基于多鉴别器生成对抗网络的时间序列生成模型

陆彦辉[1,2]，柳寒[1,2]，李航[2]，朱光旭[2]
（1. 郑州大学电气与信息工程学院，河南 郑州 450001；2. 深圳市大数据研究院，广东 深圳 518115）

摘 要：针对时间序列的隐私性和连续性导致时间序列数据集在收集过程中存在收集代价昂贵和数据缺失等问题，提出了一种基于循环神经网络的多鉴别器生成对抗网络模型，该模型能够利用小规模数据集合成得到与真实数据相似分布的时间序列数据集。多鉴别器包含时域、频域、时频域和自相关 4 种鉴别器，能够充分识别时间序列不同维度下的特征。在实验中，通过损失函数的收敛分析、主成分分析和误差分析，分别从定性和定量的角度对模型进行性能评估。结果表明，所提模型和其他参考模型相比具有更好的性能。

关键词：生成对抗网络；时间序列；傅里叶变换；自相关函数；机器学习
中图分类号：TP18
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022205

Time series generation model based on multi-discriminator generative adversarial network

LU Yanhui[1,2], LIU Han[1,2], LI Hang[2], ZHU Guangxu[2]
1. College of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China
2. Shenzhen Research Institute of Big Data, Shenzhen 518115, China

Abstract: Aiming at the problems of expensive collection cost and missing data due to the privacy and continuity of time series data set, a multi-discriminator generative adversarial network model based on recurrent neural network was proposed, which could synthesize time series dataset that were approximately distributed with real data of a small scale dataset. Multi-discriminator included four discriminators in time domain, frequency domain, time-frequency domain and autocorrelation. Different discriminators could effectively recognize the features of the time series in different domains. In the experiment, the convergence of loss function, principal component analysis and error analysis were performed to evaluate the performance of the model from qualitative and quantitative perspectives. The experimental results show that the proposed model has better performance than other reference models.

Keywords: generative adversarial network, time series, Fourier transform, autocorrelation function, machine learning

0 引言

近年来，随着计算能力的提升和 5G 网络的普及，数据生成规模逐步扩大，在生产生活中的作用也日益显著。越来越多的商业公司和组织机构依赖于大数据分析得到有效的决策[1]。大数据分析中一个重要类别是分析与时间相关的数据，涉及金融、气象、石油和医学等多个领域。例如，通过分析金融时间序列来预测股票价格[2]；通过分析气候时间序列来分析植被的变化[3]；通过分析石油产量时间序列来预测石油的产量[4]；通过分析 COVID-19 随时间变化的确诊人数来预测未来的确诊人数[5]。

收稿日期：2022−07−15；**修回日期**：2022−10−09
通信作者：李航，hangdavidli@sribd.cn
基金项目：国家自然科学基金资助项目（No.62001310）；广东省基础与应用基础研究基金资助项目（No.2022A1515010109）；深圳市科技计划基础研究项目（No.JCYJ20190813170803617）
Foundation Items: The National Natural Science Foundation of China (No.62001310), The Foundation for Basic and Applied Basic Research of Guangdong Province (No.2022A1515010109), The Basic Research Project of Shenzhen Science and Technology Plan (No.JCYJ20190813170803617)

时间序列是按照一定的时间间隔持续记录一段时间的数据，它们通常包含着丰富且复杂的信息，具备较强的研究和商业价值。然而，这些数据在收集过程中存在着各种各样的问题，例如，数据往往包含隐私信息，无法进行公开传播与实验[6]；传感器数据在收集过程中存在数据缺失[7]；数据收集困难导致可用数据集过小，难以满足模型训练需求[8]。一种可行的解决方案是通过机器学习方法生成大量与真实数据相似度较高的数据，从而满足模型训练、验证等应用。

现有基于机器学习的生成模型主要包括变分自动编码器（VAE, variational auto-encoder）[9]和生成对抗网络（GAN, generative adversarial network）[10]。其中，GAN 的研究得到了广泛的关注，已有工作提出了多种 GAN 模型，可用于生成逼真的图像和视频。鉴于 GAN 在图像生成方面的优异性能，开发高质量、多样化和特殊性的时间序列数据的工作得以进一步展开。

本文采取多鉴别器对时间序列的多种特征进行鉴别，提出了多鉴别器生成对抗网络（MDGAN, multi-discriminator generative adversarial network）模型。本文主要研究工作如下。

1) 本文提出了一种新型的 MDGAN 模型，包含时域鉴别器、频域鉴别器、时频域鉴别器和自相关鉴别器，能够对生成数据进行多角度评估，进而提高生成器的合成数据质量，使合成数据更加符合真实时间序列的分布和特征。

2) 在对所提模型进行训练时，本文引入了二分类交叉熵模型，优化了原始的 GAN 损失函数，使其适配多鉴别器网络，从而提升了模型训练效果。

3) 本文采用了不同类型的数据集对模型进行横向和纵向的对照实验，验证了本文所提模型能够有效提升合成时间序列的质量。

1 相关工作

生成对抗网络最早由 Goodfellow 提出，其核心主要体现了零和博弈思想。在生成对抗网络中，同时训练生成器网络和鉴别器网络这 2 个网络。整个网络的损失函数定义为

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{\mathrm{data}}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

其中，p_{data} 表示真实数据 x 的分布，符合随机分布 p_z 的噪声 z 表示生成器的输入，$G(z)$ 表示生成器生成的合成数据，$D(\cdot)$ 表示鉴别器对数据的评价结果，E 表示数学期望。生成器致力于学习真实数据的特征，以此生成符合真实数据分布的合成数据；鉴别器致力于分辨输入是来源于真实数据还是合成数据。在训练鉴别器的过程中，希望真实数据 x 通过鉴别器的结果 $D(x)$ 更接近真实的评价，合成数据 $G(z)$ 通过鉴别器的结果 $D(G(z))$ 更接近虚假的评价。而在训练生成器的过程中，希望合成数据 $G(z)$ 通过鉴别器的结果 $D(G(z))$ 更接近真实的评价。当训练达到纳什平衡时，认为生成器的合成数据的主要特征已经符合真实数据的主要特征。

现有工作以 GAN 为基础进行了不同方面的改进。Radford 等[11]提出的深度卷积生成对抗网络（DCGAN, deep convolutional generative adversarial network）将卷积神经网络应用到 GAN 中，在网络架构上改进了原始 GAN。Arjovsky 等[12]提出的 WGAN（Wasserstein generative adversarial network）采用 Wasserstein 距离指导整个模型的训练，在鉴别器中使用权重剪枝技术。Isola 等[13]提出的基于 GAN 的 Pix2Pix 算法用于图像像素间的转换，利用条件生成对抗网络（CGAN, conditional generative adversarial network）生成图像。Zhu 等[14]提出了循环一致性生成对抗网络（CycleAN, cycle-consistent adversarial network），以 Pix2Pix 为基础，主要应用于非配对的图片生成和转换，可以实现图片的风格转换。Karras 等[15]提出了可以控制样式的 StyleGAN（style-based generator architecture for generative adversarial network），通过修改样式的特定尺度来控制图像的生成。现有工作已经将 GAN 成功应用于图像、视频以及自然语言等方向。

循环神经网络（RNN, recurrent neural network）具有独特的环状结构，很适用于处理连续时间序列[16]。然而它缺乏学习长期依赖关系的能力，而这种关系对于根据过去预测未来是至关重要的。RNN 的变体长短期记忆（LSTM, long short term memory）网络具有长时间记忆信息的能力，进而可以学习序列信息的长期依赖关系[17]。Mogren[18]提出了具有 GAN 的连续循环神经网络（C-RNN-GAN, continuous recurrent neural network with adversarial training）模型，是最早利用 RNN 的 GAN 生成连续序列数据的例子。该模型的生成器是一个 LSTM 网络，鉴别器是一个双向的 LSTM 网络，通过时间反向传播和正则化的小批

量随机梯度下降，训练生成器和鉴别器的网络参数。

Esteban 等[19]提出了循环条件生成对抗网络（RCGAN, recurrent conditional generative adversarial network）模型。它的生成器和鉴别器都采用 RNN，和 C-RNN-GAN 不同的是，RCGAN 的生成器和鉴别器的输入需要加入附加条件来控制结果。此模型的损失函数采用二分类交叉熵（BCE, binary cross entropy），能够描述真实数据与合成数据之间的关系。RCGAN 模型是很多后续工作的模型参照。

Yoon 等[20]提出了一种时间序列生成对抗网络（TimeGAN, time-series generative adversarial network），并利用了传统的无监督 GAN 训练方法和更可控的监督学习方法。具体而言，该网络能够生成具有时间动态特性的时间序列。TimeGAN 由嵌入网络、恢复网络、生成器和鉴别器 4 个网络组件组成。自动编码网络（前 2 个网络）与生成对抗网络（后 2 个网络）联合训练，嵌入网络和恢复网络负责数据到隐式特征的转换，生成对抗网络在此空间内学习数据的潜在有效特征。

TimeGAN 主要用于生成短时间序列，因为长时间序列会大大增加生成建模的维数要求，导致复杂度过高。为了解决这个问题，Ni 等[21]提出一个名为 Signature Wasserstein-1 的度量并将其作为鉴别器的评价结果，同时提出了一种新的生成器，称为条件自回归前馈神经网络，它抓住了时间序列的自回归性质，加快了训练的速度，整个模型被称为 SigWGAN（signature Wasserstein generative adversarial network）。

尽管已有工作能够实现多种类型时间序列的生成，但是上述模型也存在不足。一是原始 GAN 面临梯度消失的问题。在训练初期，生成器的合成数据与真实数据相差很大，鉴别器可以利用高置信度区分二者，但损失函数无法为生成器提供足够大的梯度，最终导致梯度消失。二是时间序列的特征提取和利用的问题。时间序列数据的特征有多方面，涉及周期性、相关性和频域的特征等。单一鉴别器能够完成对时间序列特征的鉴别，但是不具有针对性。

对于上述 2 个代表性问题，本文设计了多鉴别器的模型。多鉴别器针对时间序列的不同特征进行针对性的鉴别，在初期训练中合成数据不会因为某一项特征不明显而直接导致梯度消失，同时也有助于提高生成器合成数据的质量。

2　多鉴别器生成对抗网络模型

本文以 GAN 和 RNN 为基础提出了 MDGAN 的模型。此模型主要由 3 个部分组成，分别是数据处理、生成器和多鉴别器。多鉴别器 GAN 结构如图 1 所示。在整个模型中，生成器输出的合成数据为 $G(Z_N)$，其中 Z_N 为输入的随机噪声。合成数据经过数据处理得到 $T(G(Z_N))$，真实时间序列 X_N 经过数据处理得到 $T(X_N)$。处理后的数据通过多鉴别器进行真/假判定。最后，通过计算鉴别器的损失函数 D loss 和生成器的损失函数 G loss 分别更新鉴别器和生成器的网络参数。

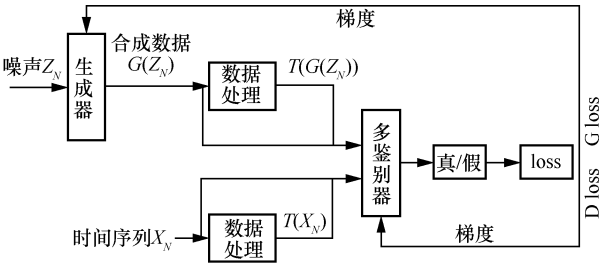

图 1　多鉴别器 GAN 结构

下面，分别介绍模型的组成部分、模型训练中的损失函数和训练方法。

2.1　数据处理

数据处理的目的是得到数据的不同特征。本文以真实时间序列的处理过程为例，介绍数据处理的流程。数据处理流程如图 2 所示。

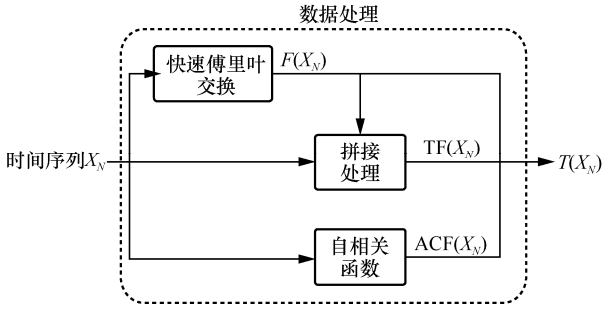

图 2　数据处理流程

真实时间序列 X_N 是一段长度为 N 的序列。序列可以描述为

$$X_N=[x(0),x(1),\cdots,x(n),\cdots,x(N-1)],\quad n=0,1,\cdots,N-1 \tag{2}$$

在数据处理的过程中，时间序列 X_N 通过傅里叶变换得到频域数据 $F(X_N)$；通过对时域和频域数据的处理和拼接得到时频域数据 $\mathrm{TF}(X_N)$；通过自

相关处理得到自相关函数 $\mathrm{ACF}(X_N)$。处理后的数据按顺序组合为 $T(X_N)$，排序方式为

$$T(X_N)=[F(X_N),\mathrm{TF}(X_N),\mathrm{ACF}(X_N)] \tag{3}$$

$T(X_N)$ 是将 3 种数据组合在一起。接下来，对式(3)中的 3 个部分分别进行介绍。

2.1.1 傅里叶变换

离散傅里叶变换（DFT, discrete Fourier transform）是信号分析最基本的方法[22]。该方法将时间序列从时间域变换到频率域，分析时间序列的频域结构与变化规律。本文对长度为 N 的时间序列 X_N 做 M 点的离散傅里叶变换。M 的取值是 2 的整数幂，且大于或等于时间序列的长度 N。X_N 的表达式为

$$X(k)=\mathrm{DFT}[x(n)]=\sum_{n=0}^{M-1}x(n)\mathrm{e}^{i\frac{2\pi}{M}nk},\quad k=0,1,\cdots,M-1 \tag{4}$$

其中，$x(n)$ 是时间序列 X_N 中的第 n 个值，$X(k)$ 是傅里叶变换后的值。在模型中使用的方法是快速傅里叶变化（FFT, fast Fourier transform）。

离散傅立叶变换后的数据是一组复数，其中一半数据和另一半数据是共轭关系。本文只取一半数据 $F(X_N)$。$F(X_N)$ 的表达式为

$$F(X_N)=\left[X(1),X(2),\cdots,X(k),\cdots,X\left(\frac{M}{2}-1\right)\right],\quad 0<k<\frac{M}{2}-1 \tag{5}$$

2.1.2 时域与频域拼接处理

傅里叶变换只反映数据在频域的特征，为了将时域和频域的特征联系在一起，常用短时傅里叶变换方法，其实质是加窗的傅里叶变换。这种方法是一种数据变形处理。但是本文希望从原始数据出发，得到一种同时包含时域数据和频域数据的形式。所以本文采取时域数据和频域数据拼接的方法分析特征。

具体的拼接方法是首先对频域数据取模后得到 $|F(X_N)|$。取模是一种对复数进行计算的方法，假设复数 $z=a+b\mathrm{i}$，复数模值计算为

$$|z|=\sqrt{a^2+b^2} \tag{6}$$

$F(X_N)$ 中的每一个值都是复数，对每一个值取模之后，本文可以得到 $|F(X_N)|$ 的表达式，即

$$|F(X_N)|=\left[|X(1)|,|X(2)|,\cdots,|X(k)|,\cdots,\left|X\left(\frac{M}{2}-1\right)\right|\right],\quad 0<k<\frac{M}{2}-1 \tag{7}$$

然后，将频域数据的模值 $|F(X_N)|$ 和时域数据 X_N 拼接的数据看作一组同时包含时域和频域特征的数据，定义为时频域数据 $\mathrm{TF}(X_N)$。时频域数据 $\mathrm{TF}(X_N)$ 的表达式为

$$\mathrm{TF}(X_N)=[X_N,|F(X_N)|] \tag{8}$$

2.1.3 自相关函数处理

自相关函数（ACF, autocorrelation function）在信号处理中经常用来分析数据并描述数据的相似性[23]。通过使用自相关函数对时间序列进行处理，进一步对数据在时域上的特征进行分析。本文将自相关函数定义为 $\mathrm{ACF}(X_N)$。离散序列的自相关函数的表达式为

$$\mathrm{ACF}(X_N)=\frac{1}{N}\sum_{n=1}^{N-m}x(n)x(n+m),\quad m=0,1,\cdots,l,l\approx\frac{N}{10} \tag{9}$$

其中，$x(n)$ 表示时间序列 X_N 中的第 n 个值，m 表示时间间隔。

2.2 生成器和鉴别器的网络结构

生成器和鉴别器的网络由 LSTM 网络构成。LSTM 网络是 RNN 的变体，一般用于与时间序列相关的任务，它由一系列结构相同的神经元构成，该神经元在每个时间步中重复使用。LSTM 的神经元内部有一个记忆状态，在处理序列数据时，输入不仅有序列数据，还有上一个时刻的记忆状态，并向下一个时刻输出当前的记忆状态。因此 LSTM 网络是处理时间序列常用的网络。

2.2.1 生成器网络

生成器的网络结构主要由 LSTM 层和全连接层构成。生成器在每个时间步的输入获取不同的随机噪声向量。随机噪声向量由标准正态分布采样得到，并通过 LSTM 网络进行计算。LSTM 网络的激活函数是 tanh 函数。全连接层将 LSTM 层的输出转换为指定的长度。生成器的网络结构如图 3 所示。

LSTM 网络的层数为 2，隐藏层的神经单元个数为 64。全连接层采用 Linear 函数进行转换，并将每个时间步的全连接层的输出组合后得到合成数据。

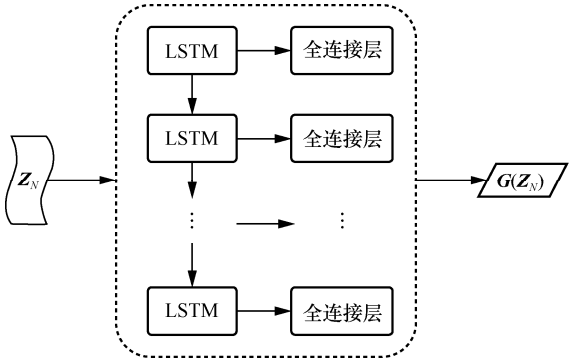

图 3　生成器的网络结构

2.2.2　鉴别器网络

鉴别器是对合成时间序列和真实时间序列的每个时间步的输出进行鉴别，最后取均值得到真/假的评价。鉴别器的网络结构如图 4 所示。

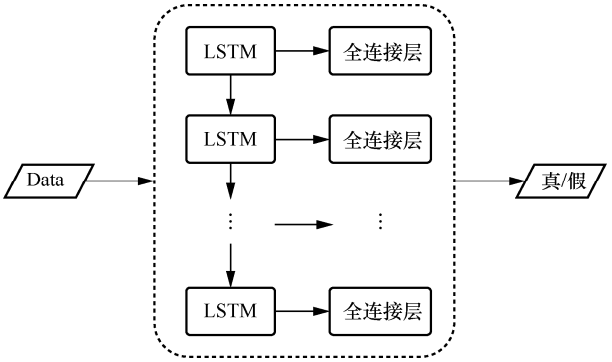

图 4　鉴别器的网络结构

Data 表示输入鉴别器网络的数据，是真实数据或合成数据以及它们的变体。鉴别器的网络结构和生成器的网络结构类似。鉴别器的全连接层使用 Sigmoid 函数，将最后的输出转化为[0,1]区间的值。输出代表鉴别器对输入的评价。本文提出的模型包含多个鉴别器，不同的数据需要通过不同的鉴别器。

合成数据和真实数据的处理过程相同，本文以真实数据的鉴别过程为例说明多鉴别器如何对数据进行鉴别。多鉴别器的处理流程如图 5 所示。

每个鉴别器网络的输出 y 的取值范围为[0,1]，将 4 个鉴别器的输出数值进行平均，定义最终结果大于或等于 0.5 的是真实数据（评价为真），小于 0.5 的是合成数据（评价为假）。因此，输出结果可表示为

$$y_D(X_N)=\frac{1}{4}(y_{D_t}(X_N)+y_{D_F}(X_N)+y_{D_{TF}}(X_N)+y_{D_{ACF}}(X_N)) \tag{10}$$

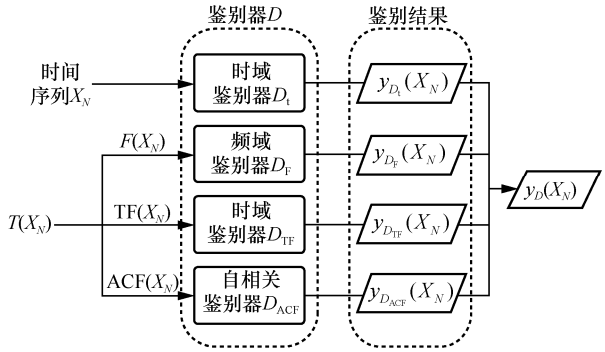

图 5　多鉴别器的处理流程

经过数据处理的数据 $T(X_N)$ 在通过频域鉴别器、时频域鉴别器和自相关鉴别器时分别提取出与之相对应的数据。将不同鉴别器的评价结果进行平均得到最终结果。

2.3　模型训练

MDGAN 模型的训练分 2 个部分介绍，第一部分介绍模型的损失函数，第二部分介绍模型的训练过程。

2.3.1　损失函数

MDGAN 模型的训练包括鉴别器和生成器 2 个部分的训练。在训练中本文使用二分类交叉熵计算损失函数。BCE 的计算式为

$$\mathrm{BCE}(\tilde{y},y)=-y\log(\tilde{y})-(1-y)\log(1-\tilde{y}) \tag{11}$$

其中，$\tilde{y}$ 表示网络预测结果，取值范围为[0,1]；y 表示标签，取值为 0 或 1。

鉴别器的目的是分辨出真实数据和合成数据。在训练中本文使用二分类交叉熵对鉴别器的预测和数据的标签进行计算。真实数据的标签为 1，合成数据的标签为 0。

越是优秀的鉴别器对真实时间序列的鉴别结果越接近 1，对合成时间序列的鉴别结果越接近 0。因此在鉴别器训练时，本文最小化数据通过鉴别器的结果与对应标签的二分类交叉熵。鉴别器的损失函数为

$$L_D=\min(L_{D_t}+L_{D_F}+L_{D_{TF}}+L_{D_{ACF}}) \tag{12}$$

因为模型有多个鉴别器，需要分别计算结果。将计算结果代入式(12)中，然后利用式(12)对 4 种鉴别器的网络参数进行更新。4 种鉴别器的计算结果分别为

$$L_{D_t}=\mathrm{BCE}(y_{D_t}(X_N),1)+\mathrm{BCE}(y_{D_t}(G(Z_N)),0) \tag{13}$$

$$L_{D_F}=\mathrm{BCE}(y_{D_F}(X_N),1)+\mathrm{BCE}(y_{D_F}(G(Z_N)),0) \tag{14}$$

$$L_{D_{TF}}=\mathrm{BCE}(y_{D_{TF}}(X_N),1)+\mathrm{BCE}(y_{D_{TF}}(G(Z_N)),0) \tag{15}$$

$$L_{D_{ACF}}=\mathrm{BCE}(y_{D_{ACF}}(X_N),1)+\mathrm{BCE}(y_{D_{ACF}}(G(Z_N)),0) \tag{16}$$

生成器的目的是随机噪声通过生成器生成与真实数据类似的合成数据。因此生成器生成的合成数据在通过鉴别器时，希望得到的评价是真实的。越是优秀的生成器生成的合成数据通过鉴别器的预测值越接近 1。因此在生成器训练时，本文最小化合成数据通过鉴别器的结果与真实标签的二分类交叉熵。生成器的损失函数为

$$L_G = \min(\mathrm{BCE}(y_{D_\mathrm{t}}(G(Z_N)),1) + \mathrm{BCE}(y_{D_\mathrm{F}}(G(Z_N)),1) + \mathrm{BCE}(y_{D_\mathrm{TF}}(G(Z_N)),1) + \mathrm{BCE}(y_{D_\mathrm{ACF}}(G(Z_N)),1)) \tag{17}$$

式(12)~式(17)中，D_t 代表时域鉴别器，D_F 代表频域鉴别器，D_TF 代表时频域鉴别器，D_ACF 代表自相关鉴别器，G 代表生成器，y_D 代表鉴别器结果，X_N 代表真实时间序列，$G(Z_N)$ 代表合成数据（1 代表真实，0 代表虚假）。

2.3.2 训练过程

在训练过程中，本文需要先对数据集进行预处理再进行训练。

数据集的预处理是先取出所有数据并进行归一化计算，然后将数据分为多个固定长度的序列进行随机组合。例如，把 10 000 个数据按 20 的固定大小分为 500 组，然后将这 500 组数据进行随机组合，目的是混合数据并使其类似于独立同分布。将预处理之后的真实时间序列分布定义为 p_r，随机噪声数据的分布 p_z 是正态分布。

在鉴别器和生成器的训练过程中，先对鉴别器进行训练，更新鉴别器参数，同时固定生成器的参数；然后对生成器进行训练，更新生成器参数，同时固定鉴别器的参数。重复上述过程。训练中对参数更新的方法采用 Adam 优化算法[24]。多鉴别器生成对抗网络生成样本算法如算法 1 所示。

算法 1 多鉴别器生成对抗网络生成样本算法

输入 批量值 m，随机噪声 z，真实样本 x，学习率 γ，鉴别器更新次数 n_d，Adam 超参 β

输出 生成器 G，鉴别器 D

初始化 生成器参数 θ_g，鉴别器参数 θ_d

1) while θ_g has not converged do

2) for $t = 0,1,\cdots,n_d$ do

3) 获取真实数据 $(x^{(1)},\cdots,x^{(m)}) \sim p_r$

4) 获取噪声数据 $(z^{(1)},\cdots,z^{(m)}) \sim p_z$

5) $\theta_d \leftarrow \mathrm{Adam}\left(\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}L_D^{(i)},\theta_d,\gamma,\beta\right)$

6) end for

7) 获取噪声数据 $\left(z^{(1)},\cdots,z^{(m)}\right) \sim p_z$

8) $\theta_g \leftarrow \mathrm{Adam}\left(\nabla_g\frac{1}{m}\sum_{i=1}^{m}L_G^{(i)},\theta_g,\gamma,\beta\right)$

9) end while

10) return G，D

3 实验结果分析

本节介绍实验使用的数据集和评价指标，通过评价指标对实验结果进行分析。在实验中，为了更好地评估模型的性能，本文进行了横向和纵向对比。纵向对比中使用 MDGAN 与频域鉴别器 GAN、自相关鉴别器 GAN、时频域鉴别器 GAN 进行比较。横向比较中使用 3 种具有代表性的时间序列生成模型与 MDGAN 进行比较，分别是 RCGAN[19]、TimeGAN[20]和 SigCWGAN[21]。

3.1 数据集

本文实验使用的数据集是地磁数据集和牛津大学金融学院股票数据集中的标准普尔 500 指数数据集。

地磁数据集共包含 11 500 条数据。该数据是由手机自带的地磁传感器收集的一段 5 min 内随手机姿态变化的地磁数据。地磁数据集经常用来分析和预测实验者使用时手机的不同姿态。

标准普尔 500 指数数据集是牛津大学金融学院收集的股票数据，包括 2000—2021 年的标准普尔 500 指数数据集，共有 5 515 条数据。每条数据包括每天的开盘价格、收盘价格和价格波动率。股票数据集经常用来分析和预测股票的趋势。

3.2 性能评估

实验中采取 3 种常用的评估方法，分别是 loss 函数收敛性、主成分分析法（PCA, principal component analysis）和误差分析，分别从定性和定量的角度说明 MDGAN 的性能。

1) loss 函数收敛性。loss 函数的收敛性主要用于评价模型的训练速度。

2) 主成分分析法。主成分分析法用于评价合成数据的分布情况，是最常用的线性降维方法。它的目标是通过某种线性投影将高维的数据映射到低维的空间中，并期望在所投影的维度上数据的信息量最大，实现使用较少的数据维度保留较多的原数据点特性。

3) 误差分析。误差分析评价合成数据的准确性。本文对合成时间序列和真实时间序列进行误差分析，并使用均方误差（MSE, mean square error）、均方根误差（RMSE, root mean squared error）、平均绝对误差（MAE, mean absolute error）和平均绝对误差百分比（MAPE,mean absolute percentage error）这 4 种误差评价指标。

3.3　纵向对比结果

在纵向对比中，本文只使用地磁数据集对模型进行比较。纵向比较的模型有 MDGAN、频域鉴别器 GAN、时频域鉴别器 GAN 和自相关鉴别器 GAN。MDGAN 中包含所有数据处理过程和对应的鉴别器，其他模型只包含一种数据处理过程和对应的鉴别器。纵向对比是为了说明多鉴别器 GAN 的合成数据比只包含一种鉴别器的 GAN 模型的合成数据更加接近真实数据。

因为数据处理方式不同，4 种模型在 loss 函数收敛性和主成分分析上的对比意义不是很重要，所以在纵向对比中本文只使用误差分析对模型合成数据的准确性进行分析。误差对比如表 1 所示。

表 1　模型误差对比

训练模型	MSE	RMSE	MAE	MAPE
MDGAN	1.175 438	1.082 374	0.863 169	7.125 314
频域鉴别器 GAN	1.330 979	1.180 982	1.147 531	8.293 083
时频域鉴别器 GAN	1.293 977	1.271 668	1.136 859	8.570 859
自相关鉴别器 GAN	1.411 853	1.310 379	1.362 662	9.135 859

从表 1 可以看出，时频域鉴别器 GAN 的误差大多略优于频域鉴别器 GAN 和自相关鉴别器 GAN 的误差。但是 MDGAN 模型的误差明显优于另外 3 种模型的误差。所以本文 MDGAN 模型生成的合成数据更加准确。

3.4　横向对比结果

3.4.1　loss 函数收敛性分析

为了对比模型的 loss 函数收敛性，本文使用地磁数据集对 MDGAN、SigCWGAN、TimeGAN 和 RCGAN 这 4 种模型进行训练，损失函数的变化如图 6 所示。其中，Sig loss 表示 SigCWGAN 模型的损失函数。

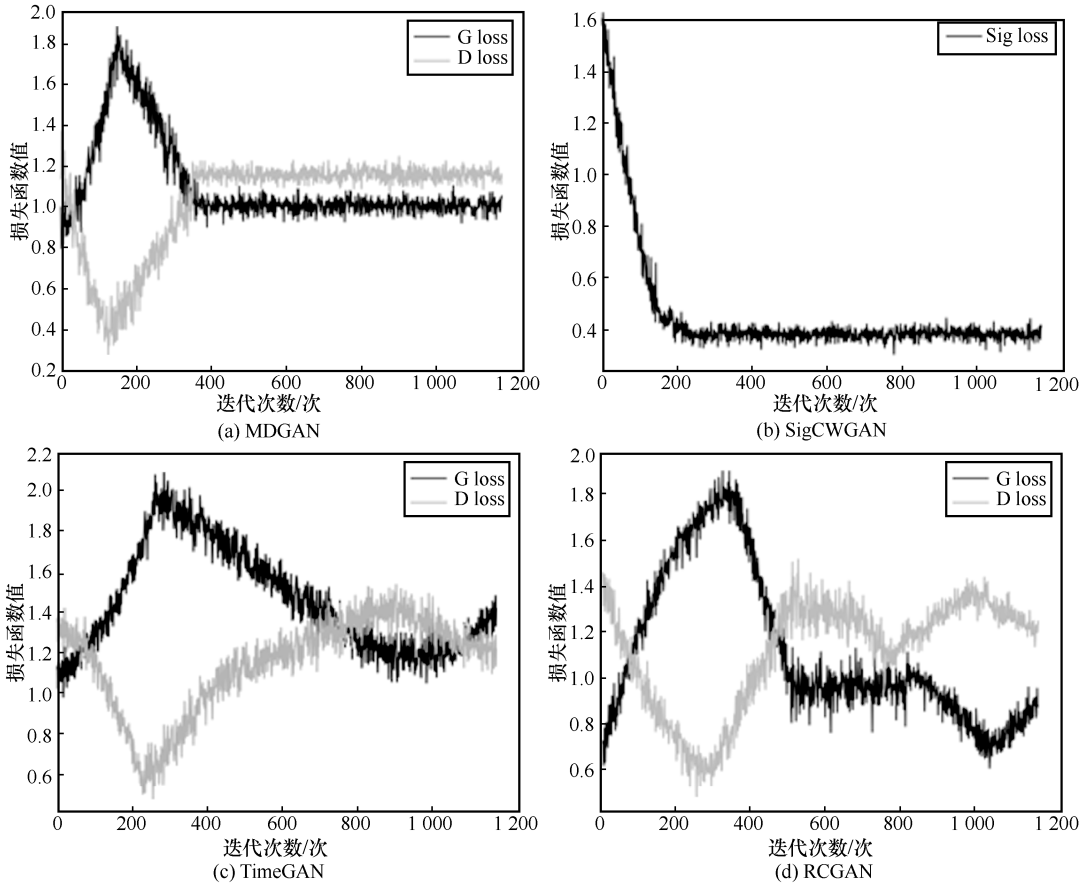

图 6　训练过程中损失函数的变化

由图 6 可以看出，TimeGAN 和 RCGAN 模型的 loss 函数在 1 000 次左右还没有趋于稳定，但是 SigCWGAN 和 MDGAN 模型的 loss 函数在 400 次左右已经趋于稳定。这是因为 TimeGAN 和 RCGAN 采用单一鉴别器，在训练过程中这 2 种模型会在生成器和鉴别器之间的博弈花费更多的时间，不如多鉴别器 GAN 的训练效率高。MDGAN 拥有多个鉴别器，在与生成器的博弈过程中会更加准确地对序列进行评价，这样有利于生成器快速地获得数据特征。而 SigCWGAN 将生成器和鉴别器的损失函数合为一个损失函数，因此会提高训练的速度。综上，本文所使用的 MDGAN 在模型训练的收敛速度上要优于 TimeGAN 和 RCGAN，与 SigCWGAN 不相上下。

3.4.2 主成分分析

为了直观地观察数据的分布，本文采用了主成分分析法将原始数据和合成数据的特征降维到二维平面，来观察数据之间的差异。

本文使用 2 个数据集进行实验，对 4 种模型进行评价。对比结果分别如图 7 和图 8 所示。合成数据覆盖部分越大，说明模型越优秀。对比 2 个数据集在 4 组模型中的实验可以看出，MDGAN 模型在 2 个数据集训练得到的合成数据分布均优于 TimeGAN、SigCWGAN 和 RCGAN 的合成数据分布。因为 MDGAN 模型采用多鉴别器对合成数据的多个特征进行鉴别，所以合成数据的分布更加接近真实数据的分布。

3.4.3 误差分析

从图 7 和图 8 中能直观看到合成数据的分布是接近真实数据数据分布的，但是不能客观地评价合成数据的好坏，因此本文对 2 个数据集的合成数据进行误差分析，分别如表 2 和表 3 所示。其中，股票数据集在预处理阶段已进行归一化处理。

表 2 地磁数据集不同模型误差对比

训练模型	MSE	RMSE	MAE	MAPE
MDGAN	1.175 438	1.082 374	0.844 753	7.047 909
TimeGAN	1.182 172	1.106 035	0.863 169	7.125 314
SigCWGAN	1.241 732	1.111 821	0.907 567	7.405 342
RCGAN	1.388 121	1.234 397	1.021 464	8.490 861

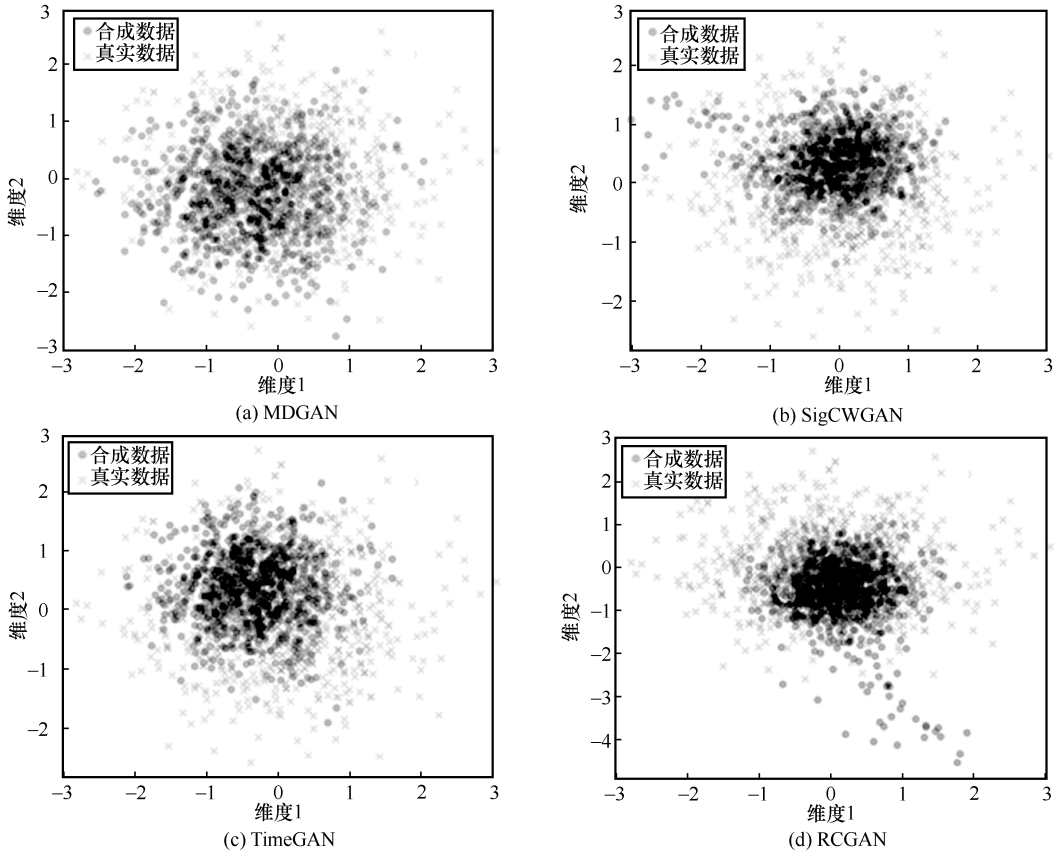

图 7 地磁数据集 PCA 可视化结果

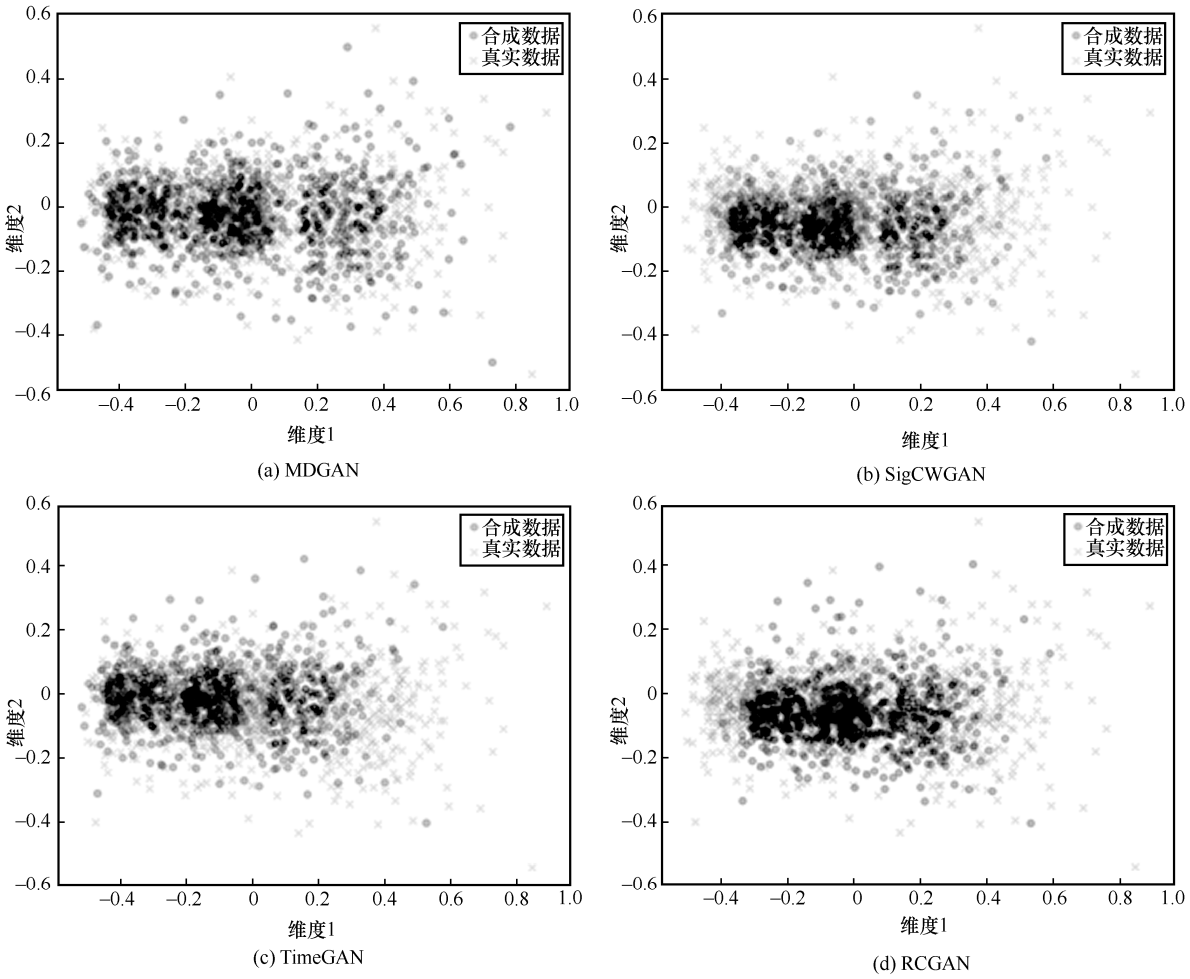

图 8　股票数据集 PCA 可视化结果

从表 2 和表 3 可以看出，MDGAN 的误差略低于 TimeGAN，但是明显低于 SigCWGAN 和 RCGAN。这说明本文所提模型的准确性要高于其他 3 种模型。

表 3　　股票数据集不同模型误差对比

训练模型	MSE	RMSE	MAE	MAPE
MDGAN	0.002 614	0.035 102	0.028 508	11.761 79
TimeGAN	0.002 981	0.051 701	0.041 101	20.810 77
SigCWGAN	0.003 176	0.050 541	0.041 165	20.214 82
RCGAN	0.004 475	0.066 354	0.054 091	28.773 71

3.4.4　总体分析

在 loss 函数收敛性方面，MDGAN 与 SigCWGAN 不相上下，明显高于 TimeGAN 和 RCGAN。在主成分分析中，MDGAN 模型合成数据的分布最接近真实数据的分布。在误差分析中，MDGAN 的误差略低于 TimeGAN，但是明显低于 SigCWGAN 和 RCGAN。

从模型的综合性能比较，本文所提 MDGAN 要略优于 SigCWGAN 和 TimeGAN，明显高于 RCGAN。

4　结束语

本文设计了基于生成对抗网络的多鉴别器时间序列生成模型，该模型采用 4 种不同的鉴别器对合成数据进行鉴别，进而更好地识别时间序列的数据特征，使生成器能够快速合成高质量的数据。实验表明，对于地磁和股票这 2 种不同类型的数据集，所提模型均能够合成出与真实数据近似度较高的数据，在模型收敛性、合成数据分布以及合成数据误差 3 个方面都保持了良好的性能。

本文所设计的 MDGAN 模型能够为一些需要大量时间序列数据集的用户提供一个获取数据的有效手段。尽管本文所提模型只通过 2 种数据集进行了实验验证，但该模型的设计思路是可以借鉴并拓展的。在

面对更加广泛的时间数据集时，可以采取针对性的特征鉴别，适当调整鉴别器的结构，使其达到复杂度和精度的最优折中。未来可进一步对特征提取的环节进行研究，使生成器输出的合成数据具有更强的可控性。

参考文献：

[1] NAEEM M, JAMAL T, DIAZ-MARTINEZ J, et al. Trends and future perspective challenges in big data[C]//Advances in Intelligent Data Analysis and Applications. Berlin: Springer, 2022: 309-325.

[2] HE H T, GAO S C, JIN T, et al. A seasonal-trend decomposition-based dendritic neuron model for financial time series prediction[J]. Applied Soft Computing, 2021, 108: 107488.

[3] ZHANG M, LIN H, LONG X R, et al. Analyzing the spatiotemporal pattern and driving factors of wetland vegetation changes using 2000—2019 time-series Landsat data[J]. Science of the Total Environment, 2021, 780: 146615.

[4] AL-SHABANDAR R, JADDOA A, LIATSIS P, et al. A deep gated recurrent neural network for petroleum production forecasting[J]. Machine Learning With Applications, 2021, 3: 100013.

[5] BALLI S. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods[J]. Chaos, Solitons & Fractals, 2021, 142: 110512.

[6] THAPA C, CAMTEPE S. Precision health data: requirements, challenges and existing techniques for data security and privacy[J]. Computers in Biology and Medicine, 2021, 129: 104130.

[7] HUANG T G, CHAKRABORTY P, SHARMA A. Deep convolutional generative adversarial networks for traffic data imputation encoding time series as images[J]. International Journal of Transportation Science and Technology, 2021: doi.org/10.1016/j.ijtst.2021.10.007.

[8] IWANA B K, UCHIDA S. An empirical survey of data augmentation for time series classification with neural networks[J]. PLoSOne, 2021, 16(7): e0254841.

[9] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. arXiv Preprint, arXiv:1312.6114, 2013.

[10] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[11] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint, arXiv: 1511.06434, 2015.

[12] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning. [S.l.]: JMLR, 2017: 214-223.

[13] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5967-5976.

[14] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2242-2251.

[15] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4396-4405.

[16] MEDSKER L R, JAIN L C. Recurrent neural networks[J]. Design and Applications, 2001, 5: 64-67.

[17] GRAVES A. Long short-term memory[J]. Supervised Sequence Labelling with Recurrent Neural Networks, 2012, 385: 37-45.

[18] MOGREN O. C-RNN-GAN: continuous recurrent neural networks with adversarial training[J]. arXiv Preprint, arXiv: 1611.09904, 2016.

[19] ESTEBAN C, HYLAND S L, RÄTSCH G. Real-valued (medical) time series generation with recurrent conditional GANs[J]. arXiv Preprint, arXiv: 1706.02633, 2017.

[20] YOON J, JARRETT D, VAN DER SCHAAR M. Time-series generative adversarial networks[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). [S.l.:s.n.], 2019: 1-11.

[21] NI H, SZPRUCH L, WIESE M, et al. Conditional Sig-Wasserstein GANs for time series generation[J]. SSRN Electronic Journal, 2020: doi.org/10.2139/ssrn.3623086.

[22] OPPENHEIM A V, WILLSKY A S, NAWAB S H, et al. Signals & systems[M]. New York: Pearson Educación, 1997.

[23] BAROT T, BURGSTEINER H, KOLLERITSCH W. Comparison of discrete autocorrelation functions with regards to statistical significance[C]//Applied Informatics and Cybernetics in Intelligent Systems. Berlin: Springer, 2020: 257-266.

[24] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv Preprint, arXiv: 1412.6980, 2014.

[作者简介]

陆彦辉（1972− ），女，河南许昌人，郑州大学教授，主要研究方向为宽带无线通信理论与系统、无线资源管理和机器学习等。

柳寒（1997− ），男，河南邓州人，郑州大学硕士生，主要研究方向为人工智能与大数据处理、大数据分析与数据挖掘。

李航（1985− ），男，河北承德人，深圳市大数据研究院副研究员，主要研究方向为无线通信与网络、机器学习等。

朱光旭（1989− ），男，广东广州人，深圳市大数据研究院副研究员，主要研究方向为边缘智能、联邦学习、通信感知一体化等。