(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109656878 A
(43)申请公布日 2019.04.19

(21)申请号 201811520690.8

(22)申请日 2018.12.12

(71)申请人 中电健康云科技有限公司
地址 610200 四川省成都市双流区东升街道成都芯谷产业园集中区内

(72)发明人 代超 徐茂 谭光鸿 吴佩军

(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 唐维虎

(51)Int.Cl.
G06F 16/11(2019.01)
G06K 9/62(2006.01)
G06N 3/08(2006.01)
G16H 10/60(2018.01)

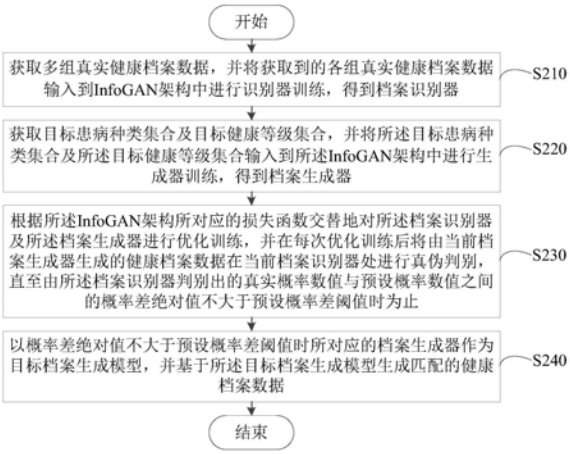
权利要求书2页 说明书9页 附图3页

(54)发明名称

健康档案数据生成方法及装置

(57)摘要

本申请提供一种健康档案数据生成方法及装置。所述方法通过多组真实健康档案数据在InfoGAN架构中训练得到档案识别器,通过目标患病种类集合及目标健康等级集合在InfoGAN架构中训练得到档案生成器,并通过InfoGAN架构的损失函数交替地对档案识别器及档案生成器进行优化训练,并在每次优化训练后基于当前档案识别器及当前档案生成器进行数据真伪判别,直至判别出的真实概率数值趋近于预设概率数值时为止的方式,从而以此时的档案生成器作为目标档案生成模型来模拟合成真实性高且数据准确度高的健康档案数据,并通过目标患病种类集合及目标健康等级集合各自的组成来确保数据多样性,便于医疗研究。



CN 109656878 A

1.一种健康档案数据生成方法,其特征在于,所述方法包括:

获取多组真实健康档案数据,并将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练,得到档案识别器,其中所述档案识别器用于判别档案数据真伪;

获取目标患病种类集合及目标健康等级集合,并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练,得到档案生成器,其中所述档案生成器用于生成健康档案数据;

根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练,并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别,直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止;

以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型,并基于所述目标档案生成模型生成匹配的健康档案数据。

2.根据权利要求1所述的方法,其特征在于,每组真实健康档案数据包括对应患者的档案记录年龄、该患者的历史患病种类,以及每个历史患病种类的历史患病次数,所述将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练,得到档案识别器的步骤包括:

针对每组真实健康档案数据,根据该组真实健康档案数据中的患者的档案记录年龄、该患者的历史患病种类及每个历史患病种类的历史患病次数,对该组真实健康档案数据进行编码,得到该组真实健康档案数据对应的每个患病种类的患病档案数据矩阵;

将该组真实健康档案数据对应的每个患病种类的患病档案数据矩阵进行矩阵耦合,得到该组真实健康档案数据所对应的档案特征矩阵;

在所述InfoGAN架构中基于得到的各组真实健康档案数据对应的档案特征矩阵进行识别器训练,得到所述档案识别器。

3.根据权利要求2所述的方法,其特征在于,所述根据该组真实健康档案数据中的患者的档案记录年龄、该患者的历史患病种类及每个历史患病种类的历史患病次数,对该组真实健康档案数据进行编码,得到该组真实健康档案数据对应的每个患病种类的患病档案数据矩阵的步骤包括:

根据所述真实健康档案数据中对应患者的档案记录年龄,计算该患者在患病时的患病时间权重;

针对该患者在所述真实健康档案数据中的每个历史患病种类,对该历史患病种类及该历史患病种类的历史患病次数进行ONE-HOT编码,得到对应的患病编码矩阵;

将每个历史患病种类所对应的患病编码矩阵与所述患病时间权重进行相乘运算,得到每个患病种类对应的所述患病档案数据矩阵。

4.根据权利要求2所述的方法,其特征在于,所述将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练,得到档案生成器的步骤包括:

根据所述目标患病种类集合及所述目标健康等级集合生成多组健康特征样本数据;

在所述InfoGAN架构中基于得到的多组健康特征样本数据进行神经网络模型训练,得到对应的档案生成器。

5.根据权利要求4所述的方法，其特征在于，所述根据所述目标患病种类集合及所述目标健康等级集合生成多组健康特征样本数据的步骤包括：

在生成每组健康特征样本数据时，从所述目标患病种类集合中随机选取至少一个目标患病种类，并从所述目标健康等级集合中随机选取至少一个目标健康等级；

将选取到的所述至少一个目标患病种类、选取到的至少一个目标健康等级，及由满足高斯分布的随机函数随机生成的噪声向量归纳于同一数据组合内，得到对应的健康特征样本数据。

6.根据权利要求4所述的方法，其特征在于，由所述档案生成器直接产生的数据为用于表示健康档案数据的格式与所述档案特征矩阵对应的健康特征矩阵，所述基于所述目标档案生成模型生成匹配的健康档案数据的步骤包括：

将基于所述目标档案生成模型生成的健康特征矩阵进行矩阵解耦，得到对应的至少一个档案状况数据矩阵，其中每个档案状况数据矩阵对应一个目标患病种类；

将得到的每个档案状况数据矩阵进行解码，并将每个档案状况数据矩阵在解码后所对应的档案数据进行归纳整理，得到对应的健康档案数据。

7.根据权利要求6所述的方法，其特征在于，所述档案数据包括对应目标患病种类的患病次数及档案记录年龄，所述将得到的每个档案状况数据矩阵进行解码的步骤包括：

将得到的所有档案状况数据矩阵进行最大公约数计算，并基于得到的最大公约数计算对应的档案记录年龄；

将每个所述档案状况数据矩阵与所述最大公约数进行相除运算，并对运算得到的每个患病编码矩阵进行ONE-HOT解码，得到每个档案状况数据矩阵所对应的目标患病种类，及每个所述目标患病种类的患病次数。

8.根据权利要求1-7中任意一项所述的方法，其特征在于，所述方法还包括：

对目标患病种类集合、目标健康等级集合、预设概率数值及预设概率差阈值进行配置。

9.一种健康档案数据生成装置，其特征在于，所述装置包括：

识别器训练模块，用于获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练，得到档案识别器，其中所述档案识别器用于判别档案数据真伪；

生成器训练模块，用于获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器，其中所述档案生成器用于生成健康档案数据；

优化训练模块，用于根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止；

数据生成模块，用于以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据。

10.根据权利要求9所述的装置，其特征在于，所述装置还包括：

参数配置模块，用于对目标患病种类集合、目标健康等级集合、预设概率数值及预设概率差阈值进行配置。

健康档案数据生成方法及装置

技术领域

[0001] 本申请涉及医疗数据生成技术领域，具体而言，涉及一种健康档案数据生成方法及装置。

背景技术

[0002] 随着人工智能技术及大数据技术的不断发展，越来越多的行业领域逐步涉及到了人工智能技术及大数据技术的应用，其中医疗领域便是众多行业领域中的一个极为重要的组成部分。而目前的医疗领域在数据获取方面存在很大不足，例如，现有的个人健康档案数据的获取过程除了花费大量时间来采集真实健康档案数据的方式之外，一般还包括通过统计分析法模拟合成健康档案数据的方式。针对这两种方式而言，前者会涉及到对患者既往病史的隐私保护，并受到医疗病例数据管制体系带来的影响而存在获取到的数据不完整、获取周期长、资源消耗大等问题，后者的实现算法复杂，在执行过程中需要大量的人工参与处理，以至于模拟合成的数据与真实数据之间的误差极大，整体的数据准确度不高，且数据多样性弱。

发明内容

[0003] 为了克服现有技术中的上述不足，本申请的目的在于提供一种健康档案数据生成方法及装置，所述健康档案数据生成方法的人工参与度极小，可模拟合成真实性高、数据准确度高且数据多样性强的健康档案数据，便于医疗研究。

[0004] 就方法而言，本申请实施例提供一种健康档案数据生成方法，所述方法包括获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练，得到档案识别器，其中所述档案识别器用于判别档案数据真伪；

[0005] 获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器，其中所述档案生成器用于生成健康档案数据；

[0006] 根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止；

[0007] 以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据。

[0008] 就装置而言，本申请实施例提供一种健康档案数据生成装置，所述装置包括：

[0009] 识别器训练模块，用于获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练，得到档案识别器，其中所述档案识别器用于判别档案数据真伪；

[0010] 生成器训练模块，用于获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器，其中所述档案生成器用于生成健康档案数据；

[0011] 优化训练模块，用于根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止；

[0012] 数据生成模块，用于以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据。

[0013] 相对于现有技术而言，本申请实施例提供的健康档案数据生成方法及装置具有以下有益效果：所述健康档案数据生成方法的人工参与度极小，可模拟合成真实性高、数据准确度高且数据多样性强的健康档案数据，便于医疗研究。所述方法在基于多组真实健康档案数据训练得到档案识别器，并基于目标患病种类集合及目标健康等级集合训练得到档案生成器后，通过交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止的方式，确保最终的经优化训练后的档案生成器所生成的健康档案数据与真实健康档案数据之间的数据相似性极强，从而模拟合成真实性高且数据准确度高的健康档案数据，并通过所述目标患病种类集合及所述目标健康等级集合各自的组成，确保基于所述方法得到的健康档案数据具有较强的数据多样性，以向医疗研究提供强大的数据支撑。

[0014] 为使本申请的上述目的、特征和优点能更明显易懂，下文特举本申请较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[0015] 为了更清楚地说明本申请实施例的技术方案，下面将对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的某些实施例，因此不应被看作是对本申请权利要求保护范围的限定，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。

[0016] 图1为本申请实施例提供的电子设备的方框示意图。

[0017] 图2为本申请实施例提供的健康档案数据生成方法的一种流程示意图。

[0018] 图3为本申请实施例提供的健康档案数据生成方法的另一种流程示意图。

[0019] 图4为本申请实施例提供的健康档案数据生成装置的一种方框示意图。

[0020] 图5为本申请实施例提供的健康档案数据生成装置的另一种方框示意图。

[0021] 图标：10-电子设备；11-存储器；12-处理器；13-通信单元；100-健康档案数据生成装置；110-识别器训练模块；120-生成器训练模块；130-优化训练模块；140-数据生成模块；150-参数配置模块。

具体实施方式

[0022] 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，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。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。

[0023] 因此，以下对在附图中提供的本申请的实施例的详细描述并非旨在限制要求保护的本申请的范围，而是仅仅表示本申请的选定实施例。基于本申请中的实施例，本领域普通技术人员在没有作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。

[0024] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

[0025] 在本申请的描述中，需要说明的是，除非另有明确的规定和限定，术语“设置”、“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本申请中的具体含义。

[0026] 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。

[0027] 请参照图1，是本申请实施例提供的电子设备10的方框示意图。在本申请实施例中，所述电子设备10可用于模拟合成真实性高、数据准确度高且数据多样性强的健康档案数据，减小人工参与度，并为医疗研究提供强大的数据支撑。其中，所述电子设备10可以是，但不限于，个人电脑(personal computer，PC)、平板电脑、个人数字助理(personal digital assistant，PDA)、移动上网设备(mobile Internet device，MID)等，所述健康档案数据用于指示患者对应的健康状况及历史患病情况。

[0028] 在本实施例中，所述电子设备10包括健康档案数据生成装置100、存储器11、处理器12及通信单元13。所述存储器11、处理器12及通信单元13各个元件相互之间直接或间接地电性连接，以实现数据的传输或交互。例如，所述存储器11、处理器12及通信单元13这些元件相互之间可通过一条或多条通讯总线或信号线实现电性连接。

[0029] 在本实施例中，所述存储器11为易失性存储器，所述存储器11可存储有InfoGAN(Information Maximizing Generative Adversarial Networks，信息最大化生成式对抗网络)架构，所述InfoGAN架构包括生成器模型及判别器模型，所述生成器模型通过数据训练可生成模拟合成满足数据获取要求的数据，所述判别器模型通过采用符合所述数据获取要求的真实数据进行数据训练，可用于对所述生成器模型生成的数据进行真伪判别。其后，所述判别器模型通过判别由所述生成器模型生成的数据相较于真实数据的真实概率数值，并在判别出的所述真实概率数值大于0.5时判定由所述生成器模型生成的数据为真，并在判别出的所述真实概率数值小于0.5时判定由所述生成器模型生成的数据为假。所述电子设备10在完成一次数据真伪判别后，通过交替地对所述判别器模型及所述生成器模型进行优化训练，以提高所述判别器模型的真伪判别能力，提高所述生成器模型的模拟合成能力。所述电子设备10在每次完成优化训练后，会针对当前的生成器模型及判别器模型进行一次

数据真伪判别，直至由优化训练后的所述生成器模型生成的数据在对应的所述判别器模型处进行数据真伪判别时的真实概率数值趋近于0.5时，停止优化训练过程，此时优化训练后的所述生成器模型所模拟合成的数据无法被所述判别器模型判别是真还是伪，所述优化训练后的所述生成器模型的模拟合成能力达到最佳。

[0030] 其中，所述电子设备10通过所述InfoGAN架构对应的损失函数进行优化训练，所述损失函数一种衡量损失和错误(这种损失与"错误地"估计有关，如费用或者设备的损失)程度的函数。所述电子设备10在交替地对所述判别器模型及所述生成器模型进行优化训练时，首先会固定生成器模型，而采用所述损失函数及真实数据对判别器模型进行优化训练更新，使该判别器模型具有当前的最佳真伪判别能力，而后在下一次优化优化训练时固定该判别器模型，而采用所述损失函数对生成器模型进行优化训练更新，提高生成器模型的模拟合成能力，使之达到当前最佳效果，从而相对地降低所述判别器模型对所述生成器模型生成的数据的真伪判别能力。

[0031] 在本实施例中，所述存储器11还用于存储预设概率数值及预设概率差阈值。所述电子设备10在进行一次数据真伪判别时，会将所述判别器模型判别出的真实概率数值与所述预设概率数值进行相减运算，并以相减得到的概率差数值的绝对值即概率差绝对值与所述预设概率差阈值进行比较，并所述概率差绝对值不大于预设概率差阈值时，判定所述真实概率数值趋近于所述预设概率数值。其中，所述预设概率数值为0.5，所述预设概率差阈值可以是0.01，也可以是0.002，还可以是0.03，具体的数值可根据需求进行不同的配置。在本实施例中，所述存储器11还可用于存储程序，所述处理器12在接收到执行指令后，可相应地执行所述程序。

[0032] 在本实施例中，所述处理器12可以是一种具有信号的处理能力的集成电路芯片。所述处理器12可以是通用处理器，包括中央处理器(Central Processing Unit，CPU)、图形处理器(Graphics Processing Unit，GPU)、网络处理器(Network Processor，NP)等。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，可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。

[0033] 在本实施例中，所述通信单元13用于通过网络建立所述电子设备10与其他用电设备之间的通信连接，并通过所述网络收发数据。例如，所述电子设备10可通过所述通信单元13从其他用电设备处获取患者的真实健康档案数据，也可通过所述通信单元13将最终训练得到的模拟合成能力达到最佳效果的生成器模型发送给其他用电设备，使所述其他用电设备直接基于得到的生成器模型模拟合成真实性高、数据准确度高且数据多样性强的健康档案数据。

[0034] 在本实施例中，所述健康档案数据生成装置100包括至少一个能够以软件或固件的形式存储于所述存储器11中或固化在所述电子设备10的操作系统中的软件功能模块。所述处理器12可用于执行所述存储器11存储的可执行模块，例如所述健康档案数据生成装置100所包括的软件功能模块及计算机程序等。所述电子设备10可通过所述健康档案数据生成装置100模拟合成真实性高、数据准确度高且数据多样性强的健康档案数据，减小人工参与度，为医疗研究提供强大的数据支撑。

[0035] 可以理解的是，图1所示的框图仅为电子设备10的一种结构组成示意图，所述电子设备10还可包括比图1中所示更多或者更少的组件，或者具有与图1所示不同的配置。图1中

所示的各组件可以采用硬件、软件或其组合实现。

[0036] 请参照图2，是本申请实施例提供的健康档案数据生成方法的一种流程示意图。在本申请实施例中，所述健康档案数据生成方法应用于上述的电子设备10，下面对图2所示的健康档案数据生成方法的具体流程和步骤进行详细阐述。

[0037] 步骤S210，获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练，得到档案识别器。

[0038] 在本实施例中，所述电子设备10可通过网络从其他用电设备处获取到多组真实健康档案数据，也可由所述电子设备10的使用人员人工导入多组真实健康档案数据，所述真实健康档案数据为现实生活中真实记录的健康档案数据。其中，每组真实健康档案数据包括对应患者的档案记录年龄、该患者的历史患病种类，以及每个历史患病种类的历史患病次数。例如，患者“张三”的档案记录年龄为24岁，患者“张三”的历史患病种类包括感冒、肺结核、肾结石三种，其中感冒的历史患病次数为40次，肺结核的历史患病次数为2次，肾结石的历史患病次数为2次。所述电子设备10在得到多组真实健康档案数据后，会将得到的多组真实健康档案数据输入到所述InfoGAN架构中的判别器模型中，并进行判别器训练，得到用于判别档案数据真伪的档案识别器。

[0039] 可选地，在本实施例中，所述将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练，得到档案识别器的步骤包括：

[0040] 针对每组真实健康档案数据，根据该组真实健康档案数据中的患者的档案记录年龄、该患者的历史患病种类及每个历史患病种类的历史患病次数，对该组真实健康档案数据进行编码，得到该组真实健康档案数据对应的每个患病种类的患病档案数据矩阵；

[0041] 将该组真实健康档案数据对应的每个患病种类的患病档案数据矩阵进行矩阵耦合，得到该组真实健康档案数据所对应的档案特征矩阵；

[0042] 在所述InfoGAN架构中基于得到的各组真实健康档案数据对应的档案特征矩阵进行识别器训练，得到所述档案识别器。

[0043] 其中，所述矩阵耦合的过程包括将患病档案数据矩阵合并为一个容置矩阵，并将该容置矩阵进行转置后得到的转置矩阵与该容置矩阵进行矩阵乘法运算。

[0044] 在本实施例的一种实施方式中，所述根据该组真实健康档案数据中的患者的档案记录年龄、该患者的历史患病种类及每个历史患病种类的历史患病次数，对该组真实健康档案数据进行编码，得到该组真实健康档案数据对应的每个患病种类的患病档案数据矩阵的步骤包括：

[0045] 根据所述真实健康档案数据中对应患者的档案记录年龄，计算该患者在患病时的患病时间权重；

[0046] 针对该患者在所述真实健康档案数据中的每个历史患病种类，对该历史患病种类及该历史患病种类的历史患病次数进行ONE-HOT编码，得到对应的患病编码矩阵；

[0047] 将每个历史患病种类所对应的患病编码矩阵与所述患病时间权重进行相乘运算，得到每个患病种类对应的所述患病档案数据矩阵。

[0048] 其中，所述患病时间权重的计算公式可用档案记录年龄与预设常数之积进行表示，所述预设常数可以是365，也可以是366，还可以是260，具体的数值可根据需求进行不同的配置。所述电子设备10在进行ONE-HOT编码时，会对当前能够从外界获取到的患病种类进

行统计排序，并对每组真实健康档案数据在统计的所有患病种类中对应存在的历史患病种类进行ONE-HOT编码。例如，统计得到的所有患病种类为n种，某组真实健康档案数据在统计的所有患病种类中对应存在的一个历史患病种类属于第三个患病种类时，该真实健康档案数据在统计的所有患病种类中对应存在的历史患病种类将被编码为[0,0,1,0,……,0]的1×n维矩阵，而当该历史患病种类所对应的历史患病次数为m次时，该历史患病种类及对应的历史患病次数将被编码为[0,0,1,0,……,0,m]的1×(n+1)维的患病编码矩阵。

[0049] 步骤S220，获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器。

[0050] 在本实施例中，所述目标患病种类集合用于表示所述电子设备10在模拟合成健康档案数据的过程中所要涉及到的患病种类的集合，所述目标健康等级集合用于表示所述电子设备10在模拟合成健康档案数据的过程中所要涉及到的健康等级的集合。其中，所述目标患病种类集合可以是感冒、骨折、静脉曲张、肺结核的集合，也可以是感冒、骨折、静脉曲张的集合，具体的集合情况可根据需求进行不同的配置；所述目标健康等级集合可以是正常、亚健康、伤残、健康等多种健康等级的集合，也可以是亚健康、伤残、健康等多种健康等级的集合，具体的集合情况可根据需求进行不同的配置。所述电子设备10在得到目标患病种类集合及目标健康等级集合后，会将得到的目标患病种类集合及目标健康等级集合输入到所述InfoGAN架构中的生成器模型中，并进行生成器训练，得到用于生成健康档案数据的档案生成器。

[0051] 可选地，在本实施例中，所述将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器的步骤包括：

[0052] 根据所述目标患病种类集合及所述目标健康等级集合生成多组健康特征样本数据；

[0053] 在所述InfoGAN架构中基于得到的多组健康特征样本数据进行神经网络模型训练，得到对应的档案生成器。

[0054] 其中，所述根据所述目标患病种类集合及所述目标健康等级集合生成多组健康特征样本数据的步骤包括：

[0055] 在生成每组健康特征样本数据时，从所述目标患病种类集合中随机选取至少一个目标患病种类，并从所述目标健康等级集合中随机选取至少一个目标健康等级；

[0056] 将选取到的所述至少一个目标患病种类、选取到的至少一个目标健康等级，及由满足高斯分布的随机函数随机生成的噪声向量归纳于同一数据组合内，得到对应的健康特征样本数据。

[0057] 步骤S230，根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止。

[0058] 在本实施例中，当所述电子设备10得到所述档案识别器及所述档案生成器后，所述电子设备10会以所述InfoGAN架构所对应的损失函数对所述档案识别器及所述档案生成器进行交替式优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在

当前档案识别器处进行真伪判别，使当前档案识别器针对由当前档案生成器生成的健康档案数据输出对应的真实概率数值，而后求出所述真实概率数值与所述预设概率数值之间的概率差绝对值，并将所述概率差绝对值与所述预设概率差阈值进行比较。若所述概率差绝对值大于所述预设概率差阈值，则进行下一次的优化训练流程；若所述概率差绝对值大于所述预设概率差阈值，则停止后续的优化训练流程，此时优化训练后的所述档案生成器所模拟合成的健康档案数据无法被当前档案判别器判别是真还是伪，所述优化训练后的所述档案生成器的模拟合成能力达到最佳效果，所述电子设备10可基于此时的所述优化训练后的所述档案生成器模拟合成真实性高且数据准确度高的健康档案数据，并通过所述目标患病种类集合及所述目标健康等级集合各自的组成，确保基于所述方法得到的健康档案数据具有较强的数据多样性，以向医疗研究提供强大的数据支撑。

[0059] 步骤S240，以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据。

[0060] 在本实施例中，当所述电子设备10得到概率差绝对值不大于预设概率差阈值时所对应的档案生成器时，所述电子设备10讲以该档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据，其中所述电子设备10可通过选取所述目标档案生成模型所对应的目标患病种类集合中的至少一个目标患病种类，和/或选取所述目标档案生成模型所对应的目标健康等级集合中至少一个目标健康等级的方式，基于该目标档案生成模型生成与被选取的目标患病种类及目标健康等级匹配的健康档案数据。

[0061] 在本实施例中，所述由所述档案生成器直接产生的数据为用于表示健康档案数据的格式与所述档案特征矩阵对应的健康特征矩阵，所述基于所述目标档案生成模型生成匹配的健康档案数据的步骤包括：

[0062] 将基于所述目标档案生成模型生成的健康特征矩阵进行矩阵解耦，得到对应的至少一个档案状况数据矩阵，其中每个档案状况数据矩阵对应一个目标患病种类；

[0063] 将得到的每个档案状况数据矩阵进行解码，并将每个档案状况数据矩阵在解码后所对应的档案数据进行归纳整理，得到对应的健康档案数据。

[0064] 其中，所述矩阵解耦的过程为上述矩阵耦合过程的反向步骤，在此就不一一赘述了。所述档案数据包括对应目标患病种类的患病次数及档案记录年龄，所述将得到的每个档案状况数据矩阵进行解码的步骤包括：

[0065] 将得到的所有档案状况数据矩阵进行最大公约数计算，并基于得到的最大公约数计算对应的档案记录年龄；

[0066] 将每个所述档案状况数据矩阵与所述最大公约数进行相除运算，并对运算得到的每个患病编码矩阵进行ONE-HOT解码，得到每个档案状况数据矩阵所对应的目标患病种类，及每个所述目标患病种类的患病次数。

[0067] 其中，计算得到的所述最大公约数即为所有档案状况数据矩阵对应的模拟患病时间权重，所述电子设备10通过将所述模拟患病时间权重除以上述的预设常数的方式，得到对应的档案记录年龄。所述ONE-HOT解码过程为上述ONE-HOT编码过程的反向步骤，在此就不一一赘述了。

[0068] 请参照图3，是本申请实施例提供的健康档案数据生成方法的另一种流程示意图。在本申请实施例中，在所述步骤S210之前，所述健康档案数据生成方法还包括步骤S209。

[0069] 步骤S209，对目标患病种类集合、目标健康等级集合、预设概率数值及预设概率差阈值进行配置。

[0070] 在本实施例中，所述电子设备10的使用人员可根据模型训练要求对所述目标患病种类集合、所述目标健康等级集合及所述预设概率差阈值进行不同的配置，并将所述预设概率数值配置为0.5。

[0071] 请参照图4，是本申请实施例提供的健康档案数据生成装置100的一种方框示意图。在本申请实施例中，所述健康档案数据生成装置100包括识别器训练模块110、生成器训练模块120、优化训练模块130及数据生成模块140。

[0072] 所述识别器训练模块110，用于获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到信息最大化生成式对抗网络InfoGAN架构中进行识别器训练，得到档案识别器，其中所述档案识别器用于判别档案数据真伪。

[0073] 在本实施例中，所述识别器训练模块110可以执行图2中的步骤S210，具体的描述可参照上文中对步骤S210的详细描述。

[0074] 所述生成器训练模块120，用于获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器，其中所述档案生成器用于生成健康档案数据。

[0075] 在本实施例中，所述生成器训练模块120可以执行图2中的步骤S220，具体的描述可参照上文中对步骤S220的详细描述。

[0076] 所述优化训练模块130，用于根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止。

[0077] 在本实施例中，所述优化训练模块130可以执行图2中的步骤S230，具体的描述可参照上文中对步骤S230的详细描述。

[0078] 所述数据生成模块140，用于以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据。

[0079] 在本实施例中，所述数据生成模块140可以执行图2中的步骤S240，具体的描述可参照上文中对步骤S240的详细描述。

[0080] 请参照图5，是本申请实施例提供的健康档案数据生成装置100的另一种方框示意图。在本申请实施例中，所述健康档案数据生成装置100还包括参数配置模块150。

[0081] 所述参数配置模块150，用于对目标患病种类集合、目标健康等级集合、预设概率数值及预设概率差阈值进行配置。

[0082] 在本实施例中，所述参数配置模块150可以执行图3中所示的步骤S209，具体的描述可参照上文中对步骤S209的详细描述。

[0083] 综上所述，在本申请实施例提供的健康档案数据生成方法及装置中，所述健康档案数据生成方法的人工参与度极小，可模拟合成真实性高、数据准确度高且数据多样性强的健康档案数据，便于医疗研究。所述方法在基于多组真实健康档案数据训练得到档案识别器，并基于目标患病种类集合及目标健康等级集合训练得到档案生成器后，通过交替地

对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止的方式，确保最终的经优化训练后的档案生成器所生成的健康档案数据与真实健康档案数据之间的数据相似性极强，从而模拟合成真实性高且数据准确度高的健康档案数据，并通过所述目标患病种类集合及所述目标健康等级集合各自的组成，确保基于所述方法得到的健康档案数据具有较强的数据多样性，以向医疗研究提供强大的数据支撑。

[0084] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

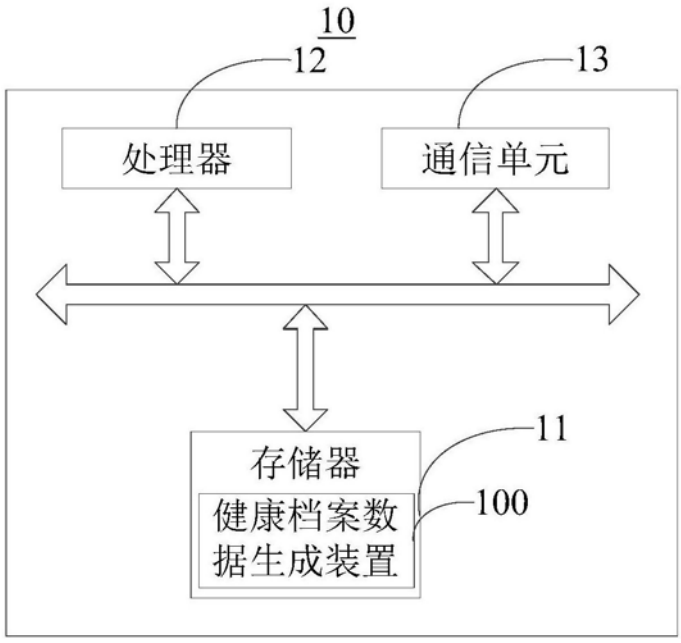

10
12
13
处理器
通信单元
存储器
11
健康档案数据生成装置
100

图1

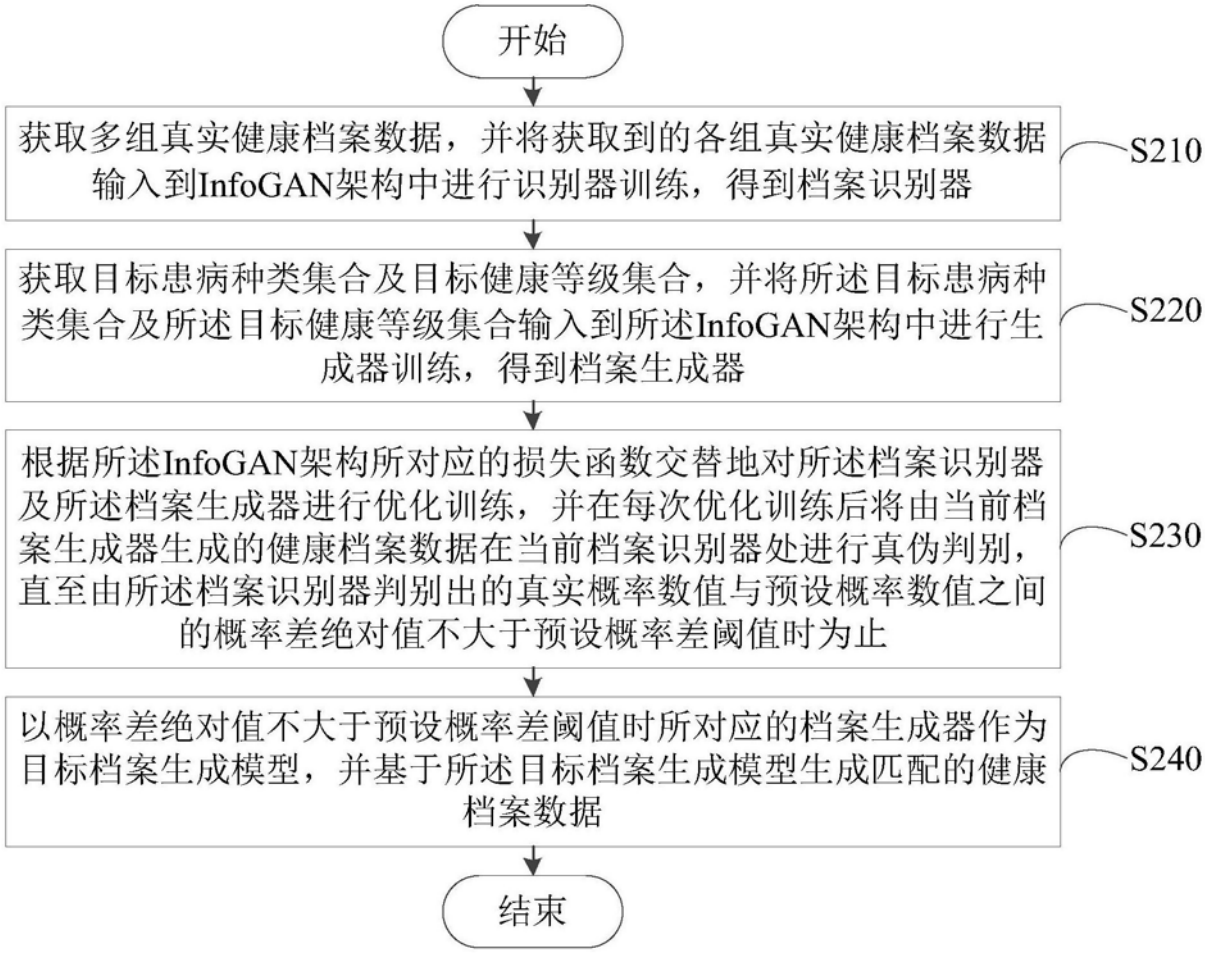

开始
获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到InfoGAN架构中进行识别器训练，得到档案识别器
S210
获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器
S220
根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止
S230
以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据
S240
结束

图2

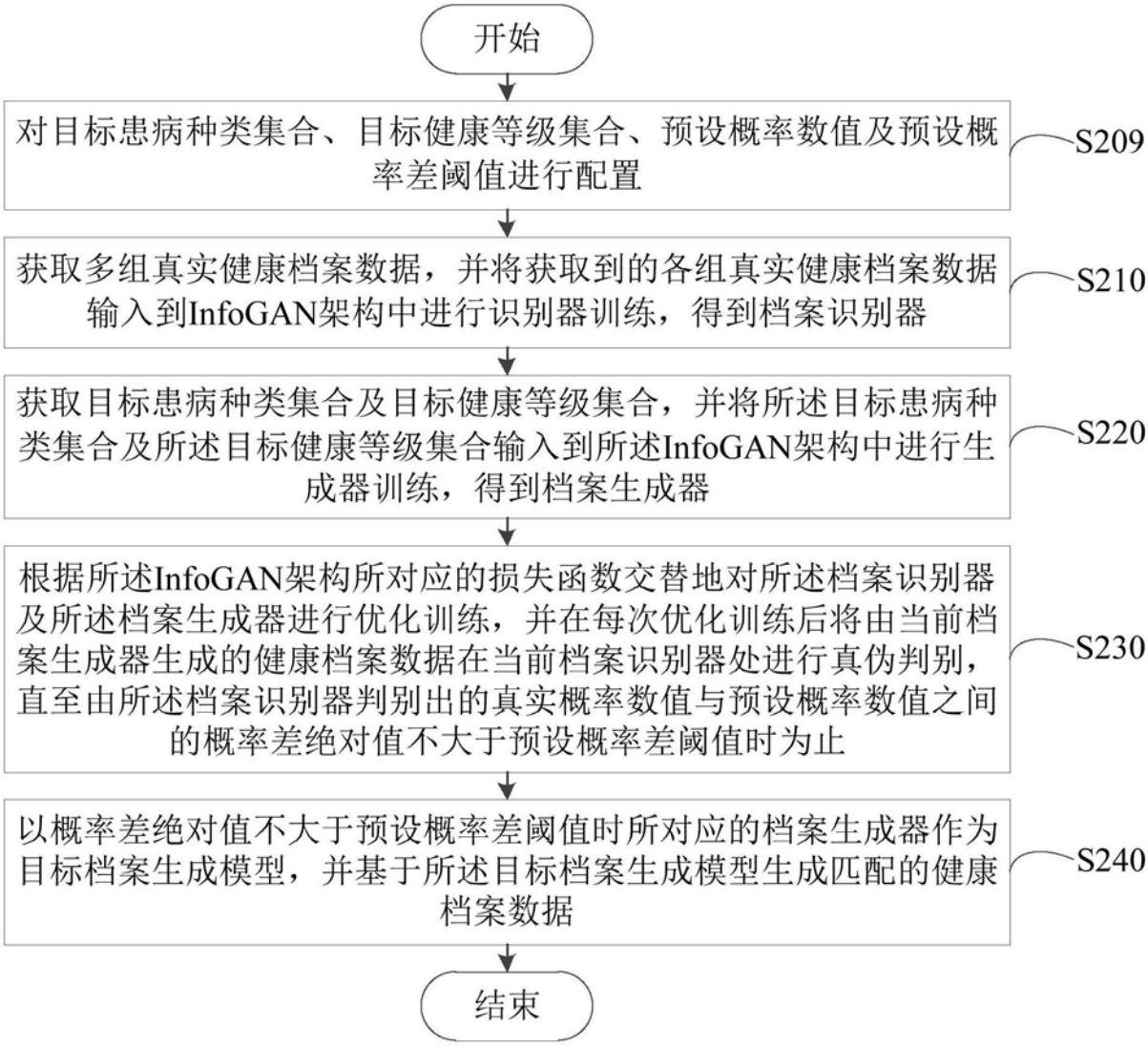

开始
对目标患病种类集合、目标健康等级集合、预设概率数值及预设概率差阈值进行配置
S209
获取多组真实健康档案数据，并将获取到的各组真实健康档案数据输入到InfoGAN架构中进行识别器训练，得到档案识别器
S210
获取目标患病种类集合及目标健康等级集合，并将所述目标患病种类集合及所述目标健康等级集合输入到所述InfoGAN架构中进行生成器训练，得到档案生成器
S220
根据所述InfoGAN架构所对应的损失函数交替地对所述档案识别器及所述档案生成器进行优化训练，并在每次优化训练后将由当前档案生成器生成的健康档案数据在当前档案识别器处进行真伪判别，直至由所述档案识别器判别出的真实概率数值与预设概率数值之间的概率差绝对值不大于预设概率差阈值时为止
S230
以概率差绝对值不大于预设概率差阈值时所对应的档案生成器作为目标档案生成模型，并基于所述目标档案生成模型生成匹配的健康档案数据
S240
结束

图3

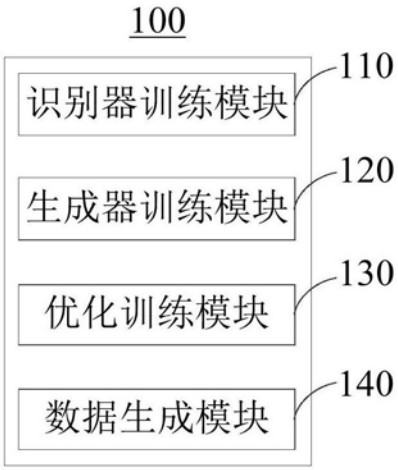

100
识别器训练模块
110
生成器训练模块
120
优化训练模块
130
数据生成模块
140

图4

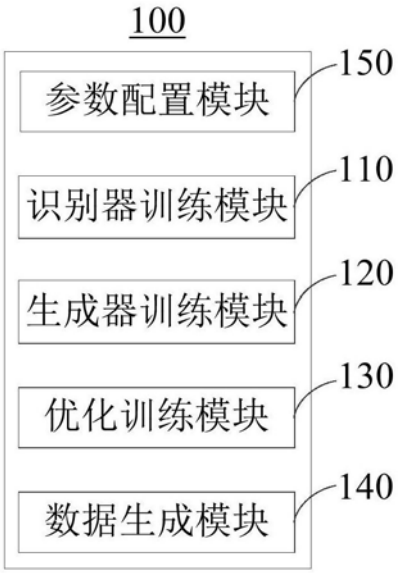

100
参数配置模块
150
识别器训练模块
110
生成器训练模块
120
优化训练模块
130
数据生成模块
140

图5