(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115982779 B
(45)授权公告日 2023.05.23

(21)申请号 202310258354.5
(22)申请日 2023.03.17
(65)同一申请的已公布的文献号
申请公布号 CN 115982779 A
(43)申请公布日 2023.04.18
(73)专利权人 北京富算科技有限公司
地址 100020 北京市朝阳区东三环中路9号19层2201
(72)发明人 赵东 卞阳 尤志强
(74)专利代理机构 北京超凡宏宇专利代理事务所(特殊普通合伙) 11463
专利代理师 周宇
(51)Int.Cl.
G06F 21/62(2013.01)
G06N 20/00(2019.01)

(56)对比文件
CN 115392480 A,2022.11.25
CN 115730182 A,2023.03.03

审查员 王东

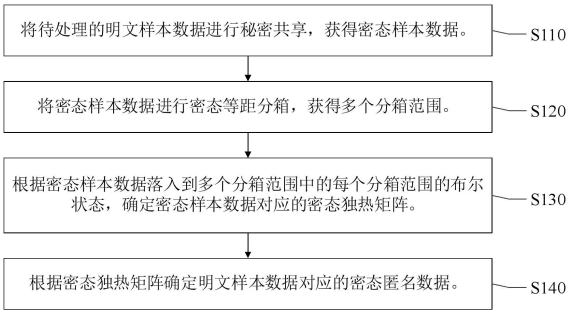
权利要求书2页 说明书12页 附图3页

(54)发明名称

一种数据匿名化方法、装置、电子设备及存储介质

(57)摘要

本申请提供一种数据匿名化方法、装置、电子设备及存储介质，本申请涉及机器学习、联邦学习和隐私保护的技术领域，用于改善联邦学习的过程中使用的模型训练数据容易泄露个人隐私的问题。该方法通过将秘密共享的密态样本数据进行密态等距分箱，并使用密态等距分箱的手段对秘密共享后的密态样本数据进行匿名化，以使数据被替换成证据权重向量，从而获得匿名化的密态匿名数据，以此来减小泄露个人隐私的风险。该数据匿名化方法主要用于联邦机器学习或者隐私集合求交等场景。



CN 115982779 B

1.一种数据匿名化方法,其特征在于,包括:

将待处理的明文样本数据进行秘密共享,获得密态样本数据;

将所述密态样本数据进行密态等距分箱,获得多个分箱范围;

根据所述密态样本数据落入到所述多个分箱范围中的每个分箱范围的布尔状态,确定所述密态样本数据对应的密态独热矩阵;

根据所述密态独热矩阵确定所述明文样本数据对应的密态匿名数据;

其中,所述根据所述密态样本数据落入到所述多个分箱范围中的每个分箱范围的布尔状态,确定所述密态样本数据对应的密态独热矩阵,包括:对所述密态样本数据落入到所述多个分箱范围中的布尔状态,获得所述密态样本数据对应的布尔矩阵;将所述布尔矩阵转换为所述密态独热矩阵;

所述根据所述密态独热矩阵确定所述明文样本数据对应的密态匿名数据,包括:计算出所述多个分箱范围中的每个分箱范围的证据权重,获得证据权重向量;根据所述密态独热矩阵和所述证据权重向量,确定所述明文样本数据对应的密态匿名数据。

2.根据权利要求1所述的方法,其特征在于,所述将所述密态样本数据进行密态等距分箱,包括:

统计出所述密态样本数据中的交集最大值和交集最小值,并根据所述交集最大值和所述交集最小值确定所述密态样本数据的样本区间;

根据预设分箱数量对所述密态样本数据的样本区间进行密态等距分箱。

3.根据权利要求1所述的方法,其特征在于,在所述根据所述密态独热矩阵确定所述明文样本数据对应的密态匿名数据之后,还包括:

使用所述密态匿名数据对机器学习模型进行联邦学习。

4.根据权利要求3所述的方法,其特征在于,所述使用所述密态匿名数据对机器学习模型进行联邦学习,包括:

将待处理的样本标签进行秘密共享,获得密态标签数据,所述样本标签是所述样本数据的类别标签;

使用所述密态匿名数据和所述密态标签数据对机器学习模型进行联邦学习。

5.根据权利要求1-4任一所述的方法,其特征在于,在所述根据所述密态独热矩阵确定所述明文样本数据对应的密态匿名数据之后,还包括:

使用秘密共享密码机制中的门限方案对所述密态匿名数据进行恢复,获得明文匿名数据;

使用所述明文匿名数据对机器学习模型进行联邦学习。

6.一种数据匿名化装置,其特征在于,包括:

样本数据获得模块,用于将待处理的明文样本数据进行秘密共享,获得密态样本数据;

分箱范围获得模块,用于将所述密态样本数据进行密态等距分箱,获得多个分箱范围;

独热矩阵确定模块,用于根据所述密态样本数据落入到所述多个分箱范围中的每个分箱范围的布尔状态,确定所述密态样本数据对应的密态独热矩阵;

匿名数据确定模块,用于根据所述密态独热矩阵确定所述明文样本数据对应的密态匿名数据;

其中,所述根据所述密态样本数据落入到所述多个分箱范围中的每个分箱范围的布尔

状态，确定所述密态样本数据对应的密态独热矩阵，包括：对所述密态样本数据落入到所述多个分箱范围中的布尔状态，获得所述密态样本数据对应的布尔矩阵；将所述布尔矩阵转换为所述密态独热矩阵；

所述根据所述密态独热矩阵确定所述明文样本数据对应的密态匿名数据，包括：计算出所述多个分箱范围中的每个分箱范围的证据权重，获得证据权重向量；根据所述密态独热矩阵和所述证据权重向量，确定所述明文样本数据对应的密态匿名数据。

7.一种电子设备，其特征在于，包括：处理器和存储器，所述存储器存储有所述处理器可执行的机器可读指令，所述机器可读指令被所述处理器执行时执行如权利要求1至5任一所述的方法。

8.一种计算机可读存储介质，其特征在于，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行如权利要求1至5任一所述的方法。

一种数据匿名化方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及机器学习、联邦学习和隐私保护的技术领域，具体而言，涉及一种数据匿名化方法、装置、电子设备及存储介质。

背景技术

[0002] 联邦机器学习(Federated Machine Learning，FML)，又被称为联邦学习(Federated Learning)、联合学习或者联盟学习，是一种机器学习框架，该机器学习框架能有效地让多个机构在满足用户隐私保护、数据安全和法律法规的要求下，进行多方使用各自的数据协作训练出一个机器学习模型。

[0003] 目前在联邦学习的过程中，虽然对原始数据进行了同态加密或安全多方计算(即秘密共享)，并使用同态加密或秘密共享后的数据进行模型训练；然而，参与联邦学习的电子设备仍然可以从同态加密后的数据解密出原始数据，或者，当秘密共享后的碎片数据足够多时，就可以还原出原始数据。进一步地，电子设备还可以从原始数据中分析出个人隐私数据，因此，目前的联邦学习的过程中使用的模型训练数据容易泄露个人隐私。

发明内容

[0004] 本申请实施例的目的在于提供一种数据匿名化方法、装置、电子设备及存储介质，用于改善联邦学习的过程中使用的模型训练数据容易泄露个人隐私的问题。

[0005] 本申请实施例提供了一种数据匿名化方法，包括：将待处理的明文样本数据进行秘密共享，获得密态样本数据；将密态样本数据进行密态等距分箱，获得多个分箱范围；根据密态样本数据落入到多个分箱范围中的每个分箱范围的布尔状态，确定密态样本数据对应的密态独热矩阵；根据密态独热矩阵确定明文样本数据对应的密态匿名数据。在上述方案的实现过程中，通过将明文样本数据秘密共享获得的密态样本数据进行密态等距分箱，并根据分箱后的密态独热矩阵来确定该明文样本数据对应的密态匿名数据，也就是说，通过使用密态等距分箱的手段对秘密共享后的密态样本数据进行匿名化，以使得数据被替换成证据权重向量，从而获得匿名化的密态匿名数据，因此使用密态匿名数据可以有效减小泄露个人隐私的风险。

[0006] 可选地，在本申请实施例中，将密态样本数据进行密态等距分箱，包括：统计出密态样本数据中的交集最大值和交集最小值，并根据交集最大值和交集最小值确定密态样本数据的样本区间；根据预设分箱数量对密态样本数据的样本区间进行密态等距分箱。在上述方案的实现过程中，通过根据预设分箱数量对密态样本数据的样本区间进行密态等距分箱，并根据分箱后的密态独热矩阵来确定该明文样本数据对应的密态匿名数据，也就是说，通过使用密态等距分箱的手段对秘密共享后的密态样本数据进行匿名化，以使得数据被替换成证据权重向量，从而获得匿名化的密态匿名数据，因此使用密态匿名数据可以有效减小泄露个人隐私的风险。

[0007] 可选地，在本申请实施例中，根据密态样本数据落入到多个分箱范围中的每个分

箱范围的布尔状态，确定密态样本数据对应的密态独热矩阵，包括：对密态样本数据落入到多个分箱范围中的布尔状态，获得密态样本数据对应的布尔矩阵；将布尔矩阵转换为密态独热矩阵。在上述方案的实现过程中，通过对密态样本数据落入到多个分箱范围中的布尔状态，获得密态样本数据对应的布尔矩阵，并将布尔矩阵转换为密态独热矩阵，该密态独热矩阵用于确定该明文样本数据对应的密态匿名数据，也就是说，通过使用密态等距分箱的手段对秘密共享后的密态样本数据进行匿名化，以使得数据被替换成证据权重向量，从而获得匿名化的密态匿名数据，因此使用密态匿名数据可以有效减小泄露个人隐私的风险。

[0008] 可选地，在本申请实施例中，根据密态独热矩阵确定明文样本数据对应的密态匿名数据，包括：计算出多个分箱范围中的每个分箱范围的证据权重，获得证据权重向量；根据密态独热矩阵和证据权重向量，确定明文样本数据对应的密态匿名数据。在上述方案的实现过程中，通过WOE算法计算出多个分箱范围中的每个分箱范围的证据权重，获得证据权重向量，并根据密态独热矩阵和证据权重向量，从而通过WOE算法可以安全有效的保护这些信息不被泄漏。

[0009] 可选地，在本申请实施例中，在根据密态独热矩阵确定明文样本数据对应的密态匿名数据之后，还包括：使用密态匿名数据对机器学习模型进行联邦学习。在上述方案的实现过程中，通过使用密态匿名数据对机器学习模型进行联邦学习，从而在不暴露敏感信息的情况下，同时又满足客户数据流通赋能的需求，有效地增加了联邦学习过程中的数据安全性。

[0010] 可选地，在本申请实施例中，使用密态匿名数据对机器学习模型进行联邦学习，包括：将待处理的样本标签进行秘密共享，获得密态标签数据，样本标签是样本数据的类别标签；使用密态匿名数据和密态标签数据对机器学习模型进行联邦学习。在上述方案的实现过程中，通过使用密态匿名数据和密态标签数据对机器学习模型进行联邦学习，从而在不暴露敏感信息的情况下，同时又满足客户数据流通赋能的需求，有效地增加了联邦学习过程中的数据安全性。

[0011] 可选地，在本申请实施例中，在根据密态独热矩阵确定明文样本数据对应的密态匿名数据之后，还包括：使用秘密共享密码机制中的门限方案对密态匿名数据进行恢复，获得明文匿名数据；使用明文匿名数据对机器学习模型进行联邦学习。在上述方案的实现过程中，通过使用秘密共享密码机制中的门限方案对密态匿名数据进行恢复，获得明文匿名数据，并使用明文匿名数据对机器学习模型进行联邦学习，从而在不暴露敏感信息的情况下，同时又满足客户数据流通赋能的需求，有效地增加了联邦学习过程中的数据安全性。

[0012] 本申请实施例还提供了一种数据匿名化装置，包括：样本数据获得模块，用于将待处理的明文样本数据进行秘密共享，获得密态样本数据；分箱范围获得模块，用于将密态样本数据进行密态等距分箱，获得多个分箱范围；独热矩阵确定模块，用于根据密态样本数据落入到多个分箱范围中的每个分箱范围的布尔状态，确定密态样本数据对应的密态独热矩阵；匿名数据确定模块，用于根据密态独热矩阵确定明文样本数据对应的密态匿名数据。

[0013] 可选地，在本申请实施例中，分箱范围获得模块，包括：样本区间确定子模块，用于统计出密态样本数据中的交集最大值和交集最小值，并根据交集最大值和交集最小值确定密态样本数据的样本区间；样本区间分箱子模块，用于根据预设分箱数量对密态样本数据的样本区间进行密态等距分箱。

[0014] 可选地，在本申请实施例中，独热矩阵确定模块，包括：布尔矩阵获得子模块，用于对密态样本数据落入到多个分箱范围中的布尔状态，获得密态样本数据对应的布尔矩阵；布尔矩阵转换子模块，用于将布尔矩阵转换为密态独热矩阵。

[0015] 可选地，在本申请实施例中，匿名数据确定模块，包括：权重向量获得子模块，用于计算出多个分箱范围中的每个分箱范围的证据权重，获得证据权重向量；密态数据确定子模块，用于根据密态独热矩阵和证据权重向量，确定明文样本数据对应的密态匿名数据。

[0016] 可选地，在本申请实施例中，数据匿名化装置，还包括：第一联邦学习模块，用于使用密态匿名数据对机器学习模型进行联邦学习。

[0017] 可选地，在本申请实施例中，第一联邦学习模块，包括：标签数据获得子模块，用于将待处理的样本标签进行秘密共享，获得密态标签数据，样本标签是样本数据的类别标签；模型联邦学习子模块，用于使用密态匿名数据和密态标签数据对机器学习模型进行联邦学习。

[0018] 可选地，在本申请实施例中，数据匿名化装置，还包括：匿名数据获得模块，用于使用秘密共享密码机制中的门限方案对密态匿名数据进行恢复，获得明文匿名数据；第二联邦学习模块，用于使用明文匿名数据对机器学习模型进行联邦学习。

[0019] 本申请实施例还提供了一种电子设备，包括：处理器和存储器，存储器存储有处理器可执行的机器可读指令，机器可读指令被处理器执行时执行如上面描述的方法。

[0020] 本申请实施例还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，该计算机程序被处理器运行时执行如上面描述的方法。

[0021] 本申请实施例的其他特征和优点将在随后的说明书阐述，并且，部分地从说明书中变得显而易见，或者通过实施本申请实施例了解。

附图说明

[0022] 为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本申请实施例中的某些实施例，因此不应被看作是对范围的限定，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他相关的附图。

[0023] 图1示出的本申请实施例提供的数据匿名化方法的流程示意图；

[0024] 图2示出的本申请实施例提供的处理明文样本数据的示意图；

[0025] 图3示出的本申请实施例提供的数据匿名化装置的结构示意图；

[0026] 图4示出的本申请实施例提供的电子设备的结构示意图。

具体实施方式

[0027] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，应当理解，本申请实施例中的附图仅起到说明和描述的目的，并不用于限定本申请实施例的保护范围。另外，应当理解，示意性的附图并未按实物比例绘制。本申请实施例中使用的流程图示出了根据本申请实施例的一些实施例实现的操作。应该理解，流程图的操作可以不按顺序实现，没有逻辑的上下文关系的步骤可以反转顺序或者同时实施。此外，本领域技术人员在本申请实施例内

容的指引下,可以向流程图添加一个或多个其他操作,也可以从流程图中移除一个或多个操作。

[0028] 另外,所描述的实施例仅仅是本申请实施例的一部分,而不是全部的实施例。通常在此处附图中描述和示出的本申请实施例的组件可以以各种不同的配置来布置和设计。因此,以下对在附图中提供的本申请实施例的详细描述并非旨在限制要求保护的本申请实施例的范围,而是仅仅表示本申请实施例的选定实施例。

[0029] 可以理解的是,本申请实施例中的“第一”、“第二”用于区别类似的对象。本领域技术人员可以理解“第一”、“第二”等字样并不对数量和执行次序进行限定,并且“第一”、“第二”等字样也并不限定一定不同。在本申请实施例的描述中,术语“和/或”仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。术语“多个”指的是两个以上(包括两个),同理,“多组”指的是两组以上(包括两组)。

[0030] 在介绍本申请实施例提供的数据匿名化方法之前,先介绍本申请实施例中所涉及的一些概念:

[0031] 隐私集合求交(Private Set Intersection,PSI),又被称为安全求交,也称为隐私保护集合交集协议,是纵向联邦学习的一部分运算过程,PSI协议允许持有各自集合的两方来共同计算两个集合的交集运算。在协议交互的最后,两方应该得到正确的交集,而且不会得到交集以外另一方集合中的任何信息。

[0032] 纵向联邦学习,是指当两个参与方的用户重叠部分很多,但是两个数据集的用户特征重叠部分比较少时的场景下,训练出联邦学习模型,具体例如:同一个地区的两个机构,一个机构有用户的消费记录,另一个机构有用户的银行记录,两个机构有很多重叠用户(即需要在不暴露交集以外的任何集合信息的情况下进行隐私集合求交,从而获得重叠用户),但是记录的数据特征是不同的,两个机构想通过加密聚合用户的不同特征来联合训练一个更强大的联邦学习模型。

[0033] 多方安全计算(Multi-Party Computation,MPC),又被称为安全多方计算(Secure Multi-Party Computation,SMPC),MPC的研究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。

[0034] 秘密共享(Secret sharing),又称秘密分享或者秘密拆分(Secret splitting),是将秘密分散到多方节点的方法,每人得到秘密的一部分,称为份额(Share)。只有当足够多种的份额结合时,才能还原出秘密原文,每个份额各自则没有用途。

[0035] 需要说明的是,本申请实施例提供的数据匿名化方法可以被电子设备执行,这里的电子设备是指具有执行计算机程序功能的设备终端或者服务器,设备终端例如:智能手机、个人电脑、平板电脑、个人数字助理或者移动上网设备等。服务器是指通过网络提供计算服务的设备,服务器例如:x86服务器以及非x86服务器,非x86服务器包括:大型机、小型机和UNIX服务器。

[0036] 下面介绍该数据匿名化方法适用的应用场景,这里的应用场景包括但不限于:在全匿踪联邦学习的隐私集合求交的过程中使用该数据匿名化方法,以两方场景进行阐述(实际可以扩展到多方场景),具体例如:使用训练数据来训练模型之前,可以使用该数据匿

名化方法对训练数据进行匿名化等，减小训练数据泄露个人隐私的风险，使得训练数据在联邦学习的流通过程中更加符合隐私相关的规定。

[0037] 请参见图1示出的本申请实施例提供的数据匿名化方法的流程示意图；该数据匿名化方法的主要思路是，通过将明文样本数据秘密共享获得的密态样本数据进行密态等距分箱，使用密态等距分箱的手段对秘密共享后的密态样本数据进行匿名化，以使得数据被替换成证据权重向量，从而获得匿名化的密态匿名数据，以此来减小泄露个人隐私的风险。上述数据匿名化方法的实施方式可以包括：

[0038] 步骤S110：将待处理的明文样本数据进行秘密共享，获得密态样本数据。

[0039] 上述步骤S110的实施方式例如：使用三大类别中的任一类别的秘密分享协议来将待处理的明文样本数据进行秘密共享，获得密态样本数据。其中，三大类别的秘密分享协议可以包括：门限秘密分享(Threshold Secret Sharing Scheme)、一般访问结构的秘密分享(General Secret Sharing Scheme) 以及介于二者之间的面向特殊访问结构的秘密分享协议等。此处以门限秘密分享类别为例进行说明，可以使用的秘密分享协议包括：基于多项式的Shamir协议、基于超平面的Blakley协议、基于中国剩余定理的Mignotee协议、Asumth-Bloom协议、Brickell协议、基于矩阵投影的秘密共享协议、Arithmetic分享协议、Boolean分享协议和Yao分享协议等。

[0040] 步骤S120：将密态样本数据进行密态等距分箱，获得多个分箱范围。

[0041] 步骤S130：根据密态样本数据落入到多个分箱范围中的每个分箱范围的布尔状态，确定密态样本数据对应的密态独热矩阵。

[0042] 步骤S140：根据密态独热矩阵确定明文样本数据对应的密态匿名数据。

[0043] 在上述的实现过程中，通过将明文样本数据秘密共享获得的密态样本数据进行密态等距分箱，并根据分箱后的密态独热矩阵来确定该明文样本数据对应的密态匿名数据，也就是说，使用密态等距分箱的手段对秘密共享后的密态样本数据进行匿名化，以使数据被替换成证据权重向量，从而获得匿名化的密态匿名数据，因此使用密态匿名数据可以有效减小泄露个人隐私的风险。

[0044] 请参见图2示出的本申请实施例提供的处理明文样本数据的示意图；可以理解的是，该数据匿名化方法可以应用在多方场景中，为了便于理解和说明，此处以两方场景进行阐述，具体例如：两方的明文样本数据可以包括第一方的明文样本数据和第二方的明文样本数据，第一方的明文样本数据可以包括：标识数据id、标签数据Y和特征数据X_{a1}、X_{a2}、X_{a3}，同理地，第二方的明文样本数据可以包括标识数据id、特征数据X_{b1}和X_{b2}。

[0045] 作为上述步骤S110的一种可选实施方式，在将待处理的明文样本数据进行秘密共享之前，还可以先对明文样本数据中的标识(ID)数据进行匿名化操作，以两方场景为例进行说明，具体例如：对明文样本数据中的标识(ID)数据进行哈希计算，获得哈希字符串，然后将哈希字符串转换为数字串(例如第一方明文数据的中的124360)，获得匿名化后的标识(ID)数据。双方的标识数据在经过哈希计算之后，就可以获得双方的哈希字符串，然后根据双方的哈希字符串对第一方的明文数据和第二方的明文数据进行秘密共享和对齐，获得秘密共享和对齐后的密态样本数据，具体例如：将双方的哈希字符串进行对比，即可获得双方的交集数据(即正样本数据)和非交集数据(即负样本数据)，图中的Storted-id相同的数据记录就是交集数据(即正样本数据)，图中的Storted-id不相同的数据记录就是非交集数据

(即负样本数据)。在具体的实践过程中,还可以增加额外的样本类型字段,该样本类型字段来标记正样本数据和负样本数据。

[0046]

排序标识(*Sorted-id*)	*Y*	*B*
<124360>	<-1>	<False>
<328492>	<1>	<True>
<328492>	<-1>	<False>
<572683>	<-1>	<False>
<748329>	<-1>	<False>
<930913>	<0>	<True>
<930913>	<-1>	<False>

[0047] 表1

[0048] 请参照上面表1,该表格示出了本申请实施例提供的匿名化后的标签数据的示意图;可选地,在将待处理的明文样本数据进行秘密共享之后,还可以先对明文样本数据中的特征(Feature)数据和标签数据进行匿名化操作,具体例如:使用非线性的标准化方法、非线性的归一化方法或者随机编码方式对明文样本数据中的特征(Feature)数据和标签数据进行匿名化操作,获得匿名化后的特征(Feature)数据。又例如:使用公式 $\langle X\rangle_j^{psi}=(B2A(\tilde{B}))\times(-9999)+\langle X\rangle_j^{psi}$ 来计算出匿名化后的特征(Feature)数据中的最大值,其中,$\langle X\rangle_j^{psi}$ 表示匿名化前的特征数据或者匿名化后的特征数据,$\tilde{B}$ 表示标签数据的真实值(例如1或者0),*B2A*表示使用Boolean分享协议和Arithmetic分享协议进行密码共享操作。请参照下面表2,该表格示出了本申请实施例提供的匿名化后的特征(Feature)数据中计算最大值时的结果。

[0049]

排序标识(*Sorted-id*)	X_{a1}	X_{a2}	X_{a3}	X_{b1}	X_{b2}	*B*
<124360>	<-9999>	<-9999>	<-9999>	<-9999>	<-9999>	<False>
<328492>	<2.5>	<3.3>	<-2>	<2.3>	<1.9>	<True>
<328492>	<-9999>	<-9999>	<-9999>	<-9999>	<-9999>	<False>
<572683>	<-9999>	<-9999>	<-9999>	<-9999>	<-9999>	<False>
<748329>	<-9999>	<-9999>	<-9999>	<-9999>	<-9999>	<False>
<930913>	<0.9>	<0.12>	<1>	<-1.2>	<-0.1>	<True>
<930913>	<-9999>	<-9999>	<-9999>	<-9999>	<-9999>	<False>

[0050] 表2

[0051] 可以理解的是,也可以使用公式 $\langle X\rangle_j^{psi}=(B2A(\tilde{B}))\times(9999)+\langle X\rangle_j^{psi}$ 计算匿名化后的特征数据中的最小值,$\langle Y\rangle_j^{psi}=(B2A(\tilde{B}))\times(-1)+\langle Y\rangle_j^{psi}$ 对标签数据进行处理,获得匿名化

后的标签数据，其中，$\langle X\rangle_j^{psi}$ 表示匿名化前的特征数据或者匿名化后的特征数据，$\langle Y\rangle_j^{psi}$ 表示匿名化前的标签数据或者匿名化后的标签数据的标签值，$\tilde{B}$ 表示标签数据的二进制值(例如1或者0)，*B2A*表示使用Boolean分享协议和Arithmetic分享协议进行密码共享操作。

[0052]

排序标识（*Sorted-id*）	X_{a1}	X_{a2}	X_{a3}	X_{b1}	X_{b2}	B
<124360>	<9999>	<9999>	<9999>	<9999>	<9999>	<False>
<328492>	<2.5>	<3.3>	<-2>	<2.3>	<1.9>	<True>
<328492>	<9999>	<9999>	<9999>	<9999>	<9999>	<False>
<572683>	<9999>	<9999>	<9999>	<9999>	<9999>	<False>
<748329>	<9999>	<9999>	<9999>	<9999>	<9999>	<False>
<930913>	<0.9>	<0.12>	<1>	<-1.2>	<-0.1>	<True>
<930913>	<9999>	<9999>	<9999>	<9999>	<9999>	<False>

[0053] 表3

[0054] 请参照上面表3，该表格示出了本申请实施例提供的统计密态样本数据中计算最小值时的结果。作为上述步骤S120的一种可选实施方式，在将密态样本数据进行密态等距分箱时，可以先确定样本区间，然后再对样本区间进行密态等距分箱，该实施方式可以包括：

[0055] 步骤S121：统计出密态样本数据中的交集最大值和交集最小值，并根据交集最大值和交集最小值确定密态样本数据的样本区间。

[0056] 可选地，可以针对特征数据部分统计出密态样本数据中的交集最大值和交集最小值，从而减小非交集部分特征数据对结果的影响，例如可以将非交集的特征数据替换为固定值，可以将统计交集最大值时，将非交集部分的特征数据全部设置为-9999，当然也可以根据数据的具体情况调整该固定值。在上述的实现过程中，通过将非交集的特征数据替换为固定值(即数据采样构造)，以使得替换后的数据即使被还原恢复成明文，也没法定位到特定自然人，同时也不泄漏交集数据大小，有效地提高了数据的安全性。

[0057] 步骤S122：根据预设分箱数量对密态样本数据的样本区间进行密态等距分箱，获得多个分箱范围。

[0058]

分割点	分箱范围	
<-9.33361816>	左（left）	右（right）
<-7.41809082>	<-9.33361816>	<-7.41809082>
<-5.50256348>	<-7.41809082>	<-5.50256348>
<-3.58703613>	<-5.50256348>	<-3.58703613>
<-1.67150879>	<-3.58703613>	<-1.67150879>
<-0.24401855>	<-1.67150879>	<-0.24401855>
<2.1595459>	<-0.24401855>	<2.1595459>
<4.07507324>	<2.1595459>	<4.07507324>
<5.99060059>	<4.07507324>	<5.99060059>
<7.90612793>	<5.99060059>	<7.90612793>
<9.82714844>	<7.90612793>	<9.82714844>

[0059] 表4

[0060] 请参照上面表4，该表格示出了本申请实施例提供的等距分箱后的分割点和分箱范围；上述步骤S121至步骤S122的实施方式例如：使用预设编程语言编译或者解释的可执行程序统计出密态样本数据中的交集最大值和交集最小值，并根据交集最大值和交集最小值确定密态样本数据的样本区间。然后，根据预设分箱数量对密态样本数据的样本区间进行密态等距分箱，获得多个分割点和多个分箱范围。其中，可以使用的编程语言例如：C、C++、Java、BASIC、JavaScript、LISP、Shell、Perl、Ruby、Python和PHP等等。

[0061]

[[<0>	<0>	<0>	<0>	<1>	<0>	<0>	<0>	<0>	<0>]
[<0>	<0>	<0>	<0>	<0>	<1>	<0>	<0>	<0>	<0>]
[<0>	<1>	<0>	<0>	<0>	<0>	<0>	<0>	<0>	<0>]
[<1>	<0>	<0>	<0>	<0>	<0>	<0>	<0>	<0>	<0>]
[<0>	<0>	<0>	<0>	<0>	<0>	<0>	<0>	<1>	<0>]
[<0>	<0>	<1>	<0>	<0>	<0>	<0>	<0>	<0>	<0>]
[<0>	<0>	<0>	<0>	<0>	<0>	<0>	<0>	<0>	<1>]]

[0062] 表5

[0063] 请参照上面表5，该表格示出了本申请实施例提供的密态独热矩阵；作为上述步骤S130的一种可选实施方式，在确定密态样本数据对应的密态独热矩阵时，可以将密态样本数据对应的布尔矩阵转换为密态独热矩阵，该实施方式可以包括：

[0064] 步骤S131：对密态样本数据落入到多个分箱范围中的布尔状态，获得密态样本数据对应的布尔矩阵。

[0065] 步骤S132：将布尔矩阵转换为密态独热矩阵。

[0066] 上述步骤S131至步骤S132的实施方式具体可以例如：使用公式

$\text{matrix}(0,1)=(Q>=\text{left})\&(Q<\text{right})$对密态样本数据落入到多个分箱范围中的布尔状态，针对密态样本数据的每个样本数据，如果该样本数据落入到多个分箱范围，该样本数据的布尔状态就赋值为1，如果该样本数据没有落入到多个分箱范围，该样本数据的布尔状态就赋值为0。针对密态样本数据的每个样本数据都进行上面的处理，就可以获得密态样本数据对应的布尔矩阵；其中，Q代表密态样本数据的每个样本数据，left代表分箱范围的最小值(即最左边的值)，right代表分箱范围的最大值(即最右边的值)，&表示两个条件均满足的逻辑符。然后，使用预设编程语言编译或者解释的可执行程序将布尔矩阵转换为密态独热矩阵，其中，可以使用的编程语言例如：C、C++、Java、BASIC、JavaScript、LISP、Shell、Perl、Ruby、Python和PHP等等。

[0067] 作为上述步骤S140的一种可选实施方式，在根据密态独热矩阵确定明文样本数据对应的密态匿名数据时，可以使用证据权重(Weights Of Evidence，WOE)向量来确定密态匿名数据，该实施方式可以包括：

[0068] 步骤S141：计算出多个分箱范围中的每个分箱范围的证据权重，获得证据权重向量。

[0069] 上述步骤S141的实施方式例如：先计算出整个标签数据的正样本量(可以表示为$Good_T$)和负样本量(可以表示为Bad_T)，以及通过密态独热矩阵计算出各个分箱范围的正样本量(可以表示为$Good_i$)和负样本量(可以表示为Bad_i)，然后，使用公式$WOE_i=ln(\frac{Bad_i}{Bad_T}/\frac{Good_i}{Good_T})=ln(\frac{Bad_i}{Bad_T})-ln(\frac{Good_i}{Good_T})$计算出多个分箱范围中的每个分箱范围的证据权重(Weight Of Evidence，WOE)，获得证据权重向量。其中，WOE_i表示第i个分箱范围的证据权重，$Good_T$表示整个标签数据的正样本量，Bad_T表示整个标签数据的负样本量，$Good_i$表示第i个分箱范围的正样本量，Bad_i表示第i个分箱范围的负样本量。

[0070] 可以使用公式
$$IV_i=(\frac{Bad_i}{Bad_T}-\frac{Good_i}{Good_T})*WOE_i=(\frac{Bad_i}{Bad_T}-\frac{Good_i}{Good_T})*ln(\frac{Bad_i}{Bad_T}/\frac{Good_i}{Good_T})$$
$$IV=\sum_{i=1}^{n}IV_i$$
来计算出信息值(Information Value，IV)，从而筛选出那些特征数据对Y标签数据的贡献大小，具体例如：针对全部特征数据中的每个特征数据，当该特征数据的IV值小于信息阈值，就删除该特征数据，以使得该特征数据不参与联邦学习。相反地，当该特征数据的IV值大于或等于信息阈值，就让该特征数据参与联邦学习。

[0071] 步骤S142：根据密态独热矩阵和证据权重向量，确定明文样本数据对应的密态匿名数据。

[0072] 上述步骤S142的实施方式例如：将密态独热矩阵中的每个值分别乘以多个分箱范围中的每个分箱范围对应的证据权重向量(即WOE向量值)，再按照行求和，从而获得样本数据对应的密态匿名数据。上述的计算过程可以使用公式表示为RESULT_WOE=($\text{matrix}_{0\text{-}1}\times\text{WOE}_i$).sum(axis=1)；其中，RESULT_WOE表示样本数据对应的密态匿名数据，$\text{matrix}_{0\text{-}1}$表示密态独热矩阵，sum(axis=1)表示按照行求和。

[0073] 作为上述数据匿名化方法的其中一种可选实施方式，在根据密态独热矩阵确定明文样本数据对应的密态匿名数据之后，还可以使用密态匿名数据进行联邦学习，该实施方

式可以包括：

[0074] 步骤S150:使用密态匿名数据对机器学习模型进行联邦学习。

[0075] 作为上述步骤S150的一种可选实施方式,在使用密态匿名数据对机器学习模型进行联邦学习时,还可以使用密态匿名数据和密态标签数据进行联邦学习,该实施方式可以包括：

[0076] 步骤S151:将待处理的样本标签进行秘密共享,获得密态标签数据,样本标签是样本数据的类别标签。

[0077] 其中,该步骤S151的实施原理和实施方式与步骤S110的实施原理和实施方式是类似的,因此,这里不再说明其实施原理和实施方式,如有不清楚的地方,可以参考对步骤S110的描述。

[0078] 步骤S152:使用密态匿名数据和密态标签数据对机器学习模型进行联邦学习。

[0079] 上述步骤S152的实施方式例如:如果第一方拥有密态标签数据,那么在第一方获得第二方的密态匿名数据之后,可以使用第二方的密态匿名数据对第一方本地存储的机器学习模型进行联邦学习。同理地,如果第二方没有密态标签数据,那么在第二方获得第一方的密态匿名数据和密态标签数据之后,可以使用第一方的密态匿名数据和密态标签数据对第二方本地存储的机器学习模型进行联邦学习。

[0080] 作为上述数据匿名化方法的其中一种可选实施方式,在根据密态独热矩阵确定明文样本数据对应的密态匿名数据之后,还可以使用明文匿名数据进行联邦学习,该实施方式可以包括：

[0081] 步骤S160:使用秘密共享密码机制中的门限方案对密态匿名数据进行恢复,获得明文匿名数据。

[0082] 步骤S170:使用明文匿名数据对机器学习模型进行联邦学习。

[0083] 上述步骤S160至步骤S170的实施方式例如:第一方在使用秘密共享密码机制中的门限方案(例如Shamir算法等等)对密态匿名数据进行恢复之后,即可获得明文匿名数据。然后,第一方可以使用明文匿名数据对第一方本地存储的机器学习模型进行联邦学习,获得联邦学习后的机器学习模型。

[0084] 请参见图3示出的本申请实施例提供的数据匿名化装置的结构示意图;本申请实施例提供了一种数据匿名化装置200,包括：

[0085] 样本数据获得模块210,用于将待处理的明文样本数据进行秘密共享,获得密态样本数据。

[0086] 分箱范围获得模块220,用于将密态样本数据进行密态等距分箱,获得多个分箱范围。

[0087] 独热矩阵确定模块230,用于根据密态样本数据落入到多个分箱范围中的每个分箱范围的布尔状态,确定密态样本数据对应的密态独热矩阵。

[0088] 匿名数据确定模块240,用于根据密态独热矩阵确定明文样本数据对应的密态匿名数据。

[0089] 可选地,在本申请实施例中,分箱范围获得模块,包括：

[0090] 样本区间确定子模块,用于统计出密态样本数据中的交集最大值和交集最小值,并根据交集最大值和交集最小值确定密态样本数据的样本区间。

[0091] 样本区间分箱子模块,用于根据预设分箱数量对密态样本数据的样本区间进行密态等距分箱。

[0092] 可选地,在本申请实施例中,独热矩阵确定模块,包括:

[0093] 布尔矩阵获得子模块,用于对密态样本数据落入到多个分箱范围中的布尔状态,获得密态样本数据对应的布尔矩阵。

[0094] 布尔矩阵转换子模块,用于将布尔矩阵转换为密态独热矩阵。

[0095] 可选地,在本申请实施例中,匿名数据确定模块,包括:

[0096] 权重向量获得子模块,用于计算出多个分箱范围中的每个分箱范围的证据权重,获得证据权重向量。

[0097] 密态数据确定子模块,用于根据密态独热矩阵和证据权重向量,确定明文样本数据对应的密态匿名数据。

[0098] 可选地,在本申请实施例中,数据匿名化装置,还包括:

[0099] 第一联邦学习模块,用于使用密态匿名数据对机器学习模型进行联邦学习。

[0100] 可选地,在本申请实施例中,第一联邦学习模块,包括:

[0101] 标签数据获得子模块,用于将待处理的样本标签进行秘密共享,获得密态标签数据,样本标签是样本数据的类别标签。

[0102] 模型联邦学习子模块,用于使用密态匿名数据和密态标签数据对机器学习模型进行联邦学习。

[0103] 可选地,在本申请实施例中,数据匿名化装置,还包括:

[0104] 匿名数据获得模块,用于使用秘密共享密码机制中的门限方案对密态匿名数据进行恢复,获得明文匿名数据。

[0105] 第二联邦学习模块,用于使用明文匿名数据对机器学习模型进行联邦学习。

[0106] 应理解的是,该装置与上述的数据匿名化方法实施例对应,能够执行上述方法实施例涉及的各个步骤,该装置具体的功能可以参见上文中的描述,此处适当省略详细描述。该装置包括至少一个能以软件或固件(firmware)的形式存储于存储器中或固化在装置的操作系统(operating system,OS)中的软件功能模块。

[0107] 请参见图4示出的本申请实施例提供的电子设备的结构示意图。本申请实施例提供的一种电子设备300,包括:处理器310和存储器320,存储器320存储有处理器310可执行的机器可读指令,机器可读指令被处理器310执行时执行如上的方法。

[0108] 本申请实施例还提供了一种计算机可读存储介质330,该计算机可读存储介质330上存储有计算机程序,该计算机程序被处理器310运行时执行如上的方法。其中,计算机可读存储介质330可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(Static Random Access Memory, 简称SRAM),电可擦除可编程只读存储器(Electrically Erasable Programmable Read-Only Memory, 简称EEPROM),可擦除可编程只读存储器(Erasable Programmable Read Only Memory, 简称EPROM),可编程只读存储器(Programmable Read-Only Memory, 简称PROM),只读存储器(Read-Only Memory, 简称ROM),磁存储器,快闪存储器,磁盘或光盘。

[0109] 需要说明的是,本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。

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

[0110] 本申请实施例提供的几个实施例中，应该理解到，所揭露的装置和方法，也可以通过其他的方式实现。以上所描述的装置实施例仅是示意性的，例如，附图中的流程图和框图显示了根据本申请实施例的多个实施例的装置、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、程序段或代码的一部分，模块、程序段或代码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。也应当注意，在有些作为替换的实现方式中，方框中所标注的功能也可以和附图中所标注的发生顺序不同。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，这主要根据所涉及的功能而定。

[0111] 另外，在本申请实施例中的各个实施例的各功能模块可以集成在一起形成一个独立的部分，也可以是各个模块单独存在，也可以两个或两个以上模块集成形成一个独立的部分。此外，在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请实施例的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0112] 以上的描述，仅为本申请实施例的可选实施方式，但本申请实施例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实施例揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请实施例的保护范围之内。

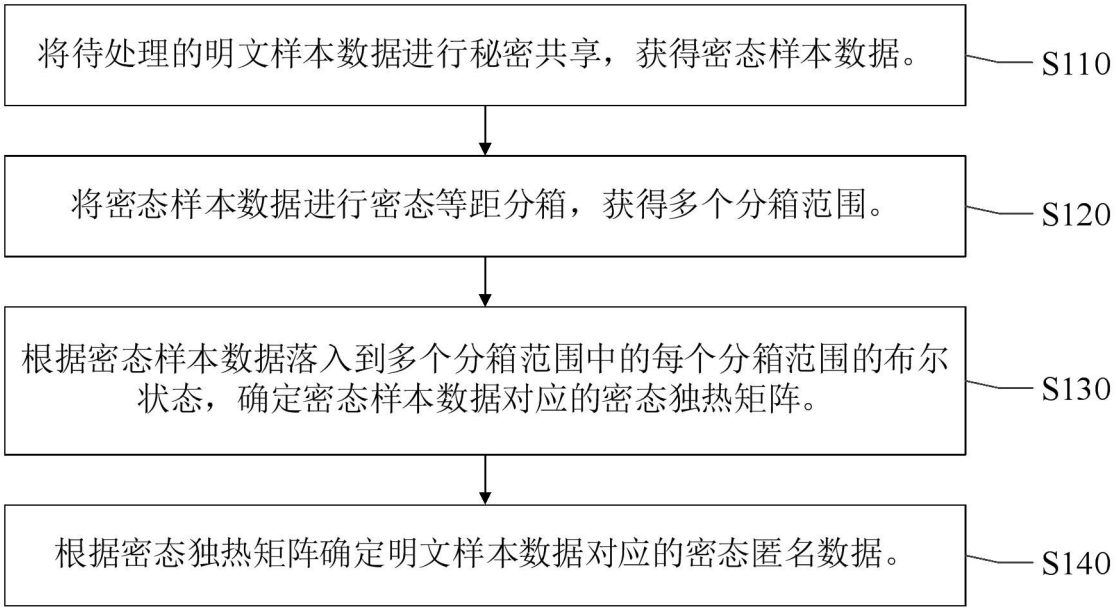

将待处理的明文样本数据进行秘密共享，获得密态样本数据。
S110
将密态样本数据进行密态等距分箱，获得多个分箱范围。
S120
根据密态样本数据落入到多个分箱范围中的每个分箱范围的布尔状态，确定密态样本数据对应的密态独热矩阵。
S130
根据密态独热矩阵确定明文样本数据对应的密态匿名数据。
S140

图1

第一方的明文样本数据

id	Y	X_{a1}	X_{a2}	X_{a3}
124360	0	1.3	5.2	3
328492	1	2.5	3.3	-2
572683	0	-1	0.5	0.2
930913	0	0.9	0.12	1

第二方的明文样本数据

id	X_{b1}	X_{b2}
328492	2.3	1.9
748329	0.89	1.41
930913	-1.2	-0.1

排序标识	Y	X_{a1}	X_{a2}	X_{a3}	X_{b1}	X_{b2}
<124360>	<0>	<0>	<0>	<0>	<0>	<0>
<328492>	<1>	<2.5>	<3.3>	<-2>	<2.3>	<1.9>
<328492>	<0>	<0>	<0>	<0>	<0>	<0>
<572683>	<0>	<0>	<0>	<0>	<0>	<0>
<748329>	<0>	<0>	<0>	<0>	<0>	<0>
<930913>	<0>	<0.9>	<0.12>	<1>	<-1.2>	<-0.1>
<930913>	<0>	<0>	<0>	<0>	<0>	<0>

秘密共享和对齐后的密态样本数据

图2

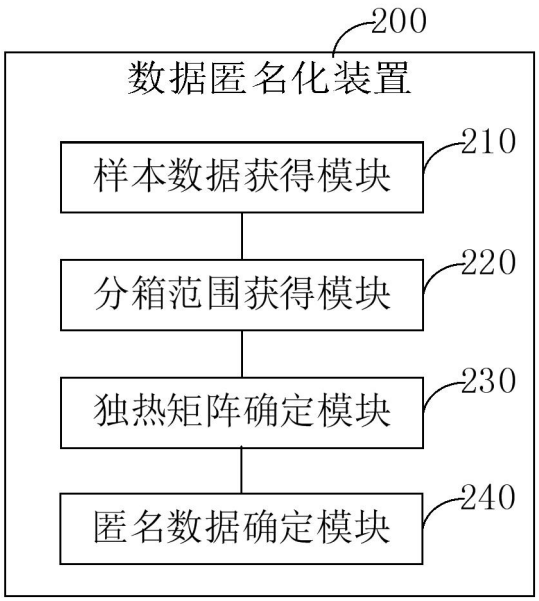

200
数据匿名化装置
210
样本数据获得模块
220
分箱范围获得模块
230
独热矩阵确定模块
240
匿名数据确定模块

图3

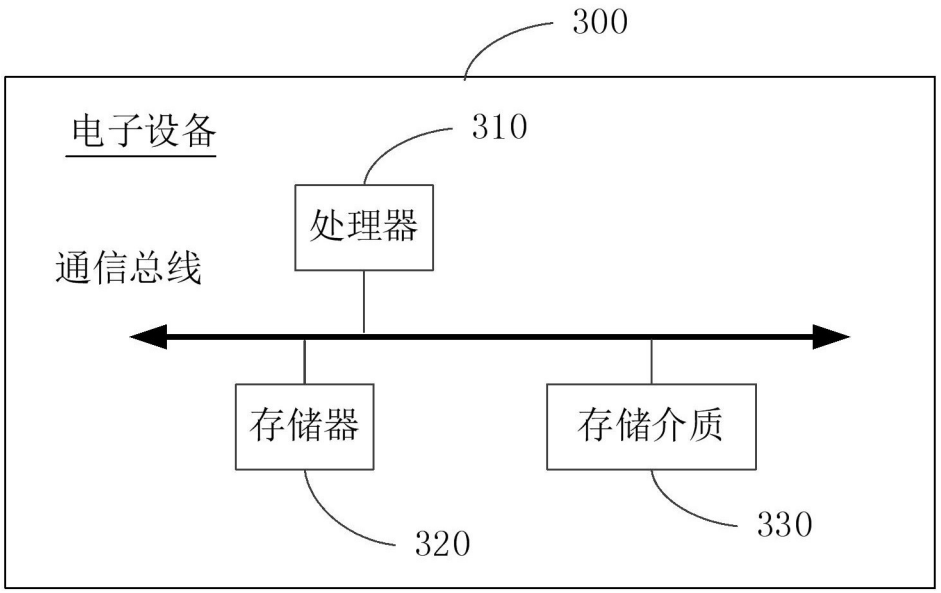

300
电子设备
310
处理器
通信总线
存储器
存储介质
320
330

图4