杨桂松, 姜艺媛, 何杏宇,等. 基于外部知识的缺失健康数据预测研究[J]. 智能计算机与应用,2025,15(12):23-30. DOI: 10.20169/j.issn.2095-2163.24040105

基于外部知识的缺失健康数据预测研究

杨桂松[1], 姜艺媛[1], 何杏宇[1,2], 傅建华[2]
(1 上海理工大学 光电信息与计算机工程学院, 上海 200093;2 上海理工大学 出版印刷与艺术设计学院, 上海 200093)

摘 要: 在大数据时代背景下,全球范围内已积累了大量的健康数据,然而,由于异构性和隐私性等因素的影响,人群健康数据的缺失现象十分严重。针对这种现象,提出了一种基于外部知识的缺失健康数据预测模型(Convolutional Neural Network - Conditional Generative Adversarial Network,CNN-CGAN)。该模型引入了外部知识,并利用卷积神经网络(Convolutional Neural Network ,CNN)进行特征提取,使用 CBAM(Convolutional Block Attention Module)注意力机制挖掘数据间相关性,最后修改条件生成对抗网络(Conditional Generative Adversarial Network ,CGAN)的损失函数来进一步发挥外部知识的作用,帮助模型完成小样本缺失数据预测。在真实的数据集上进行了对比实验,实验结果表明,CNN-CGAN 模型的均方根误差(Root Mean Square Error ,RMSE)和平均绝对误差(Mean Absolute Error ,MAE)都明显低于其他基线方法。因此,该模型通过引入外部知识来预测缺失健康数据是有效的。

关键词: 单步预测; 外部知识; 条件生成对抗网络; 数据相关性
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)12-0023-08

Research on missing health data prediction based on external knowledge

YANG Guisong[1], JIANG Yiyuan[1], HE Xingyu[1,2], FU Jianhua[2]
(1 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In the era of big data, a large amount of health data has been accumulated worldwide. However, due to the factors such as heterogeneity and privacy concerns, the phenomenon of missing data in population health data is quite severe. To solve this problem, a missing health data prediction model based on external knowledge, named Convolutional Neural Network-Conditional Generative Adversarial Network (CNN-CGAN), is proposed. This model incorporates external knowledge, utilizes Convolutional Neural Networks (CNN) for feature extraction, employs the Convolutional Block Attention Module (CBAM) attention mechanism to explore the correlations among data, and modifies the loss function of the Conditional Generative Adversarial Network (CGAN) to maximize the utilization of external knowledge and assist the model in predicting missing data with small samples. Comparative experiments were conducted on real datasets, and the results demonstrate that the CNN-CGAN model exhibits significantly lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) compared to other baseline methods. Therefore, the model is effective in predicting missing health data by incorporating external knowledge.

Key words: single step prediction; external knowledge; conditional generation adversarial network; data correlation

0 引 言

近年来,人口老龄化问题愈发严重,由于不健康的生活方式和人口的快速增长,慢性疾病和恶性疾病在中国的发病率逐渐上升,亚健康人群的数量也在急剧增加。这使得人们对群体健康状况的监控需求日益增强。《"健康是中国 2030"规划纲要》提出推进"互联网+"医疗,实施慢性病综合防控战略[1]。此外,世界卫生组织将慢性疾病列为十大基本公共卫生行为之首[2]。如今,全球已步入大数据时代,大量的数据被统计和整合,形成了不同维度的人群健康数据。通过对这些数据进行处理、分析、归纳和

基金项目: 上海市大学生创新创业训练计划(SH2023201)。
作者简介: 杨桂松(1982—),男 ,博士,副教授,主要研究方向:物联网与普适计算,大数据分析; 姜艺媛(1999—),女,硕士研究生,主要研究方向:医疗健康大数据;傅建华(2003—),男,本科生,主要研究方向:医疗健康大数据。
通信作者: 何杏宇(1984—),女,博士,副教授,主要研究方向:物联网,群智计算,大数据分析。Email:sherri_he@163.com。
收稿日期: 2024-04-01

总结,可以揭示人群健康状况的深层规律,为政府和医疗卫生事业提供数据支持,为临床提供更精确的预测模型[3]。

然而,医疗领域,数据缺失问题相当严重,构建和公开利用人口健康数据集面临着诸多挑战。在数据收集过程中,医疗机构需要在保护患者隐私的同时,付出较高的成本和开销来获取医疗数据,部分患者出于隐私保护考虑,不愿公开自己的健康数据。因此,许多人口健康数据集往往是不完整的,有缺失的,而在此基础上构建预测模型时,模型的性能可能会受到影响,导致预测的结果不准确。在医疗健康领域,决策往往基于数据分析和预测,如果数据不完整,无疑会给流行病学家、医疗卫生组织或研究人员的分析和判断带来困难,也不利于国家和政府的医疗卫生事业。

面对缺失数据的预测,现有研究通常有两种主要的处理方法:第一种方式是前期使用没有缺失的数据集进行训练,再用训练出的模型直接对缺失数据进行预测,这种方法对数据量的要求比较大,在数据量有限的情况下,模型的泛化能力可能会受到影响,导致预测结果不准确。另一种方法是在已有的完整的数据集进行模型的训练,并用此模型预测第一个缺失的数据,在预测完之后,再将这个预测的数据加入之前的数据集,重新训练模型以预测第二个缺失值,这个过程会不断重复,直至所有的缺失数据都被预测出来。这种方式训练出的模型既可以预测数据也可以补全数据,但缺点是预测误差会随着多次循环的进行而累积,这可能会导致最终结果的偏差。

因此,为了解决上述问题,本文提出了一个基于外部知识的缺失健康数据预测模型(CNN-CGAN),该模型结合了卷积神经网络(CNN)和条件生成对抗网络(CGAN)的优势,通过引入外部知识来弥补缺失数据的稀疏问题,并避免误差积累。具体来说,模型首先利用 CNN 进行特征提取,通过 CBAM 注意力机制进一步挖掘数据间的相关性。然后,结合外部知识,利用 CGAN 对缺失数据进行补全。在预测阶段,模型首先逐个补全预测年份之前的数据缺失部分,再将补全后的数据与原始数据进行整合,最后利用整合后的数据对目标年份进行预测。通过这种方式,CNN-CGAN 模型不仅能够有效地处理缺失数据,提高预测精度,还能够充分利用外部知识来增强模型的泛化能力。本文通过实验验证了该模型的有效性,并在真实数据集上取得了显著的性能提升。

1 相关工作

时序数据预测是数据分析领域的一个重要研究方向,其通过对历史数据的时间序列特征进行分析,预测未来的数值和趋势。传统的时序数据预测方法有自回归移动平均模型(Autoregressive Integrated Moving Average Model, ARIMA)[4],通过对历史数据的线性拟合来预测未来的趋势,还有多元线性回归模型(Multiple Linear Regression,MLR)[5],通过建立自变量和因变量之间的线性关系来预测未来的值。然而,传统的时序数据预测方法在处理非线性、复杂和多变的数据时,往往存在预测精度不高等问题。为了解决这一问题,近年来深度学习方法在时序数据预测中的应用得到了广泛关注。深度学习模型如循环神经网络(Recurrent Neural Network, RNN)[6]和长短时记忆网络(Long Short-Term Memory, LSTM)[7]等,具有较强的非线性拟合能力,能够捕捉数据中的复杂关系和模式。但是现实世界中,数据往往是不完整的,简单地删除所有缺少数据的样本可能会导致数据效率低下,因此有部分学者先对缺失数据进行补全,传统的补全方法有 K 近邻(K-Nearest Neighbor, KNN)[8-9]、随机森林(Random Forest,RF)[10]、MICE[11]等,一些基于深度学习的补全方法也表现出了良好的性能,如生成对抗网络(Generative Adversarial Network, GAN)[12]、自编码器(Auto-Encoder,AE)[13]。目前,基于 GAN 的补全方法 GAIN[14]是最先进的补全算法之一。此外,有的学者在生成对抗补全网络(Generative Adversarial Imputation Network,GAIN)的基础上进行改进,提出了可以综合利用疾病内部和疾病之间相关性的补全模型[15]以及可以对非传染性慢性疾病进行逐年补全的模型[16]。然而,直接使用补全后的数据进行预测又可能产生累积误差。因此,如何在保证数据完整性的同时,又避免引入额外误差,成为时序数据预测领域亟待解决的问题。

近年来,研究者们开始探索如何将外部知识融入模型中,以提高模型的推理能力和泛化能力。外部知识可以来源于多种形式,如常识知识、领域专业知识、先验知识和外部数据集等。将这些外部知识融入模型中,可以帮助模型更好地理解和处理复杂任务,提升模型的性能。给模型加入外部知识在计算机视觉、自然语言处理以及推荐系统领域得到了广泛的应用。在计算机视觉领域,有学者提出了一种利用外部知识构建可转移视觉系统的简单策略[17]。在自然语言处理领域,有学者提出了基于外部知识的视觉问题回答(Vidual Question Answering, VQA)模型,证明了外部知识的引入可以有效地提

高 VQA 模型的推理能力[18]。在推荐系统领域，有学者提出通过利用外部知识来进行短社交媒体文本的标签推荐[19]。然而，目前没有将外部知识用于缺失的人群健康数据预测领域。

因此，本文提出了一种基于外部知识的缺失数据预测模型（CNN-CGAN），该模型既弥补了缺失数据的稀疏问题，又极大程度的避免了误差积累问题。

2　基于外部知识的缺失健康数据预测模型

如图 1 所示，针对预测问题，首先，确定预测的年份，然后在外部知识的帮助下，逐个补全预测年份之前的数据缺失部分；接着，将补全后的数据与原始数据进行整合；最后，利用整合后的数据对预定的目标年份的数据进行预测。

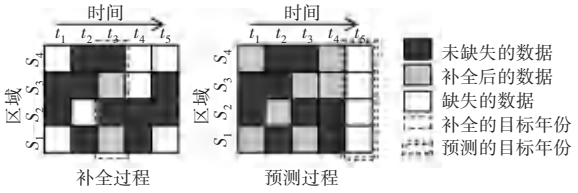


图 1　补全和预测过程说明

Fig. 1　Description of completion and prediction process

如图 2 所示，CNN-CGAN 模型采取了卷积神经网络和条件生成对抗网络相结合的网络结构，利用 CBAM 注意力机制和卷积神经网络进行特征挖掘和提取，通过给模型引入外部知识，并利用条件生成对抗网络对目标数据进行预测，各组件的具体描述如下。

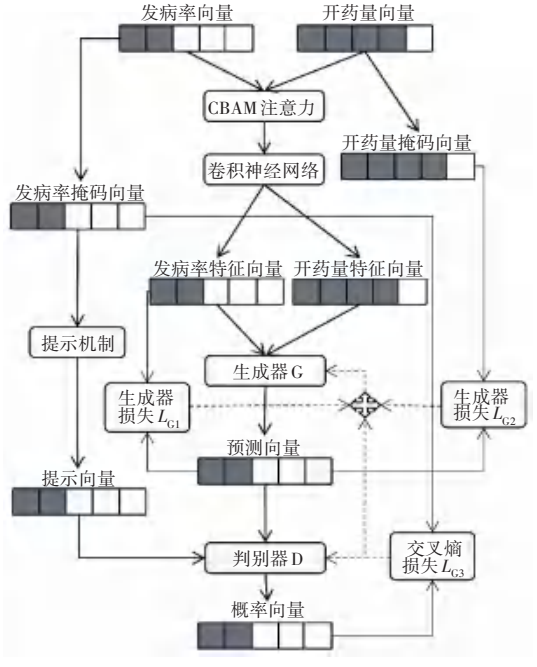


图 2　CNN-CGAN 模型架构

Fig. 2　Model architecture of CNN-CGAN

2.1　卷积神经网络

首先，对慢性疾病的发病率（糖尿病 X^1，高血压 X^2）和开药量（糖尿病 Y^1，高血压 Y^2）进行初始化操作，通过正态分布或均匀分布采样得到不同的噪声变量 $Z=(Z^1,Z^2,Z^3,Z^4)$，并用其来填补疾病的发病率和开药量中的缺失条目。然后，分别把两个疾病的发病率矩阵和开药量矩阵看作是具有两个通道的图像，输入到卷积神经网络（CNN）中进行特征提取，分别得到发病率特征矩阵 $\boldsymbol{X}'$ 和开药量特征矩阵 $\boldsymbol{Y}'$，矩阵的两个维度分别代表时间和空间，通道数量代表所研究的慢性疾病的数量。

$$\boldsymbol{X}'=C(X^1,X^2,Z^1,Z^2) \tag{1}$$

$$\boldsymbol{Y}'=C(Y^1,Y^2,Z^3,Z^4) \tag{2}$$

2.2　CBAM 注意力机制

在 CNN 之前，本文在模型中引入了 CBAM 注意力机制[20]，其结合了通道注意力和空间注意力。在特征提取过程中，通道注意力机制采用全局平均池化和全局最大池化方法来提取各通道的全局信息，并通过两个全连接层来自动学习各通道的权重系数。随后，将这两个输出结果相加，并使用 Sigmoid 函数将权重系数标准化至 0～1 的范围，以便对各通道特征进行缩放处理。完成这一步骤后，将缩放后的通道特征与原始特征相乘，以此增强关键特征对后续模型的作用力。对于空间注意力，由于卷积之后会产生多个通道，则会在每一个特征点的通道上进行最大池化和平均池化操作，将得到的两个矩阵进行拼接，并通过一个卷积层和 Sigmoid 函数来学习每个空间位置的权重。最后，将权重应用于特征图上的每个空间位置，以产生具有增强空间重要性的特征。通道注意力模块使得模型可以在综合利用疾病内部和疾病之间数据相关性的同时，沿着通道推断出注意力权重系数，为目标疾病所在的通道赋予更高的权重系数，从而重点提取输入矩阵中的有效特征，得到对应的发病率特征矩阵和开药量特征矩阵。空间注意力模块使得模型可以充分挖掘并利用不同年份的相关性，对接近目标年份的数据赋予更高系数的权重。

2.3　生成器

生成器以发病率特征矩阵 $\boldsymbol{X}'$、发病率掩码矩阵 $\boldsymbol{M}$ 和开药量特征矩阵 $\boldsymbol{Y}'$ 作为输入，其中，发病率掩码矩阵 $\boldsymbol{M}$ 跟随目标疾病的变化而变化。例如，如果选择糖尿病作为目标疾病，那么掩码矩阵就是糖尿病的发病率矩阵所生成的。根据当前选择的目标疾病可以获得相对应的发病率掩码矩阵和开药量掩码

矩阵,掩码矩阵的作用是用来表明哪些条目是存在的,哪些条目是缺失的。掩码矩阵的定义如下:

$$\boldsymbol{M}_{\mathrm{it}}=\begin{cases}1, & \text{if } X_{it} \text{ 没有缺失}\\ 0, & \text{if } X_{it} \text{ 缺失}\end{cases} \tag{3}$$

需要说明的是,在生成器网络的过程中,本文采用了特定的策略来整合不同类型的数据特征。具体来说,发病率特征矩阵 $\boldsymbol{X}'$ 和发病率掩码矩阵 $\boldsymbol{M}$ 被直接输入到网络的初始层,一次确保网络能够自始至终保留这些关键特征的完整性,而作为外部知识的开药量特征矩阵 $\boldsymbol{Y}'$ 是在生成器网络的第二层输入进去的,这样做的目的是防止模型在学习的过程中过度强化外部知识的特征,从而弱化了对发病率特征的识别和利用。

生成器 $G:\boldsymbol{X}'\times\{0,1\}^{n\times d}\rightarrow\hat{\boldsymbol{X}}$, 矩阵 $\bar{\boldsymbol{X}}$ 和 $\hat{\boldsymbol{X}}$ 的定义如下:

$$\bar{\boldsymbol{X}}=G(X',\boldsymbol{M}^1,Y') \tag{4}$$

$$\hat{\boldsymbol{X}}=\boldsymbol{M}^1\odot\boldsymbol{X}+(1-\boldsymbol{M}^1)\odot\boldsymbol{X} \tag{5}$$

其中, $\odot$ 表示按元素相乘; $\boldsymbol{X}$ 表示原始的目标疾病的发病率矩阵; $\boldsymbol{X}'$ 表示发病率的特征矩阵; $\boldsymbol{Y}'$ 表示开药量的特征矩阵; $\bar{\boldsymbol{X}}$ 对应于生成器生成的矩阵; $\hat{\boldsymbol{X}}$ 表示经过掩码矩阵替换后的矩阵,也就是本文最终想要的矩阵。

2.4 判别器

本文的研究框架中,判别器的作用和传统的生成对抗网络 GAN(Generative Adversarial Network, GAN)框架中的判别器相似,即作为生成器的对手进行博弈。然而,与传统 GAN 中判别器仅需区分生成器输出的数据为真实或虚假不同,本文中的判别器需执行更为复杂的任务。具体而言,判别器需评估填补矩阵中的数据点是由生成器产生的还是源自原始数据集,并输出一个介于 0~1 之间的概率值,以反映数据点为原始数据的可能性。该概率值越高,表明数据点越有可能是真实存在的条目。为了实现这一目标,首先,根据填补矩阵和提示矩阵,获得数据真实性的概率矩阵,再判断填补矩阵中原始数据和填补数据之间的差异。其中, $\boldsymbol{H}$ 为提示矩阵,用于控制掩码矩阵传入判别器的信息量,$\boldsymbol{H}$ 的定义不同,那么给判别器的提示信息就不同,从而可以训练出具有多个分布的生成器。最终,通过判别器的评估结果,可以选择生成器输出的最优数据分布。

2.5 优化目标

训练判别器的目的是要最大化预测发病率掩码矩阵 $\boldsymbol{M}^1$ 的概率,训练卷积神经网络和生成器的目的是要最小化判别器预测发病率掩码矩阵 $\boldsymbol{M}^1$ 的概率,二者相互对抗博弈,共同增强。将 $V(C,D,G)$ 定义如下:

$$V(C,D,G)=\mathbb{E}_{\hat{X},M^1,H}[\boldsymbol{M}^{1\mathrm{T}}\log D(\hat{\boldsymbol{X}},\boldsymbol{H})+(1-\boldsymbol{M}^1)^{\mathrm{T}}\log(1-D(\boldsymbol{X},\boldsymbol{H}))] \tag{6}$$

本文的目标和标准的 GAN 网络相同,都是最小最大化问题,公式如下:

$$\frac{\min}{C,G}\frac{\max}{D}V(C,D,G) \tag{7}$$

本文定义了两个损失函数,分别是生成器的损失和判别器的损失。

生成器的损失包含 3 个部分:第一部分是判别器的结果与发病率掩码矩阵 $\boldsymbol{M}^1$ 的交叉熵损失 L_{G1}; 第二部分是生成器生成的矩阵与发病率矩阵之间的 MSE 损失 L_{G2}; 第 3 部分是生成器的生成矩阵与开药量掩码矩阵 $\boldsymbol{M}^2$ 的 MSE 损失 L_{G3}。

$$L_{\mathrm{G1}}=-\sum_{i=0}^{n}\sum_{j=0}^{d}(1-m_{ij})\log(1-p_{ij}) \tag{8}$$

$$L_{\mathrm{G2}}=\frac{1}{n}\sum_{i=0}^{n}\sum_{j=0}^{d}(m_{ij}^1\times x_{ij}-m_{ij}^1\times\hat{y}_{ij})^2 \tag{9}$$

$$L_{\mathrm{G3}}=\frac{1}{n}\sum_{i=0}^{n}\sum_{j=0}^{d}(m_{ij}^2\times y_{ij}-m_{ij}^2\times\hat{y}_{ij})^2 \tag{10}$$

判别器的损失只包含一个部分:即判别器的结果与发病率掩码矩阵 $\boldsymbol{M}^1$ 之间的损失。

$$L_{\mathrm{D}}=-\sum_{i=0}^{n}\sum_{j=0}^{d}[m_{ij}^1\log(p_{ij})+(1-m_{ij}^1)\log(1-p_{ij})] \tag{11}$$

其中, a_{ij} 表示发病率矩阵的值; b_{ij} 表示开药量矩阵的值; m_{ij}^1 表示发病率掩码矩阵的值; m_{ij}^2 表示开药量掩码矩阵的值; $\hat{y}_{ij}$ 表示生成器生成的值; p_{ij} 表示判别器判别的值。

3 实验及分析

3.1 数据来源及预处理

本文在公开网站上收集到了 3 个主要的数据集,分别是伦敦各区慢性疾病发病率数据集、伦敦各诊所慢性疾病开药量数据集、伦敦各区人口数据集。

伦敦各区慢性疾病发病率数据集包含了伦敦 624 个选区的慢性病发病率数据(说明:对于每种疾病,发病率表示为每个登记处的患者人数与执业名单上所有患者人数的比率),其中有 30%左右的选区有缺失值。伦敦慢性疾病开药量数据集包含了伦

敦 637 家诊所 66 种药物的开药量数据，该数据集提供了对于慢性疾病药物治疗的使用情况，反映了不同药物在不同区域的分布和使用情况。伦敦各区的人口数据集包含了伦敦 624 个选区不同年份的人口数据。在进行开药量数据分析时，需要考虑人口数量对开药量的影响，因为人口规模可以影响到药物的需求量和药物的可获得性。

数据预处理是机器学习建模前的一个必要步骤，其作用是将原始数据转化为算法需要的形式，从而提高预测建模精度。本文数据预处理的步骤如下：

1）数据清洗：首先，删除数据集中重复的条目，过滤多余的数据字段，确保数据集中每个记录都是唯一的；其次，更正数据集中错误的数据类型；最后，检测原始数据中缺失的或者不合逻辑的值，将其标为缺失数据。本文使用数据缺失率描述数据的缺失程度，缺失率的计算公式为：

$$\beta = \frac{h}{l} \times 100\% \tag{12}$$

其中，h 为缺失参数的个数，l 为总参数的个数。

此外，在开药量数据集中，包含了 66 种药物的开药数据，然而，并非所有药物都与本文研究的目标疾病相关，因此，需要对药物进行过滤，筛选出与研究目标疾病直接相关的药物。

2）数据整合：开药量数据集中，诊所与选区的关系并非一一对应，一个选区中可能包含了多个诊所的开药量数据，也可能没有数据。因此，本文根据选区的邮编代码与诊所的邮编代码，对开药量数据进行了整合，使其与发病率数据分别在时间与空间上一一对应。

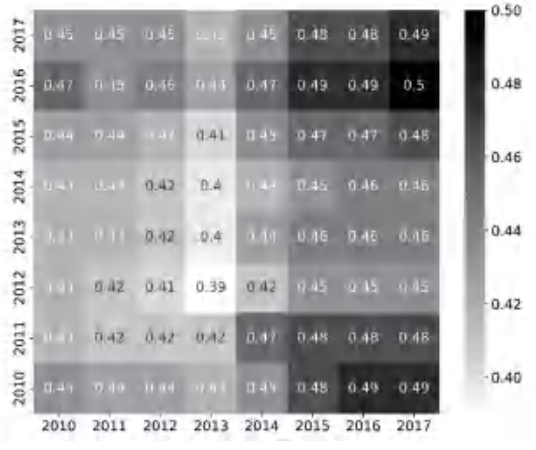
3）数据归一化：由于原始的开药量数据中，不同剂量的同种药物代码不同，因此需要对同种药物的开药量进行加权平均（例如，如果一个药物有 10 mg 和 20 mg 两种剂量，而实际使用中 10 mg 的比例是 2/3，20 mg 的比例是 1/3，则加权平均后的开药量是 10 mg 乘以 2/3 加上 20 mg 乘以 1/3），从而得到一个代表该药物总开药量的单一数值。

此外，为了消除不同选区人口数量差异的影响，需要将每年的开药量数据除以相应年份的人口数量。这样，每个选区的开药量就不再是原始的数值，而是经过归一化处理后的相对数值，这些数值反映了每万人的开药量，从而使得不同选区的开药量数据可以进行直接比较。

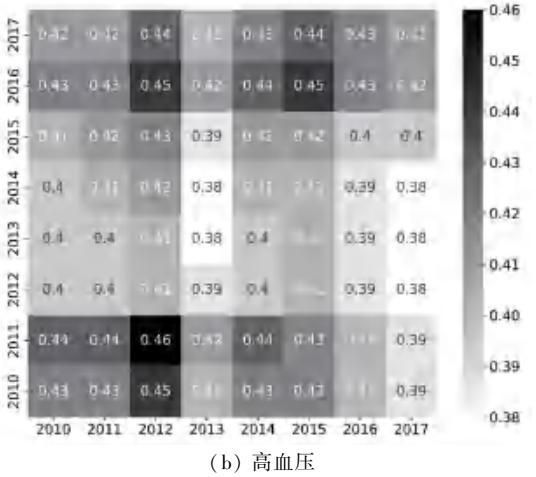
经过数据的收集与预处理，最终获得了 2010-2017 年期间伦敦 483 个选区的慢性疾病发病率数据以及开药量数据。本文选取糖尿病和高血压这两种非传染性慢性疾病作为研究对象，在数据缺失的条件下，通过引入目标疾病对应的开药量数据作为辅助，进行非传染性慢性疾病的发病率的预测。

3.2　相关性分析

在这一部分中，对发病率数据和开药量数据做了 Pearson 相关性分析。根据相关系数的定义，当相关系数值接近 1 时，表示两个变量之间存在强正相关；当相关系数值接近 -1 时，表示两个变量之间存在强负相关；而当相关系数值接近 0 时，则表示两个变量之间几乎没有线性关系。如图 3 所示，横坐标代表发病率数据，纵坐标代表开药量数据，可以直观地观察到两者之间确实存在一定的相关性，且相关系数值分布在 0.38～0.5 之间，说明糖尿病和高血压的发病率和开药量之间呈现出中等程度的正相关。这意味着，随着开药量的增加，发病率也会相应地提高，或者反过来说，发病率较高时，开药量也相对较大。证实了可以借助开药量这一外部知识来辅助预测糖尿病和高血压发病率数据。



（a）糖尿病



（b）高血压

图 3　发病率和开药量数据相关性分析

Fig. 3　Correlation analysis of prevalence and prescription data

3.3 实验环境及模型参数

本文使用的编程语言为 Python3.7,编程环境为 Tensorflow 1.14.0,模型的参数取值见表 1。

表 1 实验参数

Table 1 Experimental parameters

	参数名称	内容
卷积神经网络 CNN	网络层数	3
	激励函数	ReLU
	学习率	1E-3
生成器 G	网络层数	2
	激励函数	ReLU
	优化器	Adam
	学习率	1E-3
判别器 D	网络层数	2
	激励函数	ReLU
	优化器	Adam
	学习率	1E-3

3.4 模型评估与比较

为验证所提出的 CNN-CGAN 模型在预测方面的有效性,分别从均方根误差(Root Mean Square Error,RMSE)和平均绝对误差(Mean Absolute Error,MAE)来评估模型的性能,上述指标常用来作为机器学习模型预测结果衡量的标准,反映预测值与实验值之间的偏差,计算公式如下式。指标的数值越低,则表示模型的填补性能越好。

$$\mathrm{RMSE}=\sqrt{\frac{\sum_i (y_i-\hat{y}_i)^2}{n}} \tag{13}$$

$$\mathrm{MAE}=\frac{\sum_i |y_i-\hat{y}_i|}{n} \tag{14}$$

其中,$\hat{y}_i$ 是预测出来的结果;y_i 是原本真实的数据;n 是被预测的数目。

通过将本文提出的 CNN-CGAN 方法与其他方法进行对比,证明 CNN-CGAN 方法的优越性。对比算法如下:

(1)自回归差分移动平均模型(ARIMA):该模型通过分析数据间的自相关特性和应用差分技术来揭示数据中的隐含时间序列结构,并依据这些结构进行数据预测。

(2)循环神经网络(RNN):对于序列预测任务,如时间序列分析,RNN 可以使用最后一个时间步的隐藏状态来生成下一个时间点的预测。

(3)长短期记忆递归神经网络(LSTM):模型可以实现对信息的长期记忆和控制,多用于时间序列预测。

(4)CPH:通过对不完整的数据进行建模,提取疾病内部和不同疾病之间的时空相关性进行预测插补。

3.5 实验结果分析

本节对各种模型在真实世界数据集上进行实验。

(1)为验证 CNN-CGAN 模型对未来数据预测的有效性。本文选取 2017 年作为目标年份进行预测,首先,使用完整的 2010-2016 年的数据预测 2017 年的数据,并对比 ARIMA、RNN、LSTM、CPH 和 CNN-CGAN 的预测结果。其次,先按照 $\beta=50\%$ 丢弃数据集中的数据,然后使用 CNN-CGAN 模型将 2010-2016 年丢弃的数据依次补全(例如,将目标年份设置为 2010 年,即可得到 2010 年补全后的结果;将目标年份设置为 2016 年,即可得到 2016 年补全后的结果),再将 2010-2016 年所有补全好的值分别填入到丢弃的空白值中。最后,使用整合后的数据进行 2017 年发病率的预测,对于每种方法,在不同的缺失率下进行 10 次实验,然后取 10 次结果的平均值作为最终的 RMSE 值和 MAE 值,实验结果见表 2、表 3。

表 2 2017 年糖尿病预测结果

Table 2 Diabetes prediction results in 2017

方法	$\beta=0$(完整)		$\beta=50\%$	
	RMSE	MAE	RMSE	MAE
ARIMA	0.098 3	0.042 1	0.089 0	0.047 6
RNN	0.046 4	0.023 4	0.052 3	0.041 1
LSTM	0.016 0	0.011 9	0.017 1	0.015 3
CPH	0.012 8	0.010 4	0.023 3	0.016 2
CNN-CGAN	**0.007 6**	**0.001 8**	**0.017 0**	**0.014 0**

表 3 2017 年高血压预测结果

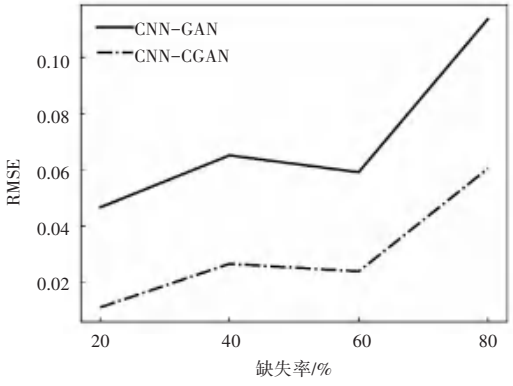
Table 3 Hypertension prediction results in 2017

方法	$\beta=0$(完整)		$\beta=50\%$	
	RMSE	MAE	RMSE	MAE
ARIMA	0.090 8	0.035 9	0.078 3	0.042 4
RNN	0.059 0	0.023 5	0.077 6	0.066 5
LSTM	0.028 1	0.014 9	0.024 4	0.018 4
CPH	0.018 1	0.015 6	0.018 9	0.015 1
CNN-CGAN	**0.012 8**	**0.002 7**	**0.013 9**	**0.006 7**

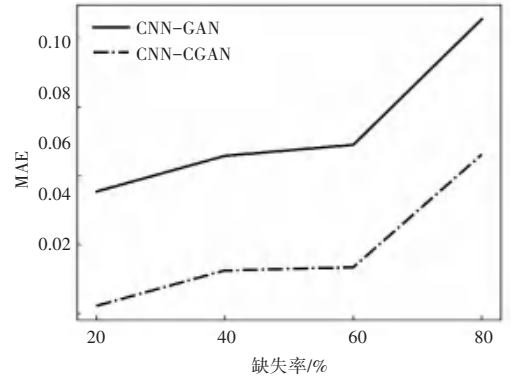
在表 2 和表 3 中可以清晰地看到预测实验的结果,以 2017 年高血压预测结果为例,在 50% 的缺失率下进行逐年补全后的结果再进行预测的 RMSE

值，与完整的数据集进行预测的 RMSE 值仅仅相差 0.001，甚至对于 ARIMA 和 LSTM 模型来说，补全后再进行预测的结果比完整数据集预测的误差值还要低。这个结果再次证明了 CNN-CGAN 模型的高准确性。此外，无论是用完整的 2010-2016 年的数据来预测，还是使用补全后的数据进行预测，CNN-CGAN 模型都展现出了良好的预测性能，并极大程度上消除了累积误差。

（2）为验证外部知识的引入对 CNN-CGAN 模型的有效性，进行了消融实验，将去除外部知识的模型称为 CNN-GAN。以 2017 年糖尿病发病率数据集，分别在 20%、40%、60%、80%的缺失率下进行 10 次预测实验，取平均值，实验结果见图 4。从图 4 中可以清晰地看到，引入外部知识的 CNN-CGAN 模型的误差要远低于没有外部知识的 CNN-GAN 模型。



（a）RMSE



（b）MAE

图 4　CNN-CGAN 和 CNN-GAN 在不同缺失率下的误差

Fig. 4　Error of CNN-CGAN and CNN-GAN at different missing rates

4　结束语

本文将疾病开药量这一外部知识引入到缺失发病率的预测模型中，以辅助模型工作。为了实现这一目标，本文在卷积神经网络 CNN 中加入了 CBAM 注意力机制，使得模型在空间和通道两个维度都能更好地进行关键数据的特征提取。此外，还修改了 CGAN 网络的损失函数，修改后的损失函数能够使得外部知识更好地辅助模型进行预测任务，从而生成更真实、更符合现实规律的数据。通过在真实的数据集上使用不同的数据预测模型进行了对比实验，结果表明，CNN-CGAN 模型在缺失数据预测方面的能力优于基线，在实际缺失数据的预测中，可以有效地利用外部知识，具有较高的准确性和可行性，有助于提高人群健康数据分析的精度和效率。

参考文献

[1] 高红，管仲军. 健康中国战略下大型公立医院高质量发展创新路径再探讨[J]. 中国行政管理，2023，39(11)：109-114.

[2] VERSCHUUREN M，OERS H V. Population health monitoring：Climbing the information pyramid [J]. Ciência & saúde coletiva，2021，26：2557-2558.

[3] PEREZ J A，POON C C Y，MERRIFIELD R D，et al. Big data for health [J]. IEEE Journal of Biomedical and Health Informatics，2015，19(4)：1193-1208.

[4] DONG S，LI R，SUN L G，et al. Short-term traffic forecast system of Beijing [J]. Transportation Research Record，2010，2193：116-123. DOI：10.3141/2193-14

[5] ZHAO J，WANG W，LIU Y，et al. A two-stage online prediction method for a blast furnace gas system and its application[J]. IEEE Transactions on Control Systems Technology，2011，19(3)：507-520.

[6] CINAR Y G，MIRISAEE H，GOSWAMI P，et al. Period-aware content attention RNNS for time series forecasting with missing values[J]. Neurocomputing，2018，312(27)：177-186.

[7] YUAN H，XU G，YAO Z，et al. Imputation of missing data in time series for air pollutants using Long Short-Term Memory Recurrent Neural Networks[C]//Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium. New York：ACM，2018：1293-1300. DOI：10.1145/3267305.3274648

[8] HASTIE T，TIBSHIRANI R，FRIEDMAN J. The elements of statistical learning. 2001 [J]. Journal of the Royal Statistical Society，2004，167(1)：192-192.

[9] QIN Y，ZHANG S，ZHANG C. Combining KNN imputation and bootstrap calibrated：Empirical likelihood for incomplete data analysis [J]. International Journal of Data Warehousing and Mining (IJDWM)，2010，6(4)：61-73. DOI：10.4018/jdwm.2010100104

[10] BPÜHLMANN P. MissForest：Non-parametric missing value imputation for mixed-type data[J]. Bioinformatics，2012，28(1)：112-118.

[11] BUUREN S V，GROOTHUIS-OUDSHOORN K. MICE：Multivariate imputation by chained equations in R[J]. Journal of Statistical Software，2011，45(3)：1-67. DOI：10.18637/jss.v045.i03

[12] ZHANG Y，ZHOU B，CAI X，et al. Missing value imputation in multivariate time series with End-to-End Generative Adversarial Networks[J]. Information Sciences，2020，551(6)：209-227.

DOI:10.1016/j.ins.2020.11.035
[13] GONDARA L, WANG K. Multiple imputation using deep denoising autoencoders[J]. arXiv preprint arXiv, 1705.02737, 2017.
[14] YOON J, JORDON J, SCHAAR M V D. GAIN: Missing data imputation using generative adversarial nets[J]. arXiv preprint arXiv, 1806.02920, 2018. DOI:10.48550/arXiv.1806.02920
[15] FENG Y, WANG J, WANG Y, et al. Completing missing prevalence rates for multiple chronic diseases by jointly leveraging both intra- and inter-disease population health data correlations [C]//Proceedings of the Web Conference. 2021:1-11. DOI:10.1145/3442381.3449811
[16] FENG Y, WANG J, WANG Y, et al. Towards sustainable compressive population health: A gan-based year-by-year imputation method[J]. ACM Transactions on Computing for Sustainable Computing, 2023, 8(3):1-24.
[17] SHEN S, LI C, HU X, et al. K-LITE: Learning transferable visual models with external knowledge[J]. arXiv preprint arXiv, 2204.09222, 2022. DOI:10.48550/arXiv.2204.09222
[18] MARINO K, RASTEGARI M, FARHADI A, et al. OK-VQA: A visual question answering benchmark requiring external knowledge[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ:IEEE, 2020:3195-3204. DOI:10.1109/CVPR.2019.00331
[19] KUMAR A S M. Hashtag recommendation for short social media texts using word-embeddings and external knowledge [J]. Knowledge and Information Systems, 2021, 63(1):1-28.
[20] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of European Conference on Computer Visions. Cham:Springer, 2018:3-19.