基于生成对抗网络技术的医疗仿真数据生成方法

向夏雨[1]，王佳慧[2]，王子睿[3]，段少明[3]，潘鹤中[1]，庄荣飞[3]，韩培义[3,4]，刘川意[3,4]

（1. 北京邮电大学网络空间安全学院，北京 100876；2. 国家信息中心信息与网络安全部，北京 100045；
3. 哈尔滨工业大学（深圳）计算机科学与技术学院，广东 深圳 518055；4. 鹏城实验室网络部，广东 深圳 518066）

摘 要：对结构化电子健康档案中行的概率分布进行建模并生成仿真数据非常困难，因为表格数据通常包含定类列，传统编码方式可能产生特征维数灾难的问题，从而使建模异常困难。针对这一问题，提出利用庞加莱球模型建模医疗分类特征的层级结构，并采用高斯耦合的生成对抗网络技术合成结构化的电子健康档案。实验表明，该方法生成的训练数据能够在保证隐私性的前提下，实现与原始数据仅相差 2%的可用性差异。

关键词：生成对抗网络；表示学习；隐私性与可用性分析；电子健康档案

中图分类号：TP309.2 数据安全

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022057

Generate medical synthetic data based on generative adversarial network

XIANG Xiayu[1], WANG Jiahui[2], WANG Zirui[3], DUAN Shaoming[3], PAN Hezhong[1], ZHUANG Rongfei[3], HAN Peiyi[3,4] LIU Chuanyi[3,4]

1. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Department of Information and Security, The State Information Center, Beijing 100045, China
3. School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
4. Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen 518066, China

Abstract: Modeling the probability distribution of rows in structured electronic health records and generating realistic synthetic data is a non-trivial task. Tabular data usually contains discrete columns, and traditional encoding approaches may suffer from the curse of feature dimensionality. Poincaré Ball model was utilized to model the hierarchical structure of nominal variables and Gaussian copula-based generative adversarial network was employed to provide synthetic structured electronic health records. The generated training data are experimentally tested to achieve only 2% difference in utility from the original data yet ensure privacy.

Keywords: generative adversarial network, representation learning, privacy-utility analysis, electronic health record

0 引言

医疗大数据的激增并不意味着数据科学家可以轻易地获取真实数据。例如，一家医院希望将电子健康档案（EHR, electronic health record）分享给一所大学用于研究，然而数据共享必须经过仔细的伦理审查，以免泄露病人的隐私[1]。此过程通常需要几个月的时间，最终还不能保证该学校可以获得批准。

为了解决上述难题，最新工作基于生成对抗网络

收稿日期：2021−12−20；**修回日期**：2022−02−17

通信作者：刘川意，liuchuanyi@hit.edu.cn

基金项目：国家重点研发计划基金资助项目（No.2016YFB0800803, No.2018YFB1004005）；国家自然科学基金资助项目（No.61872110）

Foundation Items: The National Key Research and Development Program of China (No.2016YFB0800803, No.2018YFB1004005), The National Natural Science Foundation of China (No.61872110)

（GAN, generative adversarial network）技术进行了广泛的研究，以期提供一种更安全的数据共享方式[2]。但是与 GAN 用于生成非结构化数据的成功相比，基于 GAN 的结构化仿真数据合成仍处于起步阶段。

结构化数据的混合数据类型、特征的关联关系、多模态的数据分布、不平衡的数据标签这几个独特属性对 GAN 模型的设计提出了挑战[3]。具体而言，结构化数据集往往包含各种定类特征，并且机器学习通常要求对训练的变量进行矢量表示。常规的独热编码首先将定类属性转换为多个数值模拟，然后将其用于建模[4]。

尽管独热编码技术简单易行，但其主要缺点是随着特征类别的增加，或者当数据集中存在数百万的实体时，将不可避免地产生维数灾难效应。通常对于基数为 d 的变量，其对应向量将具有 d 个维度，所形成的稀疏矩阵难以进行有效的机器学习训练[5]。其次，一位有效编码的特点是将类别分开表示为独立不相关的概念，这是因为任意 2 个向量之间的内积为零，且每个向量在欧几里得空间中彼此距离相等，这样带来的后果是消除了特征底层表示中的重要关联结构[5]。对于结构化 EHR 中的定类特征而言，前期特征编码工作[2-3]尚未针对性地解决医疗实体之间所存在的层级结构，从而导致现有的结构化仿真数据生成方案不适用于医疗数据集[6]。

因此，本文旨在研究数据中间表示学习（DIRL, data intermediate representation learning），以克服生成对抗网络用于结构化数据建模的局限性。在机器学习中，表示学习[7]可以从原始数据中自动发现特征或构建分类器所需的有效信息，这取代了过去手动的特征工程，允许程序学习相关特征并使用它们执行特定任务。直观上观察，基于同一空间的低维表示相比于独热编码更有效，这是因为特征嵌入在保留特征向量空间语义的同时，也仅仅由少量的实数点表示。

综上所述，正确表示数据是训练 GAN 的关键，通过对分类和连续变量采取合适的表示形式，并设计合理的 GAN 模型架构，可以训练出高质量的神经网络模型用于仿真数据的生成。

本文具体的贡献如下。

1) 基于表示学习的定类变量建模，利用双曲空间对大规模医学类别实体进行低维、稠密向量的映射，将庞加莱球模型与黎曼随机梯度下降优化算法用于建模特征层次关系，以此有效地保留潜在分层结构与关联关系，为 GAN 的原始数据训练提供基础。

2) 基于高斯耦合的生成对抗网络技术，利用生成模型创建近似于原始数据分布的仿真数据，首先使用高斯耦合对数据表中多元非线性的随机变量进行建模，以此捕捉不同特征之间的统计特性；随后利用优化的 WGAN（Wasserstein generative adversarial network）为分类、连续等结构化数据类型提供合成数据，通过该技术在真实 EHR 的使用受到限制时代替真实 EHR。

3) 基于隐私性和可用性指标的评估技术，利用距离的方式检验假数据的隐私性，使用最近邻对抗精度、隐私损失、散度值与差异分数定量描述与真实数据之间的差异；利用分类算法综合检验仿真数据的可用性，提出统计平均的机器学习分类指标，公正评判仿真数据相较于原始 EHR 的再入院预测效果。

最终实验表明，相较于当前结构化假数据生成的 SOTA（state-of-the-art）技术——CTGAN（conditional tabular generative adversarial network）[8]而言，本文提出的方案可以更好地表征结构化 EHR 中的分类和连续特征，最终在生成数据的可用性方面实现了超过 15%的提升，从而为隐私保护前提下医疗 EHR 的发布和挖掘提供更可靠的依据。

1 相关工作

1.1 数据脱敏

数据脱敏[9]是指对敏感信息按照预设的规则和算法进行数据变形或隐去敏感信息，从而使个人身份无法识别。

1) 传统数据脱敏技术

传统的数据脱敏技术可分为基于非数据扰乱的数据脱敏技术和基于数据扰乱的数据脱敏技术[9]。前者（例如数据抽样、去标识化）不会降低数据的真实性，可基于原始数据减少敏感细节或者对其进行部分抑制，但是会降低预测的准确性[10]；经过后者（例如数据置换、数据噪声、数据遮掩）扰乱后的数据通常是不真实的，即受到了一定程度的修改。相比于前者，基于数据扰乱的脱敏技术往往可以更好地保留原始数据的统计分布特性[11]。

本文涉及的传统数据脱敏技术为去标识化技术[12]，其定义为对相应的标识符进行直接删除的操

作。在不借助任何背景知识的情况下，该过程无法识别特定主体。

本文所用到的开源 CERNER Health Facts 数据库的糖尿病患者 EHR[13]中，所有数据在提供给数据分析师之前均已根据美国健康保险流通与责任法案（HIPAA, health insurance portability and accountability act）进行了身份去标识化处理。

2) 新型数据脱敏技术——基于生成对抗网络的仿真数据生成

传统的数据脱敏技术通常需要手工制定脱敏规则与策略，对不同场景、不同任务和海量的数据而言，该方案存在巨大的效率缺陷。

生成对抗网络[2]是一种学习数据潜在分布的无监督生成模型，通过 GAN 可以创建仿真的训练数据。在这种情况下，医院不需要发布原始 EHR，仅提供合成数据供数据科学家使用，从而可以避免敏感信息的泄露。但是使用该方法的前提是 GAN 所生成的数据应尽可能地贴近原始数据的分布，以使机器学习算法在此训练集上建模时，其隐私性和可用性均得到保障。

1.2　表示学习和生成对抗网络

数据中间表示学习是在保护数据隐私的前提下，通过有效预测任务学习特征的中间表示（嵌入）。Osia 等[14]提出了一种特征维数缩减（FDR, feature dimension reduction）技术，该技术对提取的特征进行精炼以去除多余的信息，并采用暹罗微调方法保护敏感信息免受侵害，但是并没有系统地在隐私性和可用性之间进行折中。随着 GAN 的提出[2]，学者们已经研究了几种使用 GAN 来保护数据隐私的方法，其目的是模拟攻击者和防御者之间的博弈，它们以相互冲突的可用性与隐私性为目标进行攻守。Xiao 等[15]和 Liu 等[16]设计了一种基于 GAN 的中间表示学习，该方法在保留隐私性的同时最大限度地保证了任务的实用性。这种对抗类型的机制通过模拟解码器或分类器的隐私攻击，旨在推断敏感信息；而编码器则不断试图隐藏私有信息，旨在保护隐私信息不被泄露。这种机制通过持续地学习来提高效用，最终使任务损失函数最小。然而这些工作并未针对结构化数据集进行实验和验证。Li 等[17]介绍了一个与任务无关的隐私保护数据众包框架，目的是学习一个特征提取器，使其可以从提取的中间特征里删除相应的隐私信息，将嵌入原始数据的初始信息用于机器学习下游任务。

上述解决方案背后的相同思想是利用 GAN 来混淆原始数据和特征，防止隐私泄露。但是，这些最新的嵌入机制无法准确地处理定类属性的潜在分层结构，而这正是层级相互关联医疗实体的关键特性所在[18]。

2　基于表示学习的生成对抗网络方法

基于表示学习的生成对抗网络技术是一种基于 GAN 的方法，旨在对结构化数据分布进行建模。本节基于庞加莱球模型对医疗关系实体实施数据嵌入处理，并利用黎曼随机梯度下降算法对其进行优化（2.1 节），以此保留特征潜在分层结构与关联关系，为 GAN 的原始数据训练提供基础。在 EHR 的分类特征得到有效的预处理后，本文利用高斯耦合的方式对变量的多元分布进行建模，并基于全连接网络和最新的 WGAN 技术生成近似于真实数据的仿真数据（2.2 节）。

2.1　分类特征的双曲空间数据嵌入

结构化数据中类别变量的处理通常使用独热编码，而这种方法容易带来空间爆炸的问题。受到数据中间表示学习最新进展的启发，本节将分类特征嵌入低维坐标轴中以提高空间效率，并保留其潜在的属性层次结构。

ICD-9 是《疾病和相关健康问题国际统计分类》的第 9 版，由卫生组织统一规范[19]。ICD-9 临床修改代码（ICD-9-CM）是 ICD-9 版本的更新。如表 1 所示，ICD-9-CM 代码将不同类别的疾病类型划分成不同的值域，如 390～459,785 对应循环系统疾病；250.xx 对应糖尿病等。

表 1　　ICD-9-CM 疾病类别展示

疾病类型	ICD-9 编码
循环系统疾病	390～459, 785
呼吸系统疾病	460～519, 786
消化系统疾病	520～579, 787
糖尿病	250.xx
受伤及中毒	800～999
肌肉骨骼疾病	710～739
泌尿生殖系统疾病	580～629, 788
赘生物	140～239

这种编号的医学本体通常是按照层次组织的。图 1 显示了一个示例：其中 ICD 240～279（不包含 250）是整个 ICD-9-CM 1 000 多种疾病中的一个大类别，表示“内分泌、营养和代谢性疾病以及免疫性疾病”类型。该类别中的子类别为 240～246、249～259、260～269 等，代表了不同类型的疾病种类，例如“甲状腺疾病”“其他内分泌腺疾病”和“营养缺乏症”。249.x 和 255.x 被认为是特定疾病相对应的叶节点，隶属于单个子类别（249～259），代表“患有其他昏迷的继发性糖尿病”和“肾上腺疾病”的准确疾病描述。

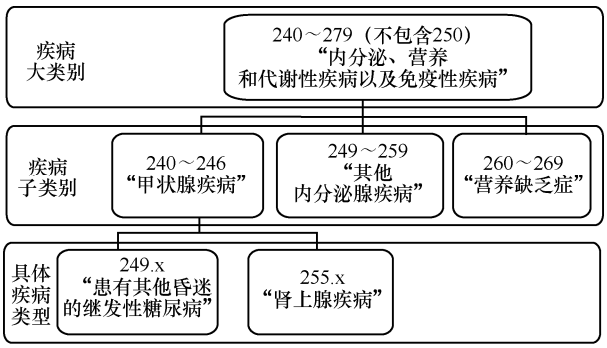


图 1　ICD-9-CM 类别层次示例

ICD-9-CM 的医学本体库较大，每个诊断代码都可以视为一个独立的特征。因此，将独热编码技术用于进一步的处理和建模是不可行的，因为它会产生巨大的稀疏矩阵。

一般而言，合适的数据中间表示可带来良好的模型性能[20]。根据 ICD-9-CM 代码的性质，在表示这些概念时需尽可能保留其分级的结构。对于文本数据而言，欧几里得空间是使用最广泛的一种数据嵌入方法。但是，对于具有明显层次的医学本体来说，双曲方法[21]更加适合，这是因为该方法可以在较低维度上保留正确的层级排名。

庞加莱球模型是类似于 n 维球体的 n 维双曲几何模型，所有点都嵌入在内。任何度量空间的特征都是基于 u 与 v 两点之间的距离。在双曲空间中，特别是对于庞加莱球模型而言，其两点之间的距离定义为

$$d(u,v)=\operatorname{arccos} h\left(1+2\frac{\|u-v\|^2}{(1-\|u\|^2)(1-\|v\|^2)}\right) \tag{1}$$

鉴于庞加莱球模型的几何性质，其可以用来对实体进行层次性的建模。对于嵌入双曲空间中的分层结构，根节点将放置在离中心较近的区域，而叶节点将被分配在靠近球体的边界，这是为了确保叶节点与其他叶节点之间保持合理的距离。

为了学习医学本体的表示，本节定义一个损失函数，旨在最小化相似本体嵌入之间的双曲线距离，并最大化不相似本体嵌入之间的双曲线距离。本节遵循文献[22]的工作，使用黎曼随机梯度下降来优化以下损失函数

$$L=\sum_{(u,v)\in S}\log\frac{\mathrm{e}^{-d_H(u,v)}}{\sum\limits_{v'\in N(u)}\mathrm{e}^{-d_H(u,v')}} \tag{2}$$

式(2)表明，任何有限树都可以嵌入有限的双曲空间中，从而近似保留实体之间的距离。本文方法利用双曲空间的特定模型，即庞加莱球模型，因为它非常适合基于梯度的优化。这使本文能够开发一种基于黎曼优化的高效算法来计算嵌入，该算法易于并行化并且可以扩展到海量 EHR。

图 2 显示了基于二维空间 Poincaré 模型的数据嵌入功能。该方法使用了数据中间表示，能够学习大规模的分类实体，并保持相似医学本体之间的数据关联嵌入。由图 2 可知，ICD-9-CM 中的不同疾病大类别已经被分开。与此同时，在对应的每个大类中，存在多个子类别；由于子类别隶属于单个大类别，故该大类别中的每个子类别仅与所在大类别之间的距离非常相近，与不同大类别之间的距离相对较远，这种情况对于其叶节点也是如此。

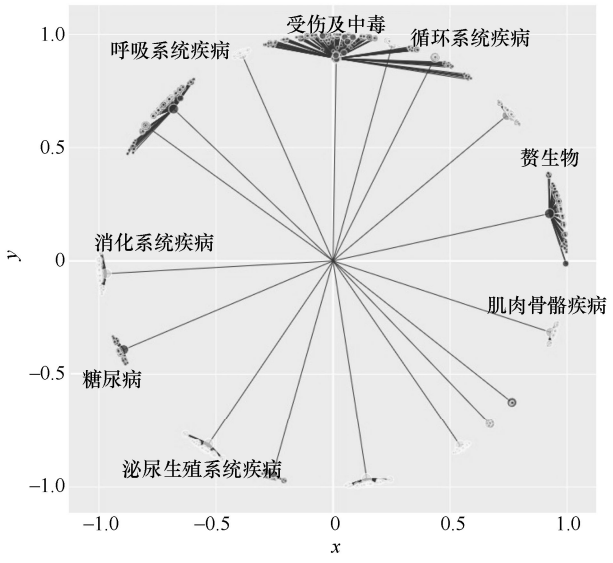


图 2　基于二维空间 Poincaré 模型的数据嵌入功能

通过本节提出的数据中间表示学习的方法，将相似的医学实体聚集在一起，并使不同类型疾病之间的

距离尽可能分开，从而保留本体的有效层级信息。

在训练神经网络模型之前，正确表示原始数据是关键环节。本节通过对定类医疗实体特征采取二维空间 Poincaré 模型的表示形式，为训练高质量的 GAN 模型提供必要的前提条件。

2.2　结构化 EHR 的仿真数据生成

本文使用了 UCI Machine Learning Repository 的 EHR 国家数据仓库，该数据库收集了美国 130 所医院的 10 年临床护理和综合交付网络的全面临床记录[13]，包括 50 类特征，例如人口统计信息、诊断结果、糖尿病药物使用清单、入院前一年的就诊次数以及代表患者和医院结果的保险信息。本文从 EHR 原始数据库中提取满足以下条件的住院信息。

1) 一次住院记录。

2) 一类糖尿病的住院，即在此期间医生将任何类型的糖尿病输入系统中作为诊断。

3) 住院时间最少一天，最多 14 天。

4) 住院期间进行了实验测试和化验检查。

5) 住院期间服用了药物。

该糖尿病数据集包含 101 766 例住院患者的病历数据、医生的专业知识、人口统计学特征（年龄、性别和种族）、诊断和住院程序（由 ICD-9-CM 进行编码）、实验室数据、药房数据、院内死亡率和医院特征等。所有数据在提供给数据分析师之前均已根据 HIPAA 进行了身份去标识化处理。

为了获得干净、唯一和经过转换的数据集进行分析，本节利用了 2 个主要的预处理步骤，如图 3 所示，其中包括数据清理和特征转换。最初的糖尿病原始数据集包含 101 766 例住院记录和 50 个数据特征。数据清理在患者记录（行）和数据变量（列）中进行，最终产生 69 990 个不同的记录和 40 个特征。

上述数据集可用于患者再入院的数据特征分析和预测，其中数据集中包含 39 个潜在的预测因素和一项结果变量，即 30 天内是否再入院。本节将再入院状态定义为具有 2 种结果：“再入院”（患者在出院后 30 天内再次住院）或“无再入院”（患者在 30 天后再次住院和没有再次住院）。

本节定义原始数据为 $\Gamma=[X;Y]$，仿真数据为 $\Gamma'=[X';Y']$，其中每个 $x_i \in X$ 和 $y_i \in Y$ 分别对应数据集中的特征和标签。通过 Γ 训练一个分类器 $f: X \to Y$（通过 Γ' 训练一个分类器 $f': X' \to Y'$），使 $x_i \in X$ 被映射至对应的预测标签 $f(x_i)$，即 0（无再入院）或 1（再入院）。

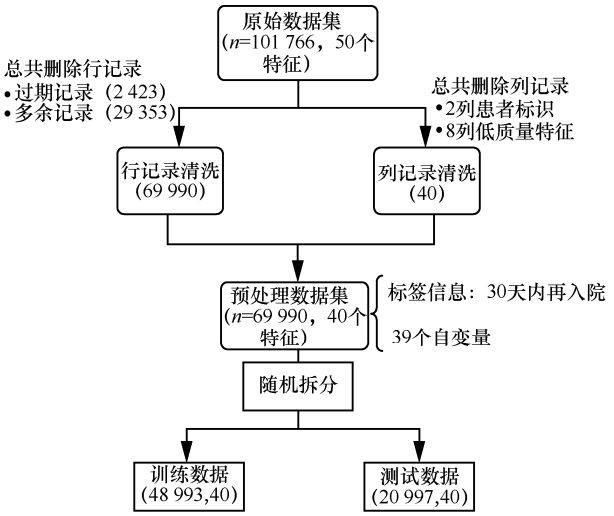


图 3　EHR 预处理流程

出院后再入院的定义是指在规定的时间段内出院患者再次住院的情况。再入院率已越来越多地被用作卫生服务研究中的一项成果指标和卫生系统的质量基准。出于伦理审查和保护原始数据隐私的考虑，本节将生成对抗网络技术用于该结构化 EHR 的仿真数据生成，最终仅提供合成的数据给数据分析师挖掘使用。

生成模型本质上是一种机器学习模型，试图研究现实世界中的数据分布，然后从学习到的分布中随机抽取样本。它符合创建仿真数据的目标——试图拟合原始数据，以便从中获取样本数据进行建模。这种机制的一个主要特征是满足了保护隐私的要求[8]。

传统的 GAN 由生成器和判别器组成，生成器的任务是创建任意数据分布的实际样本，而判别器的目标是正确区分生成器参数化的分布与真实训练数据是否相同。生成器和判别器同步进行极小化极大化博弈，因为当达到理想状态时两者处于纳什平衡，即在生成器准确拟合数据分布的情况下，判别器无法判别样本数据是否真实。

普通 GAN 的主要缺点在于没有提供控制生成数据的自主权，也没有支持生成分类数据的可能性[23]。其中，一种对 GAN 的优化是 WGAN[23]。WGAN 改进了模型训练时的稳定性，并提供了与生成数据质量相关的损失函数。经过分析发现，使用 Wasserstein 方法在生成器网络中设置对应的 softmax 输出（维数等于每个分类变量的定类值数

量），能够使 GAN 创建定类数据。

因此，本节提出一种新的仿真数据生成方法，为包含分类、连续和序数等特征类型的结构化数据集提供合成伪造数据的技术。该方案利用 WGAN，同时采用基于“合成数据仓库”[24]编码方法的新变体为变量的多元分布实施建模，具体步骤如下。

首先，通过高斯耦合对多元非线性的随机变量进行相关性的建模，以此学习原始结构化数据，从而捕捉不同特征之间的统计分布属性。当不同随机变量的边缘分布相互之间并不独立时，Copula 相较于传统方法的优势是使联合分布建模变得容易。因为简单的相关系数只能衡量线性的相关关系，不能衡量非线性的关联，所以这个时候只能利用 Copula 把不同的分布连接起来。2.1 节已经对分类特征做了数值化的预处理，并将其映射到二维数据嵌入的表示空间，这样使高斯耦合能够直接对分类数据实现操作。相较于传统“合成数据仓库”编码方法，本文提出的方法将分类数据替换至[0,1]值域[24]，提供了更好的可靠性，同时也对表格中列的分布找到一个准确的估计，为下一步生成仿真数据提供基础。

随后，在 WGAN 生成仿真数据[23]的基础上，本节寻求训练生成器模型的另一种方法，从而更好地估计给定训练数据集的数据分布，整体流程如图 4 所示。WGAN 没有使用判别器将生成的数据条目划分为真实或伪造，而是采用评价网络的方式对记录的真实性或伪造性进行评判，如算法 1 所示。这种变化是受理论论证的启发而实施的，即训练生成器应寻求使训练数据集中观察到的数据分布与所生成示例中观察到分布之间的 Wasserstein 距离最小。

算法 1 $\mathrm{WT}_{\mathrm{RAIN}}(m,\alpha_d,\alpha_g,T_d,T_g,c_p)$

输入 批量值 m，鉴别器学习率 α_d，生成器学习率 α_g，鉴别器迭代次数 T_d，生成器迭代次数 T_g，裁剪参数 c_p

输出 生成器 G，鉴别器 D

初始化 判别器的参数 $\theta_d^{(0)}$ 和生成器的参数 $\theta_g^{(0)}$

1) for $t_1=1,2,\cdots,T_g$

2) for $t_2=1,2,\cdots,T_d$

3) 获取噪声数据 $\left\{\boldsymbol{z}^{(i)}\right\}_{i=1}^{m}\sim p_z(\boldsymbol{z})$

4) 获取真实数据 $\left\{\boldsymbol{t}^{(i)}\right\}_{i=1}^{m}\sim p_{\mathrm{data}}(\boldsymbol{t})$

5) $\overline{g}_1\leftarrow\nabla_{\theta_d}\dfrac{1}{m}\sum_{i=1}^{m}[D(\boldsymbol{t}^{(i)})-D(G(\boldsymbol{z}^{(i)}))]$

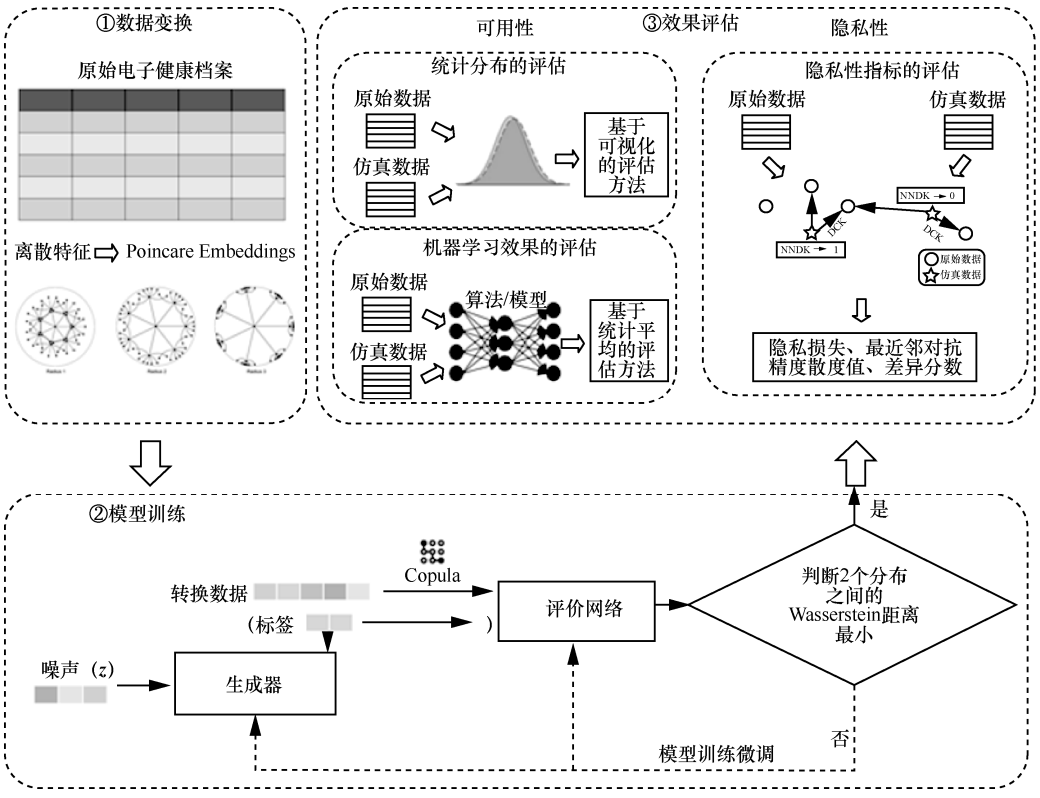


图 4 基于生成对抗网络的仿真数据生成

6) $\theta_d^{(t_2)} \leftarrow \theta_d^{(t_2-1)} + \alpha_d \mathbf{RMSProp}(\theta_d^{(t_2-1)}, \overline{g}_1)$

7) $\theta_d^{(t_2)} \leftarrow \mathbf{clip}(\theta_d^{(t_2)}, -c_p, c_p)$

8) until t_2 遍历 $1,2,\cdots,T_d$

9) end for

10) 获取噪声数据 $\{\boldsymbol{z}^{(i)}\}_{i=1}^{m} \sim p_z(\boldsymbol{z})$

11) $\overline{g}_2 \leftarrow -\nabla_{\theta_g} \frac{1}{m}\sum_{i=1}^{m} D(G(\boldsymbol{z}^{(i)}))$

12) $\theta_g^{(t_1)} \leftarrow \theta_g^{(t_1-1)} + \alpha_g \mathbf{RMSProp}(\theta_g^{(t_1-1)}, \overline{g}_2)$

13) until t_1 遍历 $1,2,\cdots,T_g$

14) end for

15) return G, D

这就意味着，通过选择合适的损失函数能够生成高质量的假数据，以此保证仿真数据的分布与真实数据的分布尽可能一致。损失函数定义为

$$L_D = -\mathrm{E}_{\boldsymbol{t}\sim p_{\text{data}}(\boldsymbol{t})}[D(\boldsymbol{t})] + \mathrm{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[D(G(\boldsymbol{z})] \tag{3}$$

$$L_D = -\mathrm{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[D(G(\boldsymbol{z})] \tag{4}$$

综上，本节针对传统处理分类特征的缺陷提出了基于双曲空间数据嵌入的方法，通过将属性投射至低维空间以稠密的向量表示，从而保留其层级结构。此外，本文提出了基于高斯耦合的改进 WGAN 用于结构化的仿真数据生成，对 EHR 中的连续、分类等属性进行拟合，提供与真实训练数据相似并能够保护隐私的仿真数据，供数据分析师分析使用。

3　实验结果分析

本节首先介绍了实验环境以及提出的评估分析指标——隐私性与可用性的衡量标准。然后检验了方法的有效性，并与当前最新工作进行了比较。最后通过消融实验验证了组件的作用，证明了本文方案的优越性。

3.1　实验环境

基于 GAN 的仿真数据生成主要使用 Tensorflow、Numpy 和 Pandas。所有机器学习建模和分析均使用 Python 3.6 中的 Sklearn 0.21 版本软件包，可用性预测使用 Lazy Predict 库。

3.2　评估指标

为了验证经过数据嵌入处理的 EHR 所生成的仿真数据效果，基于前期相关工作[25]，本节定义了隐私性和可用性两方面的评估指标。具体的分析和结果将在后续的实验部分呈现。

3.2.1　隐私性指标

考虑 2 个数据分布 PT 和 PS，其中 T 对应真实数据分布，S 对应合成数据分布。从 2 个数据集中随机抽取的样本数据分别为 $S_T = (X_T^1, Y_T^1), \cdots, (X_T^n, Y_T^n)$ 和 $S_S = (X_S^1, Y_S^1), \cdots, (X_S^n, Y_S^n)$。

直观上理解，对于 2 个数据集中的任意两点，如果基于距离远近的评估方式，假设两者之间的距离足够远，则意味着真实数据的训练集/测试集与生成数据的训练集/测试集不相同，这就表明隐私性得到了保障。

本节将辨别数据是否真实的能力通过最近邻的概念进行定义，真实数据中的一个点与仿真数据中最相近的一个点的距离为

$$d_{TS}(i) = \min_j \left\| x_T^i - x_S^j \right\| \tag{5}$$

从真实数据中同一分布中抽取的 n–1 个样本与原始分布的最近邻距离为

$$d_{TT}(i) = \min_{j, j\neq i} \left\| x_T^i - x_T^j \right\| \tag{6}$$

基于此，本节提出最近邻对抗精度的定义，如式(7)所示。

$$\mathcal{AA}_{TS} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(d_{TS}(i) > d_{TT}(i)\right) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(d_{ST}(i) > d_{SS}(i)\right) \right) \tag{7}$$

其中，$\mathbf{1}(\cdot)$为指示函数，如果判断为真，则返回结果为 1，反之为 0。对于真实数据中的任意一个点，如果它与合成数据中任意点的距离都足够远，则可以认为该点是“真阳性”，从而保证隐私未遭到泄露；同理，合成数据中任意一点都应与真实数据中的点相距足够远以便被判定为“真阴性”。所以，本节提出的最近邻对抗精度是作为区分真实数据和仿真数据对抗分类器的性能指标，如果不能辨别真实数据集与仿真数据集之间的差异，则该结果为 0.5。

更多地，隐私损失的概念源自最近邻对抗精度，旨在对真实数据的训练集/测试集与生成数据的训练集/测试集之间最近邻对抗精度的差异开展比较，如式(8)所示。

$$\text{Privacyloss} = TS_{\text{Test}} - TS_{\text{Train}} \tag{8}$$

假设真实数据的训练集/测试集与生成数据的训练集/测试集之间的最近邻对抗精度分别为 0.5，则最终隐私损失结果为 0。

额外地，本节利用散度值和差异分数进一步比

较真实数据集和合成数据集之间的差异。散度值旨在衡量真实数据分布PT和合成数据分布PS之间的距离[26]，其值域大于或等于0；生成对抗网络的目的是希望合成数据的分布尽可能地逼近甚至等于原始数据的真实概率分布，从而使散度值接近最小值。差异分数[27]定量地描述了深度学习模型与数据的吻合程度。其中，较大的数值表示生成模型对数据的拟合度较差，0表示模型的拟合度最好。在大多数情况下，给定模型的参数估计值旨在确保模型的差异函数得分最低。

值得注意的是，假设真实数据和合成数据之间存在明显差异，那么训练集和测试集的最近邻对抗精度都同时大于0.5，两者的隐私损失差值却为0。此时，需要通过可用性评估的方式分析仿真数据的好坏。

3.2.2 可用性指标

可用性评估基于糖尿病患者的EHR数据集进行再入院分类预测任务，其目的是使用各种机器学习算法评估预测性能，以便交叉验证合成数据的有效性。在再入院预测任务中，准确率是常见的基础评估方式。本节同时使用F1值进行再入院分类准确性的比较，F1值定义为精确度和召回率的调和均值，且认为精确度和召回率同等重要，相当于精确度和召回率的综合评价指标。一般而言，F1值越高，意味着模型越好，表明具有预测重新入院任务的能力，从而证明本文提出的仿真数据生成方法能够出于隐私目的生成“真实、可用”的数据。

进一步地，本文利用绝大多数分类算法综合检验合成数据的可用性，提出可用性统计平均准确率和F1值评判仿真数据相较于原始EHR的再入院预测效果。

可用性统计平均的定义是在某一给定分类任务上的平均得分，即

$$\mu=\frac{\sum X_i}{N} \tag{9}$$

其中，μ表示统计平均值，$\sum X_i$表示所有分类算法预测值的总和（例如准确率、F1值），N表示分类算法的总数。基于可用性统计平均可以实现更加公正的机器学习下游任务评测，而不依赖于单个算法的预测结果。

3.3 隐私性评估分析

本节开始将重点评估原始糖尿病EHR相比于合成糖尿病EHR的真实性，重点是隐私性−可用性的平衡。为了量化隐私保护的措施，本节验证多个指标评估其性能优劣。

本节首先对生成数据集与原始数据集之间的最近邻对抗精度和隐私损失进行了实验比较。与此同时，为了多维度地评估合成数据的质量，本节对比了不同数据量（100万、75万、50万、25万、10万、5万和2万）的仿真数据之间的表现。更多地，本节横向比较了当前结构化假数据生成的SOTA技术——CTGAN，它基于GAN来构建数据表。CTGAN的研究显示，它在85%案例中的表现优于经典的合成数据技术[8]。为公平地进行比较，本节基于CTGAN生成了不同数据量（100万、75万、50万、25万、10万、5万和2万）的仿真数据，展示最好的结果并与本文方案展开比较，如表2所示。

表2 生成数据集的最近邻对抗精度与隐私损失

数据集	训练集最近邻对抗精度	测试集最近邻对抗精度	隐私损失
100万	0.782	0.703	−0.079
75万	0.800	0.729	−0.071
50万	0.786	0.712	−0.073
25万	0.779	0.702	−0.077
10万	0.791	0.719	−0.072
5万	0.842	0.769	−0.073
2万	0.805	0.732	−0.073
No Embedding	0.909	0.820	0.089
CTGAN	0.912	0.831	0.081

从表2中可以观察到，训练集和测试集相应的最近邻对抗精度均为0.7～0.8。与此同时，若分别评估训练集与测试集，则发现训练集中的最近邻对抗精度更偏离理想精度0.5，而测试集中的数值相对更小，意味着拥有更好的表现。导致这种情况的一个很重要的原因可能是训练集中数据量较大（48 993例），而测试集中数据量较小（20 997例），数据量增大无疑给生成对抗网络合成假数据增加了难度。

对于表2中的隐私损失指标而言，100万数据量的假数据达到了最优的指标−0.079，为此可以初步推断合成数据的隐私性能并非与生成数据的数据量大小直接相关。

在此基础上，本节对经过数据嵌入的合成数据集和未经数据嵌入的合成数据进行了比较，鉴于100万数据量的（数据嵌入）假数据得到了最优结果，故原

始独热编码的合成数据（No Embedding）也生成 100 万的假数据。从表 2 中发现，未经数据嵌入的 EHR 训练集/测试集的最近邻对抗精度相对较差，这表明该仿真数据所形成的隐私保护能力相对局限。

对于 CTGAN 而言，其表现并没有优于本文提出的方法。CTGAN 训练集与测试集的最近邻对抗精度为 0.912 和 0.831，隐私损失为 0.081。一个可能的原因是对于分类数据而言，CTGAN 利用高维的独热编码和归一化的形式表示原始的一维数据，这样直接导致 GAN 更难以学习到各维度之间的关系。而本文利用表示学习保留了医疗本体的层级结构，从而可以很好地保留关联信息。

如上文所描述的特例，即便对应的隐私损失约等于 0，但是仍无法说明该数据得到保障。为证实该结论，下面对生成数据与原始数据之间的拟合效果进行可视化展示。

图 5 对原始数据与仿真数据的部分特征实施了可视化对比分析。结果表明，大部分特征都实现了较好的拟合效果。

图 6 比较了经过数据嵌入处理的 100 万生成数据（图 6(a)）与未经数据嵌入处理的 100 万生成数据（图 6(b)），并展示了它们相较于原始数据分布的区别。鉴于糖尿病患者 EHR 中存在大量具有相关性的变量特征，假如逐个对其分析，则往往是孤立不全面的。故本节采用主成分分析的方法对属性降维，以便更加直观和全面地观察原始数据与合成数据之间的差异。

从图 6 可以看到，基于数据嵌入处理的仿真数据对原始数据有一个合理的拟合；利用独热编码生成的仿真数据对原始数据的特征分布无任何拟合，对应表 2 中第 2、3 列。

上述结果表明，原始数据集得到数据嵌入后的隐私性能要优于传统独热编码的合成数据生成方案，无论是基于最近邻对抗精度，还是就整体合成数据集的隐私损失而言，从而证明了本文数据脱敏方案的优势。其重要原因在于传统采用独热编码的分类特征创建了大量冗余 0/1 属性，给生成对抗网络的拟合带来障碍，故不利于仿真数据的合成。

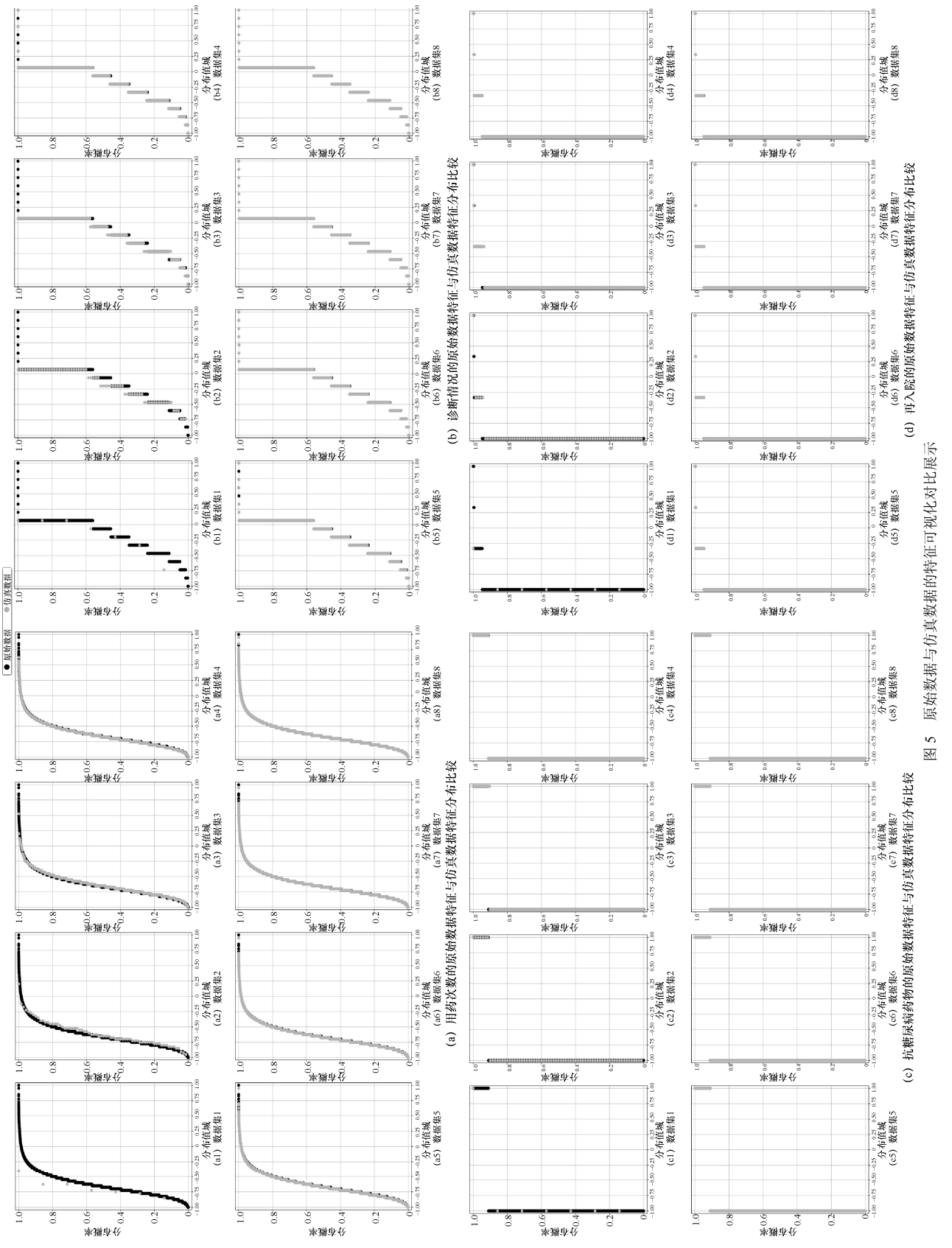
另外，值得注意的是，所有生成数据的隐私损失均接近于 0，并且 100 万最优隐私损失的差异微小到可忽略不计。这在一定程度上证明了基于改进的生成对抗网络技术对原始数据集实现了可观的隐私保护。本节引入其他相关指标进一步为仿真数据集的质量提供参考。

本节采用散度值和差异分数辅助验证生成数据的质量，如表 3 所示。实验依旧对比了不同数据量的仿真数据和未经数据嵌入处理的 100 万仿真数据的表现，并基于真实数据的训练集/测试集与生成数据的训练集/测试集的散度值、差异分数指标评估它们的性能。同样地，本节基于 CTGAN 所生成最好的仿真数据散度值和差异分数指标进行了横向对比。

就散度值而言，本节将原始训练集/测试集和生成数据的训练集/测试集进行了比较，对应值域分布为 0.17～0.24。其中，100 万、10 万和 5 万假数据集均有良好表现，这说明所生成的假数据与真实数据集之间有较好的相似性。更重要的是，若单独评估训练集与测试集，则发现绝大多数训练集中的散度值较偏离理想值 0，而测试集中的数值较小，表明其拥有相对理想的表现。该结论与表 2 中的发现吻合，即训练集中数据量较大（48 993 例），而测试集中数据量较小（20 997 例），数据量增大时无疑给生成对抗网络合成假数据增加了难度。

表 3　　**生成数据集的散度值与差异分数**

数据集	真实训练集与仿真训练集散度值	真实测试集与仿真测试集散度值	真实训练集与真实测试集差异分数	真实训练集与仿真训练集差异分数	真实测试集与仿真测试集差异分数	仿真数据集差异分数
100 万	0.186	0.189	—	2.181	2.186	1.356
75 万	0.200	0.205	—	2.186	2.212	1.312
50 万	0.203	0.205	—	2.171	2.189	1.288
25 万	0.209	0.208	2.412	2.176	2.190	1.347
10 万	0.183	0.177	—	2.188	2.194	1.344
5 万	0.183	0.178	—	2.218	2.224	1.385
2 万	0.240	0.234	—	2.190	2.194	1.234
No Embedding	5.429	5.649	3.142	3.590	5.540	3.128
CTGAN	4.839	5.069	3.142	3.300	5.350	3.584



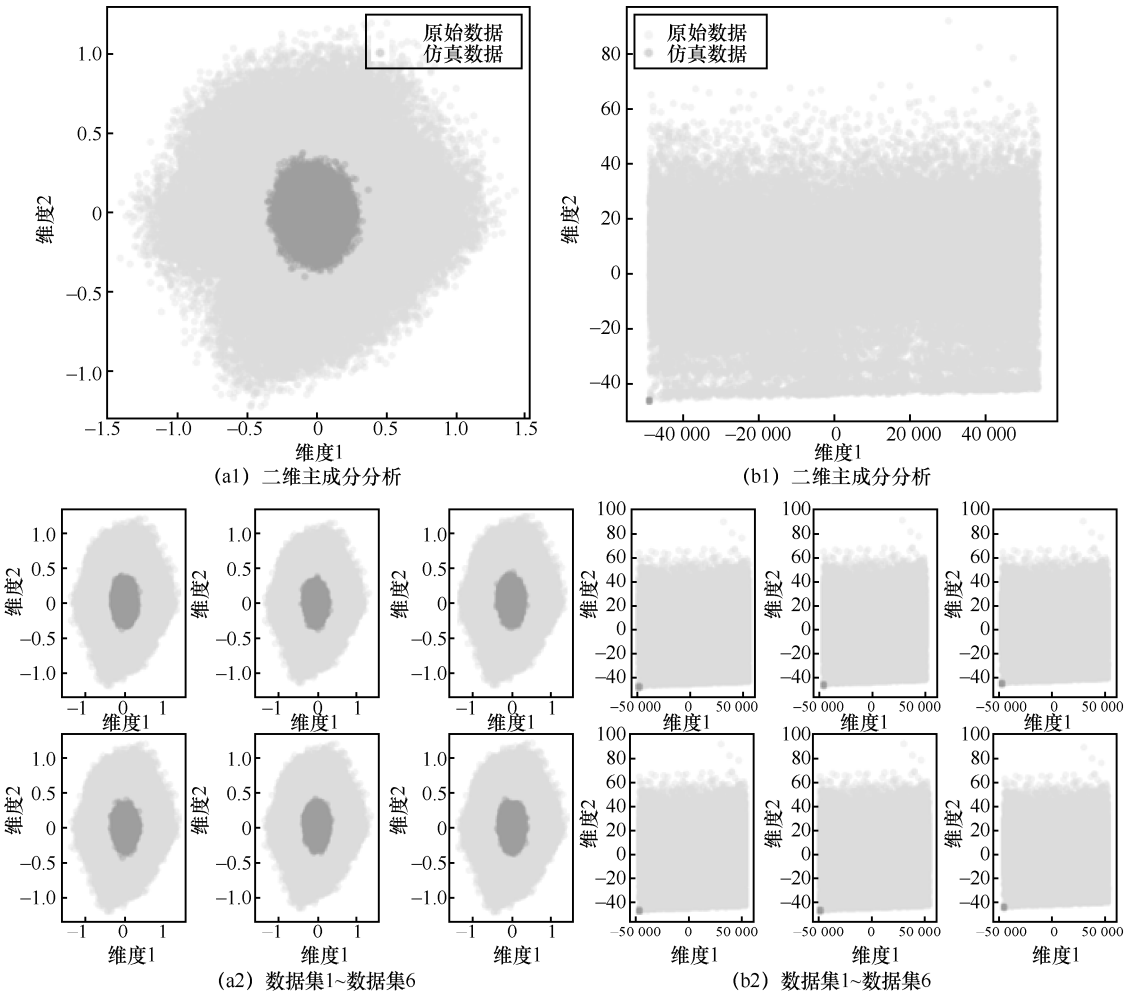
(a) 用药次数的原始数据特征与仿真数据特征分布比较

(b) 诊断情况的原始数据特征与仿真数据特征分布比较

(c) 抗糖尿病药物的原始数据特征与仿真数据特征分布比较

(d) 再入院的原始数据特征与仿真数据特征分布比较

图 5　原始数据与仿真数据的特征可视化对比展示



(a1) 二维主成分分析　(b1) 二维主成分分析

(a2) 数据集1~数据集6　(b2) 数据集1~数据集6

(a) 经过数据嵌入处理的100万仿真数据　(b) 未经过数据嵌入处理的100万仿真数据

图 6　数据嵌入效果的可视化对比展示

就差异分数而言，本节旨在通过该指标定量地描述真实数据分布和合成数据分布的相似程度。首先计算原始 EHR 中训练集与测试集之间的差异分数，得到的结果为 2.412，这也为后续的实验定下了基线。假设所生成的数据与原始数据的差异分数过大且未趋近于 0 时，则显示生成数据的质量较差。

宏观来看，通过对比合成数据的训练集与测试集，其差异分数均小于原始数据集中的数值；但通过对比两者之间的差异分数，它们均表现出如上文中测试集明显优于训练集的结果。最后，实验验证了不同合成数据集之间的区别，差异分数都已得到较好的结果，这也表明所生成的假数据内部能够保持良好的一致性。

从表 3 中同样得出与表 2 中原始数据集经过数据嵌入后的性能要优于传统独热编码的仿真数据生成方案的结论，即合成数据的训练集和测试集与原始数据集之间的散度值与差异分数较大，对其性能有相应的损耗。

此外，鉴于所使用的 EHR 是一份不平衡的数据集，CTGAN 没有真正解决训练数据不平衡而造成生成数据真实性较低的问题[8]。最后，不同合成数据集之间的差异能够维持基本的一致性。

综上所述，若单纯从隐私角度评判，本节所生成数据集具有很好的保密性质。相对于传统无数据嵌入以及最新的结构化假数据生成 CTGAN 技术而言，本文提出的方法能够将最近邻对抗精度、隐私损失、散度值以及差异分数控制在较小的范围。值得注意的是，这不完全意味着生成的数据集是理想的，因为假设该合成数据在可用性预测方面的表现异常拙劣，则仿真数据集在真实场景中也无任何实用价值可言。所

以3.4节将采用可用性的评估方式基于EHR进行再入院的预测，以此与原始数据比较。

3.4 可用性评估分析

为了研究所生成的仿真数据用于再入院预测的可用性，本节首先基于大量的分类算法训练多个机器学习模型；同时为了尽可能公正地评估合成数据的效果，本节通过使用 Scikit-Learn 中所有的分类算法以实施验证。总体来说，Scikit-Learn 中包含 26 个可用于分类任务的算法。接下来，实验通过 EHR 的训练集（48 993 例）训练 26 种机器学习模型。最后，将测试集（20 997 例）用于评估训练模型的预测能力，为数据集的可用性提供参考标准。

为了获得更好的预测精度，本节采纳网格搜索的遍历方式寻找最佳性能的超参数组合，实现模型的泛化优化。表 4 和表 5 在分类任务的背景下计算了 3 个评估指标，包括准确率、F1 值和建模所需要的时间消耗。此外，本节多维度地评估合成数据的质量，以对比生成不同数据量（100 万、75 万、50 万、25 万、10 万、5 万和 2 万）的仿真数据和原始数据之间的可用性表现差异。考虑到篇幅原因，表 4

表 4 原始数据集与合成数据集的分类算法预测

算法	原始数据集			10 万合成数据集		
	准确率	F1 值	时间消耗/s	准确率	F1 值	时间消耗/s
NearestCentroid	0.63	0.71	0.23	0.72	0.77	0.53
DecisionTreeClassifier	0.82	0.83	0.89	0.84	0.84	18.49
ExtraTreeClassifier	0.84	0.84	0.28	0.76	0.80	21.98
LabelPropagation	0.85	0.84	1 941.31	0.91	0.87	2 578.94
LabelSpreading	0.85	0.84	2 361.60	0.91	0.87	2 709.98
PassiveAggressiveClassifier	0.87	0.86	0.29	0.91	0.87	0.74
BaggingClassifier	0.91	0.87	4.69	0.79	0.81	121.11
XGBClassifier	0.91	0.87	7.02	0.77	0.80	41.74
LinearDiscriminantAnalysis	0.91	0.87	0.77	0.90	0.87	2.12
KNeighborsClassifier	0.91	0.87	78.30	0.90	0.87	675.94
QuadraticDiscriminantAnalysis	0.11	0.05	0.32	0.91	0.87	0.81
CalibratedClassifierCV	0.91	0.87	65.22	0.91	0.87	107.71
LogisticRegression	0.91	0.87	0.62	0.91	0.87	1.24
LinearSVC	0.91	0.87	17.02	0.91	0.87	25.61
RidgeClassifier	0.91	0.87	0.29	0.90	0.87	0.66
RidgeClassifierCV	0.91	0.87	0.40	0.90	0.87	1.10
DummyClassifier	0.84	0.84	0.21	0.64	0.72	0.58
GaussianNB	0.09	0.02	0.28	0.83	0.83	0.75
BernoulliNB	0.91	0.87	0.27	0.80	0.82	0.78
LGBMClassifier	0.91	0.87	0.77	0.68	0.75	3.32
SGDClassifier	0.91	0.87	0.60	0.91	0.87	1.46
ExtraTreesClassifier	0.91	0.87	8.54	0.75	0.79	0.79
AdaBoostClassifier	0.91	0.87	3.59	0.87	0.86	68.82
SVC	0.91	0.87	864.79	0.91	0.87	578.69
CheckingClassifier	0.91	0.87	0.17	0.91	0.87	0.48
RandomForestClassifier	0.91	0.87	7.47	0.28	0.35	81.50
Perceptron	0.81	0.82	0.34	0.91	0.87	0.62
可用性统计平均值	0.821	0.794	198.751	0.827	0.822	260.981

仅展示了原始数据集与 10 万合成数据集之间的再入院预测效果。

在表 4 中，F1 值是最重要的分类评估指标，其数值越大，说明合成数据能够提供更高的可用性。不同于以往的工作[3,8]，本节提出采用可用性统计平均的思想来检验所生成数据的有效性，这意味着评判方法不是单纯依赖于一两个算法表现的优劣，而是基于所有可用的分类算法，对它们求得统计平均后综合评估预测精度，以期得到公正性。

从表 4 中可以观察到，原始独热编码测试集中统计平均 F1 值为 0.794，相较于未经过任何调参优化的模型和采样算法有了明显的提升。与此同时，从单个算法来看，表 4 中的 Perceptron 和 LogisticRegression 算法的 F1 值均达到 0.8 以上。所以从这个角度总结，本节所使用的网格搜索方法对模型性能的提升有较大的帮助。

因此在表 5 中，本节基于不同数据量的生成数据与原始数据展开深入的分析比较。同时测试了 CTGAN 所生成的最好仿真数据量在 EHR 进行预测的效果。

表 5　原始数据集与合成数据集的再入院预测

数据集	准确率	F1 值	时间消耗/s
原始数据	0.821	0.794	198.75
2 万	0.751	0.752	12.60
5 万	0.758	0.743	51.20
10 万	0.827	0.821	260.98
25 万	0.434	0.441	609.49
50 万	0.302	0.310	2 008.68
75 万	0.421	0.455	4 363.73
100 万	0.486	0.499	6 637.02
CTGAN	0.646	0.667	6 637.02

已知本文所提出的假数据生成方法在 10 万数据集上实现了最优的 F1 值，即 0.821。而原始独热编码测试集中统计平均 F1 值为 0.794，其中展示了超过 2%的提升。这也意味着，所生成的假数据可以在保护隐私的同时，依然具有良好的可用性。同时对于 CTGAN 而言，其分类预测准确性比本文提出的方法的最优值下降了 15%。

结合表 4 和表 5，本节对生成数据的可用性验证进行总结，就合成数据集而言，在保证其隐私性的前提下（见 3.3 节），需额外评估其可用性以便实现对该 EHR 优劣的综合评判。

本节实验验证了相对于传统无数据嵌入以及最新的结构化假数据生成 CTGAN 技术而言，本文提出的方法在机器学习下游任务中，特别是医疗数据的建模，拥有着更好的表现。

3.5　消融实验分析

本节基于消融研究[28]实施并验证了本文提出的方法，进而明确模型中每个组件的作用。表 6 展示了消融实验组件的验证结果。

从表 6 中可以观察到，数据嵌入比独热编码的方案更有效，WGAN 比 Vanilla GAN 更适合生成仿真数据。

表 6　消融实验组件的验证结果

组件模型		F1 值
表示学习方法	数据嵌入	+2.2%
	独热编码	−3.7%
深度学习模型架构	Vanilla GAN	−7.2%
	WGAN	+1.87%

4　结束语

本文对基于生成对抗网络的仿真数据生成技术进行了研究，在此基础上提出了相应的衡量指标，验证了仿真数据集的隐私性，并比较了多个机器学习模型以此评估仿真数据集的可用性表现。

本文发现双曲空间的分类特征嵌入能够以较少的维数表示大规模医学实体之间的层次结构，避免了由于类别过多且稀疏而引起的空间爆炸问题，同时保留了属性内在的关联关系，为医疗数据的多模态问题提供了一种解决思路。

基于分类特征嵌入的生成对抗网络技术通过创造合成数据集来提供隐私保护的替代方法，从而减少了直接发布原始数据的潜在风险。本文通过保持隐私性和实用性的平衡来证明所提方法的可靠性。最终，本文希望能够通过此方式减少敏感信息泄露的可能，为数据分析师进行隐私保护下的数据挖掘提供一种更有效的途径。

参考文献：

[1] ROCHER L, HENDRICKX J M, DE MONTJOYE Y A. Estimating the success of re-identifications in incomplete datasets using generative models[J]. Nature Communications, 2019, 10(1): 1-9.

[2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.

[3] FAN J, LIU T Y, LI G L, et al. Relational data synthesis using generative adversarial networks: a design space exploration[J]. arXiv Preprint, arXiv: 2008.12763, 2020.
[4] POTDAR K, TAHER S, CHINMAY D. A comparative study of categorical variable encoding techniques for neural network classifiers[J]. International Journal of Computer Applications, 2017, 175(4): 7-9.
[5] RODRÍGUEZ P, BAUTISTA M A, GONZÀLEZ J, et al. Beyond one-hot encoding: lower dimensional target embedding[J]. Image and Vision Computing, 2018, 75: 21-31.
[6] ZHANG X, DOU D J, WU J. Learning conceptual-contextual embeddings for medical text[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9579-9586.
[7] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.
[8] XU L, SKOULARIDOU M, CUESTA-INFANTE A, et al. Modeling tabular data using conditional GAN[J]. Advances in Neural Information Processing Systems, 2019, 32: 7335-7345.
[9] AGRAWAL R, SRIKANT R. Privacy-preserving data mining[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000: 439-450.
[10] 方滨兴, 贾焰, 李爱平, 等. 大数据隐私保护技术综述[J]. 大数据, 2016, 2(1): 1-18.
FANG B X, JIA Y, LI A P, et al. Privacy preservation in big data: a survey[J]. Big Data Research, 2016, 2(1): 1-18.
[11] 李凤华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4): 1-11.
LI F H, LI H, JIA Y, et al. Privacy computing: concept, connotation and its research trend[J]. Journal on Communications, 2016, 37(4): 1-11.
[12] GARFINKEL S L. De-identification of personal information[R]. National Institute of Standards and Technology, 2015.
[13] STRACK B, DESHAZO J P, GENNINGS C, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70, 000 clinical database patient records[J]. BioMed Research International, 2014, 2014: 781670.
[14] OSIA S A, SHAHIN SHAMSABADI A, SAJADMANESH S, et al. A hybrid deep learning architecture for privacy-preserving mobile analytics[J]. IEEE Internet of Things Journal, 2020, 7(5): 4505-4518.
[15] XIAO T H, TSAI Y H, SOHN K, et al. Adversarial learning of privacy-preserving and task-oriented representations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12434-12441.
[16] LIU S C, DU J Z, SHRIVASTAVA A, et al. Privacy adversarial network[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2019, 3(4): 1-18.
[17] LI A, DUAN Y X, YANG H R, et al. TIPRDC: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 824-832.
[18] GUO C, BERKHAHN F. Entity embeddings of categorical variables[J]. arXiv Preprint, arXiv:1604.06737, 2016.
[19] SLEE V N. The international classification of diseases: ninth revision (ICD-9)[J]. Annals of Internal Medicine, 1978, 88(3): 424.
[20] CHOI E, BAHADORI M T, SEARLES E, et al. Multi-layer representation learning for medical concepts[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1495-1504.
[21] WANG X, ZHANG Y D, SHI C. Hyperbolic heterogeneous information network embedding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 5337-5344.
[22] NICKEL M, KIELA D. Poincare embeddings for learning hierarchical representations[J]. arXiv Preprint, arXiv: 1705.08039, 2017.
[23] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// Proceedings of International Conference on Machine Learning. [S.l.:S.n.], 2017: 214-223.
[24] PATKI N. The synthetic data vault: generative modeling for relational databases[D]. Cambridge: Massachusetts Institute of Technology, 2016.
[25] YALE A, DASH S, DUTTA R, et al. Privacy preserving synthetic health data[C]// Proceedings of 2019 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, [S.l.:S.n.], 2019:2-10.
[26] WEIJS S V, NOOIJEN V R, NICK V D G. Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition[J]. Monthly Weather Review, 2010, 138(9): 3387-3399.
[27] WANG W, SUN Y, HALGAMUGE S. Improving MMD-GAN training with repulsive loss function[J]. arXiv Preprint, arXiv:1812.09916, 2018.
[28] 邹福泰, 谭越, 王林, 等. 基于生成对抗网络的僵尸网络检测[J]. 通信学报, 2021, 42(7): 95-106.
ZOU F T, TAN Y, WANG L, et al. Botnet detection based on generative adversarial network[J]. Journal on Communications, 2021, 42(7): 95-106.

[作者简介]

向夏雨（1991− ），男，湖南花垣人，北京邮电大学博士生，主要研究方向为隐私保护、医疗大数据分析。

王佳慧（1983− ），女，山西大同人，国家信息中心博士生，主要研究方向为数据安全、云安全、云取证安全、大数据安全。

王子睿（2000− ），女，辽宁大连人，哈尔滨工业大学（深圳）硕士生，主要研究方向为数据安全。

段少明（1994− ），男，湖南邵阳人，哈尔滨工业大学（深圳）博士生，主要研究方向为数据安全和机器学习。

潘鹤中（1991− ），男，辽宁本溪人，北京邮电大学博士生，主要研究方向为云安全、数据安全、密码学。

庄荣飞（1992− ），男，福建泉州人，哈尔滨工业大学（深圳）博士生，主要研究方向为数据安全、机器学习安全、隐私保护。

韩培义（1992− ），男，山西吕梁人，哈尔滨工业大学（深圳）助理研究员，主要研究方向为数据安全和隐私保护。

刘川意（1982− ），男，四川乐山人，博士，哈尔滨工业大学（深圳）教授，主要研究方向为云计算与云安全、大规模存储系统、数据保护与数据安全。