(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116595562 B
(45)授权公告日 2024.07.19

(21)申请号 202310667243.X

(22)申请日 2023.06.06

(65)同一申请的已公布的文献号
申请公布号 CN 116595562 A

(43)申请公布日 2023.08.15

(73)专利权人 北京火山引擎科技有限公司
地址 100190 北京市海淀区紫金数码园4号楼13层1309

(72)发明人 林宇 蔡权伟 吴烨

(74)专利代理机构 北京世辉律师事务所 16093
专利代理师 罗利娜

(51)Int.Cl.
G06F 21/60(2013.01)
G06F 21/64(2013.01)

(56)对比文件
US 2022094519 A1,2022.03.24
CN 109525386 A,2019.03.26

审查员 张思洋

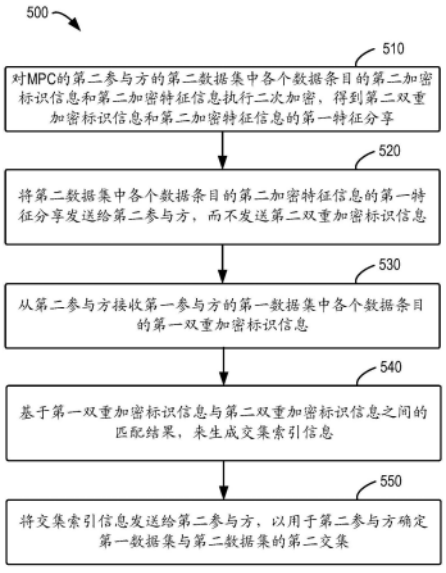
权利要求书4页 说明书24页 附图8页

(54)发明名称

数据处理方法和电子设备

(57)摘要

本公开的实施例提供了数据处理方法和电子设备。一种方法包括:在多方安全计算MPC的第一参与方,对MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密,得到第二双重加密标识信息和第二加密特征信息的第一特征分享;将第二数据集中各个数据条目的第二加密特征信息的第一特征分享发送给第二参与方,而不发送第二双重加密标识信息;从第二参与方接收第一双重加密标识信息;基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果,来生成交集索引信息,以用于第二参与方确定第一数据集与第二数据集的第二交集。



CN 116595562 B

1.一种数据处理方法,所述方法被实现在多方安全计算MPC的第一参与方,所述方法包括:

对所述MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息$Pid'_{i,j}$和第二加密特征信息$\tilde{V}_{i',j}$执行二次加密,得到第二双重加密标识信息$\widetilde{Pid}_{i,j}$和所述第二加密特征信息的第一特征分享$\left[\widetilde{V_{i',j}}\right]_0$;

将所述第二数据集中各个数据条目的所述第二加密特征信息的第一特征分享$\left[\widetilde{V_{i',j}}\right]_0$发送给所述第二参与方,而不发送所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$;

从所述第二参与方接收所述第一参与方的第一数据集中各个数据条目的第一双重加密标识信息$\widetilde{Cid}_{i,j}$;

基于所述第一双重加密标识信息$\widetilde{Cid}_{i,j}$与所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$之间的匹配结果,来生成交集索引信息,所述交集索引信息包括对所述第一数据集和所述第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,所述真索引对应的数据条目的标识信息相匹配,所述伪索引对应的数据条目的标识信息不匹配;以及

将所述交集索引信息发送给所述第二参与方,以用于所述第二参与方确定所述第一数据集与所述第二数据集的第二交集。

2.根据权利要求1所述的方法,其中对所述第二加密标识信息$Pid'_{i,j}$执行二次加密包括:

利用第一加密密钥r_c对所述第二加密标识信息$Pid'_{i,j}$执行二次加密,得到所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$,

其中所述第一加密密钥r_c还由所述第一参与方用于对所述第一数据集中各个数据条目的第一标识信息$Cid_{i,j}$执行一次加密,得到第一加密标识信息$Cid'_{i,j}$,并且

其中所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$的一次加密和所述第一加密标识信息$Cid'_{i,j}$的二次加密由所述第二参与方利用第二加密密钥r_p来执行。

3. 根据权利要求2所述的方法,其中对所述第一数据集中各个数据条目的第一特征信息$u_{i,j}$进行加密包括:

将所述第一数据集中各个数据条目的第一特征信息$u_{i,j}$按顺序划分为至少一个第一特征信息块$U_{i',j}$,每个第一特征信息块包括所述第一数据集中预定数目的数据条目的第一特征信息的顺序拼接,在每个第一特征信息块中的两个相邻数据条目之间填充有预定信息;以及

对所述至少一个第一特征信息块$U_{i',j}$进行加密,得到所述至少一个第一特征信息块$U_{i',j}$各自的第一加密特征信息$\tilde{U}_{i',j}$。

4.根据权利要求3所述的方法,其中所述预定信息为零值,和/或

其中每个第一特征信息块中的所述预定数目的数据条目的第一加密标识信息$Cid'_{i,j}$用于索引该第一特征信息块$U_{i',j}$。

5.根据权利要求1所述的方法,其中所述第二数据集中各个数据条目的所述第二加密特征信息$\tilde{V}_{i',j}$包括从所述第二数据集划分的至少一个第二特征信息块$V_{i',j}$的第二加密特征信息$\tilde{V}_{i',j}$,每个第二特征信息块$V_{i',j}$是通过将所述第二数据集中各个数据条目的第二特征信

息按顺序划分得到的,每个第二特征信息块$V_{i',j}$包括所述第二数据集中预定数目的数据条目的第二特征信息的顺序拼接,在每个第二特征信息块中的两个相邻数据条目之间填充有预定信息,并且

其中对所述第二加密特征信息$\tilde{V}_{i',j}$执行二次加密包括:

生成所述第二数据集中各个数据条目对应的第二特征分享$\gamma_{i,j}$;

将所述第二数据集中各个数据条目对应的第二特征分享按顺序划分,得到所述第二加密特征信息$\tilde{V}_{i',j}$的至少一个特征分享块$\left[V_{i',j}\right]_1$,每个特征分享块包括所述第二数据集中预定数目的数据条目对应的第二特征分享的顺序拼接,在每个特征分享块中的两个相邻第二特征分享之间填充有所述预定信息;以及

基于所述至少一个特征分享块$\left[V_{i',j}\right]_1$来对所述第二加密特征信息$\tilde{V}_{i',j}$执行同态加法处理,得到所述第二加密特征信息$\tilde{V}_{i',j}$的所述第一特征分享$\left[\widetilde{V_{i',j}}\right]_0$。

6.根据权利要求1所述的方法,其中所述第一数据集中各个数据条目的第一加密特征信息的第一特征分享$\left[\widetilde{U_{i',j}}\right]_0$与所述第一双重加密标识信息$\widetilde{Cid}_{i,j}$一起从所述第二参与方被接收到,所述方法还包括:

缓存所述第二数据集的所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$和所述第二加密特征信息的第二特征分享$\left[\widetilde{V_{i',j}}\right]_0$,所述第二加密特征信息被划分为所述第一特征分享$\left[\widetilde{V_{i',j}}\right]_0$和所述第二特征分享$\left[\widetilde{V_{i',j}}\right]_0$;

解密所述第一加密特征信息的所述第一特征分享$\left[\widetilde{U_{i',j}}\right]_0$,得到第一解密特征信息的第一特征分享$\left[U_{i',j}\right]_0$;以及

缓存所述第一数据集的所述第一双重加密标识信息$\widetilde{Cid}_{i,j}$和所述第一解密特征信息的第一特征分享$\left[U_{i',j}\right]_0$。

7.根据权利要求1所述的方法,其中所述第一双重加密标识信息$\widetilde{Cid}_{i,j}$包括多个类型分别对应的多个第一双重加密标识符,所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$包括所述多个类型分别对应的多个第二双重加密标识符,并且其中生成所述交集索引信息包括:

基于所述多个类型的优先级来确定所述匹配结果,所述匹配结果的所述确定包括:

通过比较所述第一双重加密标识信息$\widetilde{Cid}_{i,j}$中第一类型对应的第一双重加密标识符与在所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$中所述第一类型对应的第二双重加密标识符,来确定第一匹配结果,

如果所述第一匹配结果指示所述第一数据集和所述第二数据集中标识信息相匹配的至少一对数据条目,从所述第一双重加密标识信息$\widetilde{Cid}_{i,j}$和所述第二双重加密标识信息$\widetilde{Pid}_{i,j}$分别过滤掉相匹配的所述至少一对数据条目的双重加密标识信息,得到过滤后的第一双重加密标识信息和过滤后的第二双重加密标识信息;以及

通过比较过滤后的所述第一双重加密标识信息中第二类型对应的第一双重加密标识

符与在过滤后的所述第二双重加密标识信息中所述第二类型对应的第二双重加密标识符来确定第二匹配结果,所述第二类型的优先级低于所述第一类型的优先级。

8.根据权利要求1所述的方法,还包括:

在所述第一参与方处,基于所述交集索引信息来生成所述第一数据集与所述第二数据集的第一交集,所述第一交集包括所述交集索引信息中的所述真索引对应的至少一对数据条目和所述伪索引对应的至少一对数据条目;

设置针对所述第一交集中每对数据条目的匹配标记,所述真索引对应的至少一对数据条目的匹配标记为指示匹配,所述伪索引对应的至少一对数据条目的匹配标记为指示不相匹配;

与所述第二参与方一起,利用所述第一交集和所述第二交集来执行所述MPC,得到所述第一交集中每对数据条目的候选计算结果;以及

至少基于所述候选计算结果和针对所述第一交集中每对数据条目的匹配标记来确定所述MPC的目标计算结果。

9.根据权利要求8所述的方法,其中所述真索引对应的一对数据条目的匹配标志位被设置为1,所述伪索引对应的一对数据条目的匹配标志位被设置为0,并且其中确定所述目标计算结果包括:

基于所述第一交集中每对数据条目的候选计算结果与所述第一交集中每对数据条目的匹配标记的相乘运算,生成所述目标计算结果。

10.根据权利要求8所述的方法,其中针对所述第二交集中每对数据条目的匹配标记被设置为指示不匹配,并且所述目标计算结果的确定还基于针对所述第二交集中每对数据条目的匹配标记。

11.一种数据处理方法,所述方法被实现在多方安全计算MPC的第二参与方,所述方法包括:

对从所述MPC的第一参与方接收到的第一数据集中各个数据条目的第一加密标识信息$Cid'_{i,j}$和第一加密特征信息$\tilde{U}_{i',j}$执行二次加密,得到第一双重加密标识信息$\widetilde{Cid}_{i,j}$和所述第一加密特征信息的第一特征分享$\left[\widetilde{U_{i',j}}\right]_0$;

向所述第一参与方至少发送所述第一数据集中各个数据条目的第一双重加密标识信息$\widetilde{Cid}_{i,j}$;

从所述第一参与方接收针对所述第二参与方的第二数据集中各个数据条目的第二加密特征信息的第一特征分享$\left[\widetilde{V_{i',j}}\right]_0$,而未接收到所述第二数据集中各个数据条目的第二双重加密标识信息$\widetilde{Pid}_{i,j}$;

从所述第一参与方接收交集索引信息,所述交集索引信息包括对所述第一数据集和所述第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,所述真索引对应的所述至少一对数据条目的标识信息相匹配;以及

基于所述交集索引信息来确定所述第一数据集与所述第二数据集中的第二交集,所述第二交集包括所述交集索引信息中的所述真索引对应的至少一对数据条目和所述伪索引对应的至少一对数据条目。

12.根据权利要求11所述的方法,还包括:

设置针对所述第二交集中每对数据条目的匹配标记,以指示该对数据条目的标识信息不相匹配;

与所述第一参与方一起,利用所述第二交集和所述第一参与方确定的第一交集来执行所述MPC,得到所述第二交集中每对数据条目的候选计算结果;以及

至少基于所述候选计算结果和针对所述第二交集中每对数据条目的匹配标记来确定所述MPC的目标计算结果。

13.根据权利要求12所述的方法,其中所述第二交集中每对数据条目的匹配标记被设置为0,并且

其中所述第一数据集中所述真索引对应的一对数据条目的匹配标志位被设置为1,所述伪索引对应的一对数据条目的匹配标志位被设置为0。

14.根据权利要求13所述的方法,其中确定所述目标计算结果包括:

基于所述第二交集中每对数据条目的候选计算结果与所述第二交集中每对数据条目的匹配标记的相乘运算,生成所述目标计算结果。

15.一种电子设备,包括:

至少一个处理单元;以及

至少一个存储器,所述至少一个存储器被耦合到所述至少一个处理单元并且存储用于由所述至少一个处理单元执行的指令,所述指令在由所述至少一个处理单元执行时使所述电子设备执行根据权利要求1至10任一项所述的方法或根据权利要求11至14中任一项所述的方法。

数据处理方法和电子设备

技术领域

[0001] 本公开的示例实施例总体涉及计算机领域,特别地涉及数据处理方法、装置、设备和计算机可读存储介质。

背景技术

[0002] 近年来,由于用户隐私、数据安全、合法合规、商业竞争等因素,很难合法和合规的将分散的数据源整合到一起进行计算、分析和学习。在这样的背景下,基于多方安全计算(Secure Multi-party Computation,MPC)的解决方案迅速发展起来,在不需要将分散数据源集中在一起的情况下就可以联合多个分散的数据源进行联合计算、联合数据分析和联合机器学习。MPC旨在解决一组互不信任的参与方在保护数据安全的情况下执行协同计算的问题,为数据需求方提供不泄露原始数据前提下的多方协同计算能力。多方安全计算可以用于支持安全的数据合作和融合应用,在数据不出域、合法合规的前提下联合多方数据源进行计算和分析。

发明内容

[0003] 在本公开的第一方面,提供了一种数据处理方法。该方法被实现在多方安全计算MPC的第一参与方,该方法包括:对MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密,得到第二双重加密标识信息和第二加密特征信息的第一特征分享;将第二数据集中各个数据条目的第二加密特征信息的第一特征分享发送给第二参与方,而不发送第二双重加密标识信息;从第二参与方接收第一参与方的第一数据集中各个数据条目的第一双重加密标识信息;基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果,来生成交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的数据条目的标识信息相匹配,伪索引对应的数据条目的标识信息不匹配;以及将交集索引信息发送给第二参与方,以用于第二参与方确定第一数据集与第二数据集的第二交集。

[0004] 在本公开的第二方面,提供了一种数据处理方法。该方法被实现在多方安全计算MPC的第二参与方,该方法包括:对从MPC的第一参与方接收到的第一数据集中各个数据条目的第一加密标识信息和第一加密特征信息执行二次加密,得到第一双重加密标识信息和第一加密特征信息的第一特征分享;向第一参与方至少发送第一数据集中各个数据条目的第一双重加密标识信息;从第一参与方接收针对第二参与方的第二数据集中各个数据条目的第二加密特征信息的第一特征分享,而未接收到第二数据集中各个数据条目的第二双重加密标识信息;从第一参与方接收交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的至少一对数据条目的标识信息相匹配;以及基于交集索引信息来确定第一数据集与第二数据集中的第二交集,第二交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引

对应的至少一对数据条目。

[0005] 在本公开的第三方面,提供了一种数据处理装置。该装置被实现在多方安全计算MPC的第一参与方,该装置包括:二次加密模块,被配置为对MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密,得到第二双重加密标识信息和第二加密特征信息的第一特征分享。该装置还包括第一发送模块,被配置为将第二数据集中各个数据条目的第二加密特征信息的第一特征分享发送给第二参与方,而不发送第二双重加密标识信息。该装置还包括第一接收模块,被配置为从第二参与方接收第一参与方的第一数据集中各个数据条目的第一双重加密标识信息。该装置还包括交集索引确定模块,被配置为基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果,来生成交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的数据条目的标识信息相匹配,伪索引对应的数据条目的标识信息不匹配。该装置还包括第二发送模块,被配置为将交集索引信息发送给第二参与方,以用于第二参与方确定第一数据集与第二数据集的第二交集。

[0006] 在本公开的第四方面,提供了一种数据处理装置。该装置被实现在多方安全计算MPC的第二参与方,该装置包括:二次加密模块,被配置为对从MPC的第一参与方接收到的第一数据集中各个数据条目的第一加密标识信息和第一加密特征信息执行二次加密,得到第一双重加密标识信息和第一加密特征信息的第一特征分享。该装置还包括第一发送,被配置为向第一参与方至少发送第一数据集中各个数据条目的第一双重加密标识信息。该装置还包括第一接收模块,被配置为从第一参与方接收针对第二参与方的第二数据集中各个数据条目的第二加密特征信息的第一特征分享,而未接收到第二数据集中各个数据条目的第二双重加密标识信息。该装置还包括第二接收模块,被配置为从第一参与方接收交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的至少一对数据条目的标识信息相匹配。该装置还包括第二交集确定模块,被配置为基于交集索引信息来确定第一数据集与第二数据集中的第二交集,第二交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目。

[0007] 在本公开的第五方面,提供了一种电子设备。该设备包括至少一个处理单元;以及至少一个存储器,至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令。指令在由至少一个处理单元执行时使电子设备执行第一方面的方法。

[0008] 在本公开的第六方面,提供了一种电子设备。该设备包括至少一个处理单元;以及至少一个存储器,至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令。指令在由至少一个处理单元执行时使电子设备执行第二方面的方法。

[0009] 在本公开的第七方面,提供了一种计算机可读存储介质。该计算机可读存储介质上存储有计算机程序,计算机程序可由处理器执行以实现第一方面的方法。

[0010] 在本公开的第八方面,提供了一种计算机可读存储介质。该计算机可读存储介质上存储有计算机程序,计算机程序可由处理器执行以实现第二方面的方法。

[0011] 应当理解,本内容部分中所描述的内容并非旨在限定本公开的实施例的关键特征或重要特征,也不用于限制本公开的范围。本公开的其他特征将通过以下的描述而变得容

易理解。

附图说明

[0012] 结合附图并参考以下详细说明,本公开各实施例的上述和其他特征、优点及方面将变得更加明显。在附图中,相同或相似的附图标记表示相同或相似的元素,其中:

[0013] 图1示出了能够在其中实现本公开的实施例的示例环境的示意图;

[0014] 图2示出了根据本公开的一些实施例的用于数据处理的多方信令流的流程图;

[0015] 图3示出了根据本公开的一些实施例的用于交集匹配的示例的示意图;

[0016] 图4示出了根据本公开的一些实施例的基于示例数据集的数据处理信令流的流程图;

[0017] 图5示出了根据本公开的一些实施例的在第一参与方处实现的数据处理方法的流程图;

[0018] 图6示出了根据本公开的一些实施例的在第二参与方处实现的数据处理方法的流程图;

[0019] 图7示出了根据本公开的一些实施例的在第一参与方处实现的数据处理装置的示意性结构框图;

[0020] 图8示出了根据本公开的一些实施例的在第二参与方处实现的数据处理装置的示意性结构框图;以及

[0021] 图9示出了可以实施本公开的一个或多个实施例的电子设备的框图。

具体实施方式

[0022] 下面将参照附图更详细地描述本公开的实施例。虽然附图中示出了本公开的某些实施例,然而应当理解的是,本公开可以通过各种形式来实现,而且不应该被解释为限于这里阐述的实施例,相反,提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0023] 在本公开的实施例的描述中,术语“包括”及其类似用语应当理解为开放性包含,即“包括但不限于”。术语“基于”应当理解为“至少部分地基于”。术语“一个实施例”或“该实施例”应当理解为“至少一个实施例”。术语“一些实施例”应当理解为“至少一些实施例”。下文还可能包括其他明确的和隐含的定义。

[0024] 在本文中,除非明确说明,“响应于A”执行一个步骤并不意味着在“A”之后立即执行该步骤,而是可以包括一个或多个中间步骤。

[0025] 可以理解的是,本技术方案所涉及的数据(包括但不限于数据本身、数据的获得、使用、存储或删除)应当遵循相应法律法规及相关规定的要求。

[0026] 首先对本公开的实施例中涉及的名词进行简要介绍。

[0027] 秘密分享(secret share):通过某种运算将一个数据值拆分为多份的加密方式。例如,加法秘密分享可以将数据值x拆分为$x=x_1+x_2$两个秘密分享值。

[0028] 多方安全计算(MPC):指的是存在N个参与方P1,P2,…,PN,其中参与方Pi拥有输入数据Xi,在不向任何其他参与方泄露自己的输入数据的前提下,N个参与方共同计算一个函数f(X1,X2,…,XN)。在运算中通过应用密码学(如同态加密)、秘密分享、差分隐私等安全机

制,可以确保输入数据的安全性。例如,多个参与方输入数据的秘密分享值,可以计算指定算术运算、逻辑运算,输出的运算结果仍为秘密分享的形式。

[0029] 椭圆曲线密钥交换(Elliptic Curve Diffie-Hellman key Exchange,ECDH):两个参与方通过椭圆曲线加密算法实现密钥交换。

[0030] 同态加密(Homomorphic Encryption,HE):是实现多方安全计算的方法之一。同态加密允许对密文进行特定形式的代数运算得到仍然是密文空间里的运算结果。经过加密后的数据,可以通过同态加法、乘法运算等,在不解密数据的情况下计算得到新的密文,新密文解密后可以得到经过相应同态运算的数据。也就是说,在密文空间的运算等价于在明文空间里的运算。因此,利用同态加密技术,可以在加密的数据上进行运算,而在整个运算过程中无需对数据进行解密。

[0031] 图1示出了本公开的实施例能够在其中实现的示例环境100的示意图。环境100涉及基于MPC协议的安全计算。出于示例性目的,示出了参与方110(在本文中有时称为第一参与方、参与方C、或C方)和参与方120(也称为第二参与方、参与方P、或P方)。参与方110具有自己的数据集112,参与方120具有自己的数据集122。在MPC运算中,两个参与方期望在确保各自数据集的数据安全的情况下进行指定运算。

[0032] 数据集112和数据集122中的每个数据集可以包括一个或多个数据条目,每个数据条目包括标识信息和特征信息。每个数据条目的标识信息可以包括一个或多个标识类型对应的标识符(ID),特征信息可以包括一个或多个特征类型对应的特征。标识信息部分用于标识或区分特征信息部分。举例而言,对于记录广告投放情况的数据集,标识信息的类型可以包括广告投放平台标识和广告投放用户标识,特征信息的类型可以包括广告是否被点击、广告被观看时长、广告是否被收藏等。

[0033] 在一些实现中,数据集112和数据集122的标识信息可以包括一个或多个相同标识类型,例如都包括广告投放平台标识和广告投放用户标识。在一些实现中,数据集112和数据集122的特征信息可以包括一个或多个相同特征类型,或者可以包括完全不同的特征类型。

[0034] 在图1中,参与方110或参与方120均可以对应于任意类型的具有计算能力的一个或多个电子设备,包括终端设备或服务端设备。终端设备可以是任意类型的移动终端、固定终端或便携式终端,包括移动手机、台式计算机、膝上型计算机、笔记本计算机、上网本计算机、平板计算机、媒体计算机、多媒体平板、个人通信系统(PCS)设备、个人导航设备、个人数字助理(PDA)、音频/视频播放器、数码相机/摄像机、定位设备、电视接收器、无线电广播接收器、电子书设备、游戏设备或者前述各项的任意组合,包括这些设备的配件和外设或者其任意组合。服务端设备例如可以包括计算系统/服务器,诸如大型机、边缘计算节点、云环境中的计算设备,等等。

[0035] 应当理解,仅出于示例性的目的描述环境100的结构和功能,而不暗示对于本公开的范围的任何限制。例如,虽然图1为示出,在一些情况下,MPC运算还可以涉及更多参与方,每个参与方可以具有各自的数据集。

[0036] 在MPC运算中,有时需要确定多个参与方的数据集之间的交集匹配。举例来说,多个参与方各输入一个数据集,在不泄漏双方交集的条件下,确定多个数据集的交集。这里的交集指的是两个数据集中标识信息相匹配的数据条目。在一些实现中,通过交集匹配,可以

确定合并两个数据集中由相同标识符所索引的不同特征信息。在一些实现中,在获得匹配标识信息的情况下,还可以生成交集中的数据条目的特征信息的密码分享,用于后续的MPC运算。

[0037] 在一些交集匹配方案中,基于双重椭圆曲线密钥交换(ECDH)技术来生成双方并集的匿名标识(ID),其中双方的交集部分会被映射为相同的匿名ID。之后,双方通过匿名ID执行MPC协议完成后续计算。然而这类协议所生成的结果是双方的并集,当双方数据量不平衡时,并集的规模很大,而实际有意义的交集部分很少,导致后续MPC计算协议会产生很大的额外开销。

[0038] 另一些方案基于MPC协议,以秘密分享的形式匹配出双方的交集ID,并同时生成双方特征的秘密分享。然而,这类方案对通信条件要求比较高,且难以实现多ID的匹配。而且,当某个数据集中包含重复ID时,计算开销较大。

[0039] 当前,期望能够提供通信和计算方面高效,且能确保交集信息的安全性的交集匹配方案。

[0040] 根据本公开的示例实施例,提供了一种用于数据处理的改进方案。根据该方案,对于MPC中具有第一数据集的第一参与方和具有第二数据集的第二参与方,第一参与方获得第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息,并执行二次加密,得到第二双重加密标识信息和第二加密特征信息的特征分享。第一参与方将第二加密特征信息的特征分享发送给第二参与方(P),而不发送第二双重加密标识信息。第一参与方从第二参与方接收第一数据集中各个数据条目的第一双重加密标识信息。这样,第一参与方能够基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果,生成针对第一数据集和第二数据集的交集索引信息。

[0041] 交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的数据条目的标识信息相匹配,伪索引对应的数据条目的标识信息不匹配。通过该交集索引信息,第一参与方能够准确获知第一数据集和第二数据集中的交集规模。当然,由于第二数据集的标识信息和特征信息均是加密过的,具体标识信息和特征信息并不会被泄露给第一参与方。

[0042] 此外,由于交集索引信息包含伪索引,在第一参与方将交集索引信息提供给第二参与方后,第二参与方基于该交集索引信息所确定的交集可能不是准确的。这可以支持不向第二参与方泄露两个数据集的交集规模,在需要向第二参与方隐藏交集规模的实际应用中具有重要意义。

[0043] 在本公开的一些实施例中,还提出了在包含伪索引的交集索引信息基础上,第一参与方和第二参与方各自确定出第一交集和第二交集,并基于第一交集和第二交集来执行MPC运算。由于第一参与方能够获知真实的标识信息匹配状况,而第二参与方不能确定真实的标识信息匹配状况,因此在MPC运算中可以通过对第一交集和第二交集分别设置匹配标记来完成正确的MPC运算。

[0044] 在本公开的一些实施例中,还提出了对第一数据集和第二数据集中的特征信息的高效加密的方式。在本公开的一些实施例中,还提出了支持多ID匹配的方案。在标识信息包括多个类型的标识符的情况下,能够针对类型的标识符来确定交集索引信息。而且,在本公开的实施例中,还可以支持在数据集中包含重复元素,例如数据条目的相同标识信息重复

出现的情况下,仍然可以执行数据集的交集匹配。

[0045] 以下将继续参考附图描述本公开的一些示例实施例。

[0046] 图2示出了根据本公开的一些实施例的用于数据处理的多方信令流200的流程图。为便于讨论,将参考图1的环境100来信令流200。信令流200涉及参与方110和参与方120。

[0047] 在信令流200中,参与方110对参与方120的数据集122中各个数据条目的加密标识信息和加密特征信息执行二次加密(230)。类似的,参与方120对参与方110的数据集112中各个数据条目的加密标识信息和加密特征信息执行二次加密(232)。参与方110和参与方120可以通过多种方式获得对方数据集的加密标识信息和加密特征信息。

[0048] 在一些实施例中,在初始阶段,参与方110可以对自己数据集112中各个数据条目的标识信息和特征信息执行一次加密(210),得到加密标识信息和加密特征信息(在图2中分别被标记为“加密标识信息1”和“加密特征信息1”)。参与方120可以对自己数据集122中各个数据条目的标识信息和特征信息执行一次加密(212),得到加密标识信息和加密特征信息(在图2中分别被标记为“加密标识信息2”和“加密特征信息2”)。

[0049] 参与方110可以将数据集112的加密标识信息和加密特征信息发送(220)发送给参与方120,以由参与方120执行二次加密(232)。类似的,参与方120可以将数据集122的加密标识信息和加密特征信息发送(222)发送给参与方110,以由参与方110执行二次加密(230)。

[0050] 举例而言,为便于理解,假设数据集112或数据集122中的每个数据条目的标识信息包括一个或多个类型的标识符(ID),并且特征信息包括一个或多个类型的特征。假设数据集112和数据集122的数据条目均包括k个ID;数据集112包括n_c个数据条目,每个数据条目包括m_c个特征;数据集122包括n_p个数据条目,每个数据条目包括m_p个特征。

[0051] 这样,数据集112可以被表示为二维矩阵的形式,表示为$\{(Cid_{i,0},\ldots,Cid_{i,k-1},u_{i,0},\ldots,u_{i,m_c})\}_{i\in[n_c]}$(在此使用[$n_c$]的形式表示范围[0,$n_c$),下同),其中$Cid_{i,0}$指的是第i个数据条目的标识符$ID_0$.....$Cid_{i,0}$指的是第i个数据条目的标识符$ID_{k-1}$,$u_{i,0}$指的是第i个数据条目的特征0,$u_{i,m_c}$指的是第i个数据条目的特征$m_c$。类似的,数据集122可以被表示为$\{(Pid_{i,0},\ldots,Pid_{i,k-1},v_{i,0},\ldots,v_{i,m_p})\}_{i\in[n_p]}$。

[0052] 在一些实施例中,考虑到后续对数据集112和122中的标识信息和特征信息的加密需要,在初始化阶段,参与方110和参与方120可以各自确定要使用的加密方式以及密钥。

[0053] 在一些实施例中,数据集112和122中的标识信息可以基于椭圆曲线加密算法来实现,并且参与方110和120可以通过椭圆曲线密钥交换来实现密钥交换。举例来说,参与方110可以随机选择椭圆曲线加密密钥r_c;并且参与方120可以随机选择椭圆曲线加密密钥r_p。在其他实施例中,对标识信息的加密还可以基于其他任何适当的加密算法,只要参与方110和120各自选择用于对标识信息进行加密的密钥。

[0054] 在一些实施例中,数据集112和122中的特征信息可以基于同态加密(HE)算法来加密。经过同态加密后的特征信息,可以通过同态加法、乘法运算,在不解密特征信息的情况下计算得到新的密文。新密文解密后可以得到经过相应同态运算的特征信息。在一些实施例中,通过应用同态加密,可以在支持后续基于数据集112和122中加密后的特征信息来支持MPC计算。参与方110和120可以选择任何适当的同态加密算法,其中一个示例是Paillier

同态加密。在一些实施例中,参与方110可以生成随机的同态加密的公钥和私钥,即(pk_c,sk_c),并且参与方110可以将公钥pk_c发送给参与方120。类似的,参与方120可以生成随机的同态加密的公钥和私钥,即(pk_p,sk_p),并且参与方120可以将公钥pk_p发送给参与方110。

[0055] 在上述初始化阶段确定加密方式后,参与方110和120可以执行各自数据集的加密标识信息和加密特征信息的交换。在一些实施例中,加密标识信息和加密特征信息的交换可以由任一方触发。在一些实施例中,如果参与方120是可多次调用的客户端,而参与方110是服务器侧,可以由参与方120先发起请求,将数据集122的加密标识信息和加密特征信息发送给参与方110。在一些实施例中,在接收到请求后,参与方110可以根据客户端120的数据集122的大小(即其中的数据条目的数目)来确定是否需要在数据集112中填充伪数据条目。应当理解,参与方110和参与方120在不同应用场景中可能对应于不同实体,两者的交集匹配可以基于任意原因、由任一方或由双方协商来触发。

[0056] 在一些实施例中,在对数据集112和数据集122的一次加密阶段(即,信令流200中的一次加密210和212),参与方110和参与方120可以使用ECDH技术生成加密标识信息,并使用HE加密技术来加密特征信息,并将各自的加密标识信息和加密特征信息发送给对方。

[0057] 在一些实施例中,在加密之前,参与方110可以对数据集112中的各个数据条目进行乱序处理。可选地或附加地,参与方120可以对数据集122中的各个数据条目进行乱序处理。

[0058] 在一些实施例中,在加密标识信息时,参与方110可以使用第一加密密钥,例如椭圆曲线加密密钥r_c来对数据集112中各个数据条目的标识信息$Cid_{i,j}$进行加密,得到数据集112的加密标识信息,即$Cid'_{i,j}=r_c \cdot H(Cid_{i,j})$。这样,数据集112中各个数据条目的标识信息被随机化。类似的,参与方120可以使用第二加密密钥,例如椭圆曲线加密密钥r_p来对数据集122中各个数据条目的标识信息$Pid_{i,j}$进行加密,得到数据集122的加密标识信息,即$Pid'_{i,j}=r_p \cdot H(Pid_{i,j})$。这样,数据集112中各个数据条目的标识信息被随机化。在加密过程中,$H: \{0,1\}^* \rightarrow \mathbb{G}$,是将任意输入映射为椭圆曲线点的哈希函数。当然,如上文提及的,对标识信息的加密还可以基于任何其他适当加密算法。

[0059] 在加密特征信息时,参与方110可以使用适当加密算法,例如使用Paillier加密来加密数据集112的特征信息。例如,参与方110可以利用初始化阶段生成的密钥sk_c来对数据集112中各个数据条目的特征信息执行同态加密。类似地,参与方120可以使用适当加密算法,例如使用Paillier加密来加密数据集122的特征信息。例如,参与方120可以利用初始化阶段生成的密钥sk_c来对数据集122中各个数据条目的特征信息执行同态加密。

[0060] 在一些实施例中,为了提高特征信息的加密效率,还提出了批加密的方式。举例来说,假设批加密的批大小为预定数目(表示为B),即每次加密B个数据条目。参与方110将数据集112中各个数据条目的特征信息$u_{i,j}$按顺序划分为至少一个特征信息块,每个特征信息块包括数据集112中B个数据条目的特征信息的顺序拼接,此外,在每个特征信息块中的两个相邻数据条目之间填充有预定信息,以用于将各个数据条目彼此分离。例如,参与方110可以将B个数据条目的特征信息拼接或编码为:$U_{i',j}=u_{i,j}||0||u_{i+1,j}||0||\cdots||u_{i+B-1,j}$。各个特征信息可以按比特拼接,用||表示。在两个相邻数据条目之间填充的预定信息可以是比特值0。当然,也可以是任何其他预设符号或预设值。

[0061] 通过编码特征信息快,需要加密的特征信息块$U_{i',j}$的数目为$\lceil \frac{n_c}{B} \rceil$,其中$\lceil \rceil$表示向上

取整。$\lceil \frac{n_c}{B} \rceil$小于数据集112中的数据条目的数目$n_c$。经过编码后，由于数目减小，每个特征信息块$U_{i',j}$中所包括的B个数据条目的加密标识信息$Cid'_{i,j}$用于索引该特征信息块$U_{i',j}$。数据集112中，各个特征信息块被表示为$\{U_{i',j}\}_{i' \in [\frac{n_c}{B}], j \in [m_c]}$，对应的标识信息被表示为$\{(Cid'_{i,j}, \ldots, Cid'_{i+B-1,j})\}_{i \in [\frac{n_c}{B}], j \in [m_c]}$。

[0062] 参与方110接着对至少一个特征信息块$U_{i',j}$进行加密，得到至少一个特征信息块$U_{i',j}$各自的加密特征信息$\tilde{U}_{i',j}$。在一些实施例中，参与方110可以使用同态加密，例如Paillier加密，利用公钥pk_c来加密特征信息块$U_{i',j}$，得到$\tilde{U}_{i',j} = Enc(U_{i',j}, pk_c)$。

[0063] 类似地，为了提高特征信息的加密效率，参与方120也可以采用批加密的方式来加密数据集122中各个数据条目的特征信息。在一些实施例中，参与方120执行批加密时所采用的批大小可以与参与方110所采用的批大小一样，即每次加密B个数据条目。参与方120将数据集122中各个数据条目的特征信息$v_{i,j}$按顺序划分为至少一个特征信息块，每个特征信息块包括数据集122中B个数据条目的特征信息的顺序拼接，此外，在每个特征信息块中的两个相邻数据条目之间填充有预定信息，以用于将各个数据条目彼此分离。例如，参与方120可以将B个数据条目的特征信息拼接或编码为：$V_{i',j} = v_{i,j}||0||v_{i+1,j}||0||\cdots||v_{i+B-1,j}$。各个特征信息可以按比特拼接，用||表示。在两个相邻数据条目之间填充的预定信息可以是比特值0。当然，也可以是任何其他预设符号或预设值。

[0064] 通过编码特征信息快，需要加密的特征信息块$V_{i',j}$的数目为$\lceil \frac{n_p}{B} \rceil$，其中$\lceil\ \rceil$表示向上取整。$\lceil \frac{n_p}{B} \rceil$小于数据集122中的数据条目的数目$n_p$。经过编码后，由于数目减小，每个特征信息块$V_{i',j}$中所包括的B个数据条目的加密标识信息$Pid'_{i,j}$用于索引该特征信息块$V_{i',j}$。数据集122中，各个特征信息块被表示为$\{\tilde{V}_{i',j}\}_{i' \in [\frac{n_p}{B}], j \in [m_p]}$，对应的标识信息被表示为$\{(Pid'_{i,j}, \ldots, Pid'_{i+B-1,j})\}_{i \in [\frac{n_p}{B}], j \in [m_p]}$。

[0065] 参与方120接着对至少一个特征信息块$V_{i',j}$进行加密，得到至少一个特征信息块$V_{i',j}$各自的加密特征信息$\tilde{V}_{i',j}$。在一些实施例中，参与方120可以使用同态加密，例如Paillier加密，利用公钥pk_p来加密特征信息块$\tilde{V}_{i',j}$，得到$\tilde{V}_{i',j} = Enc(V_{i',j}, pk_p)$。

[0066] 在数据集112的标识信息和特征信息完成一次加密后，参与方110将数据集112的加密标识信息$Cid'_{i,j}$和加密特征信息$\tilde{U}_{i',j}$发送给参与方120。在数据集122的标识信息和特征信息完成一次加密后，参与方120将数据集122的加密标识信息$Pid'_{i,j}$和加密特征信息$\tilde{V}_{i',j}$发送给参与方120。

[0067] 在接收到数据集122的加密标识信息$Pid'_{i,j}$和加密特征信息$\tilde{V}_{i',j}$后，参与方110对参与方120的数据集122的加密标识信息$Pid'_{i,j}$和加密特征信息$\tilde{V}_{i',j}$执行二次加密(230)。类似的，在接收到数据集112的加密标识信息$Cid'_{i,j}$和加密特征信息$\tilde{U}_{i',j}$后，参与方120对参与方110的数据集112的加密标识信息$Cid'_{i,j}$和加密特征信息$\tilde{U}_{i',j}$执行二次加密(232)。

[0068] 在参与方110对参与方120的数据集122的加密标识信息$Pid'_{i,j}$和加密特征信息

$\tilde{V}_{i',j}$执行二次加密时,参与方110可以对数据集122中的加密标识信息$Pid'_{i,j}$和加密特征信息$\tilde{V}_{i',j}$执行乱序处理。举例来说,参与方110可以通过随机置换来调整数据集122中的加密标识信息$Pid'_{i,j}$和加密特征信息$\tilde{V}_{i',j}$的顺序。这样的乱序处理可以进一步防止在后续计算交集时双方根据特征信息的顺序推断出原始数据集的情况。在一些示例中,参与方110可以生成$\left[0,\lceil\frac{n_p}{B}\rceil\right)$范围内的随机置换$\pi_c$,其指示对每个加密标识信息$Pid'_{i,j}$及其加密特征信息$\tilde{V}_{i',j}$的位置置换。然后,参与方110可以按批大小B将随机置换π_c作用于接收到的参与方120的加密标识信息$\{(Pid'_{i,j},\ldots,Pid'_{i+B-1,j})\}_{i\in[\frac{n_p}{B}],j\in[m_p]}$和加密特征信息$\{\tilde{V}_{i',j}\}_{i'\in[\frac{n_p}{B}],j\in[m_p]}$来打乱数据顺序。然后,参与方110可以对调整顺序后的加密标识信息$Pid'_{i,j}$和加密特征信息$\tilde{V}_{i',j}$执行二次加密。

[0069] 在一些实施例中,在对接收到的数据集122中的加密标识信息$Pid'_{i,j}$执行二次加密时,参与方110可以再次使用第一加密密钥(例如,用于对数据集112中的标识信息执行一次加密的椭圆曲线加密密钥r_c)来对数据集122中的加密标识信息$Pid'_{i,j}$执行二次加密,得到数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}=r_c r_p\cdot H(Pid_{i,j})$。

[0070] 在一些实施例中,在对接收到的数据集122中的加密特征信息$\tilde{V}_{i',j}$执行二次加密时,参与方110可以通过特征分享的方式来生成数据集122中的加密特征信息$\tilde{V}_{i',j}$的第一特征分享。考虑到数据集122中的加密特征信息$\tilde{V}_{i',j}$是通过批加密方式被按批次编码后加密的,在对加密特征信息$\tilde{V}_{i',j}$执行二次加密时,参与方110可以生成数据集122中各个数据条目对应的第二特征分享$\{\gamma_{i,j}\}_{i\in n_p,j\in m_p}$,并将特征分享按相同的批大小B进行编码,例如将特征分享的负值按相同的批大小B进行编码。参与方110可以将特征分享$\{\gamma_{i,j}\}_{i\in n_p,j\in m_p}$按顺序划分为至少一个特征分享块$[V_{i',j}]_1$,每个特征分享块包括B个数据条目对应的特征分享的顺序拼接,即$[V_{i',j}]_1=-\gamma_{i,j}||0||-\gamma_{i+1,j}||0||\cdots||-\gamma_{i+B-1,j}$,其中两个相邻特征分享之间填充有预定信息,例如比特值0。

[0071] 然后,参与方110基于至少一个特征分享块$[V_{i',j}]_1$来对接收到的数据集122的加密特征信息$\tilde{V}_{i',j}$执行同态加法处理,得到加密特征信息$\tilde{V}_{i',j}$的第一特征分享$[\widetilde{V_{i',j}}]_0=Add\left(\tilde{V}_{i',j'}[V_{i',j}]_1\right)$。通过特征分享方式,数据集122中的数据条目的加密特征信息$\tilde{V}_{i',j}$被拆分为第一特征分享$[\widetilde{V_{i',j}}]_0$和第二特征分享$[V_{i',j}]_1$两个部分,这两部分之和等于加密特征信息$\tilde{V}_{i',j}$。

[0072] 类似地,在参与方120对参与方110的数据集112的加密标识信息$Cid'_{i,j}$和加密特征信息$\tilde{U}_{i',j}$执行二次加密时,参与方120可以对数据集112中的加密标识信息$Cid'_{i,j}$和加密特征信息$\tilde{U}_{i',j}$执行乱序处理。举例来说,参与方120可以通过随机置换来调整数据集112中的加密标识信息$Cid'_{i,j}$和加密特征信息$\tilde{U}_{i',j}$的顺序。这样的乱序处理可以进一步防止在后续计算交集时双方根据特征信息的顺序推断出原始数据集的情况。在一些示例中,参与方120可以生成$\left[0,\lceil\frac{n_c}{B}\rceil\right)$范围内的随机置换$\pi_p$,其指示对每个加密标识信息$Cid'_{i,j}$及其加密特征

信息$\widetilde{U}_{i',j}$的位置置换。然后,参与方120可以按批大小B将随机置换π_p作用于接收到的参与方110的加密标识信息$\{(Cid'_{i,j'}\ \ldots,\ Cid'_{i+B-1,j})\}_{i\in[\frac{n_c}{B}],j\in[m_c]}$和加密特征信息$\{\widetilde{U}_{i',j}\}_{i'\in[\frac{n_c}{B}],j\in[m_c]}$来打乱数据顺序。然后,参与方120可以对调整顺序后的加密标识信息$Cid'_{i,j}$和加密特征信息$\widetilde{U}_{i',j}$执行二次加密。

[0073] 在一些实施例中,在对接收到的数据集112中的加密标识信息$Cid'_{i,j}$执行二次加密时,参与方120可以再次使用第二加密密钥(例如,用于对数据集122中的标识信息执行一次加密的椭圆曲线加密密钥r_p)来对数据集112中的加密标识信息$Cid'_{i,j}$执行二次加密,得到数据集112的双重加密标识信息$\widetilde{Cid}_{i,j}=r_p\,r_c\cdot H(Cid_{i,j})$。

[0074] 在一些实施例中,在对接收到的数据集112中的加密特征信息$\widetilde{U}_{i',j}$执行二次加密时,与以上关于参与方110所描述的类似,参与方120也可以通过特征分享的方式来生成数据集112中的加密特征信息$\widetilde{U}_{i',j}$的第一特征分享。考虑到数据集112中的加密特征信息$\widetilde{V}_{i',j}$是通过批加密方式被按批次编码后加密的,在对加密特征信息$\widetilde{U}_{i',j}$执行二次加密时,参与方120可以生成数据集112中各个数据条目对应的第二特征分享$\{\delta_{i,j}\}_{i\in[n_c],j\in[m_c]}$,并将特征分享按相同的批大小B进行编码,例如将特征分享的负值按相同的批大小B进行编码。参与方120可以将特征分享$\{\delta_{i,j}\}_{i\in[n_c],j\in[m_c]}$按顺序划分为至少一个特征分享块$[u_{i',j}]_1$,每个特征分享块包括B个数据条目对应的特征分享的顺序拼接,即$[U_{i',j}]_1=-\delta_{i,j}||0||-\delta_{i+1,j}||0||\cdots||-\delta_{i+B-1,j}$,其中两个相邻特征分享之间填充有预定信息,例如比特值0。

[0075] 然后,参与方120基于至少一个特征分享块$[U_{i',j}]_1$来对接收到的数据集112的加密特征信息$\widetilde{U}_{i',j}$执行同态加法处理,得到加密特征信息$\widetilde{U}_{i',j}$的第一特征分享$[\widetilde{U_{i',j}}]_0=Add\left(\widetilde{U}_{i',j'}[U_{i',j}]_1\right)$。通过秘密分享方式,数据集112中的数据条目的加密特征信息$\widetilde{U}_{i',j}$被拆分为第一特征分享$[\widetilde{U_{i',j}}]_0$和第二特征分享$[U_{i',j}]_1$两个部分,这两部分之和等于加密特征信息$\widetilde{U}_{i',j}$。

[0076] 在二次加密后,参与方120至少将数据集112的双重加密特征信息$\widetilde{Cid}_{i,j}$发送(242)给参与方110。在一些实施例中,参与方120在发送双重加密特征信息$\widetilde{Cid}_{i,j}$同时还发送数据集112的加密特征信息的第一特征分享$[\widetilde{U_{i',j}}]_0$,其中双重加密特征信息$\widetilde{Cid}_{i,j}$是用于标识对应的加密特征信息的第一特征分享$[\widetilde{U_{i',j}}]_0$。

[0077] 参与方110将数据集122的加密特征信息的第一特征分享$[\widetilde{V_{i',j}}]_0$发送(250)给参与方120,而不发送双重加密标识信息$\widetilde{Pid}_{i,j}$。这样,参与方120将无法获得数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$。如后续将描述的,数据集112和122的双重加密标识信息用于确定两个数据集的交集索引信息,通过避免向参与方120提供数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$,可以避免向参与方120透露两个数据集的真实交集规模。

[0078] 以上讨论了在一次加密和二次加密阶段参与方110和参与方120的信息交换。

[0079] 经过信息交换后，在参与方110侧，可以缓存数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$和加密特征信息的第二特征分享$[V_{i',j}]_1$。此外，参与方110还从参与方120接收到数据集112的双重加密特征信息$\widetilde{Cid}_{i,j}$和加密特征信息的第一特征分享$[U_{i',j}]_1$。在一些实施例中，参与方110可以使用密钥sk_c解密所接收到的加密特征信息的第一特征分享$[\widetilde{U_{i',j}}]_0$，得到数据集112的解密特征信息的第一特征分享$[U_{i',j}]_0 = Dec\left([\widetilde{U_{i',j}}]_0, sk_c\right)$。然后，参与方110可以缓存数据集112的双重加密标识信息$\widetilde{Cid}_{i,j}$和解密特征信息的第一特征分享$[U_{i',j}]_0$。这样，参与方110所缓存的数据包括参与方120的数据集122的<双重加密标识信息$\widetilde{Pid}_{i,j}$，第二特征分享$[V_{i',j}]_1$>以及自己的数据集112的<双重加密标识信息$\widetilde{Cid}_{i,j}$，第一特征分享$[U_{i',j}]_0$>。

[0080] 经过信息交换后，在参与方120侧，可以缓存数据集112的双重加密特征信息$\widetilde{Cid}_{i,j}$和加密特征信息的第二特征分享$[U_{i',j}]_1$。此外，参与方120还从参与方110接收到数据集122的加密特征信息的第一特征分享$[\widetilde{V_{i',j}}]_0$。在一些实施例中，参与方120可以使用密钥sk_p解密所接收到的加密特征信息的第一特征分享$[\widetilde{V_{i',j}}]_0$，得到数据集122的解密特征信息的第一特征分享$[V_{i',j}]_0 = Dec\left([\widetilde{V_{i',j}}]_0, sk_p\right)$。然后，参与方110可以缓存数据集122的解密特征信息的第一特征分享$[V_{i',j}]_0$。这样，参与方120所缓存的数据包括参与方110的数据集112的<双重加密特征信息$\widetilde{Cid}_{i,j}$，第二特征分享$[U_{i',j}]_1$>以及自己的数据集122的<第一特征分享$[V_{i',j}]_0$>。

[0081] 接下来，参与方110基于数据集112的双重加密标识信息$\widetilde{Cid}_{i,j}$和数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$来执行(260)交集匹配。在本公开的实施例中，两个数据集的交集匹配指的是找出两个数据集中标识信息相匹配(或相同)的数据条目。参与方110基于数据集112的双重加密标识信息$\widetilde{Cid}_{i,j}$与数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$之间的匹配结果，来生成交集索引信息，以指示数据集112和数据集122中的哪些数据条目的标识信息相匹配。如前文所述，数据集112的双重加密标识信息$\widetilde{Cid}_{i,j}$分别由参与方110使用第一加密密钥r_c和参与方120使用第二加密密钥r_p进行加密，即$\widetilde{Cid}_{i,j} = r_p r_c \cdot H(Cid_{i,j})$，而数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$分别由参与方120使用第二加密密钥$r_p$和参与方110使用第一加密密钥$r_c$进行加密，即$\widetilde{Pid}_{i,j} = r_c r_p \cdot H(Pid_{i,j})$。如果数据集112中的某个数据条目的标识信息与数据集122中某个数据条目的标识信息相匹配，那么经两个密钥r_c和r_p加密后，这两个数据条目的标识信息仍然相匹配。因此，可以在不透露实际标识信息的基础上，由参与方110执行标识信息的匹配与否的判断。

[0082] 由于标识信息没有执行批加密，因此数据集112和数据集122的双重标识信息包括各自数据集中各个数据条目对应的双重加密信息。所生成的交集索引信息包括对数据集112和数据集122中的至少一对数据条目的真索引和至少一对数据条目的伪索引，真索引对应的数据条目的标识信息相匹配(例如，双重加密标识信息$\widetilde{Cid}_{k,j}$与$\widetilde{Pid}_{i,j}$相等)，伪索引对应的数据条目的标识信息不匹配(例如，双重加密标识信息$\widetilde{Cid}_{k,j}$与$\widetilde{Pid}_{i,j}$不相等)。通过确定交集索引信息，参与方110能够获知数据集112和数据集122中标识信息真实匹配的数据条目，由此参与方110可以基于匹配结果来确定数据集112和数据集122的第一交集。

[0083] 在一些实施例中,参与方110还可以执行多标识符的交集匹配。例如,如果数据集112和数据集122的标识信息包括多个类型分别对应的标识符,那么相应地双重加密标识信息$\widetilde{Cid}_{i,j}$包括多个类型分别对应的多个双重加密标识符,双重加密标识信息$\widetilde{Pid}_{i,j}$包括多个类型分别对应的多个双重加密标识符。在生成交集索引信息时,参与方110可以基于多个类型的优先级来确定双重加密标识信息$\widetilde{Cid}_{i,j}$和双重加密标识信息$\widetilde{Pid}_{i,j}$之间的匹配结果。使用多个标识符匹配交集的逻辑为:对于每个标识符的类型,查找出匹配结果,然后从双重加密标识信息中筛除掉这个标识符类型的匹配结果后,再使用下一个类型的标识符进行匹配。采取这个匹配逻辑是因为在业务实践中这个逻辑比较常用,在业务中通常会指定匹配标识符的优先级进行匹配,匹配到的交集不再使用更低优先级的标识符再次匹配,以免产生额外的重复交集组合。

[0084] 举例来说,按照多个类型的优先级,参与方110可以首先比较双重加密标识信息$\widetilde{Cid}_{i,j}$中第一类型对应的双重加密标识符与在双重加密标识信息$\widetilde{Pid}_{i,j}$中第一类型对应的双重加密标识符,来确定第一匹配结果。第一类型可以是多个类型中优先级最高的类型。参与方110可以基于第一匹配结果来生成交集索引信息中的对应的真索引和/或伪索引。

[0085] 在第一类型的比较完成后,如果第一匹配结果指示数据集112和数据集122中标识信息相匹配的至少一对数据条目,参与方110从双重加密标识信息$\widetilde{Cid}_{i,j}$和双重加密标识信息$\widetilde{Pid}_{i,j}$中分别过滤掉相匹配的至少一对数据条目的双重加密标识信息,得到过滤后的双重加密标识信息$\widetilde{Cid}_{i,j}$和过滤后的双重加密标识信息$\widetilde{Pid}_{i,j}$。然后,参与方110可以对标识信息中的第二类型进行比较,其中第二类型的优先级低于第一类型的优先级。通过比较过滤后的双重加密标识信息$\widetilde{Cid}_{i,j}$中第二类型对应的双重加密标识符与在过滤后的双重加密标识信息$\widetilde{Pid}_{i,j}$中第二类型对应的双重加密标识符,可以确定第二匹配结果。参与方110可以进一步基于第二匹配结果来生成交集索引信息中的对应的真索引和/或伪索引。

[0086] 在第二类型的比较完成后,如果第二匹配结果指示数据集112和数据集122中标识信息相匹配的至少一对数据条目,那么类似地,参与方110可以从过滤后的双重加密标识信息$\widetilde{Cid}_{i,j}$和过滤后的双重加密标识信息$\widetilde{Pid}_{i,j}$再次过滤掉相匹配的至少一对数据条目的双重加密标识信息,并且进行后续类型的标识符的匹配。

[0087] 为便于理解,举例来说,假设参与方110的数据集112和参与方120的数据集122中具有i、j、k三个类型的标识符。数据集112中三个数据条目的双重加密标识符分别是[(i0,j2,k4),(i1,j3,k5),(i2,j4,k6)],数据集122中三个数据条目的双重加密标识符分别[(i0,j2,k4),(i8,j3,k1),(i9,j2,k2)]。按照i、j、k的优先级匹配,使用第一类型i的标识符进行匹配得到的匹配结果是<(i0,j2,k4),(i0,j2,k4)>,并匹配出对应的特征信息。过滤掉匹配结果后,数据集112中剩余的双重加密标识符是[(i1,j3,k5),(i2,j4,k6)],数据集122中剩余的双重加密标识符是[(i8,j3,k1),(i9,j2,k2)]。使用第二类型j的标识符进行匹配得到的匹配结果是<(i1,j3,k5),(i8,j3,k1)>。过滤掉匹配结果后,数据集112中剩余的双重加密标识符是[(i2,j4,k6)],数据集122中剩余的双重加密标识符是[(i9,j2,k2)]。第三类型k的标识符无匹配交集。

[0088] 注意到,参与方110的双重加密信息(i0,j2,k4)和参与方120的双重加密标识信息

(i9,j2,k2)在第二类型的标识符k处是相同的,但是由于(i0,j2,k4)在第一轮被筛除了,因此不会匹配出交集对<(i0,j2,k4),(i9,j2,k2)>。

[0089] 在参与方110会得到所有匹配交集的双方索引。为了不泄露交集规模,参与方110会额外生成虚假交集索引并填充到交集索引信息中,使得交集索引的数量为$\min(n_c,n_p)$。

[0090] 在一些实施例中,对于双重加密标识信息中任一类型的标识符,参与方110可以逐一遍历数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$,以确定数据集112中是否存在与之相匹配双重加密标识信息。如果存在,参与方110可以生成一个真索引,例如<k,i>,表示数据集112中第k个数据条目与数据集122中第i个数据条目的标识信息相匹配。否则,参与方110生成伪索引(也称为虚假索引),例如<k′,i>,其中k′是从[0,n_c)范围内随机挑选的值。当然,也可以反过来,参与方110通过逐一遍历数据集112的双重加密标识信息$\widetilde{Cid}_{i,j}$来生成交集索引信息。此外,参与方110还会记录哪些索引是真实的,哪些是虚假的。

[0091] 根据上述方式,在数据集112和数据集122中数据条目不匹配,或者标识信息包括重复元素时,均能够实现交集匹配。

[0092] 为便于理解,图3示出了根据本公开的一些实施例的用于交集匹配的示例的示意图。如图3所示,参与方110缓存有数据集112的双重加密标识信息和特征分享310,以及数据集122的双重加密标识信息和特征分享320。在进行交集匹配时,对于数据集112中的双重加密标识信息[rp][rc]a,参与方110确定数据集122的双重加密标识信息中没有与之匹配的双重加密标识信息,因此生成交集索引信息330中的伪索引3,其中3是随机选择的值,指示数据集112中第0个数据条目(伪)匹配数据集122的第3个数据条目。交集索引信息330中索引的位置与数据集112中数据条目的位置相对应。对于数据集112中的双重加密标识信息[rp][rc]c,参与方110确定数据集122中第1个数据条目和第2个数据条目的双重加密标识信息[rc][rp]c均与之相匹配,因此生成交集索引信息330中的真索引[1,2],指示数据集112中第1个数据条目匹配数据集122的第1个数据条目和第2个数据条目。

[0093] 以此类推,对于数据集112中的双重加密标识信息[rp][rc]b和[rp][rc]e,均未找到与之匹配的数据集122中的双重加密标识信息,因此生成交集索引信息330中的伪索引0和-1。对于数据集112中的双重加密标识信息[rp][rc]c,参与方110确定数据集122中第1个数据条目和第2个数据条目的双重加密标识信息[rc][rp]c均与之相匹配,因此生成交集索引信息330中的真索引[1,2],指示数据集112中第4个数据条目匹配数据集122的第1个数据条目和第2个数据条目。在交集索引信息330中,参与方110会记录哪些索引是真实的,哪些是虚假的。

[0094] 在这样的匹配流程中,当某个相同的标识信息在参与方110处存在x个数据条目,在参与方120处存在y条数据条目时,总共会生成x·y条真实交集索引对。这也符合数据集查询,例如结构化查询语言SQL中数据集的内连接(inner join)的语义。

[0095] 参与方110将交集索引信息发送(270)给参与方120,以用于参与方120确定数据集112和数据集122的第二交集。例如,参与方110可以将匹配的数据条目对的索引,例如<k,i>、<k′,i>等发送给参与方120。或者,参与方110可以按数据集122的双重加密标识信息$\widetilde{Pid}_{i,j}$的顺序,将所确定的数据集112中的真实或虚假匹配的索引发送给参与方120。在图3的示例中,交集索引信息330被发送给参与方120。这样,参与方120从包含真索引和伪索引的交集

索引信息中可以确定出两个数据集的交集,但无法确定其中哪个或哪些数据条目是真实匹配的。

[0096] 在一些实施例中,返回参考图2,在获得交集索引信息后,参与方110可以基于交集索引信息来生成(280)数据集112与数据集122的第一交集。第一交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目。如前所述,参与方110处缓存有参与方120的数据集122的<双重加密标识信息$\widetilde{Pid}_{i,j}$,第二特征分享$[V_{i',j}]_1$>以及自己的数据集112的<双重加密标识信息$\widetilde{Cid}_{i,j}$,第一特征分享$[U_{i',j}]_0$>。参与方110可以使用交集索引信息中(真实或虚假)索引<a,b>来关联参与方110和参与方120的特征分享。这样,参与方110可以得到$\left\langle \{[U_{a,j}]_0\}_{j\in[m_c]}, \{[V_{b,j}]_1\}_{j\in[m_p]} \right\rangle$。如图3的示例所示,参与方110可以基于交集索引信息330来生成第一交集340。

[0097] 类似地,参与方120可以基于交集索引信息来生成(282)数据集112与数据集122的第二交集。第二交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目。如前所述,参与方120缓存有参与方110的数据集112的<双重加密特征信息$\widetilde{Cid}_{i,j}$,第二特征分享$[U_{i',j}]_1$>以及自己的数据集122的<第一特征分享$[V_{i',j}]_0$>。参与方120可以使用交集索引信息中(真实或虚假)索引<a,b>来关联参与方110和参与方120的特征分享。这样,参与方120可以得到$\left\langle \{[U_{a,j}]_1\}_{j\in[m_c]}, \{[V_{b,j}]_0\}_{j\in[m_p]} \right\rangle$。

[0098] 注意,参与方110所确定的第一交集和参与方120所确定的第二交集中,包括的均是双重加密标识信息,以及特征信息的特征分享。数据集112和数据集122中的真实标识信息和特征信息并未被透露给对方。

[0099] 由于参与方110知道交集索引信息所指示的真索引和假索引,参与方110可以设置针对第一交集中每对数据条目的匹配标记,其中真索引对应的至少一对数据条目的匹配标记为指示匹配,伪索引对应的至少一对数据条目的匹配标记为指示不相匹配。在一些实施例中,真索引对应的一对数据条目的匹配标志位被设置为1,所述伪索引对应的一对数据条目的匹配标志位被设置为0。

[0100] 例如,参与方110可以在第一交集中额外设置匹配标记列,其中记录每对数据条目的匹配标记(也称为is_real标志位),用于标识该条数据条目是真实交集还是虚假填充的交集。参与方110会根据填充的实际情况将真实匹配交集的is_real标志位设置为1,虚假匹配交集的is_real标志位设置为0。

[0101] 在参与方120处,参与方120也类似设置第二交集的匹配标记。由于参与方120不能确定交集索引信息中所指示的哪些索引是伪索引,参与方120可以将所有数据条目的is_real标志位均设置为指示不相匹配,例如均设置为0。

[0102] 继续参考图2,参与方110可以与参与方120一起,利用第一交集和第二交集来执行MPC(290)。由于第一交集和第二交集中的数据条目还包括标识信息不匹配的数据条目,利用第一交集和第二交集来执行MPC后可以得到候选计算结果。参与方110可以基于所确定的第一交集中每对数据条目的候选计算结果和针对第一交集的匹配标记来确定MPC的目标计算结果。例如,如果第一交集的匹配标记中真索引对应的一对数据条目的匹配标志位被设

置为1,伪索引对应的一对数据条目的匹配标志位被设置为0,参与方110可以基于第一交集中每对数据条目的候选计算结果与第一交集中每对数据条目的匹配标记的相乘运算,生成目标计算结果。

[0103] 类似地,参与方120可以基于所确定的第二交集中每对数据条目的候选计算结果和针对第二交集的匹配标记来确定MPC的目标计算结果。如果第二交集中每对数据条目的匹配标记被设置为0,参与方120可以基于第二交集中每对数据条目的候选计算结果与第二交集中每对数据条目的匹配标记的相乘运算,生成目标计算结果。

[0104] 因此,虽然第一交集和第二交集均不是真实交集结果,但在MPC运算后,通过调用MPC乘法将输出的候选计算结果与is_real标记位相乘,即可保留真实交集运算结果。

[0105] 为更便于理解,图4示出了根据本公开的一些实施例的基于示例数据集的数据处理信令流400的流程图。图4的信令流400可以被认为是图2的信令流的一个示例。在图4中,给出了数据集112和122的具体示例,以参考该具体示例来描述各个加密阶段以及求交阶段。

[0106] 如图4所示,在一次加密阶段,参与方120执行操作405,包括对数据集122中的标识信息和特征信息进行乱序处理;对标识信息进行随机哈,即使用第二加密密钥rp对标识信息执行一次加密;以及对特征信息执行一次加密(例如,使用P方密钥执行同态加密)。参与方120在消息1中将数据集122的加密标识信息和加密特征信息(<[rp]ID,Enc(特征2)>)406发送给参与方110。可以看出,在消息1中,数据集122中数据条目被乱序,且标识信息和特征信息被加密。

[0107] 类似地,在一次加密阶段,参与方110执行操作410,包括对数据集112中的标识信息和特征信息进行乱序处理;对标识信息进行随机哈,即使用第一加密密钥rc对标识信息执行一次加密;以及对特征信息执行一次加密(例如,使用C方密钥执行同态加密)。参与方110在消息2中将数据集112的加密标识信息和加密特征信息(<[rp]ID,Enc(特征1)>)412发送给参与方120。可以看出,在消息2中,数据集112中数据条目被乱序,且标识信息和特征信息被加密。

[0108] 在二次加密阶段,参与方110执行操作415,包括对接收到的参与方120的数据集122的加密标识信息和加密特征信息(<[rp]ID,Enc(特征2)>)执行乱序处理;利用第一加密密钥rc对加密标识信息[rp]ID执行二次加密,得到双重加密标识信息[rc][rp]ID;以及对加密特征信息Enc(特征2)执行秘密拆分,得到第一特征分享和第二特征分享。参与方110缓存数据集122的双重加密标识信息[rc][rp]ID以及加密特征信息的第二特征分享418,并在消息3.1中将数据集122的加密特征信息的第一特征分享416发送给参与方120。

[0109] 类似的,在二次加密阶段,参与方120执行操作420,包括对接收到的参与方110的数据集112的加密标识信息和加密特征信息(<[rc]ID,Enc(特征1)>)执行乱序处理;利用第二加密密钥rp对加密标识信息[rc]ID执行二次加密,得到双重加密标识信息[rp][rc]ID;以及对加密特征信息Enc(特征1)执行秘密拆分,得到第一特征分享和第二特征分享。参与方120缓存数据集112的双重加密标识信息[rp][rc]ID以及加密特征信息的第二特征分享428,并在消息3.2中将数据集112的双重加密特征信息[rp][rc]ID和加密特征信息的第一特征分享426发送给参与方110。

[0110] 这样,参与方110缓存数据集122的双重加密标识信息[rc][rp]ID以及加密特征信

息的第二特征分享418,以及从参与方120接收到的数据集112的双重加密特征信息[rp][rc]ID和加密特征信息的第一特征分享426。参与方120缓存数据集112的双重加密标识信息[rp][rc]ID以及加密特征信息的第二特征分享428,以及从参与方110接收到的数据集122的加密特征信息的第一特征分享416。

[0111] 在交集匹配阶段,参与方110可以执行操作340,包括利用缓存的数据集122的双重加密标识信息[rc][rp]ID和数据集112的双重加密特征信息[rp][rc]ID来求交,并随机选择伪索引;按照交集索引信息所指示的匹配数据条目的索引和消息3.2中数据条目的顺序,生成交集索引信息434。参与方110将交集索引信息434在消息4中发送给参与方120。此外,参与方110还基于交集索引信息434来生成第一交集442,并且还生成第一交集242中各个数据条目的匹配标记(is_real标记位)。第一交集442包括数据集112的双重加密特征信息[rp][rc]ID,参与方110缓存的数据集112和数据集122的特征分享,以及is_real标记位。

[0112] 在接收到参与方110的交集索引信息434,参与方120可以执行操作445,生成第二交集436,并且还可以设置第二交集的匹配标记,得到第二交集452。第二交集452包括参与方120缓存的数据集112的双重加密特征信息[rp][rc]ID,数据集112和数据集122的特征分享,以及is_real标记位。

[0113] 参与方110可以根据在交集索引信息中填充的实际情况将真实交集的is_real标记位设置为1,虚假交集的is_real标记位设置为0。而参与方120则将所有的is_real标记位均设置为0。在基于第一交集和第二交集执行MPC运算时,可以调用MPC乘法将MPC运算的候选结果与is_real标记位相乘,即可保留真实交集运算结果。

[0114] 根据本公开的实施例,可以支持在不暴露双方数据集的真实信息情况下,不向其中一个参与方泄露交集的规模,使另一参与方能够获知真实交集规模。此外,在一些实施例中,还可以支持获得MPC协议所需的特征分享,以进行MPC运算。在一些实施例中,在交集匹配时,可以支持使用多个标识符,按照优先级进行匹配。而且,在加密过程中,可以通过批加密的方式显著提高加密效率。在整个交互过程中两个参与方中的内存占用小,双方无需缓存同态加密的密文。

[0115] 图5示出了根据本公开的一些实施例的在第一参与方处实现的数据处理方法500的流程图。方法500例如可以被实现图1的参与方110。为便于讨论,将参考图1的环境100来描述方法500。

[0116] 在框510,参与方110对MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密,得到第二双重加密标识信息和第二加密特征信息的第一特征分享。

[0117] 在框520,参与方110将第二数据集中各个数据条目的第二加密特征信息的第一特征分享发送给第二参与方,而不发送第二双重加密标识信息。

[0118] 在框530,参与方110从第二参与方接收第一参与方的第一数据集中各个数据条目的第一双重加密标识信息。

[0119] 在框540,参与方110基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果,来生成交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的数据条目的标识信息相匹配,伪索引对应的数据条目的标识信息不匹配。

[0120] 在框550,参与方110将交集索引信息发送给第二参与方,以用于第二参与方确定第一数据集与第二数据集的第二交集。

[0121] 在一些实施例中,在接收第一双重加密标识信息之前,方法500还包括:对第一数据集中各个数据条目的第一标识信息和第一特征信息进行加密,得到第一加密标识信息和第一加密特征信息;以及向第二参与方发送第一加密标识信息和第一加密特征信息。第一加密标识信息用于由第二参与方生成第一双重加密标识信息,第一加密特征信息用于由第二参与方生成第一加密特征信息的第一特征分享。

[0122] 在一些实施例中,对第一标识信息进行加密包括:利用第一加密密钥对第一数据集中各个数据条目的第一标识信息进行加密,得到第一加密标识信息;其中第一数据集中各个数据条目的第一双重加密标识信息是由第二参与方利用第二加密密钥对第一加密标识信息进行加密后生成。

[0123] 在一些实施例中,对第二加密标识信息执行二次加密包括:利用第一加密密钥对第二加密标识信息执行二次加密,得到第二双重加密标识信息。第一加密密钥还由第一参与方用于对第一数据集中各个数据条目的第一标识信息执行一次加密,得到第一加密标识信息,并且第二双重加密标识信息的一次加密和第一加密标识信息的二次加密由第二参与方利用第二加密密钥来执行。

[0124] 在一些实施例中,对第一特征信息进行加密包括:生成用于同态加密的第一公钥和第一私钥;将第一公钥发送给第二参与方;以及利用第一公钥对第一数据集中各个数据条目的第一特征信息执行同态加密,得到第一加密特征信息。

[0125] 在一些实施例中,对第一特征信息进行加密包括:将第一数据集中各个数据条目的第一特征信息按顺序划分为至少一个第一特征信息块,每个第一特征信息块包括第一数据集中预定数目的数据条目的第一特征信息的顺序拼接,在每个第一特征信息块中的两个相邻数据条目之间填充有预定信息;以及对至少一个第一特征信息块进行加密,得到至少一个第一特征信息块各自的第一加密特征信息。

[0126] 在一些实施例中,预定信息为零值。在一些实施例中,每个第一特征信息块中的预定数目的数据条目的第一加密标识信息用于索引该第一特征信息块。

[0127] 在一些实施例中,对第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密包括:通过随机置换来调整第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息的顺序;以及对调整顺序后的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密。

[0128] 在一些实施例中,第二数据集中各个数据条目的第二加密特征信息包括从第二数据集划分的至少一个第二特征信息块的第二加密特征信息,每个第二特征信息块是通过将第二数据集中各个数据条目的第二特征信息按顺序划分得到的,每个第二特征信息块包括第二数据集中预定数目的数据条目的第二特征信息的顺序拼接,在每个第二特征信息块中的两个相邻数据条目之间填充有预定信息。

[0129] 在一些实施例中,对第二加密特征信息执行二次加密包括:生成第二数据集中各个数据条目对应的第二特征分享;将第二数据集中各个数据条目对应的第二特征分享按顺序划分,得到第二加密特征信息的至少一个特征分享块,每个特征分享块包括第二数据集中预定数目的数据条目对应的第二特征分享的顺序拼接,在每个特征分享块中的两个相邻

第二特征分享之间填充有预定信息;以及基于至少一个特征分享块来对第二加密特征信息执行同态加法处理,得到第二加密特征信息的第一特征分享。

[0130] 在一些实施例中,第一数据集中各个数据条目的第一加密特征信息的第一特征分享与第一双重加密标识信息一起从第二参与方被接收到。在一些实施例中,方法500还包括:缓存第二数据集的第二双重加密标识信息和第二加密特征信息的第二特征分享,第二加密特征信息被划分为第一特征分享和第二特征分享;解密第一加密特征信息的第一特征分享,得到第一解密特征信息的第一特征分享;以及缓存第一数据集的第一双重加密标识信息和第一解密特征信息的第一特征分享。

[0131] 在一些实施例中,第一双重加密标识信息包括多个类型分别对应的多个第一双重加密标识符,第二双重加密标识信息包括多个类型分别对应的多个第二双重加密标识符。生成交集索引信息包括:基于多个类型的优先级来确定匹配结果,匹配结果的确定包括:通过比较第一双重加密标识信息中第一类型对应的第一双重加密标识符与在第二双重加密标识信息中第一类型对应的第二双重加密标识符,来确定第一匹配结果;如果第一匹配结果指示第一数据集和第二数据集中标识信息相匹配的至少一对数据条目,从第一双重加密标识信息和第二双重加密标识信息分别过滤掉相匹配的至少一对数据条目的双重加密标识信息,得到过滤后的第一双重加密标识信息和过滤后的第二双重加密标识信息;以及通过比较过滤后的第一双重加密标识信息中第二类型对应的第一双重加密标识符与在过滤后的第二双重加密标识信息中第二类型对应的第二双重加密标识符来确定第二匹配结果,第二类型的优先级低于第一类型的优先级。

[0132] 在一些实施例中,生成交集索引信息还包括:对于多个类型中的给定类型,通过至少遍历第二双重加密标识信息中给定类型对应的第二双重加密标识符来生成交集索引信息的一部分,其中生成包括:如果第一双重加密标识信息中给定类型对应的第一双重加密标识符与在第二双重加密标识信息中给定类型对应的第二双重加密标识符之间的给定匹配结果指示第一数据集中的第一数据条目与第二数据集中的第二数据条目的标识信息相匹配,生成交集索引信息中的第一真索引,以用于索引第一数据集中的第一数据条目和第二数据集中的第二数据条目;如果给定匹配结果指示第一数据集中的任何数据条目的标识信息均不匹配第二数据集中的第二数据条目的标识信息,生成交集索引信息中的第一伪索引,以用于索引第一数据集中的随机数据条目和第二数据集中的第二数据条目。

[0133] 在一些实施例中,方法500还包括:在第一参与方处,基于交集索引信息来生成第一数据集与第二数据集的第一交集,第一交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目;设置针对第一交集中每对数据条目的匹配标记,真索引对应的至少一对数据条目的匹配标记为指示匹配,伪索引对应的至少一对数据条目的匹配标记为指示不相匹配;与第二参与方一起,利用第一交集和第二交集来执行MPC,得到第一交集中每对数据条目的候选计算结果;以及至少基于候选计算结果和针对第一交集中每对数据条目的匹配标记来确定MPC的目标计算结果。

[0134] 在一些实施例中,真索引对应的一对数据条目的匹配标志位被设置为1,伪索引对应的一对数据条目的匹配标志位被设置为0。在一些实施例中,确定目标计算结果包括:基于第一交集中每对数据条目的候选计算结果与第一交集中每对数据条目的匹配标记的相乘运算,生成目标计算结果。

[0135] 在一些实施例中,针对第二交集中每对数据条目的匹配标记被设置为指示不匹配,并且目标计算结果的确定还基于针对第二交集中每对数据条目的匹配标记。

[0136] 图6示出了根据本公开的一些实施例的在第二参与方处实现的数据处理方法的流程图。方法600例如可以被实现图1的参与方120。为便于讨论,将参考图1的环境100来描述方法600。

[0137] 在框610,参与方120对从MPC的第一参与方接收到的第一数据集中各个数据条目的第一加密标识信息和第一加密特征信息执行二次加密,得到第一双重加密标识信息和第一加密特征信息的第一特征分享。

[0138] 在框620,参与方120向第一参与方至少发送第一数据集中各个数据条目的第一双重加密标识信息。

[0139] 在框630,参与方120从第一参与方接收针对第二参与方的第二数据集中各个数据条目的第二加密特征信息的第一特征分享,而未接收到第二数据集中各个数据条目的第二双重加密标识信息。

[0140] 在框640,参与方120从第一参与方接收交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的至少一对数据条目的标识信息相匹配。

[0141] 在框650,参与方120基于交集索引信息来确定第一数据集与第二数据集中的第二交集,第二交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目。

[0142] 在一些实施例中,方法600还包括:设置针对第二交集中每对数据条目的匹配标记,以指示该对数据条目的标识信息不相匹配;与第一参与方一起,利用第二交集和第一参与方确定的第一交集来执行MPC,得到第二交集中每对数据条目的候选计算结果;以及至少基于候选计算结果和针对第二交集中每对数据条目的匹配标记来确定MPC的目标计算结果。

[0143] 在一些实施例中,第二交集中每对数据条目的匹配标记被设置为0。在一些实施例中,第一数据集中真索引对应的一对数据条目的匹配标志位被设置为1,伪索引对应的一对数据条目的匹配标志位被设置为0。

[0144] 在一些实施例中,确定目标计算结果包括:基于第二交集中每对数据条目的候选计算结果与第二交集中每对数据条目的匹配标记的相乘运算,生成目标计算结果。

[0145] 图7示出了根据本公开的一些实施例的在第一参与方处实现的数据处理装置700的示意性结构框图。装置700可以被实现为或者被包括在参与方110中。装置700中的各个模块/组件可以由硬件、软件、固件或者它们的任意组合来实现。

[0146] 如图所示,装置700包括二次加密模块710,被配置为对MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密,得到第二双重加密标识信息和第二加密特征信息的第一特征分享。装置700还包括第一发送模块720,被配置为将第二数据集中各个数据条目的第二加密特征信息的第一特征分享发送给第二参与方,而不发送第二双重加密标识信息。

[0147] 装置700还包括第一接收模块730,被配置为从第二参与方接收第一参与方的第一数据集中各个数据条目的第一双重加密标识信息。

[0148] 装置700还包括交集索引确定模块740,被配置为基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果,来生成交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的数据条目的标识信息相匹配,伪索引对应的数据条目的标识信息不匹配。

[0149] 装置700还包括第二发送模块750,被配置为将交集索引信息发送给第二参与方,以用于第二参与方确定第一数据集与第二数据集的第二交集。

[0150] 在一些实施例中,装置700还包括一次加密模块,被配置为在接收第一双重加密标识信息之前,对第一数据集中各个数据条目的第一标识信息和第一特征信息进行加密,得到第一加密标识信息和第一加密特征信息;以及第三发送模块,被配置为向第二参与方发送第一加密标识信息和第一加密特征信息。第一加密标识信息用于由第二参与方生成第一双重加密标识信息,第一加密特征信息用于由第二参与方生成第一加密特征信息的第一特征分享。

[0151] 在一些实施例中,一次加密模块包括第一密钥加密模块,被配置为利用第一加密密钥对第一数据集中各个数据条目的第一标识信息进行加密,得到第一加密标识信息;其中第一数据集中各个数据条目的第一双重加密标识信息是由第二参与方利用第二加密密钥对第一加密标识信息进行加密后生成。

[0152] 在一些实施例中,二次加密模块710包括:第一密钥二次加密模块,被配置为利用第一加密密钥对第二加密标识信息执行二次加密,得到第二双重加密标识信息。第一加密密钥还由第一参与方用于对第一数据集中各个数据条目的第一标识信息执行一次加密,得到第一加密标识信息,并且第二双重加密标识信息的一次加密和第一加密标识信息的二次加密由第二参与方利用第二加密密钥来执行。

[0153] 在一些实施例中,一次加密模块包括:密钥生成模块,被配置为生成用于同态加密的第一公钥和第一私钥;公钥发送模块,被配置为将第一公钥发送给第二参与方;以及公钥加密模块,被配置为利用第一公钥对第一数据集中各个数据条目的第一特征信息执行同态加密,得到第一加密特征信息。

[0154] 在一些实施例中,一次加密模块包括:信息块划分模块,被配置为将第一数据集中各个数据条目的第一特征信息按顺序划分为至少一个第一特征信息块,每个第一特征信息块包括第一数据集中预定数目的数据条目的第一特征信息的顺序拼接,在每个第一特征信息块中的两个相邻数据条目之间填充有预定信息;以及信息块加密模块,被配置为对至少一个第一特征信息块进行加密,得到至少一个第一特征信息块各自的第一加密特征信息。

[0155] 在一些实施例中,预定信息为零值。在一些实施例中,每个第一特征信息块中的预定数目的数据条目的第一加密标识信息用于索引该第一特征信息块。

[0156] 在一些实施例中,二次加密模块710包括:随机置换模块,被配置为通过随机置换来调整第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息的顺序;以及乱序后加密模块,被配置为对调整顺序后的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密。

[0157] 在一些实施例中,第二数据集中各个数据条目的第二加密特征信息包括从第二数据集划分的至少一个第二特征信息块的第二加密特征信息,每个第二特征信息块是通过将第二数据集中各个数据条目的第二特征信息按顺序划分得到的,每个第二特征信息块包括

第二数据集中预定数目的数据条目的第二特征信息的顺序拼接,在每个第二特征信息块中的两个相邻数据条目之间填充有预定信息。

[0158] 在一些实施例中,二次加密模块710包括:特征分享生成模块,被配置为生成第二数据集中各个数据条目对应的第二特征分享;特征分享块划分模块,被配置为将第二数据集中各个数据条目对应的第二特征分享按顺序划分,得到第二加密特征信息的至少一个特征分享块,每个特征分享块包括第二数据集中预定数目的数据条目对应的第二特征分享的顺序拼接,在每个特征分享块中的两个相邻第二特征分享之间填充有预定信息;以及同态加法加密模块,被配置为基于至少一个特征分享块来对第二加密特征信息执行同态加法处理,得到第二加密特征信息的第一特征分享。

[0159] 在一些实施例中,第一数据集中各个数据条目的第一加密特征信息的第一特征分享与第一双重加密标识信息一起从第二参与方被接收到。在一些实施例中,装置700还包括:第一缓存模块,被配置为缓存第二数据集的第二双重加密标识信息和第二加密特征信息的第二特征分享,第二加密特征信息被划分为第一特征分享和第二特征分享;第一解密模块,被配置为解密第一加密特征信息的第一特征分享,得到第一解密特征信息的第一特征分享;以及第二缓存模块,被配置为缓存第一数据集的第一双重加密标识信息和第一解密特征信息的第一特征分享。

[0160] 在一些实施例中,第一双重加密标识信息包括多个类型分别对应的多个第一双重加密标识符,第二双重加密标识信息包括多个类型分别对应的多个第二双重加密标识符。交集索引确定模块740包括基于优先级的匹配模块,被配置为:基于多个类型的优先级来确定匹配结果,匹配结果的确定包括:通过比较第一双重加密标识信息中第一类型对应的第一双重加密标识符与在第二双重加密标识信息中第一类型对应的第二双重加密标识符,来确定第一匹配结果;如果第一匹配结果指示第一数据集和第二数据集中标识信息相匹配的至少一对数据条目,从第一双重加密标识信息和第二双重加密标识信息分别过滤掉相匹配的至少一对数据条目的双重加密标识信息,得到过滤后的第一双重加密标识信息和过滤后的第二双重加密标识信息;以及通过比较过滤后的第一双重加密标识信息中第二类型对应的第一双重加密标识符与在过滤后的第二双重加密标识信息中第二类型对应的第二双重加密标识符来确定第二匹配结果,第二类型的优先级低于第一类型的优先级。

[0161] 在一些实施例中,交集索引确定模块740还包括遍历确定模块,被配置为:对于多个类型中的给定类型,通过至少遍历第二双重加密标识信息中给定类型对应的第二双重加密标识符来生成交集索引信息的一部分,其中生成包括:如果第一双重加密标识信息中给定类型对应的第一双重加密标识符与在第二双重加密标识信息中给定类型对应的第二双重加密标识符之间的给定匹配结果指示第一数据集中的第一数据条目与第二数据集中的第二数据条目的标识信息相匹配,生成交集索引信息中的第一真索引,以用于索引第一数据集中的第一数据条目和第二数据集中的第二数据条目;如果给定匹配结果指示第一数据集中的任何数据条目的标识信息均不匹配第二数据集中的第二数据条目的标识信息,生成交集索引信息中的第一伪索引,以用于索引第一数据集中的随机数据条目和第二数据集中的第二数据条目。

[0162] 在一些实施例中,装置700还包括:第一交集生成模块,被配置为在第一参与方处,基于交集索引信息来生成第一数据集与第二数据集的第一交集,第一交集包括交集索引信

息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目;标记设置模块,被配置为设置针对第一交集中每对数据条目的匹配标记,真索引对应的至少一对数据条目的匹配标记为指示匹配,伪索引对应的至少一对数据条目的匹配标记为指示不相匹配;MPC运算模块,被配置为与第二参与方一起,利用第一交集和第二交集来执行MPC,得到第一交集中每对数据条目的候选计算结果;以及目标结果确定模块,被配置为至少基于候选计算结果和针对第一交集中每对数据条目的匹配标记来确定MPC的目标计算结果。

[0163] 在一些实施例中,真索引对应的一对数据条目的匹配标志位被设置为1,伪索引对应的一对数据条目的匹配标志位被设置为0。在一些实施例中,目标结果确定模块被配置:基于第一交集中每对数据条目的候选计算结果与第一交集中每对数据条目的匹配标记的相乘运算,生成目标计算结果。

[0164] 在一些实施例中,针对第二交集中每对数据条目的匹配标记被设置为指示不匹配,并且目标计算结果的确定还基于针对第二交集中每对数据条目的匹配标记。

[0165] 图8示出了根据本公开的一些实施例的在第二参与方处实现的数据处理装置800的示意性结构框图。装置800可以被实现为或者被包括在参与方120中。装置800中的各个模块/组件可以由硬件、软件、固件或者它们的任意组合来实现。

[0166] 如图所示,装置800包括二次加密模块810,被配置为对从MPC的第一参与方接收到的第一数据集中各个数据条目的第一加密标识信息和第一加密特征信息执行二次加密,得到第一双重加密标识信息和第一加密特征信息的第一特征分享。装置800还包括第一发送模块820,被配置为向第一参与方至少发送第一数据集中各个数据条目的第一双重加密标识信息。

[0167] 装置800还包括第一接收模块830,被配置为从第一参与方接收针对第二参与方的第二数据集中各个数据条目的第二加密特征信息的第一特征分享,而未接收到第二数据集中各个数据条目的第二双重加密标识信息。

[0168] 装置800还包括第二接收模块840,被配置为从第一参与方接收交集索引信息,交集索引信息包括对第一数据集和第二数据集中的至少一对数据条目的真索引和至少一对数据条目的伪索引,真索引对应的至少一对数据条目的标识信息相匹配。

[0169] 装置800还包括第二交集确定模块850,被配置为基于交集索引信息来确定第一数据集与第二数据集中的第二交集,第二交集包括交集索引信息中的真索引对应的至少一对数据条目和伪索引对应的至少一对数据条目。

[0170] 在一些实施例中,装置800还包括:标记设置模块,被配置为设置针对第二交集中每对数据条目的匹配标记,以指示该对数据条目的标识信息不相匹配;MPC运算模块,被配置为与第一参与方一起,利用第二交集和第一参与方确定的第一交集来执行MPC,得到第二交集中每对数据条目的候选计算结果;以及目标结果确定模块,被配置为至少基于候选计算结果和针对第二交集中每对数据条目的匹配标记来确定MPC的目标计算结果。

[0171] 在一些实施例中,第二交集中每对数据条目的匹配标记被设置为0。在一些实施例中,第一数据集中真索引对应的一对数据条目的匹配标志位被设置为1,伪索引对应的一对数据条目的匹配标志位被设置为0。

[0172] 在一些实施例中,目标结果确定模块,被配置为:基于第二交集中每对数据条目的候选计算结果与第二交集中每对数据条目的匹配标记的相乘运算,生成目标计算结果。

[0173] 图9示出了可以实施本公开的一个或多个实施例的电子设备900的框图。应当理解,图9所示出的电子设备900仅仅是示例性的,而不应当构成对本文所描述的实施例的功能和范围的任何限制。图9所示出的电子设备900可以用于实现图1的参与方110或参与方120,图7所述的装置700,或图8所述的装置800。

[0174] 如图9所示,电子设备900是通用计算设备的形式。电子设备900的组件可以包括但不限于一个或多个处理器或处理单元910、存储器920、存储设备930、一个或多个通信单元940、一个或多个输入设备950以及一个或多个输出设备960。处理单元910可以是实际或虚拟处理器并且能够根据存储器920中存储的程序来执行各种处理。在多处理器系统中,多个处理单元并行执行计算机可执行指令,以提高电子设备900的并行处理能力。

[0175] 电子设备900通常包括多个计算机存储介质。这样的介质可以是电子设备900可访问的任何可以获得的介质,包括但不限于易失性和非易失性介质、可拆卸和不可拆卸介质。存储器920可以是易失性存储器(例如寄存器、高速缓存、随机访问存储器(RAM))、非易失性存储器(例如,只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、闪存)或它们的某种组合。存储设备930可以是可拆卸或不可拆卸的介质,并且可以包括机器可读介质,诸如闪存驱动、磁盘或者任何其他介质,其可以能够用于存储信息和/或数据(例如用于训练的训练数据)并且可以在电子设备900内被访问。

[0176] 电子设备900可以进一步包括另外的可拆卸/不可拆卸、易失性/非易失性存储介质。尽管未在图9中示出,可以提供用于从可拆卸、非易失性磁盘(例如“软盘”)进行读取或写入的磁盘驱动和用于从可拆卸、非易失性光盘进行读取或写入的光盘驱动。在这些情况中,每个驱动可以由一个或多个数据介质接口被连接至总线(未示出)。存储器920可以包括计算机程序产品925,其具有一个或多个程序模块,这些程序模块被配置为执行本公开的各种实施例的各种方法或动作。

[0177] 通信单元940实现通过通信介质与其他电子设备进行通信。附加地,电子设备900的组件的功能可以以单个计算集群或多个计算机器来实现,这些计算机器能够通过通信连接进行通信。因此,电子设备900可以使用与一个或多个其他服务器、网络个人计算机(PC)或者另一个网络节点的逻辑连接来在联网环境中进行操作。

[0178] 输入设备950可以是一个或多个输入设备,例如鼠标、键盘、追踪球等。输出设备960可以是一个或多个输出设备,例如显示器、扬声器、打印机等。电子设备900还可以根据需要通过通信单元940与一个或多个外部设备(未示出)进行通信,外部设备诸如存储设备、显示设备等,与一个或多个使得用户与电子设备900交互的设备进行通信,或者与使得电子设备900与一个或多个其他电子设备通信的任何设备(例如,网卡、调制解调器等)进行通信。这样的通信可以经由输入/输出(I/O)接口(未示出)来执行。

[0179] 根据本公开的示例性实施例,提供了一种计算机可读存储介质,其上存储有计算机可执行指令,其中计算机可执行指令被处理器执行以实现上文描述的方法。根据本公开的示例性实施例,还提供了一种计算机程序产品,计算机程序产品被有形地存储在非瞬态计算机可读介质上并且包括计算机可执行指令,而计算机可执行指令被处理器执行以实现上文描述的方法。

[0180] 这里参照根据本公开实现的方法、装置、设备和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框

图中各方框的组合,都可以由计算机可读程序指令实现。

[0181] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理单元,从而生产出一种机器,使得这些指令在通过计算机或其他可编程数据处理装置的处理单元执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0182] 可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上,使得在计算机、其他可编程数据处理装置或其他设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其他可编程数据处理装置、或其他设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0183] 附图中的流程图和框图显示了根据本公开的多个实现的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0184] 以上已经描述了本公开的各实现,上述说明是示例性的,并非穷尽性的,并且也不限于所公开的各实现。在不偏离所说明的各实现的范围和精神的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实现的原理、实际应用或对市场中的技术的改进,或者使本技术领域的其他普通技术人员能理解本文公开的各个实施例。

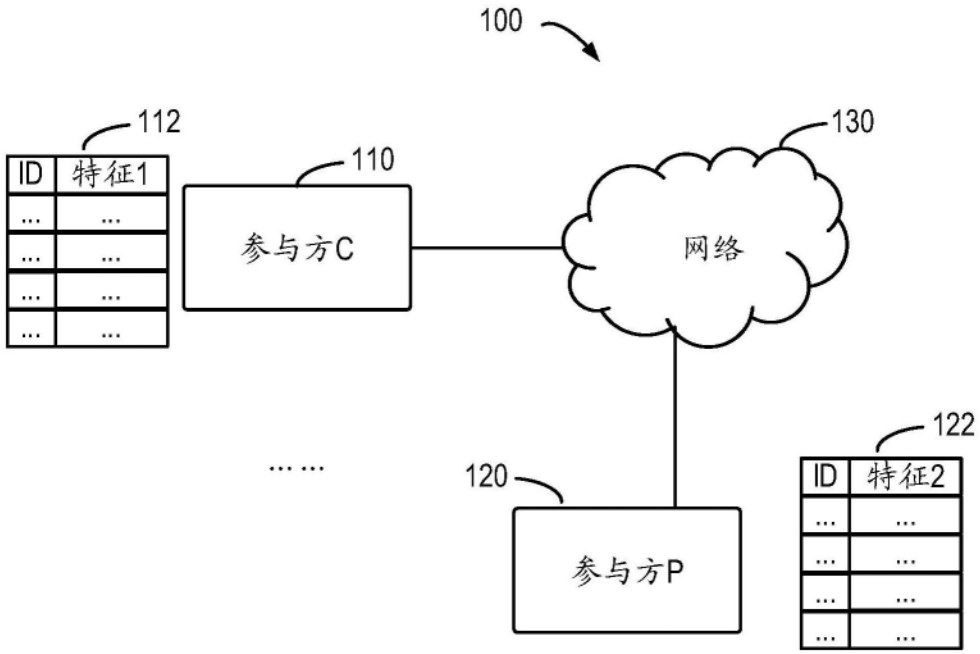


图1

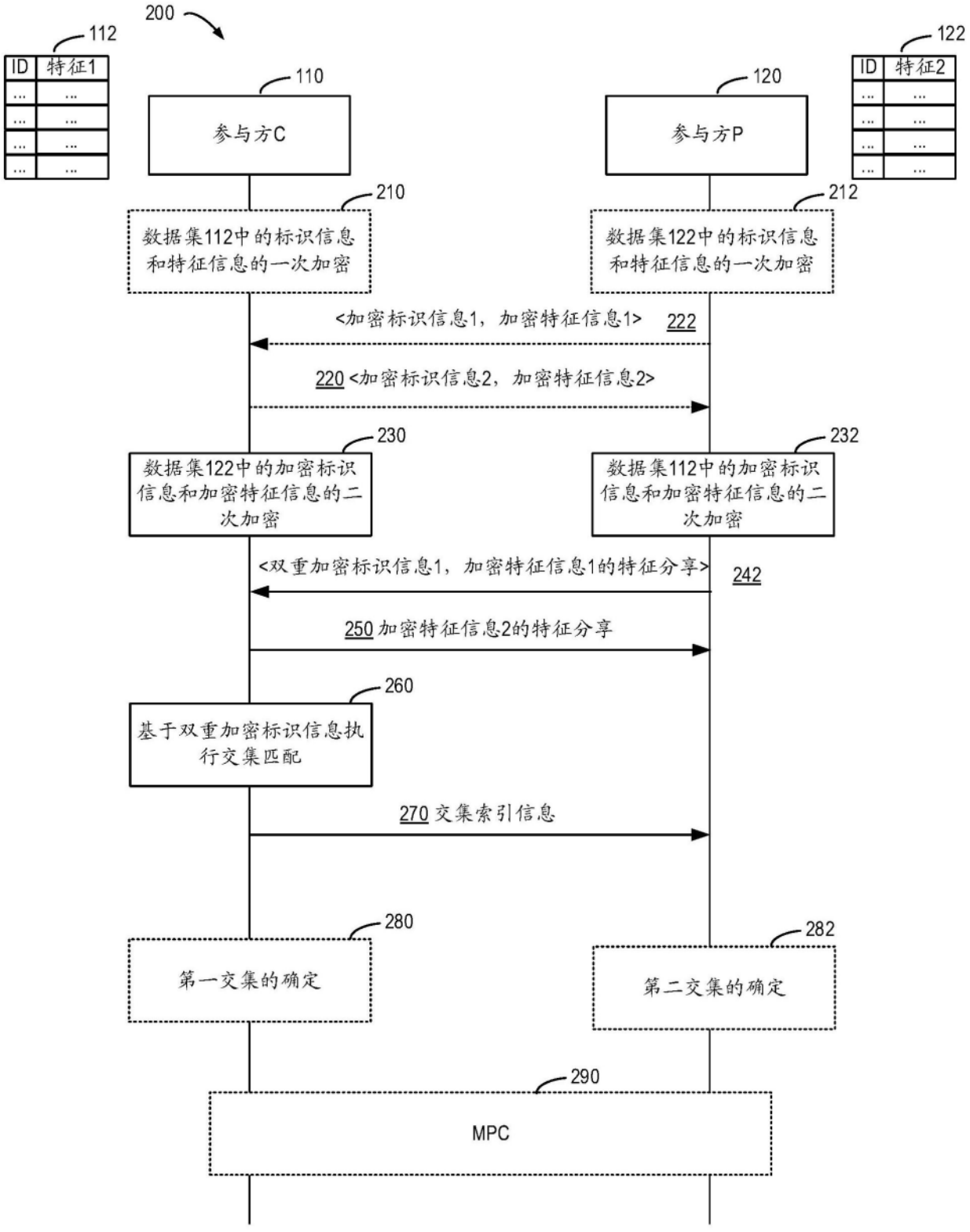

200
112
ID
特征1
122
ID
特征2
110
参与方C
120
参与方P
210
数据集112中的标识信息和特征信息的一次加密
212
数据集122中的标识信息和特征信息的一次加密
<加密标识信息1，加密特征信息1> 222
220 <加密标识信息2，加密特征信息2>
230
数据集122中的加密标识信息和加密特征信息的二次加密
232
数据集112中的加密标识信息和加密特征信息的二次加密
<双重加密标识信息1，加密特征信息1的特征分享> 242
250 加密特征信息2的特征分享
260
基于双重加密标识信息执行交集匹配
270 交集索引信息
280
第一交集的确定
282
第二交集的确定
290
MPC

图2

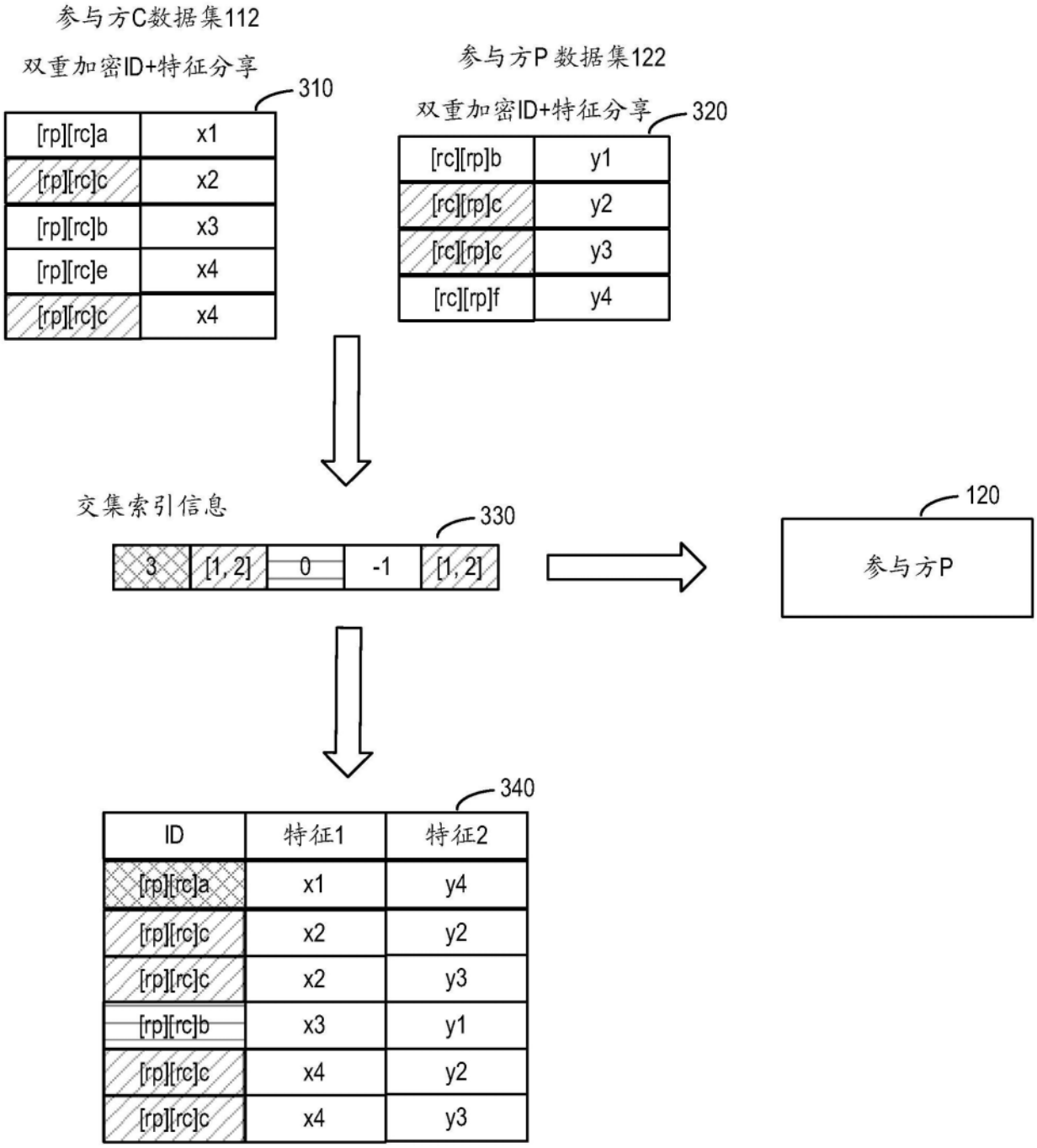

参与方C数据集112
双重加密ID+特征分享
310
[rp][rc]a x1
[rp][rc]c x2
[rp][rc]b x3
[rp][rc]e x4
[rp][rc]c x4
参与方P 数据集122
双重加密ID+特征分享
320
[rc][rp]b y1
[rc][rp]c y2
[rc][rp]c y3
[rc][rp]f y4
交集索引信息
330
3 [1, 2] 0 -1 [1, 2]
120
参与方P
340
ID 特征1 特征2
[rp][rc]a x1 y4
[rp][rc]c x2 y2
[rp][rc]c x2 y3
[rp][rc]b x3 y1
[rp][rc]c x4 y2
[rp][rc]c x4 y3

图3

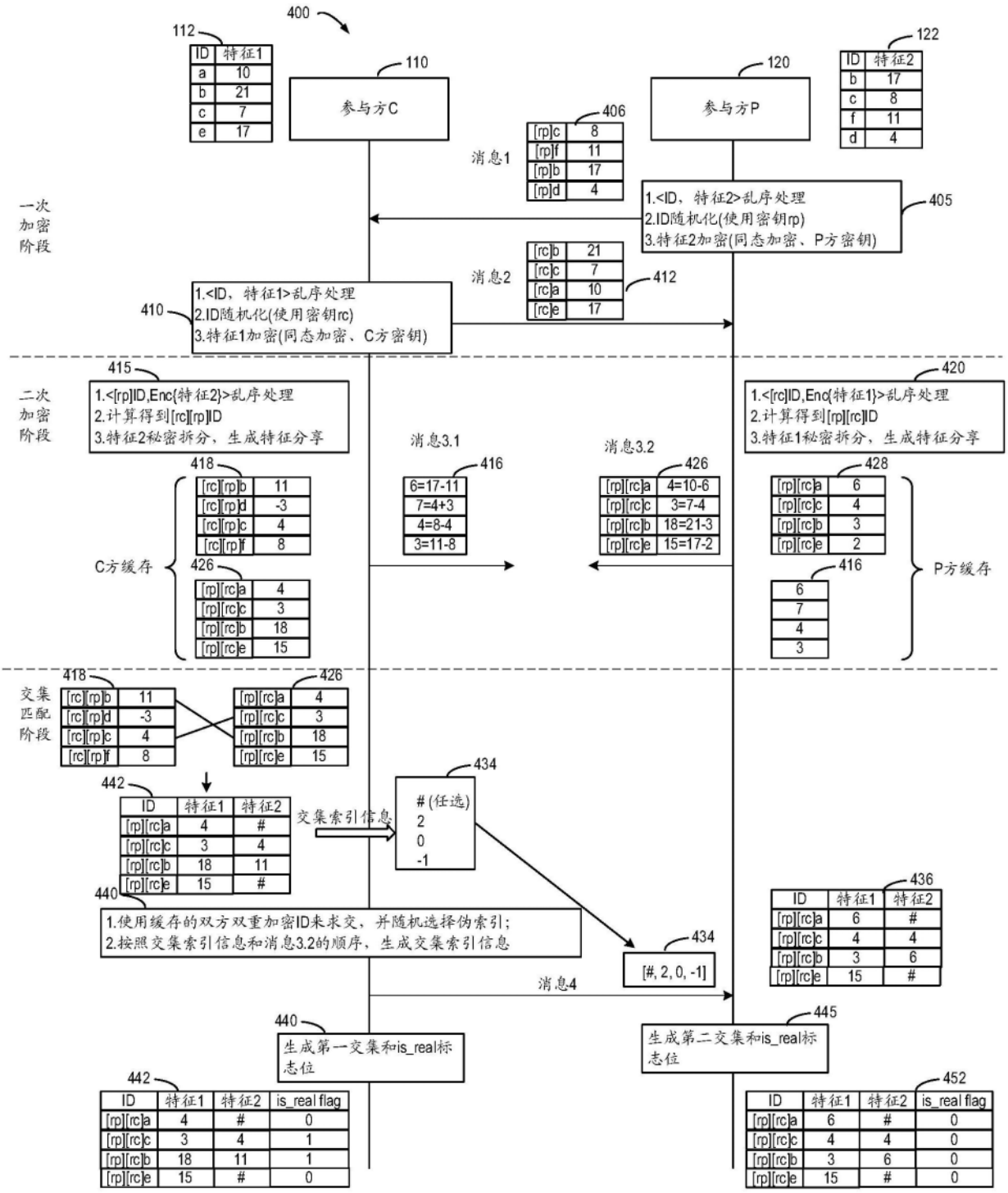

400
112
ID 特征1
a 10
b 21
c 7
e 17
110
参与方C
120
参与方P
122
ID 特征2
b 17
c 8
f 11
d 4
406
[rp]c 8
[rp]f 11
[rp]b 17
[rp]d 4
消息1
一次加密阶段
1.<ID，特征2>乱序处理
2.ID随机化(使用密钥rp)
3.特征2加密(同态加密、P方密钥)
405
[rc]b 21
[rc]c 7
[rc]a 10
[rc]e 17
消息2
412
410
1.<ID，特征1>乱序处理
2.ID随机化(使用密钥rc)
3.特征1加密(同态加密、C方密钥)
二次加密阶段
415
1.<[rp]ID,Enc{特征2}>乱序处理
2.计算得到[rc][rp]ID
3.特征2秘密拆分，生成特征分享
420
1.<[rc]ID,Enc{特征1}>乱序处理
2.计算得到[rp][rc]ID
3.特征1秘密拆分，生成特征分享
消息3.1
416
6=17-11
7=4+3
4=8-4
3=11-8
消息3.2
426
[rp][rc]a 4=10-6
[rp][rc]c 3=7-4
[rp][rc]b 18=21-3
[rp][rc]e 15=17-2
418
[rc][rp]b 11
[rc][rp]d -3
[rc][rp]c 4
[rc][rp]f 8
C方缓存
426
[rp][rc]a 4
[rp][rc]c 3
[rp][rc]b 18
[rp][rc]e 15
428
[rp][rc]a 6
[rp][rc]c 4
[rp][rc]b 3
[rp][rc]e 2
416
6
7
4
3
P方缓存
交集匹配阶段
418
[rc][rp]b 11
[rc][rp]d -3
[rc][rp]c 4
[rc][rp]f 8
426
[rp][rc]a 4
[rp][rc]c 3
[rp][rc]b 18
[rp][rc]e 15
442
ID 特征1 特征2
[rp][rc]a 4 #
[rp][rc]c 3 4
[rp][rc]b 18 11
[rp][rc]e 15 #
交集索引信息
434
#(任选)
2
0
-1
440
1.使用缓存的双方双重加密ID来求交，并随机选择伪索引；
2.按照交集索引信息和消息3.2的顺序，生成交集索引信息
434
[#, 2, 0, -1]
436
ID 特征1 特征2
[rp][rc]a 6 #
[rp][rc]c 4 4
[rp][rc]b 3 6
[rp][rc]e 15 #
消息4
440
生成第一交集和is_real标志位
445
生成第二交集和is_real标志位
442
ID 特征1 特征2 is_real flag
[rp][rc]a 4 # 0
[rp][rc]c 3 4 1
[rp][rc]b 18 11 1
[rp][rc]e 15 # 0
452
ID 特征1 特征2 is_real flag
[rp][rc]a 6 # 0
[rp][rc]c 4 4 0
[rp][rc]b 3 6 0
[rp][rc]e 15 # 0

图4

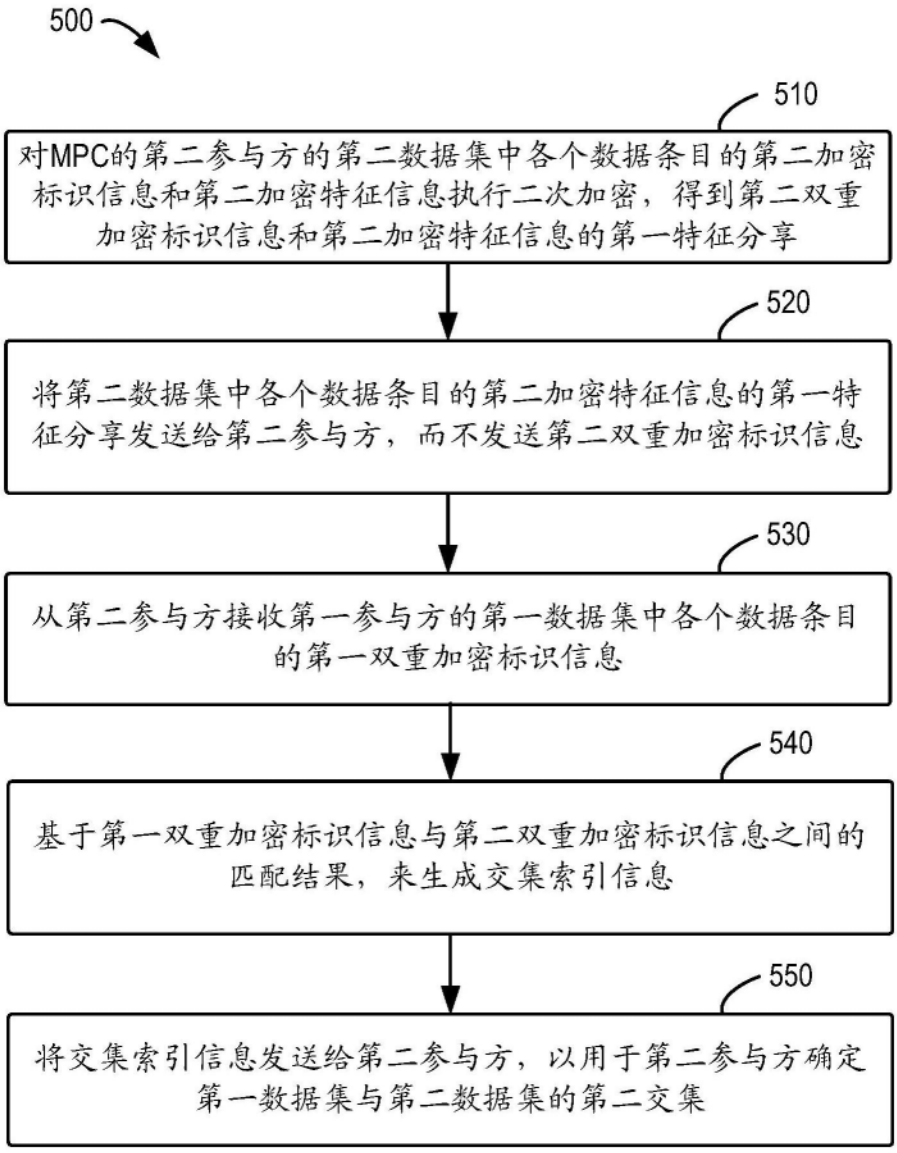

500
510
对MPC的第二参与方的第二数据集中各个数据条目的第二加密标识信息和第二加密特征信息执行二次加密，得到第二双重加密标识信息和第二加密特征信息的第一特征分享
520
将第二数据集中各个数据条目的第二加密特征信息的第一特征分享发送给第二参与方，而不发送第二双重加密标识信息
530
从第二参与方接收第一参与方的第一数据集中各个数据条目的第一双重加密标识信息
540
基于第一双重加密标识信息与第二双重加密标识信息之间的匹配结果，来生成交集索引信息
550
将交集索引信息发送给第二参与方，以用于第二参与方确定第一数据集与第二数据集的第二交集

图5

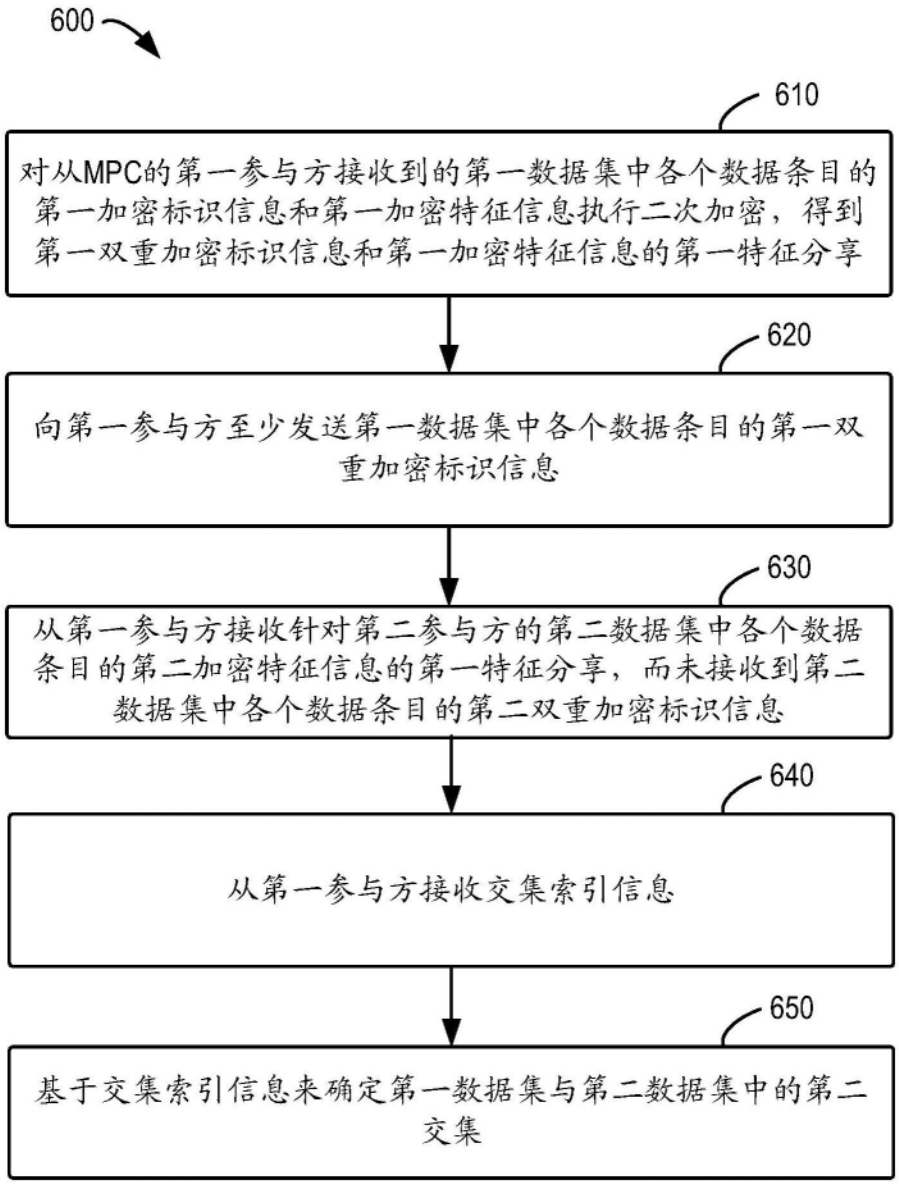

600
610
对从MPC的第一参与方接收到的第一数据集中各个数据条目的第一加密标识信息和第一加密特征信息执行二次加密，得到第一双重加密标识信息和第一加密特征信息的第一特征分享
620
向第一参与方至少发送第一数据集中各个数据条目的第一双重加密标识信息
630
从第一参与方接收针对第二参与方的第二数据集中各个数据条目的第二加密特征信息的第一特征分享，而未接收到第二数据集中各个数据条目的第二双重加密标识信息
640
从第一参与方接收交集索引信息
650
基于交集索引信息来确定第一数据集与第二数据集中的第二交集

图6

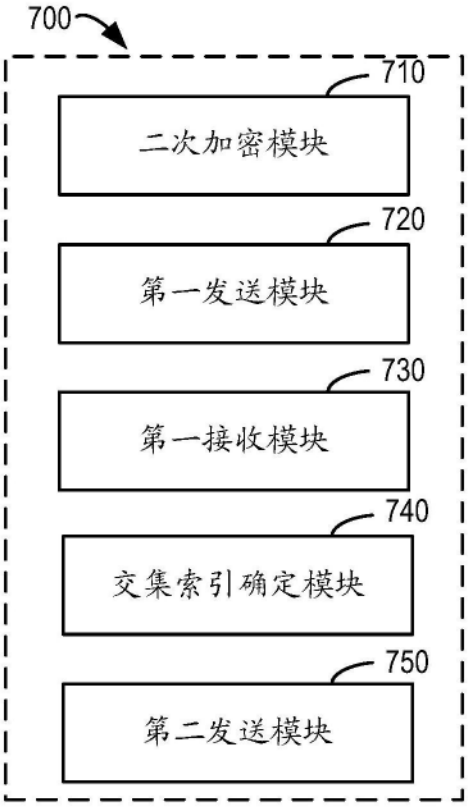

700
710
二次加密模块
720
第一发送模块
730
第一接收模块
740
交集索引确定模块
750
第二发送模块

图7

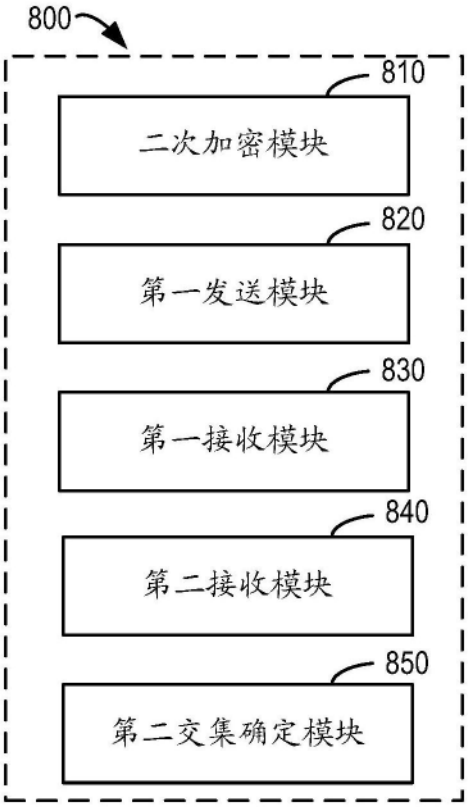

800
810
二次加密模块
820
第一发送模块
830
第一接收模块
840
第二接收模块
850
第二交集确定模块

图8

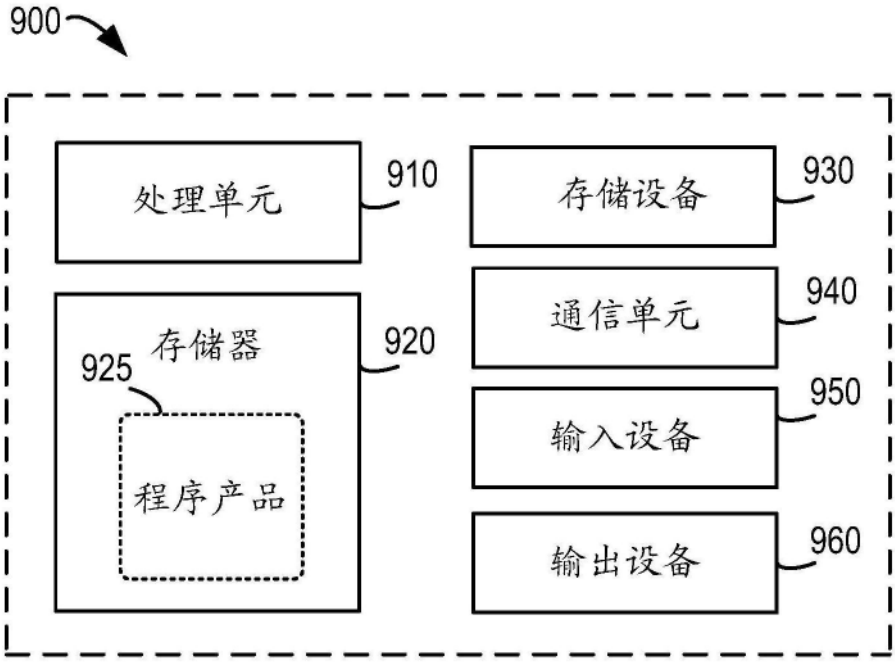

900
处理单元
910
存储设备
930
存储器
925
920
程序产品
通信单元
940
输入设备
950
输出设备
960

图9