

用于保护数据的方法、设备、装置和介质

5 相关申请的交叉引用

本申请要求申请号为 202210119229.1，题为“用于保护数据的方法、设备、装置和介质”、申请日为 2022 年 2 月 08 日的中国发明专利申请的优先权，通过引用的方式将该申请整本并入本文。

10 技术领域

本公开的示例实施例总体涉及计算机领域，特别地涉及用于保护数据的方法、设备、装置和计算机可读存储介质。

背景技术

- 15 随着人工智能技术的发展，机器学习技术已经被用于各行各业。可以基于预先采集的训练数据来训练用于完成特定功能的处理模型。然而，某些训练数据可能会涉及用户隐私和/或其他敏感数据。此时，如何以更为安全的方式采集训练数据并且将其用于训练处理模型，成为一个研究热点。目前已经提出了在具有数据的标签信息的第一方与
- 20 不具有标签信息的第二方之间联合训练模型的技术方案，然而已有技术方案的性能并不理想，因而不能有效地防止敏感信息泄漏。

发明内容

根据本公开的示例实施例，提供了一种用于保护数据的方案。

- 25 在本公开的第一方面，提供了一种用于保护数据的方法。该方法包括：由第一设备获取由第二设备基于样本数据、根据第二模型生成的特征表示。第一设备具有针对样本数据的标签信息。第一设备和第二设备用于联合训练第一设备处的第一模型和第二设备处的第二模型。该方法还包括：由第一设备基于特征表示，根据第一模型，生成

针对样本数据的预测标签；以及由第一设备基于特征表示、标签信息和预测标签，来确定用于训练第一模型和第二模型的总损失值。

在本公开的第二方面，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设备执行以下动作：由电子设备获取由第二设备基于样本数据、根据第二模型生成的特征表示，电子设备具有针对样本数据的标签信息，电子设备和第二设备用于联合训练电子设备处的第一模型和第二设备处的第二模型；由电子设备基于特征表示，根据第一模型，生成针对样本数据的预测标签；以及由电子设备基于特征表示、标签信息和预测标签，来确定用于训练第一模型和第二模型的总损失值

在本公开的第三方面，提供了一种用于保护数据的装置，该装置包括获取模块，被配置为获取由第二设备基于样本数据、根据第二模型生成的特征表示，该装置具有针对样本数据的标签信息，该装置和第二设备用于联合训练装置处的第一模型和第二设备处的第二模型；标签预测模块，被配置为基于特征表示，根据第一模型，生成针对样本数据的预测标签；以及总损失值确定模块，被配置为基于特征表示、标签信息和预测标签，来确定用于训练第一模型和第二模型的总损失值。

在本公开的第四方面，提供了一种计算机可读存储介质。介质上存储有计算机程序，程序被处理器执行时实现第一方面的方法。

应当理解，本发明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。本公开的其它特征将通过以下的描述而变得容易理解。

附图说明

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：

图 1 示出了本公开的实施例能够在其中实现的示例环境的示意图；

图 2 示出了根据本公开的一些实施例的模型联合训练的示例架构的示意图；

5 图 3 示出了根据本公开的一些实施例的数据保护的结果的示意图；

图 4 示出了根据本公开的一些实施例的用于保护数据的过程的流程图；

图 5 示出了根据本公开的一些实施例的用于保护数据的装置的框图；以及

10 图 6 示出了能够实施本公开的多个实施例的设备的框图。

具体实施方式

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
15 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。

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。

如本文中所使用的，术语“模型”可以从训练数据中学习到的相应的输入与输出之间的关联，从而在训练完成后可以针对给定的输入，
25 生成对应的输出。模型的生成可以基于机器学习技术。深度学习是一种机器学习算法，通过使用多层处理单元来处理输入和提供相应输出。神经网络模型是基于深度学习的模型的一个示例。在本文中，“模型”也可以被称为“机器学习模型”、“学习模型”、“机器学习网络”

或“学习网络”，这些术语在本文中可互换地使用。如本文所使用的，属于“确定模型的参数”或类似表达是指确定模型的参数的值（又称为参数值），包括具体值、取值集合或取值范围等。

“神经网络”是一种基于深度学习的机器学习网络。神经网络能够处理输入并且提供相应输出，其通常包括输入层和输出层以及在输入层与输出层之间的一个或多个隐藏层。在深度学习应用中使用的神经网络通常包括许多隐藏层，从而增加网络的深度。神经网络的各个层按顺序相连，从而前一层的输出被提供作为后一层的输入，其中输入层接收神经网络的输入，而输出层的输出作为神经网络的最终输出。神经网络的每个层包括一个或多个节点（也称为处理节点或神经元），每个节点处理来自上一层的输入。

通常，机器学习大致可以包括三个阶段，即训练阶段、测试阶段和应用阶段（也称为推理阶段）。在训练阶段，给定的模型可以使用大量的训练数据进行训练，不断迭代更新模型的参数值，直到模型能够从训练数据中获取一致的满足预期目标的推理。通过训练，模型可以被认为能够从训练数据中学习从输入到输出之间的关联（也称为输入到输出的映射）。训练后的模型的参数值被确定。在测试阶段，将测试输入应用到训练后的模型，测试模型是否能够提供正确的输出，从而确定模型的性能。在应用阶段，模型可以被用于基于训练得到的参数值，对实际的输入进行处理，确定对应的输出。

在机器学习领域，可以利用包括大量训练数据的训练数据集来训练预测模型，进而使得预测模型完成期望的功能。随着数据量的增加，数据通常分散存储在不同的存储位置，诸如不同的企业方处。随着数据监管政策越来越严格，以及数据隐私保护问题越来越受到重视，传统的集中式机器学习技术难以确保对数据隐私的保护。

目前已经提出了分布式机器学习技术。例如，不同的参与方可以具有各自的机器学习模型，各个参与方之间可以发送模型参数和推送数据。同时，包括用户个人的敏感信息等的敏感数据仅在某一参与方处存储，而不会被传输到其他参与方。

例如，在在线广告产业中，媒体侧（被动参与方）可以向用户推送广告。如果用户在媒体侧观看了广告，从而在广告主侧（主动参与方）进行了诸如购买、注册或信息需求等行为，则在广告主侧可以对该广告相关联的数据添加标签以标注用户的上述行为。广告主侧可以利用上述产生的标签信息来推断信息，诸如敏感商业信息，例如用户规模以及营收规模等。可以利用训练数据来训练预测模型，进而使得预测模型可以进行诸如用户规模推断以及营收规模推断等等。然而，上述标签信息对于广告主很重要，是需要被保护的敏感数据。广告主侧不希望标签信息泄露，例如不希望媒体侧窃取标签信息。因此，需要在获取训练数据期间，避免敏感数据的泄露。

目前已经提出了在具有数据的标签信息的第一方与不具有标签信息的第二方之间联合训练模型的技术方案。在本文中，具有数据的标签信息的第一方也被称为“主动参与方”，而不具有数据的标签信息的第二方也被称为“被动参与方”。在联合训练模型（也称为模型的联邦学习）的方案中，数据的标签信息仅存储在主动参与方处，而不会被传输到被动参与方处。主动参与方和被动参与方之间可以传输不敏感的数据和/或模型参数等，以对主动参与方处的模型和被动参与方处的模型分别进行训练。例如，主动参与方可以确定用于更新预测模型的梯度信息，并且向被动参与方发送该梯度信息。被动参与方可以利用该梯度信息进行模型训练，而无需获取敏感数据，例如标签信息。

然而，攻击方可以使用谱聚类等技术，来从不具备标签信息的数据信息中预测标签信息。为了避免标签信息被攻击方获取（例如，预测），即为了保护敏感信息，目前已经提出了一些关系标签信息保护的方案。例如，可以使用梯度深度泄露（DLG）算法以及改进的 DLG（iDLG）算法来保护标签信息。

具体地，DLG 算法发现被动参与方可以从主动参与方共享的梯度里面推断出相应的用户标签信息。DLG 首先产生随机的伪（dummy）训练输入数据和标签信息，这些信息通过前向传播会产生一个伪梯度。

通过最小化伪梯度与真实梯度之间的差异,DLG 可以推断出训练输入以及其对应的标签信息。iDLG 是对 DLG 的改进。iDLG 发现 DLG 并不能高质量地推断主动参与方的标签信息。在二分类的任务以及使用交叉熵作为损失函数的设定下, iDLG 发现标签信息和损失对最后一层 logit 的梯度符号上存在相关性。iDLG 利用上述发现则可以对参数的梯度进行分析来推断用于训练样本的标签信息。DLG 和 iDLG 均需要被动参与方拥有整个联邦学习各个参与方的模型结构。这在现实中难以达到。

此外,目前已经提出了向梯度信息添加噪声从而防止敏感信息泄漏的方案。然而,针对攻击方能力非常强的场景,攻击方可能能够基于由样本数据得到的特征表示来预测出与真实标签信息非常接近的标签信息。在这种情形下,即使通过向梯度信息添加噪声防止攻击方通过梯度信息预测出标签信息,仍然无法避免攻击方根据由样本数据得到的特征表示来推断出标签信息。

综上所述,目前的数据保护方案存在着诸多缺陷,难以获得令人满意的结果。因此,期望能够以更有效的方式来防止标签信息泄露以及保护敏感数据。

根据本公开的实施例,提供了一种用于保护数据的方案,旨在解决上述问题以及其他潜在问题中的一个或多个。在该方案中,由(主动参与方的)第一设备根据(被动参与方的)第二设备的第二模型基于样本数据生成的特征表示,利用第一模型来预测针对样本数据的预测标签。第一设备进而基于该特征表示、该样本数据的真实的标签信息以及所得到的预测标签,来确定总损失值。第一设备进而可以基于该总损失值来调整第一设备处的第一模型的参数。此外,第一设备还可以使第二设备基于与总损失值相关联的梯度来调整第二设备处的第二模型的参数。

在该方案中,不需要获知各个参与方的模型(即,各个设备处的模型)的结构。第一设备与第二设备之间仅共享由第二设备生成的特征表示以及与总损失值相关联的梯度,而不共享诸如标签信息等敏感

信息。

此外，本方案通过利用总损失值来调整第一模型和/或第二模型的参数，而总损失值与特征表示、标签信息和预测标签相关联。以这种方式，能够通过设置合适的损失函数来确定适当的总损失值，以降低特征表示与标签信息之间的相关性。以这样，能够避免被动参与方（或第二设备）基于针对样本数据的特征表示来对数据的标签信息进行预测，从而保护敏感数据。

在下文中，将首先参见图 1 描述根据本公开的一个示例性实现方式的示例环境。

10

示例环境

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。在图 1 的环境 100 中，期望训练和应用第一设备 120 处的第一模型 140 和/或第二设备 130 处的第二模型 150。第一设备 120 可以是主动参与方的设备。在本文中，第一设备 120 也可以被称为主动参与方设备，并且第一模型 140 也可以被称为主动参与方模型。第二设备 130 可以是被动参与方的设备。在本文中，第二设备 130 也可以被称为被动参与方设备，并且第二模型 150 也可以被称为被动参与方模型。第一设备 120 与第二设备 130 可以通过网络进行交互，以接收或发送数据或消息等。在一些实施例中，第一模型 140 和第二模型 150 可以被配置为执行相同的或不同的功能或任务，例如预测任务、分类任务等。

总体而言，环境 100 包括联合训练系统 110 以及可选的应用系统 160。在图 1 的示例实施例以及下文将会描述的一些示例实施例中，联合训练系统 110 被配置为利用训练数据集 101 来在第一设备 120 和第二设备 130 处对第一模型 140 和第二模型 150 进行联合训练。训练数据集 101 可以包括样本数据 102-1、102-2……、102-N 以及相关关联的标签信息 104-1、104-2……、104-N，其中 N 为大于等于 1 的整数。为便于讨论，样本数据 102-1、102-2……、102-N 等可以统称或单独

25

称为样本数据 102，标签信息 104-1、104-2……、104-N 等可以统称或单独称为标签信息 104。

每个样本数据 102 可以具有相对应的标签信息 104。标签信息 104 可以包括用户对于样本数据 102 的处理过程的敏感数据。以广告主侧
5 为主动参与方而媒体侧为被动参与方的实施例，用户在媒体侧点击广告后，在广告主侧可能会发生处理行为。例如，如果用户受到媒体侧广告影响而在广告主侧发生购买、注册或者信息需求等行为，则表示在广告注册发生了处理行为（也称为发生了转化）。

在一些实施例中，可以使用正例标签 1 来表示某个样本数据发生了上述转化行为，即表示该样本数据为正例。如果没有发生转化，则用负例标签 0 来表示该样本数据为负例。应当理解，在一些实施例中，也可以采取其他的方式来对各个样本数据添加标签。例如，针对购买、注册、信息需求的行为，分别对样本数据添加 3、2、1 的标签，而对于没有发生转化则同样使用 0 作为标签。应当理解，上述列举的标签
15 信息的示例仅仅是示例性的，而无意限制本公开的范围。

在一些实施例中，上述描述的标签信息 104 仅可以被第一设备 120 所获知。在一些实施例中，第一设备 120 可以基于标签信息 104 来用于推断诸如广告主的敏感商业信息，比如用户以及营收规模。标签信息 104 是主动参与方（即，第一设备 120）需要保护的重要信息。因此，不希望第二设备 130 获取标签信息 104。即，第二设备 130 可以不具有标签信息 104。
20

在联合训练前，第一模型 140 和第二模型 150 的参数值可以被初始化。经过联合训练，第一模型 140 和第二模型 150 的参数值被更新和确定。在联合训练完成后，经训练的第一模型 170 和经训练的第二模型 180 具有训练后的参数值。基于这样的参数值，经训练的第一模型 170 和经训练的第二模型 180 能够用于各种不同的处理任务，诸如预测用户规模等。在一些实施例中，第一模型 140 和第二模型 150 可以是任意适当的神经网络模型。例如，在正负标签信息的示例中，第一模型 140 和第二模型 150 可以是用于二分类的模型。第一模型 140
25

和第二模型 150 可以分别包括隐层、logit 层、softmax 层等。

环境 100 可选地包括应用系统 160。应用系统 160 可以被配置为利用经训练的第一模型 170 和经训练的第二模型 180 来执行针对源数据 162 和/或源标签信息 164 的各种处理任务。在一些实施例中，第一设备 120 或经训练的第一模型 170 具有源标签信息 164。与之相比，第二设备 130 或经训练的第二模型 180 不具有源标签信息 164。应用系统 160 还可以包括未示出的其他模型或操作，以完成相应的推理或其他任务。

在图 1 中，联合训练系统 110 和应用系统 160 可以是任何具有计算能力的系统，例如各种计算设备/系统、终端设备、服务器等。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、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。服务器包括但不限于大型机、边缘计算节点、云环境中的计算设备，等等。

应当理解，图 1 示出的环境中的部件和布置仅是示例，适于用于实现本公开所描述的示例实施例的计算系统可以包括一个或多个不同的部件、其他部件和/或不同的布置方式。例如，虽然被示出为是分离的，但联合训练系统 110 和应用系统 160 中的两项或更多项可以集成在相同系统或设备。本公开的实施例在此方面不受限制。

以下将继续参考附图，分别描述模型联合训练的示例实施例。

模型训练架构

图 2 示出了根据本公开的一些实施例的模型联合训练的示例架构 200 的示意图。图 2 的架构 200 可以被实现在图 1 的联合训练系统 110 中。架构 200 中的各个步骤可以由硬件、软件、固件或者它们的任意组合来实现。在架构 200 中，可以利用第一设备 120 和第二设备 130 两者来以分布式方式对第一模型 140 和第二模型 150 执行训练。

图 2 左侧示出了在第一设备 120 侧的处理过程，并且图 2 的右侧

示出了在第二设备 130 侧的处理过程。在一些实施例中,第二模型 150 可以对样本数据 102 进行编码以得到特征表示 212。可以基于目前已知的和/或将在未来开发的多种编码器对其进行编码以得到特征表示 212。第二设备 130 可以将特征表示 212 传输至第一模型 140,以供后
 5 续联合训练第一模型 140 和第二模型 150。应理解,虽然图 2 中仅示出了一个样本数据 102,但在一些实施例中,可以使用一组样本数据或者一批次的样本数据 $X \in \mathcal{R}^{n \times d}$ 。在一些实施例中,针对一批次样本数据的特征表示可以被表示为 $f(X)$ 。在本文中,特征表示也被称为中间层编码(例如, embedding)或者中间编码(例如, embedding)。

10 在一些实施例中,第二设备 130 可以具有攻击性。即,第二设备 130 可能会尝试利用特征表示 212 来推断或预测标签信息 104。例如,第二设备 130 可以使用谱攻击(spectral attack),例如 2 聚类离群点检测算法来预测针对特征表示 212 的标签信息。

在一些实施例中,正例样本数据可以被表示为 $\mathcal{F}^+$, 负例样本数
 15 据可以被表示为 $\mathcal{F}^-$ 。可以使用以下来针对特征表示 212 进行标签预测:

$$|\langle f(X) - \mu_F, v \rangle| \quad (1)$$

其中, $f(X)$ 表示特征表示 212, μ_F 表示样本数据 F (包括正例样本数据和负例样本数据)的中值,并且 v 表示 F 的协方差矩阵的顶
 20 奇异向量。

以一批次的样本数据 $X \in \mathcal{R}^{n \times d}$ 为例,第二设备 130 可以计算该批次的样本数据的经验值中值和协方差矩阵,来估计 μ_F 和 v 。根据所估计得到的 μ_F 和 v ,可以利用式(1)来针对该批次中的每个样本
 数据计算得分 $|\langle f(X) - \mu_F, v \rangle|$ 。第二设备 130 可以基于所计算出的
 25 的各个得分,来将所有的 $f(X)$ 划分为两个聚类。

在一些实施例中,第二设备 130 还可以使用先验指示或者经验值

来对所得到的两个聚类分别确定标签。例如，如果当前使用的数据集是不平衡的并且以负例为主，则可以将负例标签（例如，0）分配给较大的聚类，并且将正例标签（例如，1）分配给较小的聚类。

5 在一些实施例中，第二设备 130 可以使用任意宽度和深度的网络结构来增强自己侧模型的能力。在一些情形下，第二设备 130 可以具有很强的攻击能力，其可以使 $f(X)$ 和第一设备 120 的标签信息预测能力一致。

以上描述了在一些情形下作为攻击方的第二设备 130 根据特征表示 212 对标签信息 104 进行推断的过程。在一些实施例中，不希望第二设备 130 能够基于特征表示 212 对标签信息 104 进行推断。对此，10 可以通过使用以下详细描述的第一模型 140 与第二模型 150 的联合训练过程来避免第二设备 130 对标签信息 104 的推断。也就是说，通过以下所描述的对第一模型 140 与第二模型 150 的联合训练过程，避免用户的敏感信息，诸如标签信息的泄露。

15 如图 2 所示，第一设备 120 基于特征表示 212，根据第一模型 140，生成针对样本数据 102 的预测标签 222。在本文中，预测标签 222 可以是第一模型 140 的 logit 输出的结果，也被表示为 $h(f(X))$ 。第一设备 120 被配置为基于特征表示 212、标签信息 104 和预测标签 222，确定用于训练第一模型 140 和第二模型 150 的总损失值 230。

20 在一些实施例中，第一设备 120 可以基于标签信息 104 和预测标签 222，来确定第一损失。第一损失可以用来表示预测标签 222 与标签信息 104 的差别程度。例如，可以将标签信息 104 与预测标签 222 的交叉熵确定为第一损失。在本文中，第一损失可以被表示为 L_c 。在这一示例中，第一损失的值越大，则表示预测标签 222 与标签信息 25 104 的差别程度越大。

在一些实施例中，第一设备 120 可以基于特征表示 212 和标签信息 104，来确定第二损失。第二损失可以用来表示特征表示 212 与标签信息 104 的相关程度。例如，在一些实施例中，第一设备 120 可以将特征表示 212 与标签信息 104 的距离相关系数确定为第二损失。利

用距离相关系数确定第二损失可以被表示为：

$$L_d = DCOR(Y, F(X)) \quad (2)$$

其中 L_d 表示第二损失， Y 表示标签信息 104， $F(X)$ 表示特征表示 212，函数 $DCOR()$ 表示距离相关系数函数。应理解，可以使用
5 任何适当的距离相关系数函数来确定第二损失 L_d 。

附加地或备选地，还可以对上述第二损失 L_d 进行取对数处理，以得到经对数处理的第二损失。在上述示例中，第二损失的值越大，则表示特征表示 212 与标签信息 104 越相关。

在一些实施例中，第一设备 120 可以基于第一损失和第二损失，
10 来确定总损失值 230。例如，第一设备 120 可以将第一损失与第二损失的加权和确定为总损失值 230。例如，总损失值 230 可以利用以上式 (3) 来确定：

$$L = L_c + \alpha_d L_d \quad (3)$$

其中， L (也可表示为 l) 表示总损失值 230， L_c 表示第一损失， L_d
15 表示第二损失， α_d 是大于等于零的数值， α_d 表示第二损失的权重。

在一些实施例中，第二损失的权重 α_d 可以是预设的值，例如 0.5 或者其他介于 0 与 1 之间的适当的值。备选地，可以在联合训练过程中根据训练过程的结果而动态调整权重 α_d 。应当理解，上文所描述的损失函数是示意性的，而不是限制性的。

20 在一些实施例中，第一设备 120 可以通过最小化总损失值 230 来确定第一模型 140 的参数。例如，第一设备 120 可以通过从总损失值 230 与至少一个其他总损失值中确定最小总损失值，来确定第一模型 140 的参数。至少一个其他总损失值是由第一设备 120 先前或者之后确定的。

25 在一些实施例中，第一设备 120 可以至少基于式 (3) 的损失函数来进行迭代更新，从而确定第一模型 140 的参数值。例如，如果迭代收敛或者迭代达到预定次数，则对第一模型 140 的训练完成。以这

样，可以基于总损失值 230 来确定第一模型 140 的参数。

通过使用以上所描述的总损失值，可以在保证预测标签 222 的准确性的同时，减小特征表示 212 与标签信息 104 之间的相关性。即，使得特征表示 212 与标签信息 104 解耦。以这样，能够避免第二设备
5 130 或者其他攻击方基于特征表示 212 对标签信息 104 进行推断或预测。用户的敏感信息，诸如标签信息从而得以保护。

在确定了总损失值 230 之后，第一设备 120 还可以利用梯度确定模块 250，基于总损失值 230 来确定梯度 240。例如，第一设备 120 可以确定总损失值 230 相对于特征表示 212 的梯度 240。在一些实施
10 例中，可以通过以下的式（4）来确定梯度 240：

$$g = \frac{\partial l}{\partial f(X)} \tag{4}$$

其中， g 表示梯度 240， l 表示总损失值 230， $f(X)$ 表示特征表示 212。

在一些实施例中，第一设备 120 可以将所确定的梯度 240 发送给
15 第二设备 130，以使得第二设备 130 根据梯度 240 来确定第二模型 150 的参数。例如，第二设备 130 响应于接收到梯度 240，可以基于链式法则（chain rule）来更新第二模型 150 的参数。

通过利用以上所描述的联合训练过程，能够避免第二设备 130 利用特征表示 212 对标签信息 104 进行推断。以这样，本公开的技术方
20 案可以在确保第一模型和第二模型的性能的情况下更好地保护敏感数据。此外，本公开的技术方案不需要获知各个模型的结构即可对模型进行联合训练，因此容易实现。

示例结果

25 图 3 示出了根据本公开的一些实施例的数据保护的结果的示意图。结果 300 示出了未使用本公开的技术方案对数据进行保护时，第二设备 130 利用诸如谱攻击在不同层（layer）上的攻击 AUC（Area Under Curve，曲线下面积）。攻击 AUC 或者简称为 AUC 可以被用来表示

攻击方（例如，第二设备 130）区分正负标签的能力。接近 0.5 的攻击 AUC 被认为近似于随机猜测，也就是说可以认为接近 0.5 的攻击 AUC 不具危险性，即数据是安全的。

在结果 300 中，曲线 302 示出了测试 AUC（即第一设备 120 的 AUC），曲线 302、304、306、308、312 和 314 分别示出了攻击方（即，第二设备 130）的层 5 的 AUC、层 4 的 AUC、层 3 的 AUC、层 2 的 AUC 以及层 1 的 AUC。根据结果 300 可以看出，攻击方（例如，第二设备 130）的攻击能力非常强大。特别是在攻击方具有超过 4 的层数时，攻击能力接近模型的效用（utility）。

结果 320 示出了使用本公开的技术方案对数据进行保护时，第二设备 130 利用诸如谱攻击在不同层（layer）上的攻击 AUC。在结果 320 中，权重 α_d 被设置为 0.002。在结果 320 中，曲线 322 示出了测试 AUC（即第一设备 120 的 AUC），曲线 324、326、328、332 和 334 分别示出了攻击方（即，第二设备 130）的层 5 的 AUC、层 4 的 AUC、层 3 的 AUC、层 2 的 AUC 以及层 1 的 AUC。

类似地，结果 340 示出了使用本公开的技术方案对数据进行保护时，第二设备 130 利用诸如谱攻击在不同层（layer）上的攻击 AUC。在结果 340 中，权重 α_d 被设置为 0.005。在结果 340 中，曲线 342 示出了测试 AUC（即第一设备 120 的 AUC），曲线 344、346、348、352 和 354 分别示出了攻击方（即，第二设备 130）的层 5 的 AUC、层 4 的 AUC、层 3 的 AUC、层 2 的 AUC 以及层 1 的 AUC。

通过结果 320 和结果 340 可以看出，尽管攻击方（例如，第二设备 130）具有很强的攻击能力，但是使用本公开的数据保护方案，能够有效的使 AUC 接近于 0.5 这一希望得到的值。特别是结果 320（权重 α_d 被设置为 0.002），各个层的 AUC 均接近于 0.5。这也表示着本方案的数据保护方法十分有效，能够很好地避免敏感数据的泄露。

结果 360 示出了不同权重 α_d 下的 AUC 曲线。结果 360 中的曲线 362、364、366、368、372 和 374 分别示出了权重 α_d 为 0.002、0.005、0.01、0.02、0.05 和 0.1 的情况下的第一设备 120 的 AUC 曲线即测试

AUC 曲线。在一些实施例中,选择较大的 α_d 可以提高对数据的保护,即数据的隐私性好,但是模型的效用将会变差。反之,选择较小的 α_d ,数据的隐私性会变差,模型的效用与香草(vanilla)模型类似,即模型的效用较好。因此,在一些实施例中,可以通过选择适当的权重 α_d ,
5 以使得在数据的隐私性与模型的效用之间两方面均能够取得较好的表现。例如,在一些实施例中,权重 α_d 可以被选择为 0.002 与 0.005 之间的数值。应理解,关于权重 α_d 的取值仅是示例性的,而无意限制本公开的范围。

通过上述结果可以看出,使用根据本公开所描述的联合训练过程,
10 能够避免第二设备 130 利用特征表示 212 对标签信息 104 进行推断。特别是即使在第二设备 130 具有很强的攻击性时仍然能够保证敏感信息的安全。此外,通过合理设置总损失值中第二损失的权重,可以在确保第一模型和第二模型的效用的情况下更好地保护敏感数据。

15 示例过程

图 4 示出了根据本公开的一些实施例的用于保护数据的过程 400 的流程图。过程 400 可以被实现在联合训练系统 110 和/或应用系统 160 处。

在框 410 处,由第一设备 120 获取由第二设备 130 基于样本数据
20 102、根据第二模型 150 生成的特征表示 212。第一设备 120 具有针对样本数据 102 的标签信息 104。第一设备 120 和第二设备 130 用于联合训练第一设备 120 处的第一模型 140 和第二设备 130 处的第二模型 150。

在一些实施例中,标签信息 104 包括用于对于样本数据 102 的处
25 理过程的敏感数据。在一些实施例中,第二设备 130 不具有标签信息 102。

在框 420 处,由第一设备 120 基于特征表示 212,根据第一模型 140,生成针对样本数据 102 的预测标签 222。在框 430 处,由第一设备 120 基于特征表示 212、标签信息 104 和预测标签 222,来确定用

于训练第一模型 140 和第二模型 150 的总损失值 230。

5 在一些实施例中，可以基于标签信息 104 和预测标签 222，来确定第一损失。第一损失表示预测标签 222 与标签信息 104 的差别程度。在一些实施例中，可以基于特征表示 212 和标签信息 104，来确定第二损失。第二损失表示特征表示 212 与标签信息 104 的相关程度。例如，在一些实施例中，可以将特征表示 212 与标签信息 104 的距离相关系数确定为第二损失。

10 在一些实施例中，可以基于第一损失和第二损失，来确定总损失值 230。例如，在一些实施例中，可以将第一损失与第二损失的加权和确定为总损失值 230。例如，可以通过式 (3) 来确定总损失值 230。

在一些实施例中，第一设备 120 可以通过从总损失值与至少一个其他总损失值中确定最小总损失值，来确定第一模型 140 的参数。至少一个其他总损失值是由第一设备 120 先前或者之后确定的。

15 在一些实施例中，可以由第一设备 120 来确定总损失值 230 相对于特征表示 212 的梯度 240。例如，可以通过利用式 (4) 来确定梯度 240。

在一些实施例中，第一设备 120 可以将所确定的梯度 240 发送给第二设备 130。第二设备 130 可以利用梯度 240 来确定第二模型 150 的参数。

20

示例装置和设备

25 图 5 示出了根据本公开的一些实施例的用于保护数据的装置 500 的框图。装置 500 可以被实现为或者被包括在联合训练系统 110 和/或应用系统 160 中。装置 500 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。

如图所示，装置 500 包括特征表示获取模块 510，被配置为获取由第二设备基于样本数据、根据第二模型生成的特征表示。装置 500 具有针对样本数据的标签信息。装置和第二设备用于联合训练装置处的第一模型和第二设备处的第二模型。

在一些实施例中，标签信息包括用户对于样本数据的处理结果的敏感数据。在一些实施例中，第二设备不具有标签信息。

装置 500 还包括预测标签生成模块 520，被配置为基于特征表示，根据第一模型，生成针对样本数据的预测标签。

5 装置 500 还包括总损失值确定模块 530，被配置为基于特征表示、标签信息和预测标签，来确定用于训练第一模型和第二模型的总损失值。

在一些实施例中，总损失值确定模块 530 包括：第一损失确定模块，被配置为基于标签信息和预测标签，来确定第一损失。第一损失
10 表示预测标签与标签信息的差别程度。

在一些实施例中，总损失值确定模块 530 还包括：第二损失确定模块，被配置为基于特征表示和标签信息，来确定第二损失。第二损失表示特征表示与标签信息的相关程度。在一些实施例中，第二损失确定模块包括：距离相关系数确定模块，被配置为将特征表示与标签
15 信息的距离相关系数确定为第二损失。

在一些实施例中，总损失值确定模块 530 第二总损失值确定模块，被配置为基于第一损失和第二损失，来确定总损失值。在一些实施例中，第二总损失值确定模块包括：加权和模块，被配置为将第一损失与第二损失的加权和确定为总损失值。

20 在一些实施例中，装置 500 还包括：最小总损失值确定模块，被配置为通过从总损失值与至少一个其他总损失值中确定最小总损失值，来确定第一模型的参数。至少一个其他总损失值是由第一设备 120 先前或者之后确定的。

在一些实施例中，装置 500 还包括：梯度确定模块，被配置为由
25 第一设备确定总损失值相对于特征表示的梯度。在一些实施例中，装置 500 还包括：发送模块，被配置为向第二设备发送梯度，以使得第二设备根据梯度来确定第二模型的参数。

图 6 示出了示出了其中可以实施本公开的一个或多个实施例的计算设备 600 的框图。应当理解，图 6 所示出的计算设备 600 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 6 所示出的计算设备 600 可以用于实现图 1 的联合训练系统 110 和/或应用系统 160。

如图 6 所示，计算设备 600 是通用计算设备的形式。计算设备 600 的组件可以包括但不限于一个或多个处理器或处理单元 610、存储器 620、存储设备 630、一个或多个通信单元 640、一个或多个输入设备 650 以及一个或多个输出设备 660。处理单元 610 可以是实际或虚拟处理器并且能够根据存储器 620 中存储的程序来执行各种处理。在多个处理器系统中，多个处理单元并行执行计算机可执行指令，以提高计算设备 600 的并行处理能力。

计算设备 600 通常包括多个计算机存储介质。这样的介质可以是计算设备 600 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 620 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 630 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）并且可以在计算设备 600 内被访问。

计算设备 600 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 6 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 620 可以包括计算机程序产品 625，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。

通信单元 640 实现通过通信介质与其他计算设备进行通信。附加地，计算设备 600 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，这些计算机能够通过通信连接进行通信。因此，计算设备 600 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。

输入设备 650 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 660 可以是一个或多个输出设备，例如显示器、扬声器、打印机等。计算设备 600 还可以根据需要通过通信单元 640 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计算设备 600 交互的设备进行通信，或者与使得计算设备 600 与一个或多个其他计算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。

根据本公开的示例性实现方式，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现方式，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。

这里参照根据本公开实施例的方法、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。

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。

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
5 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。

附图中的流程图和框图显示了根据本公开的多个实施例的系统、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施例中，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，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。
10
15

以上已经描述了本公开的各实施例，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施例。在不偏离所说明的各实施例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施例的原理、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现方式。
20

权利要求书

1. 一种用于保护数据的方法，包括：

5 由第一设备获取由第二设备基于样本数据、根据第二模型生成的特征表示，所述第一设备具有针对所述样本数据的标签信息，所述第一设备和所述第二设备用于联合训练所述第一设备处的第一模型和所述第二设备处的所述第二模型；

由所述第一设备基于所述特征表示，根据所述第一模型，生成针对所述样本数据的预测标签；以及

10 由所述第一设备基于所述特征表示、所述标签信息和所述预测标签，来确定用于训练所述第一模型和所述第二模型的总损失值。

2. 根据权利要求1所述的方法，其中确定所述总损失值包括：

基于所述标签信息和所述预测标签，来确定第一损失，所述第一损失表示所述预测标签与所述标签信息的差别程度；

15 基于所述特征表示和所述标签信息，来确定第二损失，所述第二损失表示所述特征表示与所述标签信息的相关程度；以及

基于所述第一损失和所述第二损失，来确定所述总损失值。

3. 根据权利要求2所述的方法，其中确定所述总损失值包括：

将所述第一损失与所述第二损失的加权和确定为所述总损失值。

20 4. 根据权利要求2所述的方法，其中确定所述第二损失包括：

将所述特征表示与所述标签信息的距离相关系数确定为所述第二损失。

5. 根据权利要求1所述的方法，还包括：

25 通过从所述总损失值与至少一个其他总损失值中确定最小总损失值，来确定所述第一模型的参数，所述至少一个其他总损失值是由所述第一设备先前或者之后确定的。

6. 根据权利要求1所述的方法，还包括：

由所述第一设备确定所述总损失值相对于所述特征表示的梯度；以及

向所述第二设备发送所述梯度，以使得所述第二设备根据所述梯度来确定所述第二模型的参数。

7. 根据权利要求1所述的方法，其中所述标签信息包括用户对于所述样本数据的处理过程的敏感数据。

5 8. 根据权利要求1所述的方法，其中所述第二设备不具有所述标签信息。

9. 一种电子设备，包括：

至少一个处理单元；以及

10 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以下动作：

由所述电子设备获取由第二设备基于样本数据、根据第二模型生成的特征表示，所述电子设备具有针对所述样本数据的标签信息，所述电子设备和所述第二设备用于联合训练所述电子设备处的第一模型和所述第二设备处的所述第二模型；

15

基于所述特征表示，根据所述第一模型，生成针对所述样本数据的预测标签；以及

基于所述特征表示、所述标签信息和所述预测标签，来确定用于训练所述第一模型和所述第二模型的总损失值。

20 10. 根据权利要求9所述的电子设备，其中确定所述总损失值包括：

基于所述标签信息和所述预测标签，来确定第一损失，所述第一损失表示所述预测标签与所述标签信息的差别程度；

25 基于所述特征表示和所述标签信息，来确定第二损失，所述第二损失表示所述特征表示与所述标签信息的相关程度；以及

基于所述第一损失和所述第二损失，来确定所述总损失值。

11. 根据权利要求10所述的电子设备，其中确定所述总损失值包括：

将所述第一损失与所述第二损失的加权和确定为所述总损失值。

12. 根据权利要求 10 所述的电子设备, 其中确定所述第二损失包括:

将所述特征表示与所述标签信息的距离相关系数确定为所述第二损失。

5 13. 根据权利要求 9 所述的电子设备, 其中所述动作还包括:

通过从所述总损失值与至少一个其他总损失值中确定最小总损失值, 来确定所述第一模型的参数, 所述至少一个其他总损失值是由所述电子设备先前或者之后确定的。

14. 根据权利要求 9 所述的电子设备, 其中所述动作还包括:

10 由所述第一设备确定所述总损失值相对于所述特征表示的梯度;
以及

向所述第二设备发送所述梯度, 以使得所述第二设备根据所述梯度来确定所述第二模型的参数。

15 15. 根据权利要求 9 所述的电子设备, 其中所述标签信息包括用户对于所述样本数据的处理过程的敏感数据。

16. 根据权利要求 9 所述的电子设备, 其中所述第二设备不具有所述标签信息。

17. 一种用于保护数据的装置, 包括:

20 获取模块, 被配置为获取由第二设备基于样本数据、根据第二模型生成的特征表示, 所述装置具有针对所述样本数据的标签信息, 所述装置和所述第二设备用于联合训练所述装置处的第一模型和所述第二设备处的所述第二模型;

标签预测模块, 被配置为基于所述特征表示, 根据所述第一模型, 生成针对所述样本数据的预测标签; 以及

25 总损失值确定模块, 被配置为基于所述特征表示、所述标签信息和所述预测标签, 来确定用于训练所述第一模型和所述第二模型的总损失值。

18. 一种计算机可读存储介质, 其上存储有计算机程序, 所述程序被处理器执行时实现根据权利要求 1 至 8 中任一项所述的方法。