Modeling and Simulation 建模与仿真, 2024, 13(3), 3784-3796
Published Online May 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.133345

Hans汉斯

基于外部知识辅助的人群健康数据预测方法

傅建华，何杏宇，张鑫泽，梁　涛

上海理工大学出版印刷与艺术设计学院，上海

收稿日期：2024年4月28日；录用日期：2024年5月23日；发布日期：2024年5月31日

摘　要

随着深度学习技术的发展和引入，现有人群健康数据预测方法的性能不断提高，但仍然受到数据质量问题的限制。为此，本文提出了一种基于外部知识辅助的人群健康数据预测方法。首先，该方法以与冠心病患病率相关性较强的高血压患病率数据和选区老年人口比例数据作为外部知识辅助填补冠心病患病率数据稀疏部分，对上述数据进行预处理后，构建CNN模型对高血压患病率数据和选区老年人口比例数据提取特征矩阵，并和随机噪声、部分完整的冠心病患病率数据作为CGAN模型的输入，以生成用来填补原冠心病患病率数据中稀疏部分的人工样本；然后，该方法将填补后的完整数据集通过ARIMA模型拟合得到模型特征，并输入GRU模型进行预测分析。实验结果表明，本文方法在MAE和RMSE上和KNN模型和RNN模型相差不多，但MPAE大大降低。

关键词

人群健康数据预测，深度学习，外部知识辅助

Population Health Data Prediction Method Based on External Knowledge Assistance

Jianhua Fu, Xingyu He, Xinze Zhang, Tao Liang

College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 28th, 2024; accepted: May 23rd, 2024; published: May 31st, 2024

Abstract

With the development of deep learning technologies, the performance of existing population health data prediction methods has been improved, but still suffers the limitation of data quality. In view of this, this paper proposes a population health data prediction method based on external knowledge assistance. In this method, firstly, the data of hypertension prevalence and elderly population proportion are utilized as external knowledge to fill the sparse part of coronary heart

文章引用: 傅建华, 何杏宇, 张鑫泽, 梁涛. 基于外部知识辅助的人群健康数据预测方法[J]. 建模与仿真, 2024, 13(3): 3784-3796. DOI: 10.12677/mos.2024.133345

disease prevalence, due to their strong correlation, their feature matrixes are extracted via the CNN model and input into the CGAN model, with the complete coronary heart disease prevalence data and random noise part, to generate artificial samples; Further, the complete data set after filling is fitted by the ARIMA model to obtain the model features, and input into the GRU model for prediction analysis. The experiment results show that the proposed method has similar MAE and RMSE to RNN and KNN models, but less MPAE than them.

Keywords

Population Health Data Prediction, Deep Learning, External Knowledge Assistance

1. 引言

随着科技的快速发展和全球健康意识的提升，人群健康数据预测方法的研究逐渐成为了公共卫生、医疗科技等领域的研究热点。人群健康数据预测可以帮助人们提前预警潜在的健康风险，还能辅助相关政策制定、优化医疗资源配置[1]。人群健康数据预测早期主要依赖传统统计模型分析数据趋势和周期性来预测未来健康状态，如指数平滑模型和 ARIMA 模型[2] [3]。然而，统计模型在处理大量高维数据时存在局限性，容易出现过拟合现象，且模型的泛化能力有待提高，难以捕捉数据中的深层次特征[4]。

然而随着机器学习技术的发展，近年来深度学习技术在人群健康数据预测领域取得了显著的应用进展[5]，例如卷积神经网络(CNN)、循环神经网络(RNN)、长短期记忆网络(LSTM)等模型的应用[6]，能够自动提取数据特征，处理非线性关系，一定程度上缓解了过拟合问题，大幅提升预测准确性[7]。然而，深度学习模型需要大量的数据资源支持，而且可解释性有待提高。为此，研究者尝试融合不同的模型，形成混合模型，例如，将 ARIMA 与 LS-SVM 结合，形成 ARIMA-LS-SVM 模型[8]。虽然预测模型本身在不断优化和进步，但仍然受到数据质量问题的限制。为了提高数据质量，常用的数据填补方法有删除法和替换法[9]，但这些传统方法的效果并不理想，于是深度学习模型也被引入到数据填补技术中[10]，旨在捕捉数据复杂结构以精准填补缺失值，但在背景知识单一情况下填补效果仍然受限。

为此，本文将提出一种基于外部知识辅助的人群健康数据预测方法，利用外部知识来解决稀疏数据对预测模型的限制问题。具体地，在本文的预测方法中，首先，本文以与冠心病患病率相关性较强的高血压患病率数据和选区老年人口比例数据作为外部知识辅助填补冠心病患病率数据稀疏部分，对上述数据进行预处理后，构建 CNN 模型对高血压患病率数据和选区老年人口比例数据提取特征矩阵，并和随机噪声、部分完整的冠心病患病率数据作为 CGAN 模型的输入，以生成用来填补原冠心病患病率数据中稀疏部分的人工样本。然后，本文将填补后的完整数据集通过 ARIMA 模型拟合得到模型特征，并输入 GRU 模型进行预测分析。

2. 相关工作

2.1. 人群健康数据预测方法

随着大数据和人工智能技术的迅猛发展，人群健康数据预测已成为全球研究领域的热点。这一技术融合为洞察疾病发生趋势、预防控制和治疗提供了前所未有的机遇。在人群健康数据预测方法研究中，

统计模型，特别是 ARIMA 模型，在其中扮演着重要角色。该模型能够利用历史健康数据，如疾病发病率、医疗资源利用情况等，进行趋势、周期性和随机性的建模和预测。例如，文献[11]成功建立 ARIMA 预测模型，表明了该模型能够较好应用于流感样病例预测预警，为疫情防控提供科学依据。文献[12]采用 ARIMA 模型，对流感数据进行原始序列预处理、模型识别、参数估计和统计建模，成功预测流感发病趋势。

然而，ARIMA 模型对数据质量的要求较高，对缺失值和异常值敏感，处理复杂非线性关系时存在局限性，这在实际应用中可能导致预测结果的不准确[13]。为了解决这一问题，使用机器学习技术来辅助处理非线性关系是有必要的。文献[14]将病例人数作为输出变量建立的 BP 神经网络模型具有良好的预测效果。文献[15]提出利用 SVM 算法建立的联合模型具有最佳效能。近年来，深度学习技术在人群健康数据预测领域取得了显著进展，神经网络在数据预测中有较大的优势。文献[16]在结直肠癌的研究中，利用神经网络模型，成功识别了“炎癌转化”过程的关键基因，并预测了防治中药的效果。文献[17]成功的将卷积深度神经学习网络应用于白细胞的识别，较好地提供有关人类健康和疾病的有价值的信息。

由于单一模型在人群健康数据预测方法研究中各有其局限性。为了克服这些缺点，研究者结合多种模型进行集成学习或混合建模，以充分利用不同模型的优势，提高预测精度和稳定性。文献[18]建立了 ARIMA-SVM 模型，精确预测了海南省肺结核发病数，为肺结核的预测预警提供了新思路。文献[19]利用 SARIMA 模型的预测值加上 SVM 模型残差预测值建立得到的 SARIMA-SVM 组合模型具有更高的预测精度，更适用于全国丙肝月发病率的预测。

2.2. 人群健康数据预测中的数据填补方法

人群健康数据预测中的数据填补研究是一个既关键又复杂的领域。由于数据来源的多样性和复杂性，数据缺失成为了一个普遍存在的问题，对人群健康数据预测的准确性和可靠性产生了直接影响。目前，数据填补方法的研究已经取得了显著进展。传统的数据填补方法主要包括删除法、哑变量调整法、条件均数填补法、热平台填补法、多重填补法[20]。除了上述的传统数据填补方法，近年来还涌现出许多新的技术，如决策树[21]和随机森林[22]，利用模型学习数据的内在规律，对缺失值进行预测和填补，提高了数据完整性。此外，多重插补方法[23]，也常用于连续数据的填补。

随着机器学习和深度学习技术的进一步发展，新的数据填补方法更加复杂且有效。比如贝叶斯网络方法[24]对统计假设并不严格，结合了概率论与图论的优势，使其更适合在临床研究中应用。同时，基于 Stacking 集成学习策略[25]，将 KNN、决策树和 SVR 三种不同学习机制的模型作为基学习器，集成构建新的强学习器，用于缺失数据的填补。但是，这些新的数据填补技术在背景知识单一情况下仍然存在效果不理想的问题，而引入外部知识辅助是解决该问题的有效途径。

3. 基于外部知识辅助的人群健康数据预测方法

本文提出基于外部知识辅助的人群健康数据预测模型，总体框架如图 1 所示，由基于 CNN-CGAN 模型的外部知识辅助的数据填补模块和基于 ARIMA-GRU 模型的人群健康数据预测模块两个模块组成。

在基于 CNN-CGAN 模型的外部知识辅助的数据填补模块中，以与冠心病患病率相关性较强的高血压患病率数据和选区老年人口比例数据作为外部知识辅助填补冠心病患病率数据稀疏部分。首先对高血压患病率数据和选区老年人口比例数据进行数据预处理，使数据满足 CNN 模型输入要求，然后，构建一个 CNN 模型分别对两份数据提取特征矩阵。以 CNN 模型提取出来的特征矩阵、随机噪声和部分完整的冠心病患病率数据作为 CGAN 模型的输入，训练并生成可靠的人工样本以填补冠心病患病率数据集中的稀疏部分。

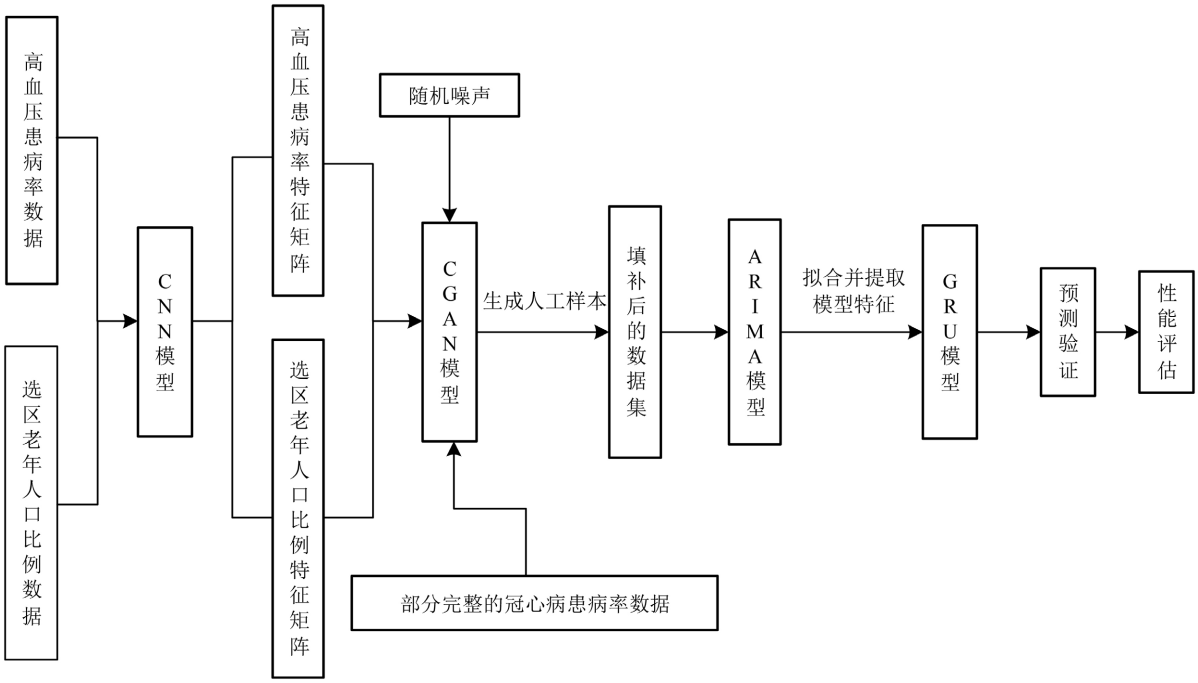

Figure 1. The structure of prediction model
图 1. 预测模型结构图

在基于 ARIMA-GRU 模型的人群健康数据预测模块中，本文先构建一个 ARIMA 模型来对填补后的冠心病患病率数据集拟合并提取模型特征，并作为 GRU 模型的输入对提前划分好的测试集进行预测，最后对实验结果进行性能评估。

3.1. 基于 CNN-CGAN 模型的外部知识辅助数据填补方法

3.1.1. CNN 模型提取特征矩阵

如图 2 所示，本文采用 CNN 模型以提取数据的特征矩阵，CNN 模型在提取数据特征矩阵时，首先接收原始数据，然后通过多个卷积层使用不同的卷积核提取局部特征，每个卷积核生成一个特征图。输入层到卷积层的函数为：

$$\alpha^l = \sigma\left(\alpha^{l-1} * W^l + b^l\right) \tag{1}$$

其中，$\sigma(\cdot)$为激活函数，采用 Relu 函数加入非线性特性，增加特征表示能力；l 为网络层数；“*”为卷积操作；α^{l-1}为我们提前准备好的高血压患病率数据和选区老年人口比例数据；α^l 为卷积层提取的特征图；b 为偏置；W 为权重。

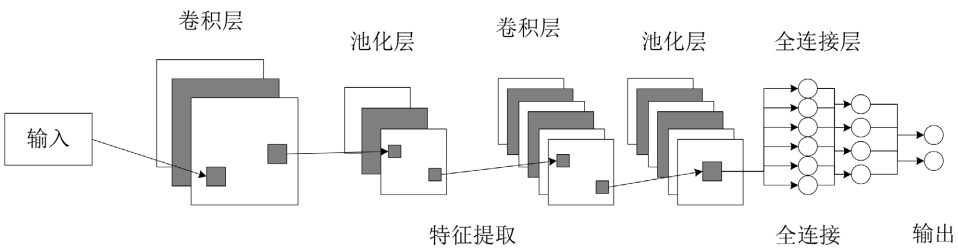

Figure 2. The basic structure of the CNN model
图 2. CNN 模型的基本结构

Table 1. CNN network parameter settings
表 1. CNN 网络参数设置

参数名称	卷积核尺寸	卷积核数量	卷积核大小	输出尺寸
输入层				1200 × 1 × 1
卷积层 1	3 × 1	6	1	1200 × 1 × 6
池化层 1	2 × 1	6	2	600 × 1 × 6
卷积层 2	3 × 1	12	1	600 × 1 × 12
池化层 2	2 × 1	12	2	300 × 1 × 12
全连接层				6000 × 1
Softmax	6	1		6

之后，池化层降低特征图的维度，减少计算量并保留重要特征。池化层的输出函数：

$$\alpha^{l} = p_{pool}\left(\alpha^{l-1}\right) \tag{2}$$

其中，$p_{pool}(\cdot)$ 表示池化层操作，在本模型中选用最大池化，以凸显特征。本文以经过两层卷积层和两层池化层处理提取出的特征图作为特征矩阵以及下一部分 CGAN 模型的输入。

全连接层将特征图展平并进行高级特征学习，最后输出层给出最终的分类或预测结果。通过不断训练，CNN 模型能够逐渐优化特征提取过程，提高任务性能。相关参数如表 1 所示，本文 CNN 模型借鉴了经典 AlexNet 网络结构思想，即卷积核数量随着卷积层的深入而变多，由于输入信号维度大于二维图像某一维度，所以常规的二维卷积核大小 3 × 3，5 × 5，7 × 7 不适用于该网络结构，所以经过多次实验，本文所涉及的 CNN 网络结构包含 1 个输入层、2 个卷积层、2 个池化层、1 个全连接层和 Softmax 分类器。各卷积层分别包括 6 个 3 × 1 的卷积核和 12 个 3 × 1 的卷积核。各池化层分别采用 2 × 1 和 2 × 1 的小窗口进行特征降维。Softmax 分类器输出 6 种识别概率向量。

3.1.2. CGAN 模型生成人工样本

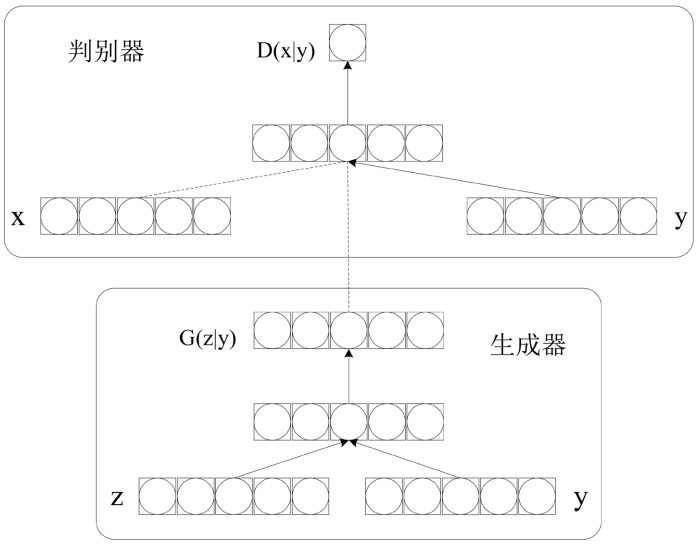

Figure 3. Frame diagram of CGAN model
图 3. CGAN 模型的框架图

如图 3 所示，CGAN 模型由生成器 G 和判别器 D 两个部分组成，其中 x 代表我们提前准备好的部分完整且连续的冠心病患病率真实数据，y 代表外部条件，即我们通过 CNN 模型提取得到的两个关于高血压患病率数据和选区老年人口比例数据的特征矩阵，z 代表随机噪声。

$$G\left(z,c_1,c_2\right)=\arg\min_G\max_D E\left[\log\left(D\left(G\left(z,c_1,c_2\right),c_1,c_2\right)\right)\right]+E\left[\log\left(1-D\left(G\left(z,c_1,c_2\right),c_1,c_2\right)\right)\right] \tag{3}$$

$$D\left(x,c_1,c_2\right)=\arg\min_G\max_D E\left[\log\left(D\left(G\left(z,c_1,c_2\right)\right)\right)\right]+E\left[\log\left(1-D\left(G\left(z,c_1,c_2\right),c_1,c_2\right)\right)\right] \tag{4}$$

判别器的损失函数基于它对真实和生成样本正确分类的能力，目标是最大化对真实样本的判断和对生成样本的判断之间的差距。通过反向传播算法计算损失函数的梯度，更新生成器和判别器的网络参数。这个过程会重复多次，每次迭代都会使得生成器更好地理解如何根据条件变量生成高质量的样本，同时判别器也会变得更加精确定位真实样本。损失函数公式如下：

$$\min_G\max_D V\left(D,G\right)=E_{X\sim P_{data}(x)}\log\left(D\left(x\middle|y\right)\right)+E_{Z\sim P_z(z)}\log\left(1-D\left(G\left(z\middle|y\right)\middle|y\right)\right) \tag{5}$$

其中，$V\left(D,G\right)$ 表示判别器 D 和生成器 G 之间的价值函数，P_{data} 为实际分布规律，P_z 为人工样本的分布规律。表 2 为该模型的相关参数和大体结构，本文在池化层依旧选择最大池化以凸显特征，并选择用 Adam 优化器对生成器和判别器进行优化，同时插入多个 Dropout 层以增强模型的鲁棒性，防止过拟合。

Table 2. CGAN network parameter settings
表 2. CGAN 网络参数设置

组成部分	结构	激活函数	优化器	Dropout 层
判别器	卷积层	Rdu		1
	Maxpool 池化层		Adam	0
	全链接层	Rdu		1
生成器	全链接层			1
	卷积层 1	Rdu	Adam	1
	卷积层 2	Rdu		0

3.2. 基于 ARIMA-GRU 模型的人群健康数据预测方法

3.2.1. ARIMA 模型拟合

如图 4 所示，在通过 CNN-CGAN 模型得到填补过的数据集后，本文先通过观察自相关函数(ACF)和偏自相关函数(PACF)图以更深入地了解数据的自相关性和偏自相关性模式。这些图形为本文提供了直观的视觉信息，帮助本文确定 ARIMA 模型的参数(p, d, q)。最后，使用极大似然估计法和确定的参数(p, d, q)对时间序列数据进行 ARIMA 模型拟合并使用这些参数来计算模型残差。这种方法通过最大化样本数据的似然函数，得到模型参数的估计值。通过不断迭代和优化，本文能够找到最合适的参数组合，使得模型能够更好地拟合实际数据。

ARIMA 模型公式：ARIMA (p, d, q)

$$Y_t=\varphi_1Y_{t-1}+\varphi_2Y_{t-2}+\cdots+\varphi_pY_{t-p}+e_t-\theta_1e_{t-1}-\theta_2e_{t-2}-\cdots-\theta_qe_{t-q} \tag{6}$$

其中，Y_t 代表的是每个选区的冠心病患病率数据，e_t 代表的是当前序列的随机扰动值，p 是自回归模型的阶数，d 是使序列平稳的差分次数，q 是滑动平均模型的阶数。

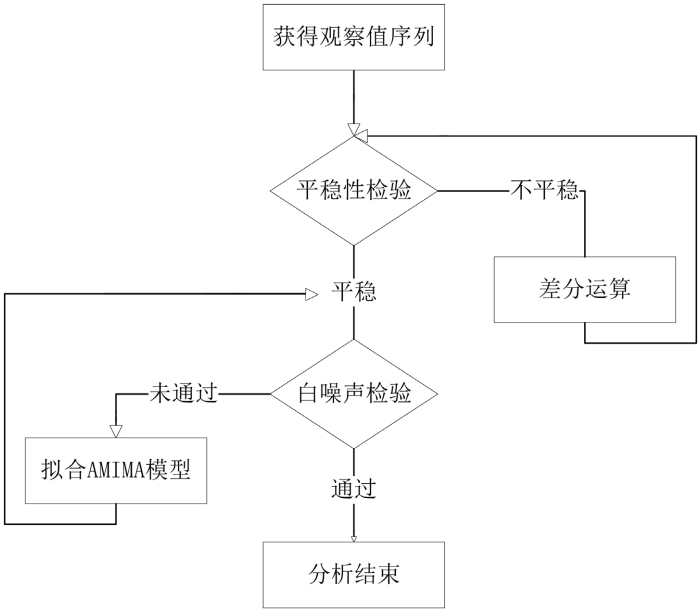

Figure 4. Framework diagram of ARIMA model
图 4. ARIMA 模型的框架图

3.2.2. GRU 模型预测

本文首先将准备好的残差数据划分为训练集、验证集和测试集，以便于模型训练和评估。然后本文构建一个 GRU 模型(如图 5 所示)并使用训练集数据来训练 GRU 模型。在训练过程中，监控验证集上的性能，以避免过拟合。训练完成后，本文使用测试集评估 GRU 模型的预测性能并以均方误差(MSE)、均方根误差(RMSE)、MAPE (平均绝对百分比误差)作为评估模型的准确性的指标。具体参数设置如表 3 所表示。

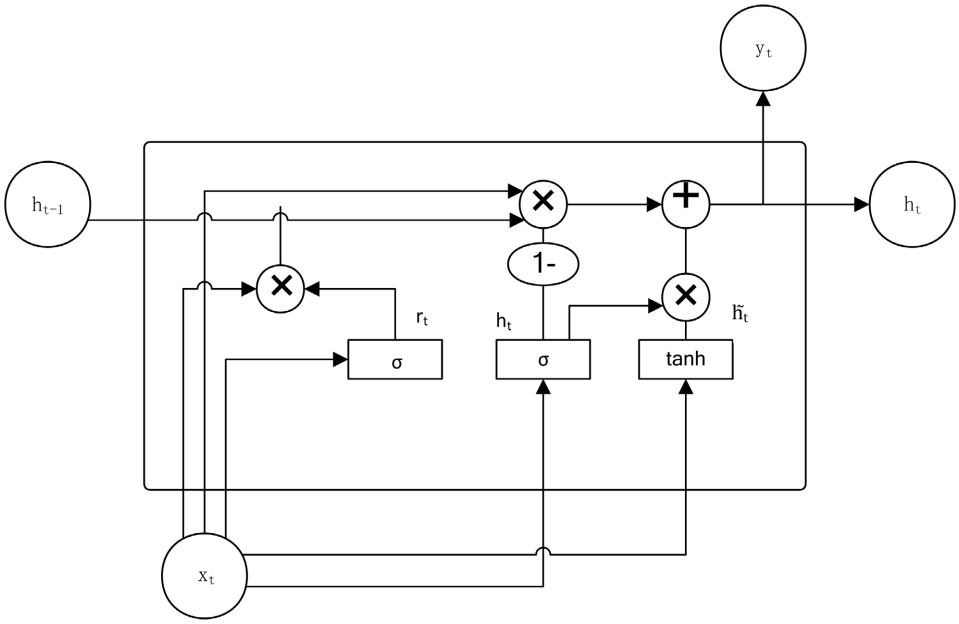

Figure 5. Frame diagram of GRU model
图 5. GRU 模型的框架图

Table 3. GRU model parameter settings
表 3. GRU 模型参数设置

参数名称	CNN-CGAN 模型
Batch rate (批处理大小)	128
Learning_epochs (学习率)	0.0003
Num_epochs (迭代次数)	100
Input_shape (输入大小)	(224, 224, 3)
Num_class (输出大小)	3
学习器	Adam
预处理	归一化
输入层	Input 函数
编码层	ConvID、MaxPoolingID 函数
预测层	人群患病率
池化层	1

公式(7)中 $\tilde{h}_t$ 代表候选隐藏状态包含了前面序列选区的患病率数据，tanh 代表双曲正切函数，同时作为激活函数控制信息的遗忘，$W_{\tilde{h}}$ 为权重矩阵，$\left[r_t * h_{t-1}, x_t\right]$ 是一个向量，它将前一序列的隐藏层状态 h_{t-1} 与当前输入 x_t 进行了元素级的乘法操作。公式(8)定义了遗忘门 $\left(u_t\right)$ 的计算方式。σ 是 sigmoid 激活函数，W_u 是权重矩阵，同样地，$\left(h_{t-1}, x_t\right)$ 是将前一序列的隐藏层状态 h_{t-1} 与当前输入 x_t 进行拼接形成的向量。遗忘门控制着保留或丢弃前一隐藏状态 h_{t-1} 中的信息。公式(9)定义了输入门 $\left(r_t\right)$ 的计算方式，与遗忘门类似，也是通过 sigmoid 激活函数和权重矩阵 W_r 来确定当前应该更新哪些信息。方程(10)定义了最终的隐藏层状态 h_t 的更新。这里，$1-u_t$ 可以看作是保留前一隐藏状态 h_{t-1} 的比例，$u_{t*}\tilde{h}_t$ 是根据候选隐藏状态 $\tilde{h}_t$ 和输入门 r_t 更新的信息。两者相加即得到了新的隐藏层状态 h_t。

$$\tilde{h}_t = \tanh\left(W_{\tilde{h}} \cdot \left[r_t * h_{t-1}, x_t\right]\right) \tag{7}$$

$$u_t = \sigma\left(W_u \cdot \left[h_{t-1}, x_t\right]\right) \tag{8}$$

$$r_t = \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{9}$$

$$h_t = \left(1-u_t\right) * h_{t-1} + u_t * \tilde{h}_t \tag{10}$$

4. 实验结果与分析

4.1. 数据集介绍

本文从英国国家统计局(Office for National Statistics)官方网站提前收集了 2018 年伦敦慢性疾病患病率，分别是选区老年人口占比，选区高血压患病率，选区冠心病患病率构建预测的数据集(80%样本为训练集，10%样本为验证集，10%样本为测试集)，其数据集参数如表 4 所示。

Table 4. Data set parameter settings
表 4. 数据集参数设置

	选区老年人口占比	选区高血压患病率	选区冠心病患病率
数据量	1500	1500	1500

4.2. 评估指标

每个模型对人群患病率预测结果的平均绝对误差(MAE)、均方根误差(RMSE)以及平均绝对百分比误差(MAPE)对比，评价使用新旧数据集时该模型的预测能力，以得出导入外部知识进行预测是否能提高模型的预测能力。MAE 是预测值与实际值之间差异的平均绝对值。相较于其他两种指标，MAE 对异常值不敏感，因为它只考虑了绝对差异，不受方差影响。RMSE 是预测值与实际值之间差异的平方的平均值的平方根。与 MAE 相比，RMSE 对大误差更敏感，因为误差被平方了。这使得 RMSE 更适合于对大误差感兴趣的情况。MAPE 是预测值与实际值之间的绝对百分比差异的平均值。MAPE 以百分比的形式表示误差，因此可以更好地理解误差相对于实际值的大小。

$$\mathrm{MAE}\left(y,\hat{y}\right)=\frac{\left(\sum_{i=1}^{n}\left|y-\hat{y}\right|\right)}{n} \tag{11}$$

$$\mathrm{RMSE}\left(y,\hat{y}\right)=\sqrt{\frac{\left(\sum_{i=1}^{n}\left(y-\hat{y}\right)^{2}\right)}{n}} \tag{12}$$

$$\mathrm{MAPE}\left(y,\hat{y}\right)=\frac{\left(\sum_{i=1}^{n}\left|\frac{\left(y_i-\hat{y}_i\right)}{y_i}\right|\right)}{n}*100\% \tag{13}$$

其中 y 是冠心病患病率第 i 个样本的实际值，$\hat{y}$ 是其对应的预测值，通过比较依据新旧数据集得到的两个结果的准确性判断依据外部知识辅助是否可以提高模型的预测能力。

4.3. 实验结果对比

目前数据填补主流算法使用 KNN 模型与 RNN 模型，为体现本文提出外部知识辅助的 CNN-CGAN 模型进行数据填补的优越性，本文采用相同的数据集和预测模型，比较本文算法和 KNN 模型以及 RNN 模型的填补后预测效果。

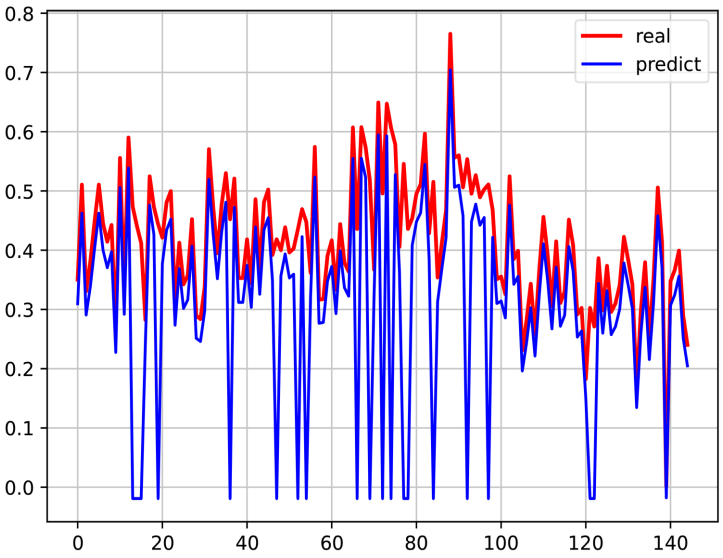

Figure 6. Comparison of predicted and true values without data filling
图 6. 不填补数据时预测值与真实值对比

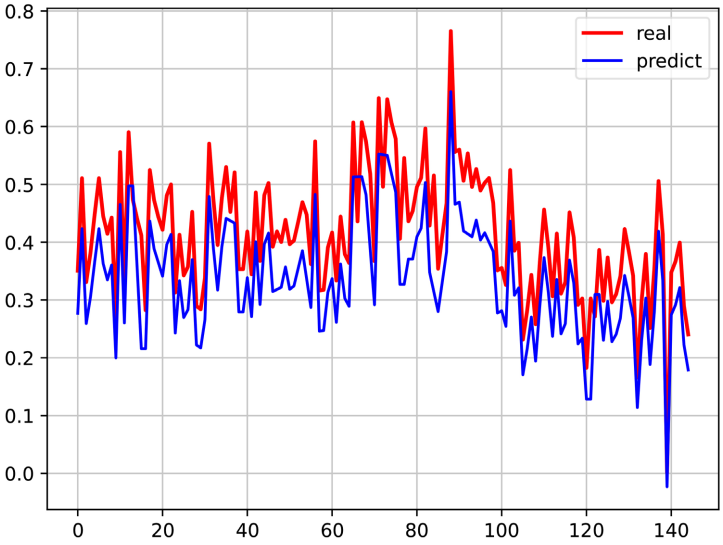

Figure 7. Comparison of predicted and true values under KNN model
图 7. KNN 模型下预测值与真实值对比

图 6 表明了不填补人群健康数据的预测值和实际值的对比效果，可以看出，患病率的预测值不能很好地跟上实际患病率的变化趋势，对于患病率的规律变化学习较差，在人群分布地区较少的数据点未能很好的达到预测效果。图 7 描绘了采用 KNN 模型填补人群健康数据的预测值和实际值的对比效果，预测值大体和实际值保持一致的变化趋势，且波动较不填补数据的模型更小，但在转折点难与原数据一致。图 8 描绘了采用 RNN 模型填补数据预测值和真实值的对比，相较于 KNN 模型和不填补数据，采用 RNN 模型填补预测准确度更高，预测趋势与真实值相对一致，但细节部分仍有不足，尤其在转折处预测值过于平缓。图 9 显示了使用 CNN-CGAN 模型填补数据后的预测值和实际值的对比效果，CNN-CGAN 模型预测健康数据趋势与实际较为一致，且在人群稀少或波动大处优于 KNN 和 RNN 模型，能够有效预测人群健康数据。

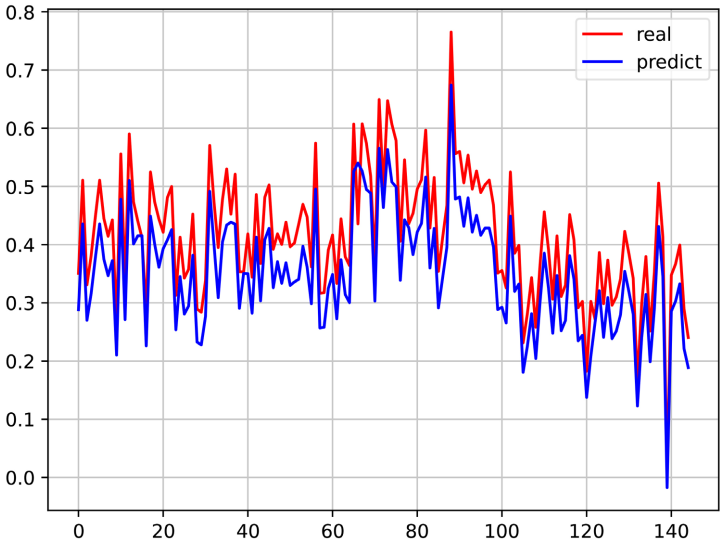

Figure 8. Comparison of predicted and true values under the RNN model
图 8. RNN 模型下预测值与真实值对比

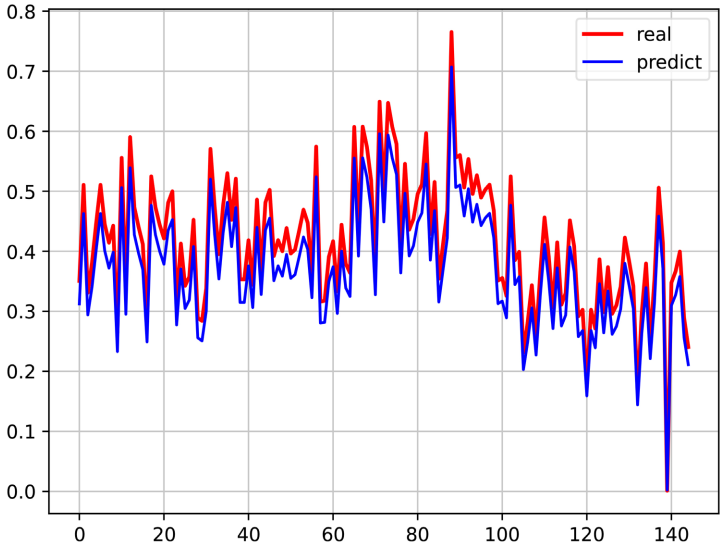

Figure 9. Comparison of predicted and true values under CNN-CGAN model
图 9. CNN-CGAN 模型下预测值与真实值对比

表 5 是每个模型对人群患病率预测结果的平均绝对误差(MAE)、均方根误差(RMSE)以及平均绝对百分比误差(MAPE)对比。由表 5 可知，针对人群健康数据的预测实验中，CNN-CGAN 模型在测试集上经过最后 10 次测试得到的 MAE 值和 RMSE 值，相较于使用 KNN 模型填补数据虽有提升，但幅度并不显著，然而 MAPE 值有极大的减小；对比于 RNN 模型填补数据，三个指标表现出来的结果表明 CNN-CGAN 模型在本数据集中的表现完全优于 RNN 模型。这表明相较于独立的 KNN 模型或 RNN 模型，CNN-CGAN 模型填补的数据集质量更好。因此，CNN-CGAN 模型在填补数据集稀疏部份方面表现出更高的准确性，尤其在捕捉数据特征或趋势方面做得更好。

Table 5. Comparison of MAE, RMSE and MAPE values in models
表 5. 模型中 MAE、RMSE、MAPE 值对比

算法名称	MAE	RMSE	MAPE
不填补数据	0.3662	0.4051	477.976
KNN	0.0110	0.0130	23.434
RNN	0.0179	0.0187	12.739
CNN-CGAN	0.0126	0.0139	6.671

5. 讨论

本文研究数据质量对健康数据预测模型的限制问题，提出了一种基于外部知识辅助的人群健康数据预测方法。本文以与冠心病患病率相关性较强的高血压患病率数据和选区老年人口比例数据作为外部知识辅助填补冠心病患病率数据稀疏部分，构建 CNN 模型对高血压患病率数据和选区老年人口比例数据提取特征矩阵，并和随机噪声、部分完整的冠心病患病率数据作为 CGAN 模型的输入，以生成用来填补原冠心病患病率数据中稀疏部分的人工样本。最后，本文将填补后的完整数据集通过 ARIMA 模型拟合得到模型特征，并输入 GRU 模型进行预测分析。实验结果显示，本文方法在 MAE 和 RMSE 上和 KNN 模

型和 RNN 模型相差不多，但 MPAE 大大降低。为了取得更好的预测效果，本文将在后续的模型应用中进一步对预测模型本身的结构和参数进行优化，以提升其预测性能。

基金项目

国家自然科学基金项目(No.61802257)；上海市大学生创新创业训练计划项目(SH2023201)。

参考文献

[1] 蔺洁, 李霞, 刘佳. 基于指数平滑模型的克拉玛依市流感样病例预测分析[J]. 疾病预防控制通报, 2021, 36(6): 8-11.

[2] 耿利彬, 杨育松, 王娅琼, 等. ARIMA 模型在流感样病例发病预测中的应用[J]. 首都公共卫生, 2021, 15(1): 45-47.

[3] 陈宝, 丘美娇, 林尤斌, 等. SARIMA 模型在海南某医院流感样病例预测中的可行性分析[J]. 南昌大学学报(医学版), 2022, 62(2): 75-78, 99.

[4] 王燕. 应用时间序列分析[M]. 北京: 中国人民大学出版社, 2015.

[5] 杜垚强, 杨叶晓青, 徐怡琳, 等. 股骨骨折手术患者临床输血机器学习预测模型的构建分析[C]//中国输血协会. 中国输血协会第十一届输血大会会议论文汇编(2022.09 大连)——信息化专题. 2022: 2.

[6] 余璟璐, 江丽莉, 裴立红. 空腹血糖正常人群中 2 型糖尿病及糖耐量受损的 ANN 预测模型研究[J]. 中国卫生检验杂志, 2023, 33(16): 1971-1974.

[7] 严虹, 刘国烨, 李砚, 等. 深度学习在检验医学中的研究与应用[J]. 中华检验医学杂志, 2019, 42(12): 1063-1066.

[8] Singh, S., Parmar, K.S., Singh Makkhan, S.J., Kaur, J., Peshoria, S. and Kumar, J. (2020) Study of ARIMA and Least Square Support Vector Machine (LS-SVM) Models for the Prediction of SARS-CoV-2 Confirmed Cases in the Most Affected Countries. *Chaos, Solitons and Fractals*, **139**, Article ID: 110086. https://doi.org/10.1016/j.chaos.2020.110086

[9] 张婷婷. 面向失衡数据集的数据缺失问题研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2017.

[10] 符祥远. 基于深度学习的交通流数据填补及预测[D]: [硕士学位论文]. 西安: 长安大学, 2023.

[11] 史继新, 张文增, 冀国强, 等. ARIMA 模型在流感样病例预测预警中的应用[J]. 首都公共卫生, 2010, 4(1): 12-16.

[12] 戴皓云, 周楠, 任香, 等. 基于 ARIMA 模型各亚型流行性感冒流行特征与趋势预测[J]. 疾病监测, 2022, 37(10): 1338-1345.

[13] 杨真真, 谢艳秋, 靳旭东, 等. 基于 ARIMA 时间序列模型的传染病发展趋势预测——以 COVID-19 为例[J]. 中国科技信息, 2021(3): 70-72.

[14] 李申龙, 王振平, 卢国群, 等. 基于时间序列和机器学习预测尘肺病发展趋势研究[J]. 中国煤炭, 2023, 49(10): 68-73.

[15] 谌典, 周畅, 张奥懿, 等. 基于临床、超声特征及影像组学构建机器学习模型预测慢性肾脏病患者肾功能损伤程度[J]. 中国医学影像技术, 2024, 40(4): 575-579.

[16] 刘洋, 曹赛雅, 冯月娇, 等. 应用机器学习和神经网络模型识别结直肠癌“炎癌转化”过程的关键基因及防治中药预测[J]. 中草药, 2023, 54(19): 6386-6399.

[17] Shahin, A.I., Guo, Y.H., Amin, K.M. and Sharawi, A.A. (2017) White Blood Cells Identification System Based on Convolutional Deep Neural Learning Networks. *Computer Methods and Programs in Biomedicine*, **168**, 69-80. https://doi.org/10.1016/j.cmpb.2017.11.015

[18] 杨美涛, 王彦丁, 李志强, 等. ARIMA-SVM 组合模型在肺结核发病趋势预测中的应用[J]. 现代预防医学, 2023, 50(11): 1921-1926.

[19] 侯文涛, 张飞扬, 张瑞杰, 等. 基于 SARIMA-SVM 组合模型的丙型肝炎发病率预测研究[J]. 数学的实践与认识, 2022, 52(3): 140-146.

[20] 王睿. 胃食管反流病流行病学调查及其缺失数据的处理方法研究[D]: [博士学位论文]. 上海: 第二军医大学, 2009.

[21] 岳根霞, 刘金花, 刘峰. 基于决策树算法的医疗大数据填补及分类仿真[J]. 计算机仿真, 2021, 38(1): 451-454, 459.

[22] 程亮. 基于随机森林的基坑监测数据填补对比研究[J]. 城市地质, 2021, 16(4): 466-473.

[23] 解东方. 心血管病流行病调查中缺失数据填补方法的比较及模拟研究[D]: [博士学位论文]. 北京: 北京协和医学院, 2014.

[24] 刘癸壬. 贝叶斯网络在急性冠脉综合征死亡风险评估中的应用[D]: [硕士学位论文]. 南京: 东南大学, 2023.

[25] 吴世彬. 基于 Stacking 集成学习的医疗数据填补方法研究[D]: [硕士学位论文]. 武汉: 华中农业大学, 2023.