其他参与方持有的特征。通过该方法，能够在保护数据隐私的前提下，对各参与方持有的特征数据中存在共线性的数据进行筛选与剔除，得到不存在线性关系的训练数据，从而在
10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500 505 510 515 520 525 530 535 540 545 550 555 560 565 570 575 580 585 590 595 600 605 610 615 620 625 630 635 640 645 650 655 660 665 670 675 680 685 690 695 700 705 710 715 720 725 730 735 740 745 750 755 760 765 770 775 780 785 790 795 800 805 810 815 820 825 830 835 840 845 850 855 860 865 870 875 880 885 890 895 900 905 910 915 920 925 930 935 940 945 950 955 960 965 970 975 980 985 990 995 1000 1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100 1105 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200 1205 1210 1215 1220 1225 1230 1235 1240 1245 1250 1255 1260 1265 1270 1275 1280 1285 1290 1295 1300 1305 1310 1315 1320 1325 1330 1335 1340 1345 1350 1355 1360 1365 1370 1375 1380 1385 1390 1395 1400 1405 1410 1415 1420 1425 1430 1435 1440 1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500 1505 1510 1515 1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595 1600 1605 1610 1615 1620 1625 1630 1635 1640 1645 1650 1655 1660 1665 1670 1675 1680 1685 1690 1695 1700 1705 1710 1715 1720 1725 1730 1735 1740 1745 1750 1755 1760 1765 1770 1775 1780 1785 1790 1795 1800 1805 1810 1815 1820 1825 1830 1835 1840 1845 1850 1855 1860 1865 1870 1875 1880 1885 1890 1895 1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020 2025 2030 2035 2040 2045 2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100 2105 2110 2115 2120 2125 2130 2135 2140 2145 2150 2155 2160 2165 2170 2175 2180 2185 2190 2195 2200 2205 2210 2215 2220 2225 2230 2235 2240 2245 2250 2255 2260 2265 2270 2275 2280 2285 2290 2295 2300 2305 2310 2315 2320 2325 2330 2335 2340 2345 2350 2355 2360 2365 2370 2375 2380 2385 2390 2395 2400 2405 2410 2415 2420 2425 2430 2435 2440 2445 2450 2455 2460 2465 2470 2475 2480 2485 2490 2495 2500 2505 2510 2515 2520 2525 2530 2535 2540 2545 2550 2555 2560 2565 2570 2575 2580 2585 2590 2595 2600 2605 2610 2615 2620 2625 2630 2635 2640 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700 2705 2710 2715 2720 2725 2730 2735 2740 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800 2805 2810 2815 2820 2825 2830 2835 2840 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890 2895 2900 2905 2910 2915 2920 2925 2930 2935 2940 2945 2950 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000 3005 3010 3015 3020 3025 3030 3035 3040 3045 3050 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100 3105 3110 3115 3120 3125 3130 3135 3140 3145 3150 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200 3205 3210 3215 3220 3225 3230 3235 3240 3245 3250 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300 3305 3310 3315 3320 3325 3330 3335 3340 3345 3350 3355 3360 3365 3370 3375 3380 3385 3390 3395 3400 3405 3410 3415 3420 3425 3430 3435 3440 3445 3450 3455 3460 3465 3470 3475 3480 3485 3490 3495 3500 3505 3510 3515 3520 3525 3530 3535 3540 3545 3550 3555 3560 3565 3570 3575 3580 3585 3590 3595 3600 3605 3610 3615 3620 3625 3630 3635 3640 3645 3650 3655 3660 3665 3670 3675 3680 3685 3690 3695 3700 3705 3710 3715 3720 3725 3730 3735 3740 3745 3750 3755 3760 3765 3770 3775 3780 3785 3790 3795 3800 3805 3810 3815 3820 3825 3830 3835 3840 3845 3850 3855 3860 3865 3870 3875 3880 3885 3890 3895 3900 3905 3910 3915 3920 3925 3930 3935 3940 3945 3950 3955 3960 3965 3970 3975 3980 3985 3990 3995 4000 4005 4010 4015 4020 4025 4030 4035 4040 4045 4050 4055 4060 4065 4070 4075 4080 4085 4090 4095 4100 4105 4110 4115 4120 4125 4130 4135 4140 4145 4150 4155 4160 4165 4170 4175 4180 4185 4190 4195 4200 4205 4210 4215 4220 4225 4230 4235 4240 4245 4250 4255 4260 4265 4270 4275 4280 4285 4290 4295 4300 4305 4310 4315 4320 4325 4330 4335 4340 4345 4350 4355 4360 4365 4370 4375 4380 4385 4390 4395 4400 4405 4410 4415 4420 4425 4430 4435 4440 4445 4450 4455 4460 4465 4470 4475 4480 4485 4490 4495 4500 4505 4510 4515 4520 4525 4530 4535 4540 4545 4550 4555 4560 4565 4570 4575 4580 4585 4590 4595 4600 4605 4610 4615 4620 4625 4630 4635 4640 4645 4650 4655 4660 4665 4670 4675 4680 4685 4690 4695 4700 4705 4710 4715 4720 4725 4730 4735 4740 4745 4750 4755 4760 4765 4770 4775 4780 4785 4790 4795 4800 4805 4810 4815 4820 4825 4830 4835 4840 4845 4850 4855 4860 4865 4870 4875 4880 4885 4890 4895 4900 4905 4910 4915 4920 4925 4930 4935 4940 4945 4950 4955 4960 4965 4970 4975 4980 4985 4990 4995 5000 5005 5010 5015 5020 5025 5030 5035 5040 5045 5050 5055 5060 5065 5070 5075 5080 5085 5090 5095 5100 5105 5110 5115 5120 5125 5130 5135 5140 5145 5150 5155 5160 5165 5170 5175 5180 5185 5190 5195 5200 5205 5210 5215 5220 5225 5230 5235 5240 5245 5250 5255 5260 5265 5270 5275 5280 5285 5290 5295 5300 5305 5310 5315 5320 5325 5330 5335 5340 5345 5350 5355 5360 5365 5370 5375 5380 5385 5390 5395 5400 5405 5410 5415 5420 5425 5430 5435 5440 5445 5450 5455 5460 5465 5470 5475 5480 5485 5490 5495 5500 5505 5510 5515 5520 5525 5530 5535 5540 5545 5550 5555 5560 5565 5570 5575 5580 5585 5590 5595 5600 5605 5610 5615 5620 5625 5630 5635 5640 5645 5650 5655 5660 5665 5670 5675 5680 5685 5690 5695 5700 5705 5710 5715 5720 5725 5730 5735 5740 5745 5750 5755 5760 5765 5770 5775 5780 5785 5790 5795 5800 5805 5810 5815 5820 5825 5830 5835 5840 5845 5850 5855 5860 5865 5870 5875 5880 5885 5890 5895 5900 5905 5910 5915 5920 5925 5930 5935 5940 5945 5950 5955 5960 5965 5970 5975 5980 5985 5990 5995 6000 6005 6010 6015 6020 6025 6030 6035 6040 6045 6050 6055 6060 6065 6070 6075 6080 6085 6090 6095 6100 6105 6110 6115 6120 6125 6130 6135 6140 6145 6150 6155 6160 6165 6170 6175 6180 6185 6190 6195 6200 6205 6210 6215 6220 6225 6230 6235 6240 6245 6250 6255 6260 6265 6270 6275 6280 6285 6290 6295 6300 6305 6310 6315 6320 6325 6330 6335 6340 6345 6350 6355 6360 6365 6370 6375 6380 6385 6390 6395 6400 6405 6410 6415 6420 6425 6430 6435 6440 6445 6450 6455 6460 6465 6470 6475 6480 6485 6490 6495 6500 6505 6510 6515 6520 6525 6530 6535 6540 6545 6550 6555 6560 6565 6570 6575 6580 6585 6590 6595 6600 6605 6610 6615 6620 6625 6630 6635 6640 6645 6650 6655 6660 6665 6670 6675 6680 6685 6690 6695 6700 6705 6710 6715 6720 6725 6730 6735 6740 6745 6750 6755 6760 6765 6770 6775 6780 6785 6790 6795 6800 6805 6810 6815 6820 6825 6830 6835 6840 6845 6850 6855 6860 6865 6870 6875 6880 6885 6890 6895 6900 6905 6910 6915 6920 6925 6930 6935 6940 6945 6950 6955 6960 6965 6970 6975 6980 6985 6990 6995 7000 7005 7010 7015 7020 7025 7030 7035 7040 7045 7050 7055 7060 7065 7070 7075 7080 7085 7090 7095 7100 7105 7110 7115 7120 7125 7130 7135 7140 7145 7150 7155 7160 7165 7170 7175 7180 7185 7190 7195 7200 7205 7210 7215 7220 7225 7230 7235 7240 7245 7250 7255 7260 7265 7270 7275 7280 7285 7290 7295 7300 7305 7310 7315 7320 7325 7330 7335 7340 7345 7350 7355 7360 7365 7370 7375 7380 7385 7390 7395 7400 7405 7410 7415 7420 7425 7430 7435 7440 7445 7450 7455 7460 7465 7470 7475 7480 7485 7490 7495 7500 7505 7510 7515 7520 7525 7530 7535 7540 7545 7550 7555 7560 7565 7570 7575 7580 7585 7590 7595 7600 7605 7610 7615 7620 7625 7630 7635 7640 7645 7650 7655 7660 7665 7670 7675 7680 7685 7690 7695 7700 7705 7710 7715 7720 7725 7730 7735 7740 7745 7750 7755 7760 7765 7770 7775 7780 7785 7790 7795 7800 7805 7810 7815 7820 7825 7830 7835 7840 7845 7850 7855 7860 7865 7870 7875 7880 7885 7890 7895 7900 7905 7910 7915 7920 7925 7930 7935 7940 7945 7950 7955 7960 7965 7970 7975 7980 7985 7990 7995 8000 8005 8010 8015 8020 8025 8030 8035 8040 8045 8050 8055 8060 8065 8070 8075 8080 8085 8090 8095 8100 8105 8110 8115 8120 8125 8130 8135 8140 8145 8150 8155 8160 8165 8170 8175 8180 8185 8190 8195 8200 8205 8210 8215 8220 8225 8230 8235 8240 8245 8250 8255 8260 8265 8270 8275 8280 8285 8290 8295 8300 8305 8310 8315 8320 8325 8330 8335 8340 8345 8350 8355 8360 8365 8370 8375 8380 8385 8390 8395 8400 8405 8410 8415 8420 8425 8430 8435 8440 8445 8450 8455 8460 8465 8470 8475 8480 8485 8490 8495 8500 8505 8510 8515 8520 8525 8530 8535 8540 8545 8550 8555 8560 8565 8570 8575 8580 8585 8590 8595 8600 8605 8610 8615 8620 8625 8630 8635 8640 8645 8650 8655 8660 8665 8670 8675 8680 8685 8690 8695 8700 8705 8710 8715 8720 8725 8730 8735 8740 8745 8750 8755 8760 8765 8770 8775 8780 8785 8790 8795 8800 8805 8810 8815 8820 8825 8830 8835 8840 8845 8850 8855 8860 8865 8870 8875 8880 8885 8890 8895 8900 8905 8910 8915 8920 8925 8930 8935 8940 8945 8950 8955 8960 8965 8970 8975 8980 8985 8990 8995 9000 9005 9010 9015 9020 9025 9030 9035 9040 9045 9050 9055 9060 9065 9070 9075 9080 9085 9090 9095 9100 9105 9110 9115 9120 9125 9130 9135 9140 9145 9150 9155 9160 9165 9170 9175 9180 9185 9190 9195 9200 9205 9210 9215 9220 9225 9230 9235 9240 9245 9250 9255 9260 9265 9270 9275 9280 9285 9290 9295 9300 9305 9310 9315 9320 9325 9330 9335 9340 9345 9350 9355 9360 9365 9370 9375 9380 9385 9390 9395 9400 9405 9410 9415 9420 9425 9430 9435 9440 9445 9450 9455 9460 9465 9470 9475 9480 9485 9490 9495 9500 9505 9510 9515 9520 9525 9530 9535 9540 9545 9550 9555 9560 9565 9570 9575 9580 9585 9590 9595 9600 9605 9610 9615 9620 9625 9630 9635 9640 9645 9650 9655 9660 9665 9670 9675 9680 9685 9690 9695 9700 9705 9710 9715 9720 9725 9730 9735 9740 9745 9750 9755 9760 9765 9770 9775 9780 9785 9790 9795 9800 9805 9810 9815 9820 9825 9830 9835 9840 9845 9850 9855 9860 9865 9870 9875 9880 9885 9890 9895 9900 9905 9910 9915 9920 9925 9930 9935 9940 9945 9950 9955 9960 9965 9970 9975 9980 9985 9990 9995 10000 10005 10010 10015 10020 10025 10030 10035 10040 10045 10050 10055 10060 10065 10070 10075 10080 10085 10090 10095 10100 10105 10110 10115 10120 10125 10130 10135 10140 10145 10150 10155 10160 10165 10170 10175 10180 10185 10190 10195 10200 10205 10210 10215 10220 10225 10230 10235 10240 10245 10250 10255 10260 10265 10270 10275 10280 10285 10290 10295 10300 10305 10310 10315 10320 10325 10330 10335 10340 10345 10350 10355 10360 10365 10370 10375 10380 10385 10390 10395 10400 10405 10410 10415 10420 10425 10430 10435 10440 10445 10450 10455 10460 10465 10470 10475 10480 10485 10490 10495 10500 10505 10510 10515 10520 10525 10530 10535 10540 10545 10550 10555 10560 10565 10570 10575 10580 10585 10590 10595 10600 10605 10610 10615 10620 10625 10630 10635 10640 10645 10650 10655 10660 10665 10670 10675 10680 10685 10690 10695 10700 10705 10710 10715 10720 10725 10730 10735 10740 10745 10750 10755 10760 10765 10770 10775 10780 10785 10790 10795 10800 10805 10810 10815 10820 10825 10830 10835 10840 10845 10850 10855 10860 10865 10870 10875 10880 10885 10890 10895 10900 10905 10910 10915 10920 10925 10930 10935 10940 10945 10950 10955 10960 10965 10970 10975 10980 10985 10990 10995 11000 11005 11010 11015 11020 11025 11030 11035 11040 11045 11050 11055 11060 11065 11070 11075 11080 11085 11090 11095 11100 11105 11110 11115 11120 11125 11130 11135 11140 11145 11150 11155 11160 11165 11170 11175 11180 11185 11190 11195 11200 11205 11210 11215 11220 11225 11230 11235 11240 11245 11250 11255 11260 11265 11270 11275 11280 11285 11290 11295 11300 11305 11310 11315 11320 11325 11330 11335 11340 11345 11350 11355 11360 11365 11370 11375 11380 11385 11390 11395 11400 11405 11410 11415 11420 11425 11430 11435 11440 11445 11450 11455 11460 11465 11470 11475 11480 11485 11490 11495 11500 11505 11510 11515 11520 11525 11530 11535 11540 11545 11550 11555 11560 11565 11570 11575 11580 11585 11590 11595 11600 11605 11610 11615 11620 11625 11630 11635 11640 11645 11650 11655 11660 11665 11670 11675 11680 11685 11690 11695 11700 11705 11710 11715 11720 11725 11730 11735 11740 11745 11750 11755 11760 11765 11770 11775 11780 11785 11790 11795 11800 11805 11810 11815 11

03	b3	e3	f3
04	b4	e4	f4
05	b5	e5	f5
06	b6	e6	f6

根据第一参与方的数据库表，得到第一参与方持有的特征数据

$$X = \begin{bmatrix} a1 & b1 & c1 & d1 \\ a2 & b2 & c2 & d2 \\ a3 & b3 & c3 & d3 \\ a4 & b4 & c4 & d4 \\ a6 & b6 & c6 & d6 \end{bmatrix}; \text{ 根据第二参与方的数据库表，得到第二参与方持有的特征数据}$$

$$Y = \begin{bmatrix} b1 & e1 & f1 \\ b3 & e3 & f3 \\ b4 & e4 & f4 \\ b5 & e5 & f5 \\ b6 & e6 & f6 \end{bmatrix}。$$

在隐私保护的前提下，由于各参与方无法获知其他参与方存储有哪些用户的数据，因此本申请实施例第一参与方和第二参与方基于隐私保护技术确定共同用户。基于隐私保护技术确定共同用户可以实现为：第一参与方和第二参与方分别获取自身持有用户的标识；然后基于隐私保护技术，对第一参与方持有用户的标识和第二参与方持有用户的标识计算交集，得到的结果即为共同用户。利用该方法确定得到第一参与方和第二参与方持有的共同用户为 ID 为 01、03、04 和 06 的用户，记为 $S=\{01,03,04,06\}$ 。

得到共同用户 S 后，第一参与方确定参与本次训练的目标用户，如随机地从共同用户中筛选出随机个用户作为参与本次训练的目标用户，然后从自身存储的数据中获取这些目标用户的特征数据作为第一样本特征数据。并且，第一参与方将筛选出的目标用户的 ID 发送给第二参与方，第二参与方从自身存储的数据中获取这些目标用户的特征数据，作为第二样本特征数据。或者，也可以由第二参与方确定参与本次训练的目标用户，将确定的目标用户的 ID 发送给第一参与方。例如，第一参与方随机地从 S 中筛选出 ID 为 01、03 和 04 这 3 个用户作为目标用户，记为 $S1=\{01,03,04\}$ ，第一参与方从自身持有的特征数据中筛选出目标用户 $S1$ 的特征数据，得到第一样本特征数据

$$A = \begin{bmatrix} a1 & b1 & c1 & d1 \\ a3 & b3 & c3 & d3 \\ a4 & b4 & c4 & d4 \end{bmatrix}。$$

在第二参与方侧，第二参与方从自身持有的特征数据中筛选出目标用户 $S1$ 的特征

$$\text{数据，得到第二样本特征数据 } B = \begin{bmatrix} b1 & e1 & f1 \\ b3 & e3 & f3 \\ b4 & e4 & f4 \end{bmatrix}。$$

本申请实施例中通过上述步骤，在各参与方数据不出本地、实现隐私保护的前提下，

第一参与方获得第一样本特征数据，第二参与方获得第二样本特征数据。

在一些实施例中，上述图 3 所示实施例中步骤 S301 “基于第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵”可以通过以下步骤来实现：

- 5 步骤 S3011，基于第一样本特征数据，确定第一样本特征数据对应的各特征的特征数据和第一样本特征数据对应的样本个数。

仍以上述举例数据进行说明，如表 3 所示，第一样本特征数据对应的特征有：AA、BB、CC 和 DD，特征 AA 的特征数据为{a1,a3,a4}，特征 BB 的特征数据为{b1,b3,b4}，特征 CC 的特征数据为{c1,c3,c4}，特征 DD 的特征数据为{d1,d3,d4}，可见，各特征的特征数据对应 A 的一列数据。第一样本特征数据对应的样本有：{a1,b1,c1,d1}、{a3,b3,c3,d3}和{a4,b4,c4,d4}，确定样本个数 n=3。

步骤 S3012，分别计算各特征的特征数据对应的均值和标准差。

根据 $\bar{x} = \frac{x1+x3+x4}{n}$ 计算特征 AA、BB、CC 和 DD 的特征数据对应的均值，根据

$$\sigma_x = \sqrt{\frac{(x1-\bar{x})^2 + (x3-\bar{x})^2 + (x4-\bar{x})^2}{n-1}}$$

15 差，其中 $x=a,b,c,d$ ， $n=3$ 。

步骤 S3013，基于各特征的特征数据、各特征的特征数据对应的均值、各特征的特征数据对应的标准差和样本个数，确定处理后的第一样本特征数据；

本申请实施例中，可根据各特征的特征数据对应的均值 $\bar{x}$ 、各特征的特征数据对应的标准差 σ_x 和样本个数 n 对各特征的特征数据 x 进行更新处理，得到处理后的各个特征

20 数据 x。这里处理后的各个特征数据 x 的处理公式可以为 $x = \frac{x - \bar{x}}{\sqrt{n}\sigma_x}$ ，从而得到处理后的

$$\text{第一样本特征数据 } C = \begin{bmatrix} \frac{(a1-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(b1-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(c1-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(d1-\bar{d})}{\sqrt{3}\sigma_d} \\ \frac{(a3-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(b3-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(c3-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(d3-\bar{d})}{\sqrt{3}\sigma_d} \\ \frac{(a4-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(b4-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(c4-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(d4-\bar{d})}{\sqrt{3}\sigma_d} \end{bmatrix}, \text{ 其中, } \bar{x} = \frac{x1+x3+x4}{n},$$

$$\sigma_x = \sqrt{\frac{(x1-\bar{x})^2 + (x3-\bar{x})^2 + (x4-\bar{x})^2}{n-1}}, x=a,b,c,d, n=3.$$

步骤 S3014，将处理后的第一样本特征数据输入至安全计算模型，得到第一矩阵。

25 本申请实施例中各参与方基于安全多方计算协议如 SPDZ 协议或 MASCOT 等进行训练，得到训练好的安全计算模型。该训练好的模型能够进行隐私下的矩阵加减法和矩阵乘法。安全多方计算 (MPC, Secure Multi-Party Computation) 研究的主要是针对无可信第三方的情况下，如何安全地计算一个约定函数的问题。若参与联邦学习训练的参与方有 m 个，训练好的安全计算模型可以表示为 $y_{1,...,m} = f(x_{1,...,m})$ ，其中 $m \geq 2$ 。训练完成

后, 该 m 个参与方均持有该安全计算模型, 各参与方输入自身私有的输入值, 得到对应的输出值, 例如, 第 i 个参与方输入 x_i , 输出的是 y_i , 其中, $1 \leq i \leq m$ 。

本申请实施例中以两个参与方为例, 训练好的安全计算模型表示为 $y_{1,2} = f(x_{1,2})$, 第一参与方和第二参与方均持有该模型, 第一参与方将处理后的第一样本特征数据 C 输入 $y_{1,2} = f(x_{1,2})$, 得到第一矩阵 E , 该第一矩阵 $E = C^T D$, 其中 D 为处理后的第二样本特征数据。需要说明的是, 这里第一矩阵 $E = C^T D$ 仅是为了说明 E 的值等于 $C^T D$ 的值, 并非表明 E 是根据 $C^T D$ 计算得到的, 第一参与方是根据 $y_{1,2} = f(x_{1,2})$ 计算得到的 E , 第一参与方无需从第二参与方获取处理后的第二样本特征数据 D 。由 $C^T D$ 可知, 第一矩阵 E 为 4×3 的矩阵。

对应的, 在第二参与方侧, 第二参与方将处理后的第二样本特征数据 D 输入 $y_{1,2} = f(x_{1,2})$, 得到第一矩阵 E , 该第一矩阵 $E = C^T D$, 其中 C^T 为处理后的第一样本特征数据的转置矩阵。需要说明的是, 这里第一矩阵 $E = C^T D$ 仅是为了说明 E 的值等于 $C^T D$ 的值, 并非表明 E 是根据 $C^T D$ 计算得到的, 第二参与方是根据 $y_{1,2} = f(x_{1,2})$ 计算得到的 E , 第二参与方无需从第一参与方获取处理后的第一样本特征数据的转置矩阵 C^T 。

本申请实施例中, 第一矩阵 E 是基于安全多方计算协议确定的, 各参与方无法获知其他参与方的私有数据, 能够确保各参与方数据的隐私安全。

步骤 S3015, 根据处理后的第一样本特征数据和第一矩阵, 构建虚拟的特征相关矩阵。

本申请实施例中, “构建虚拟的特征相关矩阵”, 可以通过以下步骤来实现:

步骤 S30151, 根据处理后的第一样本特征数据确定第一对称矩阵。

对称矩阵 (Symmetric Matrices) 是指以主对角线为对称轴, 各元素对应相等的方形矩阵。本申请实施例中, 根据处理后的第一样本特征数据 C 构造第一对称矩阵 F , 可以

$$\text{根据处理后的第一样本数据 } C \text{ 的转置矩阵 } C^T = \begin{bmatrix} \frac{(a1-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(a3-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(a4-\bar{a})}{\sqrt{3}\sigma_a} \\ \frac{(b1-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(b3-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(b4-\bar{b})}{\sqrt{3}\sigma_b} \\ \frac{(c1-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(c3-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(c4-\bar{c})}{\sqrt{3}\sigma_c} \\ \frac{(d1-\bar{d})}{\sqrt{3}\sigma_d} & \frac{(d3-\bar{d})}{\sqrt{3}\sigma_d} & \frac{(d4-\bar{d})}{\sqrt{3}\sigma_d} \end{bmatrix} \text{ 和处理后的}$$

$$\text{第一样本数据 } C = \begin{bmatrix} \frac{(a1-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(b1-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(c1-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(d1-\bar{d})}{\sqrt{3}\sigma_d} \\ \frac{(a3-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(b3-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(c3-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(d3-\bar{d})}{\sqrt{3}\sigma_d} \\ \frac{(a4-\bar{a})}{\sqrt{3}\sigma_a} & \frac{(b4-\bar{b})}{\sqrt{3}\sigma_b} & \frac{(c4-\bar{c})}{\sqrt{3}\sigma_c} & \frac{(d4-\bar{d})}{\sqrt{3}\sigma_d} \end{bmatrix}, \text{ 确定的第一对称矩阵 } F, F \text{ 的}$$

计算公式 $F=C^TC$, F 的维度为 $4*4$ 。

步骤 S30152, 生成一个行数和列数均等于得到第一矩阵的列数的空矩阵。

在根据第一参与方持有的特征和第二参与方持有的特征构建特征相关矩阵时, 在各参与方数据不出本地的前提下, 第一参与方只能获知第二参与方持有的特征, 无法获得第二参与方持有的各特征的特征数据, 因此, 本申请实施例中, 在第一参与方构建的是一个虚拟的特征相关矩阵, 其中与第二参与方特征的特征数据相关的数据为空。

根据上述步骤 S3014 可知第一矩阵 E 为 $4*3$ 的矩阵, 其列数为 3, 因此生成的空矩阵为维度为 $3*3$ 的矩阵, 记为 G' 。与该空矩阵 G' 相对应的第二对称矩阵 G , 由第二参与方根据第二样本特征数据 B 确定的。

步骤 S30153, 根据第一对称矩阵、第一矩阵、第一矩阵的转置矩阵和空矩阵, 构建虚拟的特征相关矩阵。

根据第一对称矩阵 F 、第一矩阵 E 、第一矩阵的转置矩阵 E^T , 以及空矩阵 G' , 构建虚拟的特征相关矩阵 H , 表示为 $H = \begin{pmatrix} F & E \\ E^T & G' \end{pmatrix}$ 。其中, 第一对称矩阵 F 的维度为 $4*4$,

第一矩阵 E 的维度为 $4*3$, 第一矩阵的转置矩阵 E^T 的维度为 $3*4$, 空矩阵 G' 的维度为 $3*3$, 因此特征相关矩阵 H 的维度为 $7*7$ 。

本申请实施例提供的方法, 第一参与方基于安全多方计算协议与其他参与方训练获得安全计算模型, 能够在各参与方持有的数据不出本地、保证数据隐私的前提下, 将无法量化联邦学习中多方数据的共线性的问题, 转换为可以处理地分块矩阵行列式问题, 从而能够基于该特征相关矩阵量化联邦学习中多方数据的共线性因子, 为剔除存在共线性的训练数据、得到准确率和稳定性的联邦模型提供基础。

在一些实施例中, 上述图 3 所示实施例中步骤 S302 “基于特征相关矩阵, 确定第一样本特征数据对应的各特征的共线性量化因子”, 可以通过以下步骤来实现:

步骤 S3021, 确定特征相关矩阵的行列式。

特征相关矩阵 H 的行列式表示为 $|H|$, 本申请实施例中, 确定该特征相关矩阵的行列式可以通过步骤 S30211 至步骤 S30214 来实现:

步骤 S30211, 生成一个行列式为预设值、维度与空矩阵相同的第一随机矩阵。

在计算特征相关矩阵的行列式时, 由于各参与方的数据不能出本地, 为了确保私有数据不泄露, 本申请实施例中生成一个用于混淆原始数据的第一随机矩阵, 如此实现隐私保护。

本申请实施例中, 第一参与方生成的第一随机矩阵满足: 行列式为预设值、维度与空矩阵 G' 相同的矩阵。为方便计算, 可以将预设值设置为 1, 根据上述步骤 S30152 可知空矩阵 G' 为 $3*3$ 的矩阵, 因此生成的第一随机矩阵为行列式为 1, 维度为 $3*3$ 的矩阵, 记为 R_1 。

步骤 S30212, 将特征相关矩阵和第一随机矩阵输入至安全计算模型, 得到第二矩阵。

第一参与方和第二参与方训练好的安全计算模型表示为 $y_{1,2} = f(x_{1,2})$, 第一参与方将特征相关矩阵 H 和第一随机矩阵 R_1 输入该模型, 得到第二矩阵 J , 该第二矩阵 $J=R_1$

($D-CA^{-1}B$) R_2 , 其中 A^{-1} 为第一样本特征数据 A 的逆, D 为处理后的第二样本特征数据, R_2 为第二参与方生成的行列式为 1、维度为 $4*4$ (维度与第一对称矩阵 F 相同) 的第二随机矩阵。需要说明的是, 这里第二矩阵 $J=R_1(D-CA^{-1}B)R_2$ 仅是为了说明第二矩阵 J 的值等于 $R_1(D-CA^{-1}B)R_2$ 的值, 并非表明第二矩阵 J 是根据 $R_1(D-CA^{-1}B)R_2$ 计算得到的, 第一参与方是根据 $y_{1,2}=f(x_{1,2})$ 计算得到的 J , 第一参与方无需从第二参与方获取处理后的第二样本特征数据 D 以及第二随机矩阵 R_2 。

本申请实施例中, 第二矩阵 J 是基于安全多方计算协议确定的, 并且各参与方利用生成的随机矩阵对各自的数据进行混淆, 使得各参与方无法获知其他参与方的私有数据, 因此能够确保各参与方数据的隐私安全。

步骤 S30213, 分别计算第一对称矩阵和第二矩阵的行列式。

根据步骤 S30151 得到第一对称矩阵 F , 根据步骤 S30212 获得第二矩阵 J , 分别计算第一对称矩阵 F 和第二矩阵 J 的行列式, 得到第一对称矩阵的行列式 $|F|$ 和第二矩阵的行列式 $|J|$ 。

这里, 当生成的第一随机矩阵的行列式 $|R_1|$ 或第二随机矩阵的行列式 $|R_2|$ 不为 1 时,

计算第二矩阵的行列式 $|J|$ 时, 需要乘以 $\frac{1}{|R_1|*|R_2|}$, 即 $|J|=\frac{1}{|R_1|*|R_2|}|R_1(D-CA^{-1}B)R_2|$ 。

步骤 S30214, 将第一对称矩阵的行列式和第二矩阵的行列式相乘, 得到特征相关矩阵的行列式。

将第一对称矩阵的行列式 $|F|$ 与第二矩阵的行列式 $|J|$ 相乘, 得到特征相关矩阵的行列式 $|H|=|F|*|J|$ 。本申请实施例中, 各参与方基于安全计算模型计算得到虚拟的特征相关矩阵的行列式, 并且整个计算过程无需获取其他参与方的私有数据, 能够确保各参与方的数据安全。

步骤 S3022, 删除特征相关矩阵第 i 行、第 i 列数据, 得到特征相关矩阵对应的各余子式。

其中, $i=1, 2, \dots, m_1$, 其中, m_1 为第一样本特征数据对应的特征个数。如上举例, 第一样本特征数据 A 的特征个数为 4, $m_1=4$ 。由于第一参与方只能确定自身持有的第一样本特征数据对应的各特征的共线性量化因子, 因此这里只需得到特征相关矩阵对应的第 1 余子式、第 2 余子式、第 3 余子式和第 4 余子式。删除特征相关矩阵 H_{7*7} 的第 i 行、第 i 列的数据, 剩下的数据不改变原来的顺序, 构成维度为 $6*6$ 的矩阵, 该矩阵即为第一样本特征数据的第 i 个特征对应的第 i 余子式 H_{ii} , 例如, 当 i 为 2 时, 得到的即为特征 BB 对应的第 2 余子式 H_{22} 。

步骤 S3023, 确定各余子式的行列式。

本申请实施例中, 确定各余子式的行列式的计算方式, 可以与步骤 S3021 中确定特征相关矩阵的行列式的方式相同。例如, 确定第 i 个特征对应余子式的行列式, 可以实现为: 生成一个行列式为预设值、维度与空矩阵相同的随机矩阵; 将第 i 余子式和该随机矩阵输入至安全计算模型, 得到第 i 余子式对应的矩阵; 分别计算第一对称矩阵和该第 i 余子式对应的矩阵的行列式; 将第一对称矩阵的行列式和第 i 余子式对应的矩阵矩阵的行列式相乘, 得到第 i 余子式的行列式 $|H_{ii}|$ 。

步骤 S3024, 基于特征相关矩阵的行列式和各余子式的行列式, 确定第一样本特征数据对应的各特征的共线性量化因子。

本申请实施例中, 以方差膨胀因子为例确定共线性量化因子, 第 i 个特征的方差膨胀因子 VIF_i , 可以根据特征相关矩阵的行列式 $|H|$ 和第 i 余子式的行列式 $|H_{ii}|$ 确定, 计算公式如式 (1) 所示:

$$VIF_i = \frac{|H_{ii}|}{|H|} \quad (1);$$

其中, 当 $i=1$ 时, 得到的 VIF_1 为特征 AA 的共线性量化因子, $i=2$ 时, 得到的 VIF_2 为特征 BB 的共线性量化因子, $i=3$ 时, 得到的 VIF_3 为特征 CC 的共线性量化因子, $i=4$ 时, 得到的 VIF_4 为特征 DD 的共线性量化因子。

10 本申请实施例提供的方法, 第一参与方基于安全多方计算协议, 计算得到特征相关矩阵的行列式, 以及第一样本特征数据对应的各特征的余子式的行列式, 从而得到第一样本特征数据对应的各特征的方差膨胀因子, 为剔除存在共线性的训练数据、得到准确率和稳定性的联邦模型提供基础。

15 在一些实施例中, 上述图 3 所示实施例中步骤 S303 “基于共线性量化因子, 从第一样本特征数据对应的各特征中确定目标特征”, 可以通过以下步骤来实现:

步骤 S3031, 判断第一样本特征数据对应的各特征的共线性量化因子是否大于预设边界值。

这里判断第一样本特征数据对应的各特征的共线性量化因子与预设边界值的大小, 即判断 VIF_1 、 VIF_2 、 VIF_3 和 VIF_4 与预设边界值的大小。

20 步骤 S3032, 将共线性量化因子大于预设边界值的特征确定为目标特征。

预设边界值以 10 为例, 判断得到 VIF_1 和 VIF_2 大于 10, 则将 VIF_1 对应的特征 AA 和 VIF_2 对应的特征 BB 确定为目标特征。

25 结合上述举例, 第一参与方的目标特征 AA 为生日, 第一参与方的目标特征 BB 为年龄, 第二参与方的特征 BB 为年龄, 确定生日与年龄之间存在线性关系 (年龄等于今年减去出生年)、年龄与年龄之间存在线性关系 (相等, 即倍数为 1)。即第一参与方的目标特征 AA 的特征数据与第一参与方的特征 BB 的特征数据、以及第二参与方的特征 BB 的特征数据存在线性关系, 第一参与方的目标特征 BB 的特征数据与第一参与方的特征 AA 的特征数据、以及第二参与方的特征 BB 的特征数据存在线性关系。

得到目标特征后, 执行步骤 S304, 在第一样本特征数据中删除目标特征的特征数

30 据, 即删除第一样本特征数据 $A = \begin{bmatrix} a1 & b1 & c1 & d1 \\ a3 & b3 & c3 & d3 \\ a4 & b4 & c4 & d4 \end{bmatrix}$ 中第 1 列数据 (特征 AA 的特征数

据) 和第 2 列数据 (特征 BB 的特征数据), 得到的第一训练数据 $A' = \begin{bmatrix} c1 & d1 \\ c3 & d3 \\ c4 & d4 \end{bmatrix}$ 。

与上述确定第一参与方持有的第一样本特征数据对应的各特征中的目标特征步骤相同, 确定得到第二参与方持有的第二样本特征数据对应的各特征中的目标特征为第二

参与方的特征 BB，在第二样本特征数据中删除特征 BB 的特征数据，即删除第二样本

$$\text{特征数据 } B = \begin{bmatrix} b1 & e1 & f1 \\ b3 & e3 & f3 \\ b4 & e4 & f4 \end{bmatrix} \text{ 中第 1 列数据 (特征 BB 的特征数据), 得到的第二训练数据}$$

$$B' = \begin{bmatrix} e1 & f1 \\ e3 & f3 \\ e4 & f4 \end{bmatrix}.$$

然后，第一参与方和第二参与方根据第一训练数据 A' 和第二训练数据 B' 进行联合
5 训练，第一训练数据 A' 和第二训练数据 B' 中已经删除存在共线关系的特征数据，因此
在进行联合训练时，第一参与方和第二参与方利用不存在线性关系的第一训练数据和第
二训练数据进行联合训练，得到的联邦模型准确率较高、稳定性较好，能够提高联邦模
型的建模效果。

基于前述的实施例，本申请实施例再提供一种数据处理方法，图 4 为本申请实施例
10 提供的数据处理方法的再一种实现流程示意图，应用于图 1 所示的网络架构，如图 4 所
示，该数据处理方法包括以下步骤：

步骤 S401，第一参与方和第二参与方基于安全多方计算协议，确定第一参与方和
第二参与方持有的共同用户。

由于第一参与方和第二参与方进行联合训练，进行训练的数据需要来自相同的用户
15 才能进行联合训练，即第一参与方用于进行联合训练的第一样本特征数据，与第二参与
方用于进行联合训练的第二样本特征数据，其各自对应的 ID 是相同的。获取第一样本
特征数据时，首先需要确定各参与方持有的共同用户，然后从这些共同用户中筛选出参
与本次训练的目标用户，将这些目标用户的特征数据作为第一样本特征数据。

在隐私保护的前提下，由于各参与方无法获知其他参与方存储有哪些用户的数据，
20 因此本申请实施例中第一参与方和第二参与方基于隐私保护技术确定共同用户。基于隐
私保护技术确定共同用户可以实现为：第一参与方和第二参与方分别获取自身持有用户
的标识；然后基于隐私保护技术，对第一参与方持有用户的标识和第二参与方持有用户
的标识计算交集，得到的结果即为共同用户。

步骤 S402，第一参与方确定参与本次训练的目标用户。

25 得到共同用户后，第一参与方确定参与本次训练的目标用户，如随机地从共同用户
中筛选出随机个用户作为参与本次训练的目标用户。

步骤 S403，第一参与方将目标用户的标识发送给第二参与方。

在另一些实施例中，也可以由第二参与方确定参与本次训练的目标用户，此时步骤
S402 和步骤 S403 可以替换为：

30 步骤 S402'，第二参与方确定参与本次训练的目标用户。

步骤 S403'，第二参与方将目标用户的标识发送给第一参与方。

步骤 S404，第一参与方获取自身持有的第一样本特征数据。

第一参与方从自身存储的数据中获取目标用户的特征数据作为第一样本特征数据。

步骤 S405，第一参与方基于第一样本特征数据和预先训练好的安全计算模型，构

建虚拟的第一特征相关矩阵。

在一些实施例中，第一参与方构建第一特征相关矩阵可以实现为：基于第一样本特征数据，确定第一样本特征数据对应的各特征的特征数据和第一样本特征数据对应的样本个数；分别计算各特征的特征数据对应的均值和标准差；基于各特征的特征数据、各特征的特征数据对应的均值、各特征的特征数据对应的标准差和样本个数，确定处理后的第一样本特征数据；将处理后的第一样本特征数据输入至安全计算模型，得到第一矩阵；根据处理后的第一样本特征数据确定第一对称矩阵；生成一个行数和列数均等于第一矩阵的列数的第一空矩阵；根据第一对称矩阵、第一矩阵、第一矩阵的转置矩阵和第一空矩阵，构建虚拟的第一特征相关矩阵。

例如，第一样本特征数据为 A ；根据 $x = \frac{x - \bar{x}}{\sqrt{n}\sigma_x}$ 对 A 中各个特征数据 x 进行处理，得到处理后的第一样本特征数据为 C ；将处理后的第一样本特征数据 C 输入至预先训练好的安全计算模型 $y_{1,\dots,m} = f(x_{1,\dots,m})$ ，得到第一矩阵 E ；根据处理后的第一样本特征数据的转置矩阵 C^T 和处理后的第一样本特征数据 C ，确定第一对称矩阵 $F = C^T C$ ；第一参与方生成的第一空矩阵为 G' ；得到第一参与方构建的第一特征相关矩阵 $H = \begin{pmatrix} F & E \\ E^T & G' \end{pmatrix}$ 。

步骤 S406，第一参与方基于第一特征相关矩阵，确定第一样本特征数据对应的各特征的共线性量化因子。

在一些实施例中，第一参与方确定第一样本特征数据对应的各特征的共线性量化因子可以实现为：确定第一特征相关矩阵的行列式；删除第一特征相关矩阵第 i 行、第 i 列数据，得到第一特征相关矩阵对应的各余子式，其中， $i=1, 2, \dots, m_1$ ， m_1 为第一样本特征数据对应的特征个数；确定各余子式的行列式；基于第一特征相关矩阵的行列式和各余子式的行列式，确定第一样本特征数据对应的各特征的共线性量化因子。

本申请实施例中，共线性量化因子可以为方差膨胀因子、回归因子或相关系数。以方差膨胀因子为例，基于特征相关矩阵的行列式和各余子式的行列式，确定第一样本特征数据对应的各特征的共线性量化因子可以实现为：根据各余子式的行列式与特征相关矩阵的行列式，计算各特征的方差膨胀因子；将各特征的方差膨胀因子确定为第一样本特征数据对应的各特征的共线性量化因子。删除第一特征相关矩阵 H 的第 i 行、第 i 列的数据，得到第一样本特征数据的第 i 个特征对应的第 i 余子式 H_{ii} ，计算第一特征相关矩阵的行列式 $|H|$ 和第 i 余子式的行列式 $|H_{ii}|$ ，将第 i 余子式的行列式与第一特征相关矩阵的行列式的商 $VIF_i = \frac{|H_{ii}|}{|H|}$ ，确定为第 i 特征的方差膨胀因子，其中 $i=1, 2, \dots, m_1$ ， m_1

为第一样本特征数据对应的特征个数，如此得到第一样本特征数据对应的各特征的共线性量化因子。

其中，确定第一特征相关矩阵的行列式 $|H|$ 可以实现为：生成一个行列式为预设值、维度与第一空矩阵相同的第一随机矩阵；将第一特征相关矩阵和第一随机矩阵输入至安全计算模型，得到第二矩阵；分别计算第一对称矩阵和第二矩阵的行列式；将第一对称矩阵的行列式和第二矩阵的行列式相乘，得到第一特征相关矩阵的行列式。

同理，确定第 i 个特征对应的第 i 余子式的行列式 $|H_{ii}|$ ，可以实现为：生成一个行列式为预设值、维度与第一空矩阵相同的随机矩阵；将第 i 余子式和该随机矩阵输入至安全计算模型，得到第 i 余子式对应的矩阵；分别计算第一对称矩阵和该第 i 余子式对应的矩阵的行列式；将第一对称矩阵的行列式和第 i 余子式对应的矩阵矩阵的行列式相乘，
5 得到第 i 余子式的行列式 $|H_{ii}|$ 。

步骤 S407，第一参与方基于共线性量化因子，从第一样本特征数据对应的各特征中确定目标特征。

在一些实施例中，第一参与方根据共线性量化因子确定目标特征可以实现为：判断第一样本特征数据对应的各特征的共线性量化因子是否大于预设边界值；将共线性量化因子大于预设边界值的特征确定为目标特征。
10

步骤 S408，第一参与方从第一样本特征数据中删除目标特征的特征数据，得到第一参与方与其他参与方进行联合训练的第一训练数据。

其中，第一参与方中目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，其他特征包括第一参与方持有的除目标特征以外的特征和第二参与方持有的所有特征。
15

步骤 S409，第二参与方获取自身持有的第二样本特征数据。

第二参与方从自身存储的数据中获取目标用户的特征数据作为第二样本特征数据。

步骤 S410，第二参与方基于第二样本特征数据和预先训练好的安全计算模型，构建虚拟的第二特征相关矩阵。

在一些实施例中，第二参与方构建第二特征相关矩阵可以实现为：基于第二样本特征数据确定第二样本特征数据对应的各特征的特征数据和第二样本特征数据对应的样本个数；分别计算各特征的特征数据对应的均值和标准差；基于各特征的特征数据、各特征的特征数据对应的均值、各特征的特征数据对应的标准差和样本个数，确定处理后的第二样本特征数据；将处理后的第二样本特征数据输入至安全计算模型，得到第一矩阵；根据处理后的第二样本特征数据确定第二对称矩阵；生成一个行数和列数均等于第一矩阵的列数的第二空矩阵；根据第二对称矩阵、第一矩阵、第一矩阵的转置矩阵和第二空矩阵，构建虚拟的第一特征相关矩阵。
20
25

例如，第二样本特征数据为 B ；根据 $x = \frac{x - \bar{x}}{\sqrt{n}\sigma_x}$ 对 B 中各个特征数据 x 进行处理，

得到处理后的第二样本特征数据为 D ；将处理后的第二样本特征数据 D 输入至预先训练好的安全计算模型 $y_{1,\dots,m} = f(x_{1,\dots,m})$ ，得到第一矩阵 E ；根据处理后的第二样本特征数据的转置矩阵 D^T 和处理后的第二样本特征数据 D ，确定第二对称矩阵 $G = D^T D$ ；第二参与方生成的第二空矩阵为 F' ；得到第二参与方构建的第二特征相关矩阵 $H' = \begin{pmatrix} F' & E \\ E^T & G \end{pmatrix}$ 。
30

步骤 S411，第二参与方基于特征相关矩阵，确定第二样本特征数据对应的各特征的共线性量化因子。

在一些实施例中，第二参与方确定第二样本特征数据对应的各特征的共线性量化因子可以实现为：确定第二特征相关矩阵的行列式；删除第二特征相关矩阵第 $i+m_1$ 行、
35

第 $i+m_1$ 列数据 (因为第二特征相关矩阵 H' 的前 m_1 列为第一样本特征数据的特征, 因此在确定第二样本特征数据对应的各特征的共线性量化因子时, 需要加上 m_1), 得到第二特征相关矩阵对应的各余子式, 其中, $i=1, 2, \dots, m_2$, m_2 为第二样本特征数据对应的特征个数; 确定各余子式的行列式; 基于第二特征相关矩阵的行列式和各余子式的行列式, 确定第二样本特征数据对应的各特征的共线性量化因子。

删除第二特征相关矩阵 H' 的第 $i+m_1$ 行、第 $i+m_1$ 列的数据, 得到第二样本特征数据的第 i 个特征对应的第 i 余子式 $H'_{i+m_1, i+m_1}$, 计算第二特征相关矩阵的行列式 $|H'|$ 和第 i 余子式的行列式 $|H'_{i+m_1, i+m_1}|$, 确定第一样本特征数据对应的各特征的共线性量化因子 (以方

差膨胀因子为例) $VIF_i = \frac{|H'_{i+m_1, i+m_1}|}{|H'|}$ 。

其中, 确定第二特征相关矩阵的行列式 $|H'|$ 可以实现为: 生成一个行列式为预设值、维度与第二空矩阵相同的第二随机矩阵; 将第二特征相关矩阵和第二随机矩阵输入至安全计算模型, 得到第二矩阵; 分别计算第二对称矩阵和第二矩阵的行列式; 将第二对称矩阵的行列式和第二矩阵的行列式相乘, 得到第二特征相关矩阵的行列式。

同理, 确定第 i 个特征对应的第 i 余子式的行列式 $|H'_{i+m_1, i+m_1}|$, 可以实现为: 生成一个行列式为预设值、维度与第二空矩阵相同的随机矩阵; 将第 i 余子式和该随机矩阵输入至安全计算模型, 得到第 i 余子式对应的矩阵; 分别计算第二对称矩阵和该第 i 余子式对应的矩阵的行列式; 将第二对称矩阵的行列式和第 i 余子式对应的矩阵的行列式相乘, 得到第 i 余子式的行列式 $|H'_{i+m_1, i+m_1}|$ 。

步骤 S412, 第二参与方基于共线性量化因子, 从第二样本特征数据对应的各特征中确定目标特征。

在一些实施例中, 第二参与方根据共线性量化因子确定目标特征可以实现为: 判断第二样本特征数据对应的各特征的共线性量化因子是否大于预设边界值; 将共线性量化因子大于预设边界值的特征确定为目标特征。

步骤 S413, 第二参与方从第二样本特征数据中删除目标特征的特征数据, 得到第二参与方与其他参与方进行联合训练的第二训练数据。

其中, 第二参与方中目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系, 其他特征包括第二参与方持有的除目标特征以外的特征和第一参与方持有的所有特征。

在另一些实施例中, 在步骤 S413 之后, 第一参与方和第二参与方基于安全多方计算协议, 利用第一训练数据和第二训练数据进行联合模型, 由于第一训练数据和第二训练数据中不存在线性关系的训练数据, 能够提高训练得到的联邦模型的准确率和稳定性, 从而提高联邦模型的建模效果。

本申请实施例提供的数据处理方法, 第一参与方与第二参与方基于安全多方计算预先训练得到安全计算模型; 然后第一参与方和第二参与方分别获取第一样本特征数据和第二样本特征数据, 第一参与方基于第一样本特征数据和安全计算模型, 构建虚拟的第一特征相关矩阵, 第二参与方基于第二样本特征数据和安全计算模型, 构建虚拟的第二

特征相关矩阵；然后第一参与方基于第一特征相关矩阵确定第一样本特征数据对应的各特征的共线性量化因子，第二参与方基于第二特征相关矩阵确定第二样本特征数据对应的各特征的共线性量化因子；第一参与方由第一样本特征数据对应的各特征的共线性量化因子确定第一样本特征数据对应的各特征中的目标特征，第二参与方由第二样本特征数据对应的各特征的共线性量化因子确定第二样本特征数据对应的各特征中的目标特征，这些目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，这里的其他特征包括第一参与方和第二参与方持有的除目标特征以外的所有特征；确定目标特征后，第一参与方从第一样本特征数据中删除目标特征的特征数据，得到第一训练数据；第二参与方从第二样本特征数据中删除目标特征的特征数据，得到第二训练数据。

如此，能够在保护数据隐私的前提下，对第一参与方和第二参与方持有的特征数据中存在共线性的数据进行筛选与剔除，得到不存在线性关系的训练数据，从而在进行联合训练时，第一参与方和第二参与方利用不存在线性关系的训练数据进行联合训练，能够提高联邦模型的准确率和稳定性，提高联邦模型的建模效果。

下面，将说明本申请实施例在一个实际的应用场景中的示例性应用。

纵向联邦学习通常由不同的参与方联合训练机器学习模型。在线性模型的建模过程中，共线性的存在会显著影响线性模型的稳定和效果，因此在建模过程中需要剔除。方差膨胀因子（VIF, Variance Inflation Factor）可以很好的量化某一个特征与其他所有特征的共线性，在实际建模中非常常用。相关技术中，计算 VIF 时，通常计算方式需要将各方数据汇总到一个地方进行。但是在各方（如银行企业）的数据可能涉及个人隐私或者商业机密，直接开放给其他参与方会造成信息泄露。相关技术中针对线性联邦建模的特征筛选方案主要都是针对单列数据，缺少 VIF 这种能够刻画多列数据的共线性的方法。相关技术无法在保护数据隐私的情况下联合多方的数据计算 VIF。

本申请实施例针对两个独立参与方的场景，可以假设为两个公司之间进行数据联合建模。由于合规、隐私和商业机密等原因，各自公司的原始数据不能离开各自的公司。建模过程中交换的中间数据也不能推导或者泄露不必要的原始数据信息。两方公司各自持有多个同一批 ID 用户的不同特征数据，这些特征之间可能存在共线性，可能影响后续线性模型的建模效果。本申请实施例利用两方的隐私安全的技术方案来计算方差膨胀因子 VIF，基于此来进行特征筛选。

图 5 为本申请实施例提供的纵向联邦情形下的方差膨胀因子的计算流程示意图，图 6 为本申请实施例提供的相关矩阵的行列式的计算流程示意图，下面结合图 5 和图 6，对本申请实施例提供的纵向联邦情形下的方差膨胀因子的计算方法进行详细说明。

1) 本申请实施例提供有两个独立的参与方（分别记为 Alice 跟 Bob），各自持有同一批 ID 的不同特征的数据，分别记为矩阵 A, B（这里 A, B 具有相同的行数 n, A 具有列数 m_1 , B 具有列数 m_2 ）。

2) Alice, Bob 在本地分别对 A, B 的特征（矩阵的列）进行归一化，得到矩阵 C, D。归一化的方式为：
$$x = \frac{x - \bar{x}}{\sqrt{n\sigma_x}}$$
，x 代表一列特征， $\bar{x}$ 代表特征的均值， σ_x 代表特征的标准差。

3) 通过扩展的 SPDZ 协议，进行隐私下的矩阵加减法和矩阵乘法，计算矩阵乘积

$C^T D = E$, 计算的结果 Alice, Bob 均持有。本申请实施例中, 通过安全多方计算协议 SPDZ, 实现隐私保护下矩阵的加减法和矩阵乘法计算, 各参与方输入自身的私有数据, 得到计算结果, 无法获得其他参与方的私有数据, 使得整个计算过程不泄漏结果之外的任何额外信息, 兼具高安全性与实用计算效率。

5 4) Alice 本地计算矩阵乘法 $C^T C = F$, Bob 本地计算矩阵乘法 $D^T D = G$ 。

5) 上一步得到了皮尔逊 Pearson 相关性矩 $H = \begin{pmatrix} F & E \\ E^T & G \end{pmatrix}$ 。

注 1: 这里 Alice 持有 F 、 E 、 E^T , Bob 持有 E 、 E^T 、 G 。

根据 VIF 定义, 第 i 个特征的 $VIF_i = \frac{|H_{ii}|}{|H|}$, 这里 H_{ii} 为矩阵 H 的余子式 (即删除矩

阵的第 i 行和第 i 列剩下的矩阵), $|H|$ 表示矩阵 H 的行列式。

10 6) 假设第 i 个特征在 Alice 方 (在 Bob 方需要指标 i 加上 F_{ii} 的列数 m_1), 那么 Alice 可以删除矩阵 F 的第 i 行跟第 i 列得到矩阵 F_{ii} 。Alice 跟 Bob 都可以删除矩阵 E 的第 i 行得到 E_{i*} , 删除矩阵 E^T 的第 i 列得到 E_{*i}^T 。那么余子式 $H_{ii} = \begin{pmatrix} F_{ii} & E_{i*} \\ E_{*i}^T & G_{i+m_1, i+m_1} \end{pmatrix}$ 。

注 2: 这里 Alice 持有 F_{ii} 、 E_{i*} 、 E_{*i}^T , Bob 持有 E_{i*} 、 E_{*i}^T 、 $G_{i+m_1, i+m_1}$ 。

15 本申请实施例中, 通过巧妙的矩阵变换将无法处理的原始矩阵 H 的行列式的计算问题, 转换为可以处理的分块矩阵 F_{ii} 、 E_{i*} 、 E_{*i}^T 和 $G_{i+m_1, i+m_1}$ 的行列式计算问题。

下面解决根据分块矩阵 F_{ii} 、 E_{i*} 、 E_{*i}^T 和 $G_{i+m_1, i+m_1}$ 的行列式计算问题。

7) 根据注 1 跟注 2, 计算 $|H_{ii}|$ 与 $|H|$ 所需要解决的是同一个问题。在计算 $|H|$ 时, 下面的 M_1 、 M_2 、 M_3 、 M_4 分别对应 F 、 E 、 E^T 、 G , 在计算 $|H_{ii}|$ 时, 下面的 M_1 、 M_2 、 M_3 、 M_4 分别对应 F_{ii} 、 E_{i*} 、 E_{*i}^T 、 $G_{i+m_1, i+m_1}$ 。Alice 持有矩阵 M_1 、 M_2 、 M_3 , Bob 持有矩阵 M_2 、

20 M_3 、 M_4 , 计算行列式 $\begin{vmatrix} M_1 & M_2 \\ M_3 & M_4 \end{vmatrix}$ 。

8) 从这里开始解决 7) 的问题 (由于 Alice 无法得到 M_4 , Bob 无法得到 M_1 , 所以 7) 中行列式无法直接计算得到)。当 M_1 可逆时, $\begin{vmatrix} M_1 & M_2 \\ M_3 & M_4 \end{vmatrix} = |M_1| |M_4 - M_3 M_1^{-1} M_2|$ 。

9) Alice 本地计算 $|M_1|$ 和 $M_3 M_1^{-1} M_2$, 并本地生成行列式为 1、维度与 M_4 一致的随机矩阵 R_1 。Bob 本地生成行列式为 1 与 M_4 维度一致的随机矩阵 R_2 。

25 10) 再次利用扩展的 SPDZ 协议计算矩阵减法跟乘法并恢复结果矩阵 $J = R_1 (M_4 - M_3 M_1^{-1} M_2) R_2$ 。

由于随机矩阵的存在, 通过 J , Alice 跟 Bob 均无法恢复对方的原始矩阵。

本申请实施例中, 通过构造特殊的随机矩阵来混淆原始矩阵, 将涉及原始数据信息的矩阵的行列式问题转化具有相同行列式的随机矩阵的行列式计算问题。通过随机矩阵对原始矩阵进行混淆处理, 各参与方无法恢复其他参与方的数据, 能够对各参与方的数据进行隐私保护, 避免造成数据泄露。

11) Alice 本地计算 $|M_1||J|=|M_1||R_1||M_4-M_3M_1^{-1}M_2||R_2|$, 从而得到行列式 $\begin{vmatrix} M_1 & M_2 \\ M_3 & M_4 \end{vmatrix}$,

解决了 7) 的问题。

12) 根据 7) 至 11) 描述的方法, 依次计算 $|H_{11}|$ 、 $|H_{22}|$ 、...、 $|H_{m_1+m_2, m_1+m_2}|$ 和 $|H|$, 从而得到 $VIF_1, VIF_2, \dots, VIF_{m_1+m_2}$ 。

5 本申请实施例提供的方法, 通过巧妙的矩阵变换将无法处理的原始行列式计算问题转换为可以处理的分块矩阵行列式问题; 通过构造特殊的随机矩阵来混淆原始矩阵, 将涉及原始数据信息的矩阵的行列式问题转化具有相同行列式的随机矩阵的行列式计算问题; 通过 SPDZ 协议的巧妙配合, 整个计算过程不泄漏结果之外的任何额外信息, 兼具高安全性与实用计算效率。通过本申请实施例提供的方法, 使得方差膨胀因子 VIF 的安全计算成为可能, 进而能进行高效的特征筛选, 提高后续线性模型的整体效果; 并且
10 兼顾高安全性与实用性, 除结果外, 不泄漏不必要的数据信息, 且整体计算开销都控制在生产实用的效率范围。

下面继续说明本申请实施例提供的数据处理装置的实施为软件模块的示例性结构, 在一些实施例中, 如图 2 所示, 存储在存储器 140 的数据处理装置 70 应用于对模型进行联合训练的第二参与方, 所述数据处理装置 70 中的软件模块可以包括:
15

构建模块 71, 配置为基于所述第一参与方持有的第一样本特征数据和预先训练好的安全计算模型, 构建虚拟的特征相关矩阵, 所述安全计算模型由所述第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到;

20 第一确定模块 72, 配置为基于所述特征相关矩阵, 确定所述第一样本特征数据对应的各特征的共线性量化因子;

第二确定模块 73, 配置为基于所述共线性量化因子, 从所述第一样本特征数据对应的各特征中确定目标特征;

删除模块 74, 配置为从所述第一样本特征数据中删除所述目标特征的特征数据, 得到所述第一参与方与所述其他参与方进行联合训练的第一训练数据;

25 其中, 所述目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系, 所述其他特征包括所述第一参与方持有的除所述目标特征以外的特征和所述其他参与方持有的特征。

在一些实施例中, 所述构建模块 71, 包括:

30 第一确定子模块, 配置为基于所述第一样本特征数据, 确定所述第一样本特征数据对应的各特征的特征数据和所述第一样本特征数据对应的样本个数;

计算子模块, 配置为分别计算所述各特征的特征数据对应的均值和标准差;

第二确定子模块, 配置为基于所述各特征的特征数据、所述各特征的特征数据对应的均值、所述各特征的特征数据对应的标准差和所述样本个数, 确定处理后的第一样本特征数据;

35 输入子模块, 配置为将所述处理后的第一样本特征数据输入至所述安全计算模型, 得到第一矩阵;

构建子模块, 配置为根据所述处理后的第一样本特征数据和所述第一矩阵, 构建虚

拟的特征相关矩阵。

在一些实施例中，所述构建子模块，包括：

第一确定单元，配置为根据所述处理后的第一样本特征数据确定第一对称矩阵；

第一生成单元，配置为生成一个行数和列数均等于所述第一矩阵的列数的空矩阵；

5 构建单元，配置为根据所述第一对称矩阵、所述第一矩阵、所述第一矩阵的转置矩阵和所述空矩阵，构建虚拟的特征相关矩阵。

在一些实施例中，所述第一确定模块 72，包括：

第三确定子模块，配置为确定所述特征相关矩阵的行列式；

10 删除子模块，配置为删除所述特征相关矩阵第 i 行、第 i 列数据，得到所述特征相关矩阵对应的各余子式，其中， $i=1, 2, \dots, m_1$ ， m_1 为所述第一样本特征数据对应的特征个数；

第四确定子模块，配置为确定所述各余子式的行列式；

第五确定子模块，配置为基于所述特征相关矩阵的行列式和所述各余子式的行列式，确定所述第一样本特征数据对应的各特征的共线性量化因子。

15 在一些实施例中，所述第三确定子模块，包括：

第二生成单元，配置为生成一个行列式为预设值、维度与所述空矩阵相同的第一随机矩阵；

输入单元，配置为将所述特征相关矩阵和所述第一随机矩阵输入至所述安全计算模型，得到第二矩阵；

20 第一计算单元，配置为分别计算所述第一对称矩阵和所述第二矩阵的行列式；

第二计算单元，配置为将所述第一对称矩阵的行列式和所述第二矩阵的行列式相乘，得到所述特征相关矩阵的行列式。

在一些实施例中，所述第五确定子模块，包括：

25 第三计算单元，配置为根据所述各余子式的行列式与所述特征相关矩阵的行列式，计算各特征的方差膨胀因子；

第二确定单元，配置为将所述各特征的方差膨胀因子确定为所述第一样本特征数据对应的各特征的共线性量化因子。

在一些实施例中，所述第二确定模块 73，包括：

30 判断子模块，配置为判断所述第一样本特征数据对应的各特征的共线性量化因子是否大于预设边界值；

第六确定子模块，配置为将共线性量化因子大于预设边界值的特征确定为目标特征。

在一些实施例中，所述数据处理装置 70，还包括：

获取模块，配置为获取所述第一参与方和所述其他参与方持有的共同训练样本；

35 筛选模块，配置为从所述共同训练样本中筛选出本次训练样本；

第三确定模块，配置为将所述本次训练样本的特征数据确定为所述第一样本特征数据。

在一些实施例中，所述获取模块，包括：

获取子模块，配置为获取所述第一参与方持有数据的标识；

第七确定子模块，配置为基于加密计算，利用所述第一参与方持有数据的标识和所述其他参与方持有数据的标识，确定所述第一参与方和所述其他参与方的共同训练样本。

这里需要指出的是：以上数据处理装置实施例的描述，与上述方法描述是类似的，具有同方法实施例相同的有益效果。对于本申请数据处理装置实施例中未披露的技术细节，本领域的技术人员请参照本申请方法实施例的描述而理解。

本申请实施例提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行本申请实施例上述的数据处理方法。

本申请实施例提供一种存储有可执行指令的存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本申请实施例提供的方法，例如，如图3至图6示出的方法。

在一些实施例中，存储介质可以是 FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或 CD-ROM 等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言（包括编译或解释语言，或者声明性或过程性语言）来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

作为示例，可执行指令可以但不一定对应于文件系统中的文件，可以可被存储在保存其它程序或数据的文件的一部分，例如，存储在超文本标记语言（HTML，Hyper Text Markup Language）文档中的一个或多个脚本中，存储在专用于所讨论的程序的单个文件中，或者，存储在多个协同文件（例如，存储一个或多个模块、子程序或代码部分的文件）中。

作为示例，可执行指令可被部署为在一个计算设备上执行，或者在位于一个地点的多个计算设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个计算设备上执行。

以上所述，仅为本申请的实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

工业实用性

本申请实施例提供一种数据处理方法、装置、设备、存储介质及程序产品，所述方法包括：基于第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵；基于特征相关矩阵，确定第一样本特征数据对应的各特征的共线性量化因子；基于共线性量化因子，从第一样本特征数据对应的各特征中确定目标特征；从第一样本特征数据中删除目标特征的特征数据，得到用于进行联合训练的第一训练数据；其中，目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系。如此，在保护数

据隐私的前提下，实现对各参与方持有的特征数据中存在共线性的数据进行筛选与剔除，能够提高联合训练的到的联邦模型的准确率和稳定性，提高联邦模型的建模效果。

权利要求书

1、一种数据处理方法，应用于联邦学习的第一参与方，所述方法包括：

基于所述第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵，所述安全计算模型由所述第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到；

基于所述特征相关矩阵，确定所述第一样本特征数据对应的各特征的共线性量化因子；

基于所述共线性量化因子，从所述第一样本特征数据对应的各特征中确定目标特征；

从所述第一样本特征数据中删除所述目标特征的特征数据，得到所述第一参与方与所述其他参与方进行联合训练的第一训练数据；

其中，所述目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系，所述其他特征包括所述第一参与方持有的除所述目标特征以外的特征和所述其他参与方持有的特征。

2、根据权利要求 1 所述的方法，其中，所述基于所述第一参与方持有的第一样本特征数据和预先训练好的安全计算模型，构建虚拟的特征相关矩阵，包括：

基于所述第一样本特征数据，确定第一样本特征数据对应的各特征的特征数据和所述第一样本特征数据对应的样本个数；

分别计算所述各特征的特征数据对应的均值和标准差；

基于所述各特征的特征数据、所述各特征的特征数据对应的均值、所述各特征的特征数据对应的标准差和所述样本个数，确定处理后的第一样本特征数据；

将所述处理后的第一样本特征数据输入至所述安全计算模型，得到第一矩阵；

根据所述处理后的第一样本特征数据和所述第一矩阵，构建虚拟的特征相关矩阵。

3、根据权利要求 2 所述的方法，其中，所述根据所述处理后的第一样本特征数据和所述第一矩阵，构建虚拟的特征相关矩阵，包括：

根据所述处理后的第一样本特征数据确定第一对称矩阵；

生成一个行数和列数均等于所述第一矩阵的列数的空矩阵；

根据所述第一对称矩阵、所述第一矩阵、所述第一矩阵的转置矩阵和所述空矩阵，构建虚拟的特征相关矩阵。

4、根据权利要求 1 至 3 任一项所述的方法，其中，所述基于所述特征相关矩阵，确定所述第一样本特征数据对应的各特征的共线性量化因子，包括：

确定所述特征相关矩阵的行列式；

删除所述特征相关矩阵第 i 行、第 i 列数据，得到所述特征相关矩阵对应的各余子式，其中， $i=1, 2, \dots, m_1$ ， m_1 为所述第一样本特征数据对应的特征个数；

确定所述各余子式的行列式；

基于所述特征相关矩阵的行列式和所述各余子式的行列式，确定所述第一样本特征数据对应的各特征的共线性量化因子。

5、根据权利要求 4 所述的方法，其中，所述确定所述特征相关矩阵的行列式，包

括:

生成一个行列式为预设值、维度与所述空矩阵相同的第一随机矩阵;

将所述特征相关矩阵和所述第一随机矩阵输入至所述安全计算模型, 得到第二矩阵;

5 分别计算所述第一对称矩阵和所述第二矩阵的行列式;

将所述第一对称矩阵的行列式和所述第二矩阵的行列式相乘, 得到所述特征相关矩阵的行列式。

6、根据权利要求 4 所述的方法, 其中, 所述基于所述特征相关矩阵的行列式和所述各余子式的行列式, 确定所述第一样本特征数据对应的各特征的共线性量化因子, 包

10 括:

根据所述各余子式的行列式与所述特征相关矩阵的行列式, 计算各特征的方差膨胀因子;

将所述各特征的方差膨胀因子确定为所述第一样本特征数据对应的各特征的共线性量化因子。

15 7、根据权利要求 1 所述的方法, 其中, 所述基于所述共线性量化因子, 从所述第一样本特征数据对应的各特征中确定目标特征, 包括:

判断所述第一样本特征数据对应的各特征的共线性量化因子是否大于预设边界值;

将共线性量化因子大于预设边界值的特征确定为目标特征。

8、根据权利要求 1 所述的方法, 其中, 所述方法还包括:

20 获取所述第一参与方和所述其他参与方持有的共同训练样本;

从所述共同训练样本中筛选出本次训练样本;

将所述本次训练样本的特征数据确定为所述第一样本特征数据。

9、根据权利要求 8 所述的方法, 其中, 所述获取所述第一参与方和所述其他参与方持有的共同训练样本, 包括:

25 获取所述第一参与方持有数据的标识;

基于加密计算, 利用所述第一参与方持有数据的标识和所述其他参与方持有数据的标识, 确定所述第一参与方和所述其他参与方的共同训练样本。

10、一种数据处理装置, 应用于联邦学习的第一参与方, 所述装置包括:

30 构建模块, 配置为基于所述第一参与方持有的第一样本特征数据和预先训练好的安全计算模型, 构建虚拟的特征相关矩阵, 所述安全计算模型由所述第一参与方和联邦学习的其他参与方基于安全多方计算预先训练得到;

第一确定模块, 配置为基于所述特征相关矩阵, 确定所述第一样本特征数据对应的各特征的共线性量化因子;

35 第二确定模块, 配置为基于所述共线性量化因子, 从所述第一样本特征数据对应的各特征中确定目标特征;

删除模块, 配置为从所述第一样本特征数据中删除所述目标特征的特征数据, 得到所述第一参与方与所述其他参与方进行联合训练的第一训练数据;

其中, 所述目标特征的特征数据与其他特征中至少一个特征的特征数据存在线性关系, 所述其他特征包括所述第一参与方持有的除所述目标特征以外的特征和所述其他参

与方持有的特征。

11、根据权利要求 10 所述的装置，其中，所述构建模块，包括：

第一确定子模块，配置为基于所述第一样本特征数据，确定所述第一样本特征数据对应的各特征的特征数据和所述第一样本特征数据对应的样本个数；

5 计算子模块，配置为分别计算所述各特征的特征数据对应的均值和标准差；

第二确定子模块，配置为基于所述各特征的特征数据、所述各特征的特征数据对应的均值、所述各特征的特征数据对应的标准差和所述样本个数，确定处理后的第一样本特征数据；

10 输入子模块，配置为将所述处理后的第一样本特征数据输入至所述安全计算模型，得到第一矩阵；

构建子模块，配置为根据所述处理后的第一样本特征数据和所述第一矩阵，构建虚拟的特征相关矩阵。

12、根据权利要求 11 所述的装置，其中，所述构建子模块，包括：

第一确定单元，配置为根据所述处理后的第一样本特征数据确定第一对称矩阵；

15 第一生成单元，配置为生成一个行数和列数均等于所述第一矩阵的列数的空矩阵；

构建单元，配置为根据所述第一对称矩阵、所述第一矩阵、所述第一矩阵的转置矩阵和所述空矩阵，构建虚拟的特征相关矩阵。

13、根据权利要求 10 至 12 任一项所述的装置，其中，所述第一确定模块，包括：

第三确定子模块，配置为确定所述特征相关矩阵的行列式；

20 删除子模块，配置为删除所述特征相关矩阵第 i 行、第 i 列数据，得到所述特征相关矩阵对应的各余子式，其中， $i=1, 2, \dots, m_1$ ， m_1 为所述第一样本特征数据对应的特征个数；

第四确定子模块，配置为确定所述各余子式的行列式；

25 第五确定子模块，配置为基于所述特征相关矩阵的行列式和所述各余子式的行列式，确定所述第一样本特征数据对应的各特征的共线性量化因子。

14、根据权利要求 13 所述的装置，其中，所述第三确定子模块，包括：

第二生成单元，配置为生成一个行列式为预设值、维度与所述空矩阵相同的第一随机矩阵；

30 输入单元，配置为将所述特征相关矩阵和所述第一随机矩阵输入至所述安全计算模型，得到第二矩阵；

第一计算单元，配置为分别计算所述第一对称矩阵和所述第二矩阵的行列式；

第二计算单元，配置为将所述第一对称矩阵的行列式和所述第二矩阵的行列式相乘，得到所述特征相关矩阵的行列式。

15、根据权利要求 13 所述的装置，其中，所述第五确定子模块，包括：

35 第三计算单元，配置为根据所述各余子式的行列式与所述特征相关矩阵的行列式，计算各特征的方差膨胀因子；

第二确定单元，配置为将所述各特征的方差膨胀因子确定为所述第一样本特征数据对应的各特征的共线性量化因子。

16、根据权利要求 10 所述的装置，其中，所述第二确定模块，包括：

判断子模块，配置为判断所述第一样本特征数据对应的各特征的共线性量化因子是否大于预设边界值；

第六确定子模块，配置为将共线性量化因子大于预设边界值的特征确定为目标特征。

5 17、根据权利要求 10 所述的装置，其中，所述装置还包括：

获取模块，配置为获取所述第一参与方和所述其他参与方持有的共同训练样本；

筛选模块，配置为从所述共同训练样本中筛选出本次训练样本；

第三确定模块，配置为将所述本次训练样本的特征数据确定为所述第一样本特征数据。

10 18、一种数据处理设备，所述设备包括：

存储器，配置为存储可执行指令；

处理器，配置为执行所述存储器中存储的可执行指令时，实现权利要求 1 至 9 任一项所述的数据处理方法。

15 19、一种计算机可读存储介质，所述计算机可读存储介质上存储有可执行指令，配置为引起处理器执行时，实现权利要求 1 至 9 任一项所述的数据处理方法。

20、一种计算机程序产品，包括计算机程序，所述计算机程序被处理器执行时实现权利要求 1 至 9 任一项所述的数据处理方法。

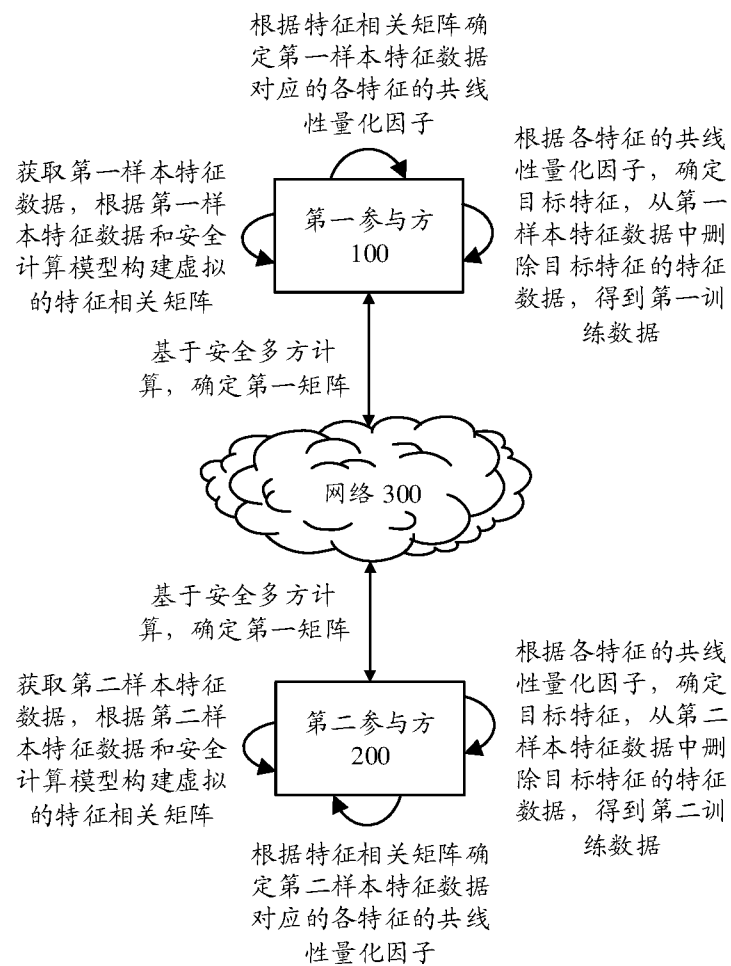


图 1

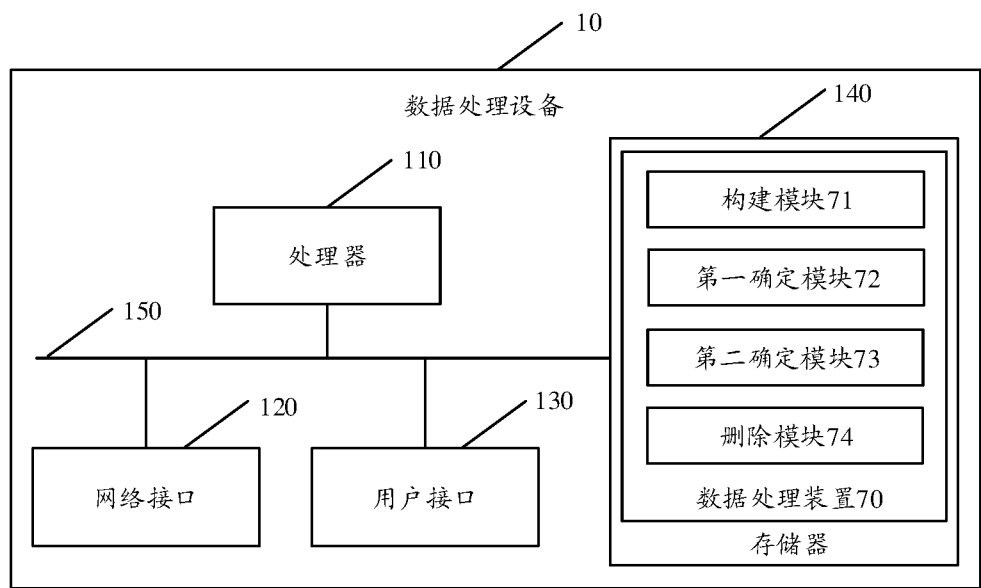


图 2

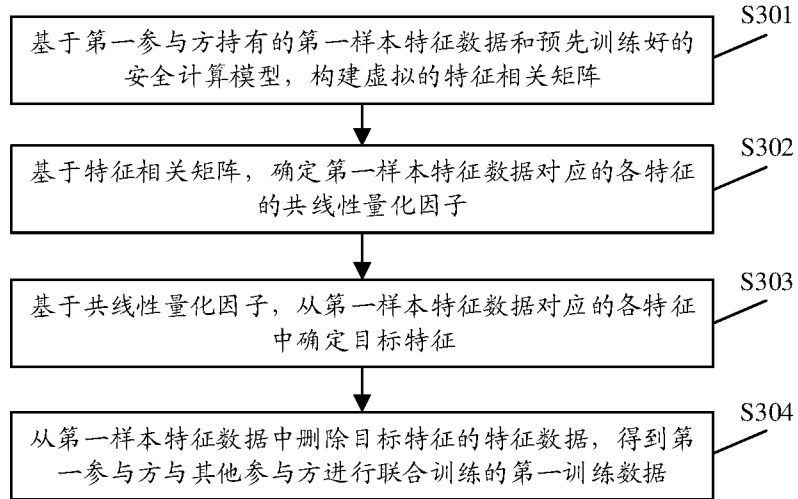


图 3

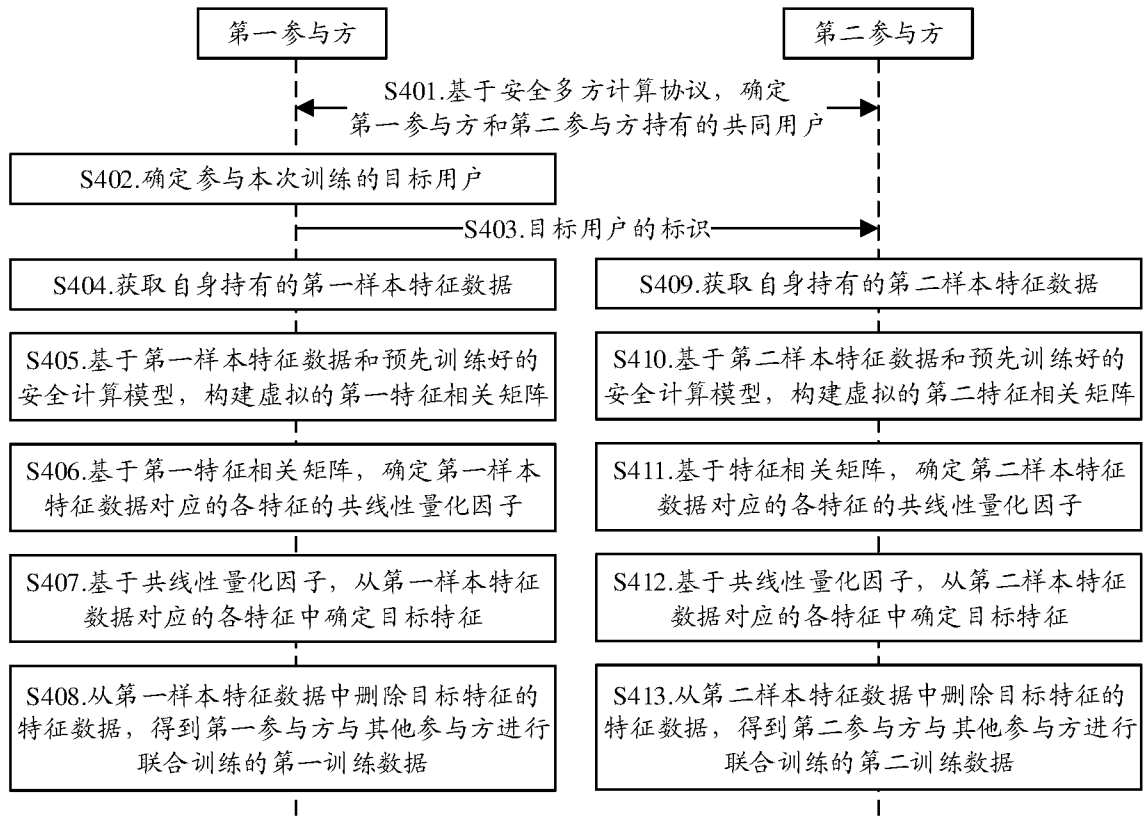


图 4

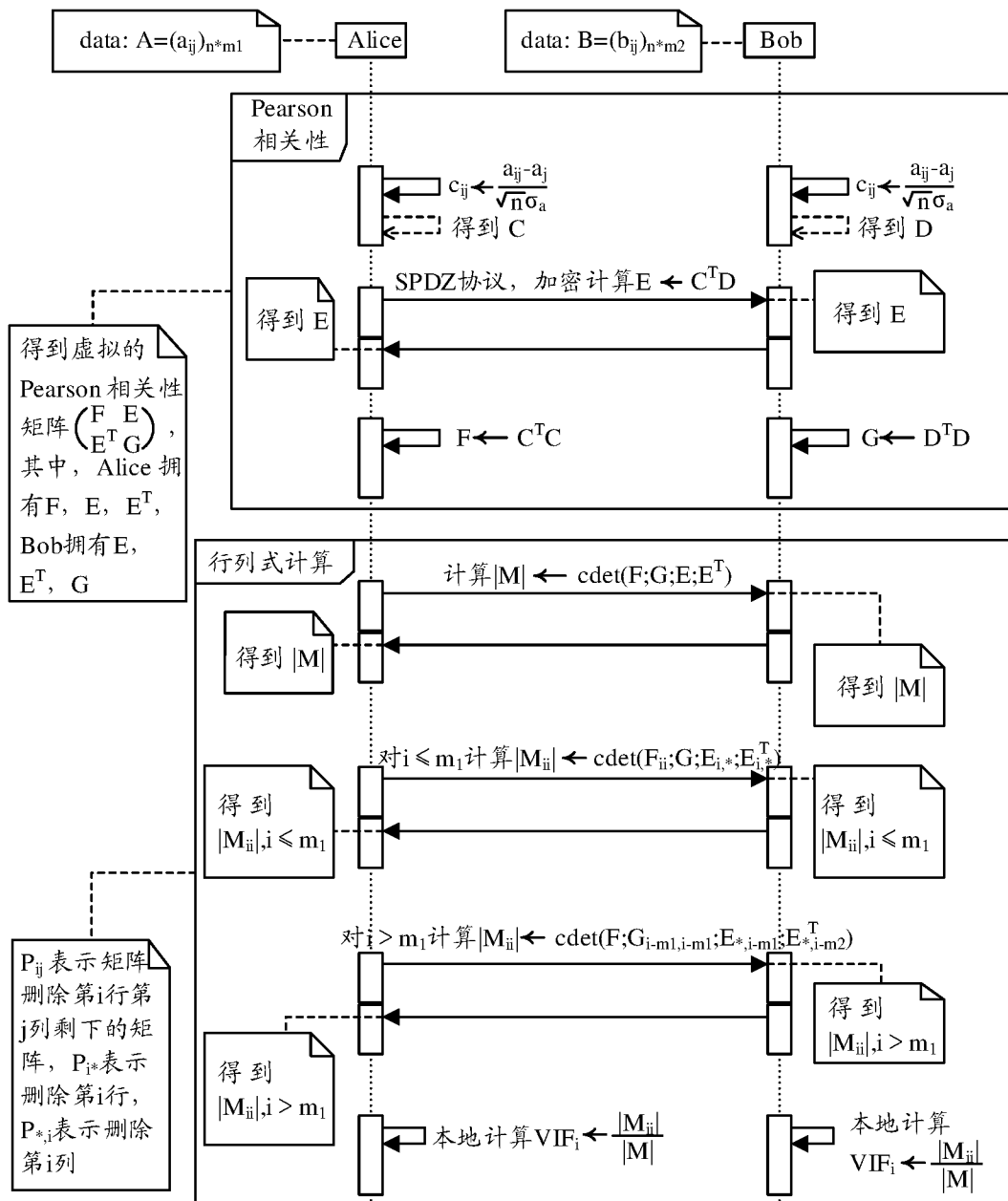


图 5

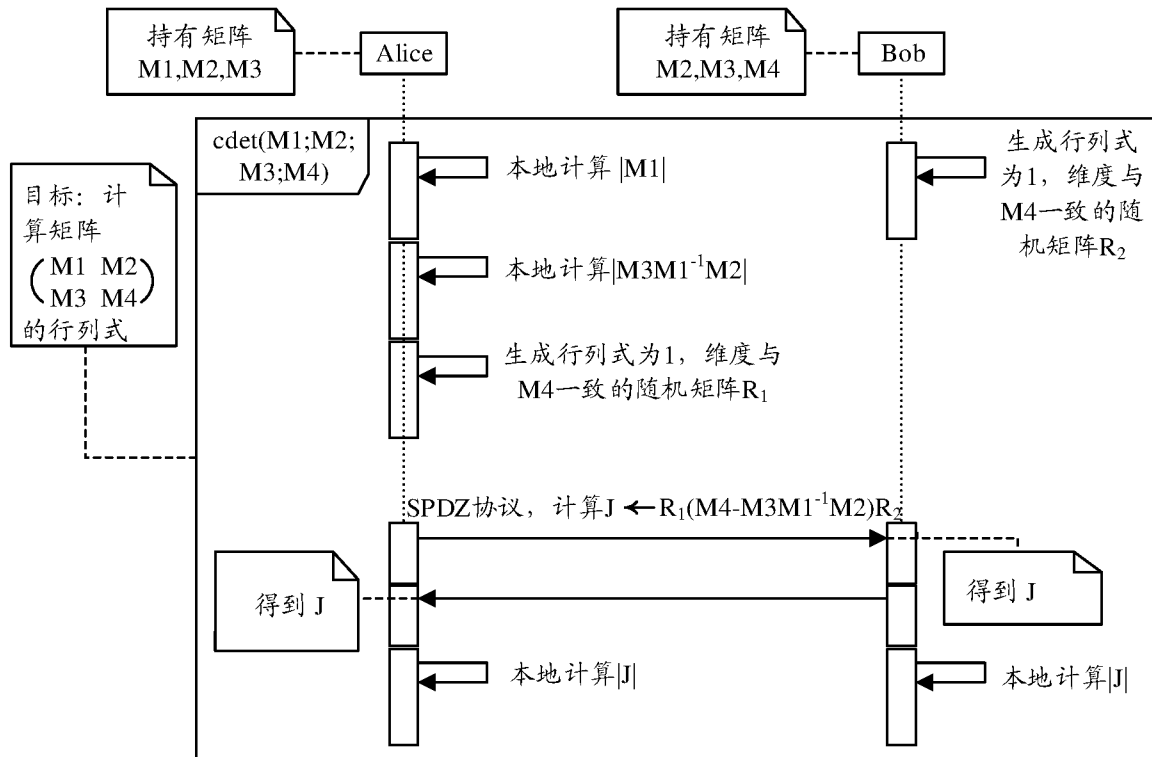


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/140955

A. CLASSIFICATION OF SUBJECT MATTER

G06N 20/20(2019.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N

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

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

CNABS, CNTXT, DWPI, SIPOABS, USTXT, EPTXT, WOTXT, CNKI, 万方, WANFANG, IEEE: 联邦, 联合, 学习, 训练, 模型, 参与, 矩阵, 计算, 共线性, 线性相关, 线性关系, 方差膨胀, 因子, 系数, 样本, 目标, 删除, 均值, 标准差, 行列式, 余子式, 加密, 秘密, 隐私, 安全, federated, learning, train, model, participate, matrix, compute, calculate, collinearity, linearly dependent, linear relationship, variance inflation, factor, VIF, coefficient, modulus, sample, target, object, delete, mean, standard deviation, determinant, cofactor, encrypt, secret, privacy, secrecy, safe

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113095514 A (SHENZHEN QIANHAI WEBANK CO., LTD.) 09 July 2021 (2021-07-09) claims 1-10, description paragraphs 1-240, and figures 1-6	1-20
X	CN 112597540 A (ALIPAY (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 02 April 2021 (2021-04-02) description, paragraphs 41-138, and figures 1-13	1-20
A	CN 110909216 A (ZHIFUBAO (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 24 March 2020 (2020-03-24) entire document	1-20
A	CN 111062487 A (ZHIFUBAO (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 24 April 2020 (2020-04-24) entire document	1-20
A	CN 111966473 A (ZHIFUBAO (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 20 November 2020 (2020-11-20) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance
 “E” earlier application or patent but published on or after the international filing date
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 “O” document referring to an oral disclosure, use, exhibition or other means
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
 “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
 “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
 “&” document member of the same patent family

Date of the actual completion of the international search

15 February 2022

Date of mailing of the international search report

04 March 2022

Name and mailing address of the ISA/CN

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

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/140955

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 111654853 A (SUOXINDA BEIJING DATA TECHNOLOGY CO., LTD. et al.) 11 September 2020 (2020-09-11) entire document	1-20
A	CN 111160573 A (ZHIFUBAO (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 15 May 2020 (2020-05-15) entire document	1-20
A	US 2016283735 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 29 September 2016 (2016-09-29) entire document	1-20
A	US 2020285984 A1 (HCL TECHNOLOGIES LIMITED) 10 September 2020 (2020-09-10) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/140955

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113095514	A	09 July 2021	None			
CN	112597540	A	02 April 2021	CN	112597540	B	01 October 2021
CN	110909216	A	24 March 2020	None			
CN	111062487	A	24 April 2020	CN	111062487	B	20 April 2021
CN	111966473	A	20 November 2020	None			
CN	111654853	A	11 September 2020	CN	111654853	B	10 November 2020
CN	111160573	A	15 May 2020	WO	2021197035	A1	07 October 2021
				CN	111160573	B	30 June 2020
US	2016283735	A1	29 September 2016	US	10395180	B2	27 August 2019
				US	2016283738	A1	29 September 2016
				US	10395182	B2	27 August 2019
US	2020285984	A1	10 September 2020	IN	201911008643	A	29 March 2019

国际检索报告

国际申请号

PCT/CN2021/140955

A. 主题的分类

G06N 20/20 (2019.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNABS, CNTXT, DWPI, SIPOABS, USTXT, EPTXT, WOTXT, CNKI, 万方, IEEE: 联邦, 联合, 学习, 训练, 模型, 参与, 矩阵, 计算, 共线性, 线性相关, 线性关系, 方差膨胀, 因子, 系数, 样本, 目标, 删除, 均值, 标准差, 行列式, 余子式, 加密, 秘密, 隐私, 安全, federated, learning, train, model, participate, matrix, compute, calculate, collinearity, linearly dependent, linear relationship, variance inflation, factor, VIF, coefficient, modulus, sample, target, object, delete, mean, standard deviation, determinant, cofactor, encrypt, secret, privacy, secrecy, safe

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 113095514 A (深圳前海微众银行股份有限公司) 2021年7月9日 (2021 - 07 - 09) 权利要求1-10, 说明书第1-240段, 附图1-6	1-20
X	CN 112597540 A (支付宝杭州信息技术有限公司) 2021年4月2日 (2021 - 04 - 02) 说明书第41-138段, 附图1-13	1-20
A	CN 110909216 A (支付宝杭州信息技术有限公司) 2020年3月24日 (2020 - 03 - 24) 全文	1-20
A	CN 111062487 A (支付宝杭州信息技术有限公司) 2020年4月24日 (2020 - 04 - 24) 全文	1-20
A	CN 111966473 A (支付宝杭州信息技术有限公司) 2020年11月20日 (2020 - 11 - 20) 全文	1-20
A	CN 111654853 A (索信达北京数据技术有限公司等) 2020年9月11日 (2020 - 09 - 11) 全文	1-20
A	CN 111160573 A (支付宝杭州信息技术有限公司) 2020年5月15日 (2020 - 05 - 15) 全文	1-20

☒ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年2月15日

国际检索报告邮寄日期

2022年3月4日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

戴自立

电话号码 (86-27)59371936

C. 相关文件		
类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	US 2016283735 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2016年9月29日 (2016 - 09 - 29) 全文	1-20
A	US 2020285984 A1 (HCL TECHNOLOGIES LIMITED) 2020年9月10日 (2020 - 09 - 10) 全文	1-20

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/140955

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	113095514	A	2021年7月9日	无	
CN	112597540	A	2021年4月2日	CN 112597540	B 2021年10月1日
CN	110909216	A	2020年3月24日	无	
CN	111062487	A	2020年4月24日	CN 111062487	B 2021年4月20日
CN	111966473	A	2020年11月20日	无	
CN	111654853	A	2020年9月11日	CN 111654853	B 2020年11月10日
CN	111160573	A	2020年5月15日	WO 2021197035	A1 2021年10月7日
				CN 111160573	B 2020年6月30日
US	2016283735	A1	2016年9月29日	US 10395180	B2 2019年8月27日
				US 2016283738	A1 2016年9月29日
				US 10395182	B2 2019年8月27日
US	2020285984	A1	2020年9月10日	IN 201911008643	A 2019年3月29日