

(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
2023 年 4 月 27 日 (27.04.2023)



(10) 国际公布号
WO 2023/065632 A1

(51) 国际专利分类号:
G06F 21/62 (2013.01)

(21) 国际申请号: PCT/CN2022/089872

(22) 国际申请日: 2022 年 4 月 28 日 (28.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111229481.X 2021 年 10 月 21 日 (21.10.2021) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。

(72) 发明人: 郑旭如(ZHENG, Xuru); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。
赵盟盟(ZHAO, Mengmeng); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。
王磊(WANG, Lei); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。

(74) 代理人: 深圳市力道知识产权代理事务所(普通合伙)(SHENZHEN LIDAO INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦一路 50 号鹏基商务时空大厦 1718, Guangdong 518000 (CN)。

(54) **Title:** DATA DESENSITIZATION METHOD, DATA DESENSITIZATION APPARATUS, DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 数据脱敏方法、数据脱敏装置、设备及存储介质

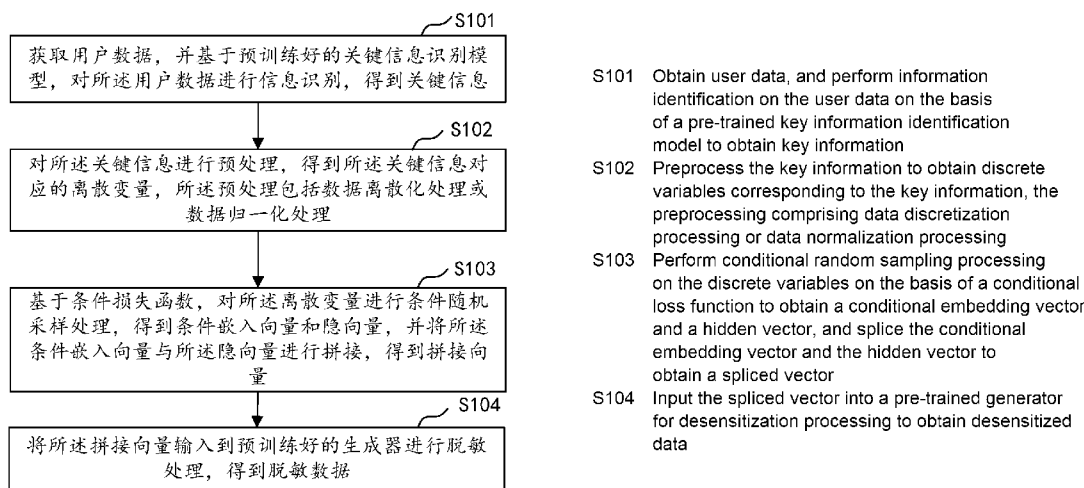


图 2

(57) **Abstract:** The present application relates to the field of artificial intelligence, in particular to a data desensitization method, a data desensitization apparatus, a device, and a storage medium. The method comprises: obtaining user data, and performing information identification on the user data on the basis of a pre-trained key information identification model to obtain key information; preprocessing the key information to obtain discrete variables corresponding to the key information, the preprocessing comprising data discretization processing or data normalization processing; performing conditional random sampling processing on the discrete variables on the basis of a conditional loss function to obtain a conditional embedding vector and a hidden vector, and splicing the conditional embedding vector and the hidden vector to obtain a spliced vector; and inputting the spliced vector into a pre-trained generator for desensitization processing to obtain desensitized data. Therefore, desensitized data cannot be easily reversely cracked, thereby ensuring that privacy data will not be leaked, and improving the security of the desensitized data.

(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, JM, 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:

一 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请涉及人工智能领域, 尤其涉及数据脱敏方法、数据脱敏装置、设备及存储介质, 所述方法包括: 获取用户数据, 并基于预训练好的关键信息识别模型, 对所述用户数据进行信息识别, 得到关键信息; 对所述关键信息进行预处理, 得到所述关键信息对应的离散变量, 所述预处理包括数据离散化处理或数据归一化处理; 基于条件损失函数, 对所述离散变量进行条件随机采样处理, 得到条件嵌入向量和隐向量, 并将所述条件嵌入向量与所述隐向量进行拼接, 得到拼接向量; 将所述拼接向量输入到预训练好的生成器进行脱敏处理, 得到脱敏数据。由此能够使脱敏数据的无法被轻易逆向破解, 从而保证了隐私数据不被泄露, 提高了脱敏数据的安全性。

数据脱敏方法、数据脱敏装置、设备及存储介质

本申请要求于 2021 年 10 月 21 日提交中国专利局、申请号为 202111229481.X、发明名称为“数据脱敏方法、数据脱敏装置、设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能领域，尤其涉及数据脱敏方法、数据脱敏装置、计算机设备及存储介质。

背景技术

大数据时代下，针对数据进行的攻击频率日益增多，攻击方式也日趋丰富。数据脱敏技术是解决数据安全风险和一种行之有效的方法。数据脱敏是指对关键信息或个人信息按照预设的规则或者变换进行数据变形，从而使得个人身份无法识别或者隐去关键信息。目前常见的结构化数据脱敏方式有基于匿名化技术或者扰乱技术的脱敏方式。

但是发明人意识到基于匿名化技术或扰乱技术的结构化数据脱敏方法中存在脱敏后数据与原数据是一对一的映射关系的问题，导致脱敏后的数据容易被逆向破解，从而被轻易还原出原数据，进而导致原数据中隐私信息的泄露，数据安全性较差。

发明内容

本申请提供了一种数据脱敏方法、数据脱敏装置、计算机设备及存储介质，旨在解决现有的脱敏方式存在容易被逆向破解导致隐私信息容易被泄露的问题。

为实现上述目的，本申请提供一种数据脱敏方法，所述方法包括：

获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

为实现上述目的，本申请还提供一种数据脱敏装置，所述数据脱敏装置包括：

关键信息提取模块，用于获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

信息处理模块，用于对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

向量拼接模块，用于基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

数据脱敏模块，用于将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

此外，为实现上述目的，本申请还提供一种计算机设备，所述计算机设备包括存储器和处理器；所述存储器，用于存储计算机程序；所述处理器，用于执行所述的计算机程序并在执行所述的计算机程序时实现如下步骤：

获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，

得到关键信息；

对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时使所述处理器实现如下步骤：

获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

本申请实施例公开的数据脱敏方法、数据脱敏装置、设备及存储介质，通过提取用户数据的关键信息以及关键信息的离散变量，从而生成拼接向量，并利用预训练好的生成器对拼接向量进行脱敏处理，得到脱敏数据，由此能够使脱敏数据的无法轻易被逆向破解，从而保证了隐私数据不被泄露，提高了脱敏数据的安全性。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，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的附图。

图1是本申请实施例提供的一种数据脱敏方法的场景示意图；

图2是本申请实施例提供的一种数据脱敏方法的流程示意图；

图3是本申请一实施例提供的一种数据脱敏装置的示意性框图；

图4是本申请一实施例提供的一种计算机设备的示意性框图。

具体实施方式

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，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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他实施例，都属于本申请保护的范围。

附图中所示的流程图仅是示例说明，不是必须包括所有的内容和操作/步骤，也不是必须按所描述的顺序执行。例如，有的操作/步骤还可以分解、组合或部分合并，因此实际执行的顺序有可能根据实际情况改变。另外，虽然在装置示意图中进行了功能模块的划分，但是在某些情况下，可以以不同于装置示意图中的模块划分。

在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

数据脱敏技术是一种可以通过对数据进行处理，从而达到降低和去除数据敏感程度的数据处理技术。采用数据脱敏技术，可以降低数据泄露的风险和危害，有效地保护用户数据的隐私。在互联网或医疗领域，用户可以通过个人数字空间存储、查看和分享个人医疗健康数据，但是个人的医疗数据在网上看病、网上购买药品、门诊预约等流程中会面临用户医疗敏感信息泄露的风险，而在医疗行业用户的数据具有极高的真实性和敏感性特点，一旦用户的个人敏感信息泄露可能会对用户本人造成潜在的生命威胁。借助数据脱敏，个人数字空间中的信息可以被用于业务相关的分析和处理，同时避免用户数据的泄露。

目前常见的结构化数据脱敏方式有基于匿名化技术或者扰乱技术的脱敏方式。常见的匿名化技术有 k-匿名、l-多样性和 t-closeness 等，它们是通过将单条记录的准标识符进行泛化后使得该数据在整个数据集中无法进行区分，从而达到脱敏的效果。基于扰乱技术则是往记录中加入噪声，比如在连续值中加入加性或者乘性噪声，从而达到脱敏的效果。

而基于匿名化技术或扰乱技术的结构化数据脱敏方法中存在脱敏后数据与原数据是一对一的映射关系的问题，导致脱敏后的数据存在被逆向的风险，而且经过脱敏的数据往往与原数据差别较大而失去了研究的价值。

为解决上述问题，本申请提供了一种数据脱敏方法，可以应用在服务器中，具体应用金融、医疗等多个领域，通过不断对生成器参数进行迭代更新，得到预训练好的生成器，提取用户数据的敏感信息，并利用预训练好的生成器对敏感信息进行脱敏处理，得到脱敏数据，由此能够使脱敏数据的无法被轻易逆向破解，从而保证了隐私数据不被泄露，提高了脱敏数据的安全性。

其中，服务器例如可以为单独的服务器或服务器集群。但为了便于理解，以下实施例将应用于服务器的数据脱敏方法进行详细介绍。

下面结合附图，对本申请的一些实施方式作详细说明。在不冲突的情况下，下述的实施例及实施例中的特征可以相互组合。

如图 1 所示，本申请实施例提供的数据脱敏方法，可以应用于如图 1 所示的应用环境中。该应用环境中包含有终端设备 110 和服务器 120，其中，终端设备 110 可以通过网络与服务器 120 进行通信。具体地，服务器 120 获取终端设备 110 发送的用户数据，服务器 120 对用户数据进行关键信息提取、信息处理以及脱敏处理后生成脱敏数据，并将该脱敏数据发送给终端设备 110，以实现数据脱敏处理。其中，服务器 120 可以是独立的服务器，也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network, CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。终端设备 110 可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。

请参阅图 2，图 2 是本申请实施例提供的一种数据脱敏方法的示意图。其中，该数据脱敏方法可以应用在服务器中，由此能够使脱敏数据的无法被轻易逆向破解，从而保证了隐私数据不被泄露，提高了脱敏数据的安全性。

如图 2 所示，该数据脱敏方法包括步骤 S101 至步骤 S104。

S101、获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息。

其中，所述用户数据为包含关键信息的数据，具体可以包括医疗数据比如病历数据、金融数据比如银行账户数据等。所述关键信息识别模型可以是基于注意力机制的预训练 BERT-CRF 模型，用于提取用户数据中的关键信息。所述关键信息是用户需要进行脱敏的信息，一般为用户的隐私信息，比如关键信息可以为病历数据中的身高、体重信息等，还可以为银行账户数据中的账户余额信息和投资信息等。需要说明的是，一切敏感信息或隐私信息均可以作为关键信息。

本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中，人工智能(Artificial Intelligence, AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

在一些实施例中，对所述用户数据进行分词处理，得到多个分词；对每个所述分词进行

特征提取，得到每个所述分词的嵌入特征；根据每个所述分词的嵌入特征进行词义预测，得到每个所述分词对应的词义；根据每个所述分词对应的词义对所述多个分词进行筛选，得到关键信息。由此可以能够准确地提取得到关键信息，提高脱敏数据生成的准确性以及安全性。

其中，所述嵌入特征为词嵌入特征，位置嵌入特征和切分嵌入特征。所述词嵌入特征是每个分词的向量表示，所述位置嵌入为每个分词位置的向量表示，所述切分嵌入特征用于区分不同的两句话。

具体地，可以基于分词算法对所述用户数据进行分词处理，得到多个分词，所述分词算法可以为正向最大匹配法、逆向最大匹配法、基于隐马尔可夫模型的分词算法、基于条件随机场的分词算法等算法。

示例性的，可以基于隐马尔可夫模型的分词算法对用户数据比如病历数据的病历文本为“患者存在尿频、易饿、焦虑、震颤等症状，疑似糖尿病，”进行分词，得到对应的多个分词比如尿频、易饿、焦虑、震颤等。

具体地，可以对每个所述分词进行特征提取，得到每个所述分词的嵌入特征，并基于词义预测模型，根据每个所述分词的嵌入特征对每个所述分词进行词义预测，得到每个分词对应的词义预测结果，并基于每个分词对应的词义预测结果对所述多个分词进行筛选，得到关键信息。由此能够最大程度地挖掘文本特征，抽取更丰富的词表示，消除传统词向量比如 Word2vec、Glove 等无法动态表示上下文信息和无法解决一词多义的缺点。从而可以快速得到每个分词与预设的标准敏感分词的相似度，进而快速得到对应的关键信息。

其中，所述词义预测模型用于预测每个分词与预设的标准敏感分词的相似程度，所述词义预测模型通过语义匹配模型与标准敏感分词数据库进行训练得到，所述语义预测模型可以包括 LSTM 匹配模型、MV-DSSM 模型、ESIM 模型等模型，所述词义预测结果为每个分词与标准敏感分词数据库中的标准敏感分词的相似度。

示例性的，比如分词包括账户信息的账户余额等分词以及股票走势信息等分词，可以对每个所述分词进行特征提取，得到每个所述分词的词嵌入特征，位置嵌入特征和切分嵌入特征，基于 LSTM 匹配模型，根据词嵌入特征，位置嵌入特征和切分嵌入特征对每个所述分词进行词义预测，得到每个分词对应的词义预测结果，并基于每个分词对应的词义预测结果将股票走势信息对应的分词筛选掉，得到关键信息。

S102、对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理。

由于关键信息一般都是连续型数据，因此需要对连续数据和离散型数据进行表示转换，即数据预处理操作，是神经网络的输入和输出的关键一步。

示例性的，当关键信息为身高、体重等信息时，则该关键信息为连续型数据，当关键信息为投资企业个数等信息时，则该关键信息为离散型数据。

其中，所述离散变量指变量值可以按一定顺序一一列举，通常以整数位取值的变量，比如职工人数、工厂数、机器台数等。具体地，所述数据归一化处理可以包括最大最小归一化处理和根据高斯混合模型进行归一化处理；所述数据离散化处理可以包括 K-bins 离散化处理和回归树离散化处理。

在一些实施例中，对所述关键信息进行最大最小归一化处理，得到所述关键信息对应的离散变量；或，通过高斯混合模型对所述关键信息进行归一化处理，得到所述关键信息对应的离散变量；或，对所述关键信息进行 K-bins 离散化处理，得到所述关键信息对应的离散变量；或，对所述关键信息进行回归树离散化处理，得到所述关键信息对应的离散变量。

具体地，若关键信息为连续型数据，可以通过最大最小线性变换被映射到[0,1]的范围内，使得利用 tanh 激活函数可以表示该连续值，得到所述关键信息对应的离散变量。

具体地，若关键信息为连续型数据，可以通过高斯混合模型对关键信息进行拟合，根据该关键信息在混合模型的高斯分量的概率来采样高斯分量，用采样出来的高斯分量对记录中的关键信息进行归一化表示。那么，关键信息即可由归一化后的表示以及高斯分量的独热编

码共同构成，从而得到所述关键信息对应的离散变量。

具体地，若关键信息为连续型数据，可以对所述关键信息进行 K-bins 离散化处理，得到所述关键信息对应的离散变量。其中，所述离散化，也可称为分箱，即按一定规则将关键信息分入各个区间中，并用独热编码表示每个区间，从而将关键信息用一个含有四个区间的分段函数进行拟合，得到所述关键信息对应的离散变量。

具体地，若关键信息为连续型数据，可以利用 CART 回归树对所述关键信息进行离散化处理，得到所述关键信息对应的离散变量。其中，CART 回归树可以预测连续型数据，其叶子节点即表示一个预测值。将关键信息的回归树或回归树集合的一系列叶子节点通过独热编码表示，即可将该关键信息转化为离散值。

需要说明的是，若关键信息为离散型数据，则无需进行数据离散化处理或数据归一化处理。

S103、基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量。

其中，所述条件损失函数为基于对抗生成网络的条件损失函数，所述基于对抗生成网络的损失函数的数据项是基于条件概率生成的，其初衷是使得数据能够按条件生成，使得同样类型的待脱敏数据与生成的脱敏数据的分布尽可能一致。然而，因为每次采样出来的条件都可能是不同的变量，因此，基于条件变量下的数据很难得到充分的训练，可以观察到生成的数据中对应的变量的值与其条件变量的值不一致。可通过对条件变量的预测，对训练过程进行约束，使得条件变量的值与生成的数据中对应的变量的值一致，即可进一步优化数据生成的效果。

具体地，所述条件嵌入向量可以随机等概率地从所述关键信息对应的多个离散变量中挑选出一个符合预设条件的离散变量，所述隐向量可以从所述关键信息对应的白噪声采样得到，所述拼接向量是通过所述条件嵌入向量与所述隐向量拼接得到的，用于作为生成器的输入。通过加入隐向量，从而改变了脱敏后数据与原数据一对一的映射关系，从而使脱敏后的数据不容易被逆向破解，得到隐私信息。

具体地，可以通过构建离散变量下各个值的概率质量分布函数，得到该离散变量的分布式表示，并对该离散变量的分布式表示进行条件随机采样处理，得到条件嵌入向量和隐向量。

示例性的，可以通过深度神经网络转换离散变量对应的白噪声，以从该离散变量的分布式表示生成隐向量。

在一些实施例中，对所述条件嵌入向量进行转化处理，得到独热编码；将所述独热编码与所述隐向量进行拼接，得到拼接向量。其中，所述独热编码即 One-Hot 编码，又称有效编码，其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候，其中只有一位有效。将所述条件嵌入向量转化为独热编码能够解决判别器不好处理属性数据的问题，同时也在一定程度上也起到了扩充向量特征的作用。

具体地，可以通过深度神经网络对条件嵌入向量进行转化处理，得到独热编码，并通过独热编码与隐向量进行拼接，得到拼接向量。由此可以得到符合生成器输入需求的拼接向量。

S104、将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

其中，所述预训练好的生成器是基于对抗生成网络训练生成的，所述脱敏数据为对待脱敏数据中关键信息进行脱敏后的数据。

在一些实施例中，获取训练数据对应的拼接向量，并将所述拼接向量输入到第一生成器进行脱敏处理，得到脱敏后的数据；基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器；根据预设的学习率和所述预训练好的判别器的参数，对所述第一生成器的参数进行多次迭代更新，得到第二生成器，并将所述第二生成器作为预训练好的生成器。由此可以通过预训练好的判别器和脱敏后的数据对第一生成器的参数进行多次迭代更新，能够生成非常真实的脱敏数据。之所以要先训练得到预训练好的判别器，再训练生成器，是因为要先拥有一个好的判别器，使得能够教好地区分出待脱敏数据和生成的脱敏数

据之后，才能够更为准确地对生成器的参数进行更新。

其中，所述训练数据为用于训练生成器参数的待脱敏数据集，所述第一生成器为预设的未经过训练的生成器，所述第二生成器为第一生成器经过多次迭代更新生成的。其中，所述第一生成器和所述第二生成器的参数不同。可以通过离散变量的分布式表示得到该离散变量的先验概率，并在先验概率中采样出参数，以此作为第一生成器的参数。具体地，可以通过随机梯度哈密顿蒙特卡洛方法对生成器和判别器进行训练，得到预训练好的生成器和预训练好的判别器。

具体地，基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器是通过将所述条件嵌入向量分别与所述脱敏后的数据和所述训练数据进行拼接，得到第一拼接数据和第二拼接数据，并根据计算第一拼接数据和第二拼接数据的相似度，并根据第一拼接数据和第二拼接数据的相似度优化损失函数，并对通过该损失函数对判别器进行梯度裁剪，得到预训练好的判别器。

示例性的，可以通过第一生成器和预设的判别器参数对判别器参数进行训练，将脱敏数据尽可能判别为假，从而调整判别器参数，进而提高判别器对于待脱敏数据的判别能力。

示例性的，可以通过第一生成器和预训练好的判别器参数的先验概率计算第二生成器的后验概率，从而使该脱敏数据要尽可能使得判别器误判其为待脱敏数据，从而调整生成器参数，能够生成真实的脱敏数据。

在一些实施例，得到第二生成器之后，基于统计信息的损失函数，对所述第二生成器进行增噪处理，得到预训练好的生成器，其中，所述第一生成器、所述参数更新后的生成器和所述预训练好的生成器的参数不同。由此可以控制脱敏数据的生成质量和脱敏程度。

其中，所述基于统计信息的损失函数可以包括基于均值的损失函数、基于方差的损失函数等。

具体地，可以对第二生成器的参数加入高斯噪声，由此可以实现多项式拟合正弦曲线，所述高斯噪声是符合高斯正态分布的误差。而高斯噪声的具体数值可以通过实验得到。

示例性的，可以对第二生成器的参数引入误差项，从而对第二生成器的参数进行修正，得到预训练好的生成器。由于误差项的存在，使得生成的脱敏数据与原数据存在一定差别，但差别不大，由此可以避免经过脱敏的数据往往与原数据差别较大而失去了研究的价值，同时也保证数据不会被轻易被逆向。

在一些实施例，得到脱敏数据之后，对所述脱敏数据的离散变量进行随机采样处理，得到目标离散变量；基于逻辑回归模型，根据所述脱敏数据的其余离散变量对所述目标离散变量进行预测，得到所述目标离散变量的预测结果；基于所述目标离散变量的预测结果调整所述预训练好的生成器的参数。由此可以通过对离散变量进行预测，从而调整生成器的参数，进而达到更好的脱敏效果。这里更好的脱敏效果是指可以防止脱敏后的数据被逆向破解，同时还能保持与原数据的关联。

其中，所述目标离散变量为从脱敏数据的多个离散变量中随机采样得到的，同时为了脱敏数据与原数据的关联离散变量，一般可以认为目标离散变量不发生改变，脱敏数据与原数据差别较小，便不会失去了研究的价值，因此需要保证目标离散变量的一致性。所述逻辑回归模型用于对离散变量进行预测。

具体地，可以利用交叉熵损失函数来判断生成的目标离散变量的预测结果与目标离散变量是否一致，从而确定所述脱敏数据的生成质量。若目标离散变量的预测结果与目标离散变量一致，则无需对所述预训练好的生成器的参数进行调整；若目标离散变量的预测结果与目标离散变量不一致，则确定目标离散变量的预测结果与目标离散变量的差值，根据所述差值调整所述预训练好的生成器的参数。由此可以确定目标离散变量的准确性，避免生成的脱敏数据使原数据差别过大。由于脱敏数据与原数据的大部分离散变量是相同的，因此去除其中一个离散变量，可以根据其余离散变量准确地预测得到该离散变量。

示例性的，若脱敏数据的目标离散变量是鞋码为43码，基于逻辑回归模型，则可以通过

脱敏数据的其余离散变量比如身高、体重等对目标离散变量进行预测，得到对于鞋码的预测结果，判断生成的鞋码的预测结果与脱敏数据的鞋码是否一致，比如若生成的鞋码的预测结果是鞋码为 40 码，则确定差值为 1 码，根据所述差值对所述预训练好的生成器的参数进行不断地迭代更新；若生成的鞋码的预测结果是鞋码为 43 码，则无需对所述预训练好的生成器的参数进行调整。

在一些实施例中，服务器还可以将用于提示用户脱敏数据已经生成的提示信息发送给终端设备。

其中，所述提示信息的方式具体可以包括应用程序（APP）或 Email、短信、聊天工具，聊天工具例如微信、qq 等。

示例性的，当脱敏数据已经生成时，服务器会发送脱敏数据已经生成的提示信息给终端设备以提醒用户。

请参阅图 3，图 3 是本申请一实施例提供的一种数据脱敏装置的示意性框图，该数据脱敏装置可以配置于服务器中，用于执行前述的数据脱敏方法。

如图 3 所示，该数据脱敏装置 200 包括：关键信息提取模块 201、信息处理模块 202、向量拼接模块 203 和数据脱敏模块 204。

关键信息提取模块 201，用于获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

信息处理模块 202，用于对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

向量拼接模块 203，用于基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

数据脱敏模块 204，用于将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据；

特征提取模块 201，还用于对所述用户数据进行分词处理，得到多个分词；对每个所述分词进行特征提取，得到每个所述分词的嵌入特征；根据每个所述分词的嵌入特征进行词义预测，得到每个所述分词对应的词义；根据每个所述分词对应的词义对所述多个分词进行筛选，得到关键信息。

信息处理模块 202，还用于对所述关键信息进行最大最小归一化处理，得到所述关键信息对应的离散变量；或，通过高斯混合模型对所述关键信息进行归一化处理，得到所述关键信息对应的离散变量；或，对所述关键信息进行 K-bins 离散化处理，得到所述关键信息对应的离散变量；或，对所述关键信息进行回归树离散化处理，得到所述关键信息对应的离散变量。

向量拼接模块 203，还用于对所述条件嵌入向量进行转化处理，得到独热编码；将所述独热编码与所述隐向量进行拼接，得到拼接向量。

生成器训练模块 205，用于获取训练数据对应的拼接向量，并将所述拼接向量输入到第一生成器进行脱敏处理，得到脱敏后的数据；基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器；根据预设的学习率和所述预训练好的判别器的参数，对所述第一生成器的参数进行多次迭代更新，得到第二生成器，并将所述第二生成器作为预训练好的生成器。

生成器训练模块 205，还用于基于统计信息的损失函数，对所述第二生成器进行增噪处理，得到预训练好的生成器，其中，所述第一生成器、所述第二生成器和所述预训练好的生成器的参数不同。

生成器训练模块 205，还用于对所述脱敏数据的离散变量进行随机采样处理，得到目标离散变量；基于逻辑回归模型，根据所述脱敏数据的其余离散变量对所述目标离散变量进行预测，得到所述目标离散变量的预测结果；基于所述目标离散变量的预测结果调整所述预训练好的生成器的参数。

需要说明的是，所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的装置和各模块、单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

本申请的方法、装置可用于众多通用或专用的计算机系统环境或配置中。例如：个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器系统、基于微处理器的系统、机顶盒、可编程的消费终端设备、网络 PC、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。

示例性的，上述的方法、装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 4 所示的计算机设备上运行。

请参阅图 4，图 4 是本申请实施例提供的一种计算机设备的示意图。该计算机设备可以是服务器。

如图 4 所示，该计算机设备包括通过系统总线连接的处理器、存储器和网络接口，其中，存储器可以包括易失性存储介质、非易失性存储介质和内存存储器。

非易失性存储介质可存储操作系统和计算机程序。该计算机程序包括程序指令，该程序指令被执行时，可使得处理器执行任意一种数据脱敏方法。

处理器用于提供计算和控制能力，支撑整个计算机设备的运行。

内存存储器为非易失性存储介质中的计算机程序的运行提供环境，该计算机程序被处理器执行时，可使得处理器执行任意一种数据脱敏方法。

该网络接口用于进行网络通信，如发送分配的任务等。本领域技术人员可以理解，该计算机设备的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

应当理解的是，处理器可以是中央处理单元（Central Processing Unit, CPU），该处理器还可以是其他通用处理器、数字信号处理器（Digital Signal Processor, DSP）、专用集成电路（Application Specific Integrated Circuit, ASIC）、现场可编程门阵列（Field-Programmable Gate Array, FPGA）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该处理器也可以是任何常规的处理器等。

其中，在一些实施方式中，所述处理器用于运行存储在存储器中的计算机程序，以实现如下步骤：获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

在一些实施例中，所述处理器还用于：对所述用户数据进行分词处理，得到多个分词；对每个所述分词进行特征提取，得到每个所述分词的嵌入特征；根据每个所述分词的嵌入特征进行词义预测，得到每个所述分词对应的词义；根据每个所述分词对应的词义对所述多个分词进行筛选，得到关键信息。

在一些实施例中，所述处理器还用于：对所述关键信息进行最大最小归一化处理，得到所述关键信息对应的离散变量；或，通过高斯混合模型对所述关键信息进行归一化处理，得到所述关键信息对应的离散变量；或，对所述关键信息进行 K-bins 离散化处理，得到所述关键信息对应的离散变量；或，对所述关键信息进行回归树离散化处理，得到所述关键信息对应的离散变量。

在一些实施例中，所述处理器还用于：对所述条件嵌入向量进行转化处理，得到独热编码；将所述独热编码与所述隐向量进行拼接，得到拼接向量。

在一些实施例中，所述处理器还用于：获取训练数据对应的拼接向量，并将所述拼接向

量输入到第一生成器进行脱敏处理，得到脱敏后的数据；基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器；根据预设的学习率和所述预训练好的判别器的参数，对所述第一生成器的参数进行多次迭代更新，得到第二生成器，并将所述第二生成器作为预训练好的生成器。

在一些实施例中，所述处理器还用于：基于统计信息的损失函数，对所述第二生成器进行增噪处理，得到预训练好的生成器，其中，所述第一生成器、所述第二生成器和所述预训练好的生成器的参数不同。

在一些实施例中，所述处理器还用于：对所述脱敏数据的离散变量进行随机采样处理，得到目标离散变量；基于逻辑回归模型，根据所述脱敏数据的其余离散变量对所述目标离散变量进行预测，得到所述目标离散变量的预测结果；基于所述目标离散变量的预测结果调整所述预训练好的生成器的参数。

本申请实施例还提供一种计算机可读存储介质，所述计算机可读存储介质可以是非易失性，也可以是易失性。所述计算机可读存储介质上存储有计算机程序，所述计算机程序中包括程序指令，所述程序指令被执行时实现本申请实施例提供的任一种数据脱敏方法。

其中，所述计算机可读存储介质可以是前述实施例所述的计算机设备的内部存储单元，例如所述计算机设备的硬盘或内存。所述计算机可读存储介质也可以是所述计算机设备的外部存储设备，例如所述计算机设备上配备的插接式硬盘，智能存储卡(Smart Media Card, SMC)，安全数字(Secure Digital, SD)卡，闪存卡(Flash Card)等。

进一步地，所述计算机可读存储介质可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序等；存储数据区可存储根据区块链节点的使用所创建的数据等。

本申请所指区块链语言模型的存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

以上所述，仅为本申请的具体实施方式，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，可轻易想到各种等效的修改或替换，这些修改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权利要求的保护范围为准。

权 利 要 求 书

1、一种数据脱敏方法，其中，所述方法包括：

获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

2、根据权利要求1所述的方法，其中，所述基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息，包括：

对所述用户数据进行分词处理，得到多个分词；

对每个所述分词进行特征提取，得到每个所述分词的嵌入特征；

根据每个所述分词的嵌入特征进行词义预测，得到每个所述分词对应的词义；

根据每个所述分词对应的词义对所述多个分词进行筛选，得到关键信息。

3、根据权利要求1所述的方法，其中，所述对所述关键信息进行预处理，得到所述关键信息对应的离散变量，包括：

对所述关键信息进行最大最小归一化处理，得到所述关键信息对应的离散变量；或，

通过高斯混合模型对所述关键信息进行归一化处理，得到所述关键信息对应的离散变量；或，

对所述关键信息进行K-bins离散化处理，得到所述关键信息对应的离散变量；或，

对所述关键信息进行回归树离散化处理，得到所述关键信息对应的离散变量。

4、根据权利要求1所述的方法，其中，所述将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量，包括：

对所述条件嵌入向量进行转化处理，得到独热编码；

将所述独热编码与所述隐向量进行拼接，得到拼接向量。

5、根据权利要求1所述的方法，其中，所述方法还包括：

获取训练数据对应的拼接向量，并将所述拼接向量输入到第一生成器进行脱敏处理，得到脱敏后的数据；

基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器；

根据预设的学习率和所述预训练好的判别器的参数，对所述第一生成器的参数进行多次迭代更新，得到第二生成器，并将所述第二生成器作为预训练好的生成器。

6、根据权利要求5所述的方法，其中，所述得到第二生成器之后，所述方法还包括：

基于统计信息的损失函数，对所述第二生成器进行增噪处理，得到预训练好的生成器，其中，所述第一生成器、所述第二生成器和所述预训练好的生成器的参数不同。

7、根据权利要求1所述的方法，其中，所述得到脱敏数据之后，所述方法还包括：

对所述脱敏数据的离散变量进行随机采样处理，得到目标离散变量；

基于逻辑回归模型，根据所述脱敏数据的其余离散变量对所述目标离散变量进行预测，得到所述目标离散变量的预测结果；

基于所述目标离散变量的预测结果调整所述预训练好的生成器的参数。

8、一种数据脱敏装置，其中，包括：

关键信息提取模块，用于获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

信息处理模块，用于对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

向量拼接模块，用于基于条件损失函数，对所述离散变量进行条件随机采样处理，得到

条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

数据脱敏模块，用于将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

9、一种计算机设备，其中，所述计算机设备包括存储器和处理器；

所述存储器，用于存储计算机程序；

所述处理器，用于执行所述的计算机程序并在执行所述的计算机程序时实现如下步骤：

获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

10、根据权利要求9所述的计算机设备，其中，所述处理器在实现所述基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息，用于实现：

对所述用户数据进行分词处理，得到多个分词；

对每个所述分词进行特征提取，得到每个所述分词的嵌入特征；

根据每个所述分词的嵌入特征进行词义预测，得到每个所述分词对应的词义；

根据每个所述分词对应的词义对所述多个分词进行筛选，得到关键信息。

11、根据权利要求9所述的计算机设备，其中，所述处理器在实现所述对所述关键信息进行预处理，得到所述关键信息对应的离散变量，用于实现：

对所述关键信息进行最大最小归一化处理，得到所述关键信息对应的离散变量；或，

通过高斯混合模型对所述关键信息进行归一化处理，得到所述关键信息对应的离散变量；或，

对所述关键信息进行K-bins离散化处理，得到所述关键信息对应的离散变量；或，

对所述关键信息进行回归树离散化处理，得到所述关键信息对应的离散变量。

12、根据权利要求9所述的计算机设备，其中，所述处理器在实现所述将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量，用于实现：

对所述条件嵌入向量进行转化处理，得到独热编码；

将所述独热编码与所述隐向量进行拼接，得到拼接向量。

13、根据权利要求9所述的计算机设备，其中，所述处理器还用于实现：

获取训练数据对应的拼接向量，并将所述拼接向量输入到第一生成器进行脱敏处理，得到脱敏后的数据；

基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器；

根据预设的学习率和所述预训练好的判别器的参数，对所述第一生成器的参数进行多次迭代更新，得到第二生成器，并将所述第二生成器作为预训练好的生成器。

14、根据权利要求9所述的计算机设备，其中，所述处理器在实现所述得到脱敏数据之后，还用于实现：

对所述脱敏数据的离散变量进行随机采样处理，得到目标离散变量；

基于逻辑回归模型，根据所述脱敏数据的其余离散变量对所述目标离散变量进行预测，得到所述目标离散变量的预测结果；

基于所述目标离散变量的预测结果调整所述预训练好的生成器的参数。

15、一种计算机可读存储介质，其中，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时使所述处理器实现如下步骤：

获取用户数据，并基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息；

对所述关键信息进行预处理，得到所述关键信息对应的离散变量，所述预处理包括数据离散化处理或数据归一化处理；

基于条件损失函数，对所述离散变量进行条件随机采样处理，得到条件嵌入向量和隐向量，并将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量；

将所述拼接向量输入到预训练好的生成器进行脱敏处理，得到脱敏数据。

16、根据权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现所述基于预训练好的关键信息识别模型，对所述用户数据进行信息识别，得到关键信息，用于实现：

对所述用户数据进行分词处理，得到多个分词；

对每个所述分词进行特征提取，得到每个所述分词的嵌入特征；

根据每个所述分词的嵌入特征进行词义预测，得到每个所述分词对应的词义；

根据每个所述分词对应的词义对所述多个分词进行筛选，得到关键信息。

17、根据权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现所述对所述关键信息进行预处理，得到所述关键信息对应的离散变量，用于实现：

对所述关键信息进行最大最小归一化处理，得到所述关键信息对应的离散变量；或，

通过高斯混合模型对所述关键信息进行归一化处理，得到所述关键信息对应的离散变量；或，

对所述关键信息进行 K-bins 离散化处理，得到所述关键信息对应的离散变量；或，

对所述关键信息进行回归树离散化处理，得到所述关键信息对应的离散变量。

18、根据权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现所述将所述条件嵌入向量与所述隐向量进行拼接，得到拼接向量，用于实现：

对所述条件嵌入向量进行转化处理，得到独热编码；

将所述独热编码与所述隐向量进行拼接，得到拼接向量。

19、根据权利要求 15 所述的计算机可读存储介质，其中，所述处理器还用于实现：

获取训练数据对应的拼接向量，并将所述拼接向量输入到第一生成器进行脱敏处理，得到脱敏后的数据；

基于所述脱敏后的数据和训练数据对预设的判别器进行训练，得到预训练好的判别器；

根据预设的学习率和所述预训练好的判别器的参数，对所述第一生成器的参数进行多次迭代更新，得到第二生成器，并将所述第二生成器作为预训练好的生成器。

20、根据权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现所述得到脱敏数据之后，还用于实现：

对所述脱敏数据的离散变量进行随机采样处理，得到目标离散变量；

基于逻辑回归模型，根据所述脱敏数据的其余离散变量对所述目标离散变量进行预测，得到所述目标离散变量的预测结果；

基于所述目标离散变量的预测结果调整所述预训练好的生成器的参数。

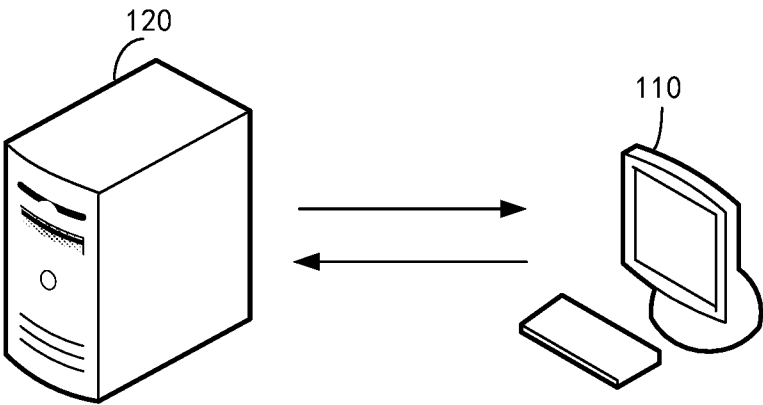


图 1

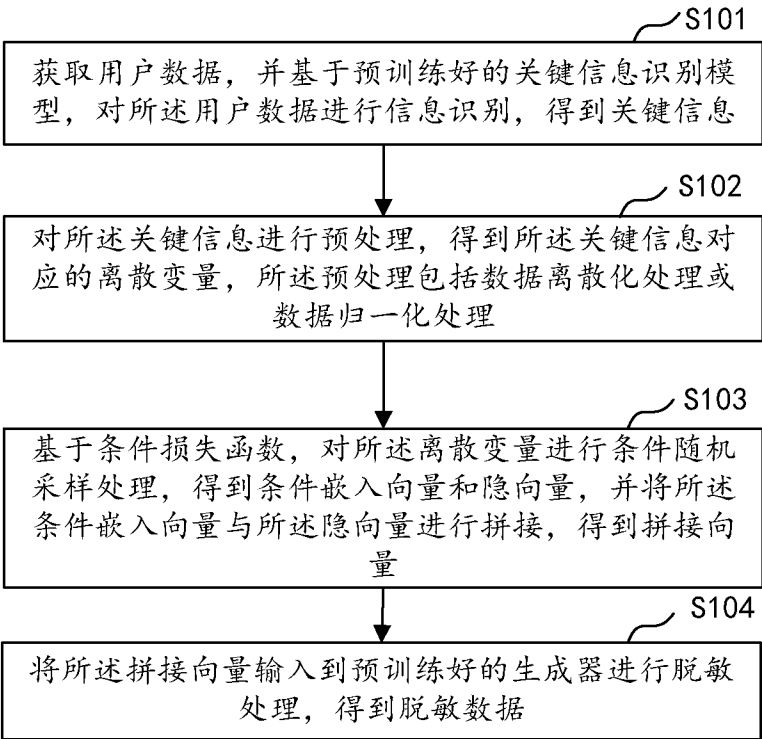


图 2

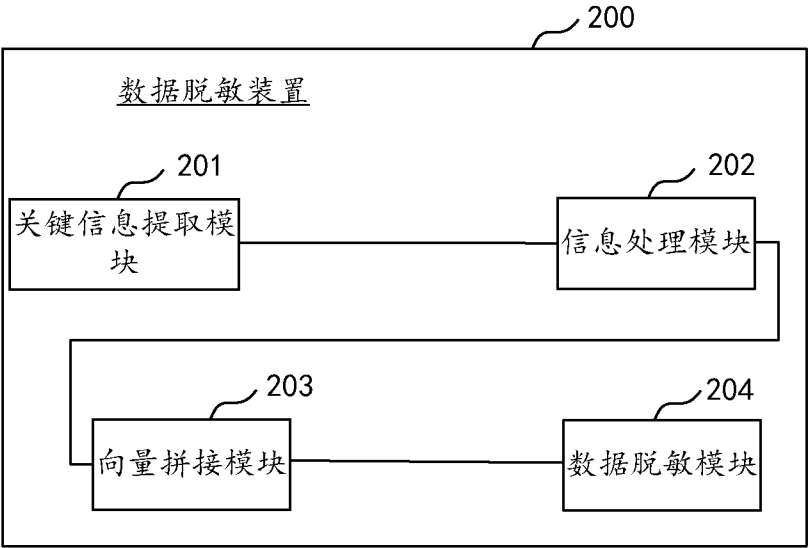


图 3

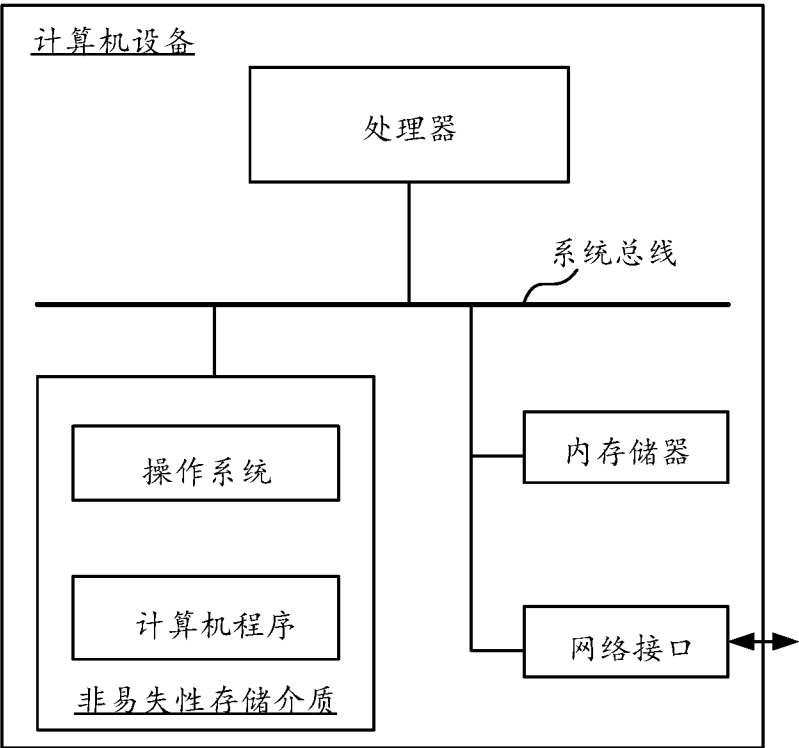


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/089872

A. CLASSIFICATION OF SUBJECT MATTER

G06F 21/62(2013.01)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)

G06F

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)

CNPAT, WPI, EPODOC, CNKI, IEEE: 脱敏, 敏感, 隐私, 识别, 分词, 提取, 筛选, 关键, 离散, 归一, 采样, 嵌入向量, 隐向量, 拼接, 独热编码, 训练, 模型, 调整, desensitization, privacy, identify, training, recognition, model, key, sample, discrete, vector, embedding, implicit, splicing

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113886885 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 04 January 2022 (2022-01-04) claims 1-10, and description, paragraphs [0024]-[0113], and figures 1-4	1-20
A	CN 111143884 A (BEIJING MEDICAL CROSS THE CLOUD TECHNOLOGY CO., LTD.) 12 May 2020 (2020-05-12) claims 9-10, and description, paragraphs [0094]-[0175]	1-20
A	CN 110188202 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 30 August 2019 (2019-08-30) entire document	1-20
A	CN 113254649 A (PING AN LIFE INSURANCE COMPANY OF CHINA, LTD.) 13 August 2021 (2021-08-13) entire document	1-20
A	WO 2021027533 A1 (PINGAN INTERNATIONAL SMART CITY TECHNOLOGY CO., LTD.) 18 February 2021 (2021-02-18) 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

29 June 2022

Date of mailing of the international search report

27 July 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
Information on patent family members

International application No.

PCT/CN2022/089872

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113886885	A	04 January 2022	None			
CN	111143884	A	12 May 2020	None			
CN	110188202	A	30 August 2019	None			
CN	113254649	A	13 August 2021	None			
WO	2021027533	A1	18 February 2021	CN	110598206	A	20 December 2019

A. 主题的分类 G06F 21/62 (2013.01) 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) G06F 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)) CNPAT, WPI, EPDOC, CNKI, IEEE: 脱敏, 敏感, 隐私, 识别, 分词, 提取, 筛选, 关键, 离散, 归一, 采样, 嵌入, 向量, 隐向量, 拼接, 独热编码, 训练, 模型, 调整, desensitization, privacy, identify, training, recognition, model, key, sample, discrete, vector, embedding, implicit, splicing																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 113886885 A (平安科技深圳有限公司) 2022年1月4日 (2022 - 01 - 04) 权利要求1-10, 说明书第[0024]-[0113]段, 附图1-4</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 111143884 A (北京懿医云科技有限公司) 2020年5月12日 (2020 - 05 - 12) 权利要求9-10, 说明书第[0094]-[0175]段</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 110188202 A (北京百度网讯科技有限公司) 2019年8月30日 (2019 - 08 - 30) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 113254649 A (中国平安人寿保险股份有限公司) 2021年8月13日 (2021 - 08 - 13) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>WO 2021027533 A1 (PINGAN INTERNATIONAL SMART CITY TECHNOLOGY CO., LTD.) 2021年2月18日 (2021 - 02 - 18) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 113886885 A (平安科技深圳有限公司) 2022年1月4日 (2022 - 01 - 04) 权利要求1-10, 说明书第[0024]-[0113]段, 附图1-4	1-20	A	CN 111143884 A (北京懿医云科技有限公司) 2020年5月12日 (2020 - 05 - 12) 权利要求9-10, 说明书第[0094]-[0175]段	1-20	A	CN 110188202 A (北京百度网讯科技有限公司) 2019年8月30日 (2019 - 08 - 30) 全文	1-20	A	CN 113254649 A (中国平安人寿保险股份有限公司) 2021年8月13日 (2021 - 08 - 13) 全文	1-20	A	WO 2021027533 A1 (PINGAN INTERNATIONAL SMART CITY TECHNOLOGY CO., LTD.) 2021年2月18日 (2021 - 02 - 18) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 113886885 A (平安科技深圳有限公司) 2022年1月4日 (2022 - 01 - 04) 权利要求1-10, 说明书第[0024]-[0113]段, 附图1-4	1-20																		
A	CN 111143884 A (北京懿医云科技有限公司) 2020年5月12日 (2020 - 05 - 12) 权利要求9-10, 说明书第[0094]-[0175]段	1-20																		
A	CN 110188202 A (北京百度网讯科技有限公司) 2019年8月30日 (2019 - 08 - 30) 全文	1-20																		
A	CN 113254649 A (中国平安人寿保险股份有限公司) 2021年8月13日 (2021 - 08 - 13) 全文	1-20																		
A	WO 2021027533 A1 (PINGAN INTERNATIONAL SMART CITY TECHNOLOGY CO., LTD.) 2021年2月18日 (2021 - 02 - 18) 全文	1-20																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
<table border="0"> <tr> <td> * 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 </td> <td> “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 </td> </tr> </table>			* 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																
* 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6月29日		国际检索报告邮寄日期 2022年7月27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 周亚楠 电话号码 86-(10)-53961530																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/089872

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	113886885	A	2022年1月4日	无	
CN	111143884	A	2020年5月12日	无	
CN	110188202	A	2019年8月30日	无	
CN	113254649	A	2021年8月13日	无	
WO	2021027533	A1	2021年2月18日	CN	110598206 A 2019年12月20日