Advances in Applied Mathematics 应用数学进展, 2024, 13(4), 1800-1809
Published Online April 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.134169

Hans汉斯

基于图协同过滤的单细胞RNA测序数据填补

李雪枫

中国地质大学(武汉)数学与物理学院，湖北 武汉

收稿日期：2024年3月28日；录用日期：2024年4月23日；发布日期：2024年4月30日

摘　要

单细胞RNA测序(Single-cell RNA Sequencing, scRNA-seq)技术能以单细胞的分辨率分析转录组数据，在生物学研究中展现出广泛的应用前景。然而技术问题会导致scRNA-seq数据存在部分基因表达缺失的情况，称之为零膨胀事件。这种情况严重阻碍了下游分析，故需要对scRNA-seq数据进行填补。本文提出了一种基于图协同过滤的单细胞RNA测序数据填补算法，为scRNA-seq分析提供了一个深度学习框架。它通过结构邻居对比的图协同过滤方法提取细胞特征表示和基因特征表示，并将两者的内积应用于零膨胀负二项分布自编码器来填补scRNA-seq数据。仿真实验结果验证了该算法在仿真数据集上的填补能力，且通过下游聚类分析实验表明该算法在公共真实数据集上细胞聚类的性能。

关键词

单细胞RNA测序，填补，图协同过滤，零膨胀负二项分布

Imputation of scRNA-seq Data Based on Graph Collaborative Filtering

Xuefeng Li

School of Mathematics and Physics, China University of Geosciences (Wuhan), Wuhan Hubei

Received: Mar. 28th, 2024; accepted: Apr. 23rd, 2024; published: Apr. 30th, 2024

Abstract

Single-cell RNA sequencing (scRNA-seq) technology can analyze transcriptome data at the single-cell level and is widely used in biology. However, technical issues can lead to missing gene expression in scRNA-seq data, which is called zero-inflation event. This situation seriously hinders downstream analysis, so it is necessary to impute the scRNA-seq data. This article proposes an imputation algorithm of scRNA-seq data based on graph collaborative filtering, providing a deep learning frame-

文章引用: 李雪枫. 基于图协同过滤的单细胞 RNA 测序数据填补[J]. 应用数学进展, 2024, 13(4): 1800-1809.
DOI: 10.12677/aam.2024.134169

work for scRNA-seq analysis. It extracts cell feature representations and gene feature representations through the graph collaborative filtering method of comparing structural neighbors, and applies the inner product of the two to the zero-inflated negative binomial distribution autoencoder to impute scRNA-seq data. The simulation experiment results have verified the imputation ability of the algorithm on the simulation dataset, and downstream clustering analysis experiments have shown the performance of the algorithm on cell clustering on public real datasets.

Keywords

Single-Cell RNA Sequencing, Imputation, Graph Collaborative Filtering, Zero-Inflated Negative Binomial Distribution

1. 引言

单细胞 RNA 测序(Single-cell RNA Sequencing, scRNA-seq) [1]技术已成为生物信息学的研究热点。scRNA-seq 技术在单细胞水平上分析转录组数据，且应用广泛[2] [3] [4]。

尽管 scRNA-seq 技术在单细胞分析中具有潜力，但其应用仍受到技术噪声和实验条件的限制。例如，由于 RNA 输入不足或细胞测序深度不足等技术或实验条件限制，可能会出现部分基因表达数据缺失的情况，即所谓的零膨胀事件或 dropout 事件。这些缺失值可能会导致重要生物学信息的丢失，并对 scRNA-seq 数据的下游分析造成阻碍。采取措施来估算或推断这些缺失值是处理 scRNA-seq 数据集的一个不可或缺的步骤，也就是 scRNA-seq 数据填补。

针对 scRNA-seq 数据的缺失值填补问题，目前研究人员已经开发了许多填补方法。例如，精确单细胞填补(Single-cell Impute, scImpute) [5]通过 Gamma-正态分布混合模型估计哪些值受到 dropout 的影响，最后通过借用其他相似细胞中相同基因的信息来估算。基于细胞 Markov 亲和图的填补(Markov Affinity-based Graph Imputation of Cells, MAGIC) [6]基于热扩散的思想并对相似细胞中的信息加权来进行估算 dropout。而自适应阈值低秩近似填补(Adaptively Thresholded Low-rank Approximation, ALRA) [7]通过观察到的基因表达矩阵进行低秩矩阵补全，再基于奇异值分解求解进而填补。除了基于统计的方法，目前针对 scRNA-seq 数据的高维度、高稀疏性，基于深度学习的方法具备有效性和高效性。深度计数自编码器去噪(Deep Count Autoencoder Network, DCA) [8]将基因表达分布建模为负二项(Negative Binomial, NB)分布或零膨胀负二项(Zero-inflated Negative Binomial, ZINB)分布，通过自编码器学习到的分布参数进而预测去噪后的基因表达矩阵。而单细胞变分推断(Single-cell Variational Inference, scVI) [9]通过变分自编码器来指定 ZINB 分布进行填补。单细胞图神经网络填补(Single-cell Graph Neural Network, scGNN) [10]通过特征自编码器学习并构建细胞图，利用图自编码器聚合细胞间关系，最后基于特征自编码器重构表达谱。

针对 scRNA-seq 数据填补问题，本文提出一种基于图协同过滤的单细胞 RNA 测序数据填补(Imputation of scRNA-seq Data based on Graph Collaborative Filtering, scGCF)算法。通过图协同过滤分别获得细胞特征表示和基因特征表示，再将细胞特征表示和基因特征表示交互，进而输入基于 ZINB 分布的自编码器重建 scRNA-seq 数据的表达谱。

2. 模型介绍

scGCF模型主要由图协同过滤框架、ZINB分布自编码器这两个部分组成，其模型示意图如图1所示。给定基因表达矩阵 $\boldsymbol{X}=\left(x_{ij}\right)_{m\times n}\in\mathbb{R}^{m\times n}$，细胞的集合 $\mathcal{C}=\{c\}$ 且 $|\mathcal{C}|=m$，基因的集合 $\mathcal{G}=\{g\}$ 且 $|\mathcal{G}|=n$，细胞基因交互矩阵是 $\boldsymbol{A}=\left(a_{ij}\right)_{m\times n}\in\mathbb{R}^{m\times n}$，且 $a_{ij}=\begin{cases}1, \text{ if } x_{ij}>0\\ 0, \text{ if } x_{ij}=0\end{cases}$，构图如下：

$$\mathcal{T}=\{\mathcal{V},\mathcal{E}\} \tag{1}$$

其中 $\mathcal{V}=\{\mathcal{C}\cup\mathcal{G}\}$ 表示节点集，$\mathcal{E}=\{(c,g)\,|\,c\in\mathcal{C},g\in\mathcal{G},\boldsymbol{A}_{cg}=1\}$ 表示边集。

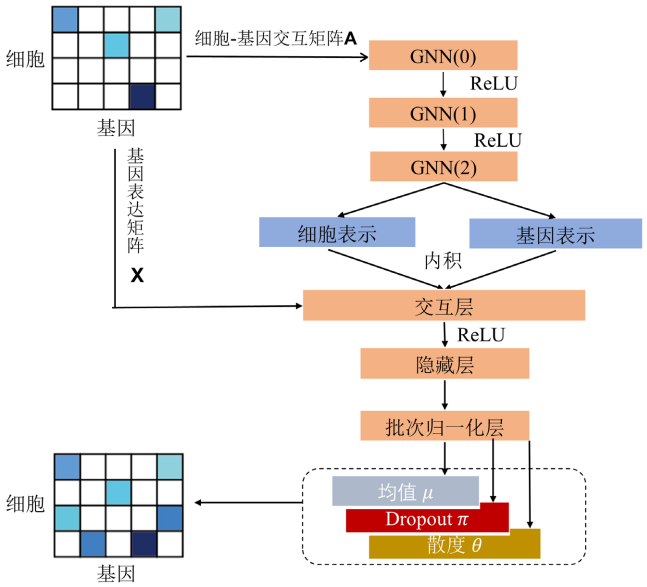

Figure 1. Schematic diagram of scGCF's model structure
图 1. scGCF 模型结构示意图

2.1. 数据预处理

针对 scRNA-seq 数据的原始计数矩阵，本文采取了预处理操作以避免低质量数据对后续分析的影响。本文移除了在极少数(少于 3 个)细胞中表达的基因以及在少数(少于 200 个)基因中表达的细胞，得到计数矩阵 $\boldsymbol{X}=\left(x_{ij}\right)_{m\times n}\in\mathbb{R}^{m\times n}$，其中 m,n 分别为过滤后的细胞和基因的数量。

对细胞 i，大小因子为：

$$s_i=\sum_{j=1}^{n}x_{ij}\Big/median_{i=1,\cdots,m}\left(\sum_{j=1}^{n}x_{ij}\right), \tag{2}$$

再作大小因子归一化来减轻不同测序深度可能带来的影响，后经 log 转换、伪计数加 1 后，得到：

$$\boldsymbol{X}'=\log\left(\boldsymbol{G}_c^{-1}\boldsymbol{X}+\boldsymbol{I}\right), \tag{3}$$

其中 $\boldsymbol{I}$ 为全一元素的矩阵，$\boldsymbol{G}_c=\mathrm{diag}\left(s_1,s_2,\cdots,s_m\right)\in\mathbb{R}^{m\times m}$ 且 $\mathrm{diag}\left(s_1,s_2,\cdots,s_m\right)$ 是一个以 $s_1,s_2,\cdots,s_m$ 作为对角线的对角矩阵。

对基因 j，定义基因因子：

$$g_j=\max_i x_{ij}', \tag{4}$$

同理全部基因因子构成的矩阵为$\boldsymbol{G}_g=\mathrm{diag}\left(g_1,g_2,\cdots,g_n\right)\in\mathbb{R}^{n\times n}$。矩阵$\boldsymbol{X}$中每个细胞作为一个样本，其中的细胞序号$i$和所有基因的序号作为输入，样本中基因表达作为标签，输入模型。矩阵$\boldsymbol{G}_c$和$\boldsymbol{G}_g$则用于最后的 scRNA-seq 数据填补。

2.2. 图协同过滤框架

对观察到的细胞和基因之间的交互进行建模，通过在图$\mathcal{T}$上应用传播和预测函数来生成细胞和基因特征表示：

$$\boldsymbol{e}_c^{(l+1)}=\mathrm{ReLU}\left(\sum_{i\in N_c}\frac{1}{\sqrt{\left|N_c\right|\left|N_i\right|}}\boldsymbol{e}_i^{(l)}\right),\boldsymbol{e}_g^{(l+1)}=\mathrm{ReLU}\left(\sum_{j\in N_g}\frac{1}{\sqrt{\left|N_g\right|\left|N_j\right|}}\boldsymbol{e}_j^{(l)}\right),\tag{5}$$

其中K是图神经网络(Graph Neural Network, GNN)的层数，$\boldsymbol{e}_c^{(0)}\in\mathbb{R}^{k\times1}$和$\boldsymbol{e}_g^{(0)}\in\mathbb{R}^{k\times1}$分别是 Xavier 初始化后的细胞$c$特征表示以及基因$g$特征表示，参数$k$是细胞(基因)特征表示中的特征维度。$\boldsymbol{e}_c^{(l+1)}$和$\boldsymbol{e}_g^{(l+1)}$分别是经过$l+1$层 GNN 后的细胞$c$特征表示及其基因$g$特征表示，且$l\left(0\le l\le K-1\right)$。$N_c$和$\left|N_c\right|$分别表示细胞$c$在$\mathcal{T}$上的邻居及其邻居个数。经过所有的 GNN 层后，采用加权和函数来得到细胞特征表示和基因特征表示：

$$\boldsymbol{e}_c=\frac{1}{K+1}\sum_{l=0}^{K}\boldsymbol{e}_c^{(l)},\boldsymbol{e}_g=\frac{1}{K+1}\sum_{l=0}^{K}\boldsymbol{e}_g^{(l)},\tag{6}$$

并采用$\xi_{cg}=\boldsymbol{e}_c^{\mathrm{T}}\cdot\boldsymbol{e}_g$来预测基因$g$在细胞$c$中的表达量。

鉴于交互图是二分图，图协同过滤框架在图上进行偶数次 GNN 的传播，会自然聚合同质结构邻居的信息。本文对每个细胞及其结构邻居作对比，将细胞本身的表示和偶数层 GNN 相应输出的表示视为一对，基于 InfoNCE 来最小化每一对之间的距离：

$$L^{\mathcal{C}}=\sum_{c\in\mathcal{C}}-\log\frac{\exp\left[\left(\boldsymbol{e}_c^{(e)}\cdot\boldsymbol{e}_c^{(0)}\right)/\tau\right]}{\sum_{\tilde{c}\in\mathcal{C}}\exp\left[\left(\boldsymbol{e}_c^{(e)}\cdot\boldsymbol{e}_{\tilde{c}}^{(0)}\right)/\tau\right]},\tag{7}$$

其中τ代表 softmax 的温度超参，e是一个偶数，$\boldsymbol{e}_c^{(e)}$和$\boldsymbol{e}_c^{(0)}$分别为第e层 GNN 和初始的细胞表示。同样的，可以得到针对基因的结构对比损失：

$$L^{\mathcal{G}}=\sum_{g\in\mathcal{G}}-\log\frac{\exp\left[\left(\boldsymbol{e}_g^{(e)}\cdot\boldsymbol{e}_g^{(0)}\right)/\tau\right]}{\sum_{\tilde{g}\in\mathcal{G}}\exp\left[\left(\boldsymbol{e}_g^{(e)}\cdot\boldsymbol{e}_{\tilde{g}}^{(0)}\right)/\tau\right]},\tag{8}$$

其中$\boldsymbol{e}_g^{(e)}$和$\boldsymbol{e}_g^{(0)}$分别为第e层 GNN 和初始的基因表示。总结构对比损失定义如下：

$$L_S=L^{\mathcal{C}}+\alpha L^{\mathcal{G}}\tag{9}$$

其中α是一个超参数，用于平衡权重。

2.3. ZINB 分布自编码器

考虑到 scRNA-seq 的计数数据表现为高度稀疏和过度分散，假设其分布为零膨胀负二项分布：

$$\mathrm{ZINB}\left(x\mid\pi,\mu,\theta\right)=\pi\cdot I_0\left(x\right)+\left(1-\pi\right)\cdot\mathrm{NB}\left(x\mid\mu,\theta\right),\tag{10}$$

其中负二项分布$\mathrm{NB}\left(x\mid\mu,\theta\right)=\frac{\Gamma\left(x+\theta\right)}{x!\Gamma\left(\theta\right)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x}$表征 scRNA-seq 数据的计数分布，$\mu,\theta$表示均

值和散度，π 表示真实的基因表达值被观测为 0 的概率，$I_0(x)$ 为示性函数。

基于 scRNA-seq 数据的 ZINB 分布特征来设置自解码器。基于 ZINB 分布的负对数似然损失函数为：

$$L_{\mathrm{ZINB}} = -\mathrm{ZINB}(x \mid \pi, \mu, \theta), \tag{11}$$

以上述函数作为损失函数进行解码来模拟 scRNA-seq 数据的分布，进而得到三个输出层 $\boldsymbol{\Pi}, \boldsymbol{M}, \boldsymbol{\Theta}$，它们分别代表 dropout 事件概率、NB 分布的均值和散度，则有下式：

$$\left\{\hat{\boldsymbol{\Pi}}, \hat{\boldsymbol{M}}, \hat{\boldsymbol{\Theta}}\right\} = \arg\max_{\boldsymbol{\Pi}, \boldsymbol{M}, \boldsymbol{\Theta}} \mathrm{ZINB}(\boldsymbol{X} \mid \boldsymbol{\Pi}, \boldsymbol{M}, \boldsymbol{\Theta}) = \arg\max_{\boldsymbol{\Pi}, \boldsymbol{M}, \boldsymbol{\Theta}} \prod_{c=1}^{m} \prod_{g=1}^{n} \mathrm{ZINB}\left(x_{cg} \mid \pi_{cg}, \mu_{cg}, \theta_{cg}\right), \tag{12}$$

其中 m、n 分别表示细胞数和基因数。

将提出的总结构邻居对比学习损失与 ZINB 分布重构损失结合，并添加 GNN 层中表示的 2-范数惩罚损失，将算法的训练损失定义为：

$$L = L_{\mathrm{ZINB}} + \lambda_1 L_S + \lambda_2 \left\|\bar{\boldsymbol{e}}\right\|_2, \tag{13}$$

其中 λ_1 和 λ_2 是控制所提出的总结构邻居对比学习损失和正则化项的权重的超参数，$\bar{\boldsymbol{e}}$ 表示 GNN 层中细胞和基因表示的参数集合。

连接 ZINB 分布自编码器后的网络架构表示如下：

$$\begin{cases} \boldsymbol{H} = \mathrm{ReLU}\left(\boldsymbol{\phi}\boldsymbol{W}^{\mathrm{T}} + \boldsymbol{b}\right), \\ \boldsymbol{B} = \mathrm{BatchNorm}\left(\boldsymbol{H}\right), \\ \boldsymbol{M} = \boldsymbol{G}_c\left[\exp\left(\mathrm{sigmoid}\left(\boldsymbol{B}\boldsymbol{W}_{\boldsymbol{M}}^{\mathrm{T}} + \boldsymbol{b}_{\boldsymbol{M}}\right) \times \boldsymbol{G}_g\right) - 1\right], \\ \boldsymbol{\Pi} = \mathrm{sigmoid}\left(\boldsymbol{B}\boldsymbol{W}_{\boldsymbol{\Pi}}^{\mathrm{T}} + \boldsymbol{b}_{\boldsymbol{\Pi}}\right), \\ \boldsymbol{\Theta} = \mathrm{softplus}\left(\left(\boldsymbol{B}\boldsymbol{W}_{\boldsymbol{\Theta}}^{\mathrm{T}} + \boldsymbol{b}_{\boldsymbol{\Theta}}\right) \times \boldsymbol{G}_g\right), \end{cases} \tag{14}$$

其中 $\boldsymbol{\phi} = \left(\xi_{cg}\right)_{m \times n} \in \mathbb{R}^{m \times n}$ 为经过 K (默认为 2)个 GNN 层后初步估计的基因表达矩阵，r 为隐藏层的神经元数。$\boldsymbol{W} \in \mathbb{R}^{r \times n}$ 和 $\boldsymbol{b} \in \mathbb{R}^{m \times r}$ 分别是从交互层到隐藏层的权重和偏置，$\boldsymbol{H} \in \mathbb{R}^{m \times r}$ 是隐藏层表示。$\mathrm{BatchNorm}(\cdot)$ 表示批次归一化(Batch Normalization, BN)层的函数，$\boldsymbol{B} \in \mathbb{R}^{m \times r}$ 是 BN 层输出的表示。$\boldsymbol{M} \in \mathbb{R}^{m \times n}$、$\boldsymbol{\Pi} \in \mathbb{R}^{m \times n}$、$\boldsymbol{\Theta} \in \mathbb{R}^{m \times n}$ 分别指 ZINB 分布的三个输出层输出，$\boldsymbol{W}_{\boldsymbol{M}} \in \mathbb{R}^{n \times r}$、$\boldsymbol{b}_{\boldsymbol{M}} \in \mathbb{R}^{m \times n}$ 分别是从神经元为 r 的 BN 层映射回神经元为 n 的参数网络层 $\boldsymbol{M}$ 时的权重和偏置。$\boldsymbol{W}_{\boldsymbol{\Pi}} \in \mathbb{R}^{n \times r}$、$\boldsymbol{b}_{\boldsymbol{\Pi}} \in \mathbb{R}^{m \times n}$ 和 $\boldsymbol{W}_{\boldsymbol{\Theta}} \in \mathbb{R}^{n \times r}$、$\boldsymbol{b}_{\boldsymbol{\Theta}} \in \mathbb{R}^{m \times n}$ 同理。指数函数和 softplus 函数分别应用于 $\boldsymbol{M}$ 和 $\boldsymbol{\Theta}$ 以确保它们的非负性，sigmoid 激活函数则限制 $\boldsymbol{\Pi}$ 的取值范围为 $[0,1]$。$\boldsymbol{M}$ 作为最终预测的基因表达矩阵。

3. 数值实验

3.1. scRNA-seq 数据集

为了评估 scRNA-seq 数据的填补效果，使用 R 包 Splatter [11]生成 scRNA-seq 仿真数据。固定仿真数据中真正零表达的比例为 35%，分别设计含有 dropout 时零值比例为 85%、90%、95%的仿真观测矩阵及其对应不含 dropout 的真实矩阵。利用这三个稀疏度不同的仿真数据集进行仿真实验。

为了测试 scGCF 在聚类分析中的能力，采用真实 scRNA-seq 数据集：PBMC 数据集[12]、Zeisel 数据集[4]、Human kidney 数据集[13]和 Adam 数据集[14]。PBMC 数据集是由 10x 基因组学平台测序得到的人类外周血单核细胞数据集，Human kidney 数据集是由 10x 基因组学平台测序得到的人类肾脏细胞数据集，两者均可从 10x 基因组学网站下载。Zeisel 数据集是由 Illumina 测序平台得到的 3005 个来自小鼠体

感皮层和海马体区域的细胞组成，Adam 数据集则由 3660 个小鼠肾脏细胞数据组成且由 Drop-seq 平台测序，两者均可在基因表达数据库下载。

3.2. 对比算法及评价指标

本文以观察到的表达矩阵作基准，记为 observed，并选择了六种对比算法 scImpute [5]、MAGIC [6]、DCA [8]、ALRA [7]、scVI [9]、scGNN [10]。scImpute 和 MAGIC 作为非深度学习对比算法。而由于算法涉及到图协同网络和 ZINB 自编码器，故选取 ALRA、DCA、scVI 和 scGNN 进行对比。所有对比算法均采用其默认参数来执行填补操作。

为了评价填补的效果，本文从仿真数据基因表达恢复实验和聚类分析实验来对 scGCF 进行评价。在填补实验中，通过 Splatter [11]包模拟得到一个设置细胞数量 m、基因量数 n、零表达值比例等参数的真实表达矩阵 $\boldsymbol{Y}=\left(y_{ij}\right)_{m\times n}$ 和对应含有 dropout 的观测基因表达矩阵 $\boldsymbol{Y}'=\left(y'_{ij}\right)_{m\times n}$。再对 $\boldsymbol{Y}'$ 进行填补得到矩阵 $\hat{\boldsymbol{Y}}=\left(\hat{y}_{ij}\right)_{m\times n}$，将计算 $\hat{\boldsymbol{Y}}$ 与真实基因表达 $\boldsymbol{Y}$ 之间的平均绝对误差(Mean Absolute Error, MAE)、均方根误差(Root Mean Squared Error, RMSE)、Pearson 相关系数(Pearson Correlation Coefficient, PCC)这三个评价指标以衡量填补后的数据在数值和结构上与真实数据的接近程度。MAE 和 RMSE 都表示填补误差，数值越小表征填补越精确，其计算公式如下：

$$\mathrm{MAE}=\frac{1}{mn}\sum_i\sum_j\left|y_{ij}-\hat{y}_{ij}\right|,\ \mathrm{RMSE}=\left(\frac{1}{mn}\sum_i\sum_j\left(y_{ij}-\hat{y}_{ij}\right)^2\right)^{\frac{1}{2}},\tag{15}$$

PCC 表示填补后数据与原始数据之间的相关性，越接近 1 则相关性越强，其计算公式如下：

$$\mathrm{PCC}=\frac{\sum_i\sum_j\left(y_{ij}-y_{mean}\right)\left(\hat{y}_{ij}-\hat{y}_{mean}\right)}{\sqrt{\sum_i\sum_j\left(y_{ij}-y_{mean}\right)^2}\sqrt{\sum_i\sum_j\left(\hat{y}_{ij}-\hat{y}_{mean}\right)^2}},\tag{16}$$

其中 y_{mean} 为真实基因表达矩阵 $\boldsymbol{Y}$ 的均值，$\hat{y}_{mean}$ 表示填补后矩阵 $\hat{\boldsymbol{Y}}$ 的均值。

聚类分析实验通过将细胞聚类后得到簇标签，将其和真实的细胞类型对比，以衡量聚类准确性。聚类分析的指标为可调节 Rand 系数(Adjusted Rand Index, ARI)和标准化互信息(Normalized Mutual Information, NMI)。ARI 的定义如下：

$$\mathrm{ARI}\left(\boldsymbol{c},\hat{\boldsymbol{c}}\right)=\left[\sum_{ij}\binom{n_{ij}}{2}-\left(\sum_i\binom{a_i}{2}-\sum_j\binom{b_j}{2}\right)\Big/\binom{r}{2}\right]\Big/\left[\frac{1}{2}\left(\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right)-\left(\sum_i\binom{a_i}{2}-\sum_j\binom{b_j}{2}\right)\Big/\binom{r}{2}\right],\tag{17}$$

其中 $\boldsymbol{N}=\left(n_{ij}\right)_{r\times k}$ 为混淆矩阵，r 为真实细胞类型数，k 为聚类簇数，且 $a_i=\sum_j n_{ij}$，$b_j=\sum_i n_{ij}$，$\boldsymbol{c},\hat{\boldsymbol{c}}$。

分别为真实细胞标签和聚类簇标签。而 NMI 的定义如下：

$$\mathrm{NMI}=\frac{2I\left(U,V\right)}{H\left(U\right)+H\left(V\right)},\tag{18}$$

其中 $I\left(U,V\right)$ 为互信息

$$I\left(U,V\right)=\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{\left|u_i\cap v_j\right|}{m}\log\frac{m\left|u_i\cap v_j\right|}{\left|u_i\right|\times\left|v_j\right|},\tag{19}$$

其中 u_i、v_j 分别为真实细胞的标签中属于第 i 类的集合和预测的簇标签中属于第 j 类的集合，而交叉熵

$H(U)$、$H(V)$的公式为：

$$H(U)=-\sum_{i=1}^{r}\frac{u_i}{m}\log\frac{u_i}{m},\ H(V)=-\sum_{j=1}^{k}\frac{v_j}{m}\log\frac{v_j}{m}, \tag{20}$$

ARI 和 NMI 的数值越接近 1 意味着聚类效果越理想。

3.3. 仿真实验

为了验证填补效果，进行了仿真数据填补实验。本文设定将观测矩阵与其真实表达矩阵对比并计算 MAE、RMSE 以及 PCC 指标，记为 observed 方法。接着运用了 scImpute、MAGIC、DCA、ALRA、scVI、scGNN 和 scGCF 这七种不同的算法，对仿真的观测矩阵进行了填补操作，将这些填补后的矩阵与对应不同稀疏性的真实表达矩阵进行了对比，同样计算了三个指标。各算法计算获得的三个指标结果见图 2。当零表达比例达到 85%、90%和 95%时，scGCF 算法展现出了显著的优势，不仅取得了更高的 PCC 指标，还实现了更低的 RMSE。特别是在零表达比例高达 90%和 95%的情况下，scGCF 算法在 MAE 上表现超越了 observed 和其他 6 种填补算法。在零值比例为 85%时 scGCF 算法的 MAE 值也仅次于 DCA。这些实验结果充分表明 scGCF 能比其他对比算法更好地恢复基因表达矩阵。

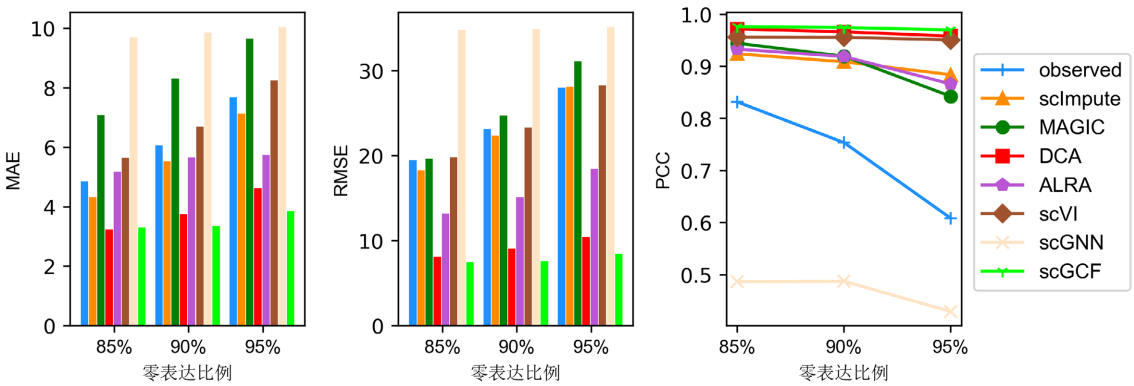

Figure 2. Comparison of metrics (a) MAE; (b) RMSE; (c) PCC among methods on simulated data
图 2. 仿真数据上各算法的指标比较(a) MAE 值；(b) RMSE 值；(c) PCC 值

3.4. 聚类分析实验

为了研究 scGCF 用于聚类分析的性能，考虑 PBMC 数据集、Zeisel 数据集、Human kidney 数据集和 Adam 数据集。聚类时采用 Louvain [15]聚类，且聚类的分辨率设置为 1，并对 ARI 和 NMI 指标进行计算。

对于 observed、scImpute、MAGIC 和 ALRA 算法，依次将观测矩阵或填补矩阵进行 PCA 将矩阵上的基因维度缩减至 50，最后进行聚类。对于 DCA、scVI、scGNN 和 scGCF 算法，将对应算法得到的细胞隐层表示用于聚类。对于四个完整的数据集(PBMC、Zeisel、Human Kidney 和 Adam)，通过 Louvain 聚类获得的指标结果如图 3 所示。在这四个数据集上 scGCF 算法取得了更高的 ARI 值。同时 scGCF 在 Zeisel 数据集上取得了第二高的 NMI 值，在其余的三个数据集上取得了更高的 NMI 值。整体上，与其他对比算法相比 scGCF 在完整数据集上取得了较好的聚类效果。

为了评估这些算法对细胞聚类的稳健性，对这些完整的数据集使用了采样策略。随机对数据集中的每类细胞采样其各自的 95%，并组合构成一个子样本集。该采样过程在一个数据集上独立执行了十次，得到了十个子样本，再将算法应用于这十个子样本聚类而得到聚类指标。每个数据集上得到的 ARI 值以

及 NMI 值分别绘制为图 4 和图 5 中的箱线图。

由图 4 可知在 Zeisel、Human kidney 和 Adam 这三个数据集上 scGCF 的 ARI 值更高且更稳定，在 PBMC 数据集的 ARI 值上 scGCF 和 observed 均高于其他对比算法。由图 5 可知，scGCF 在 PBMC 和 Adam 这两个数据集上的 NMI 值表现更好，而在 Zeisel 数据集上的 NMI 值仅低于 scGNN，在 Human kidney 数据

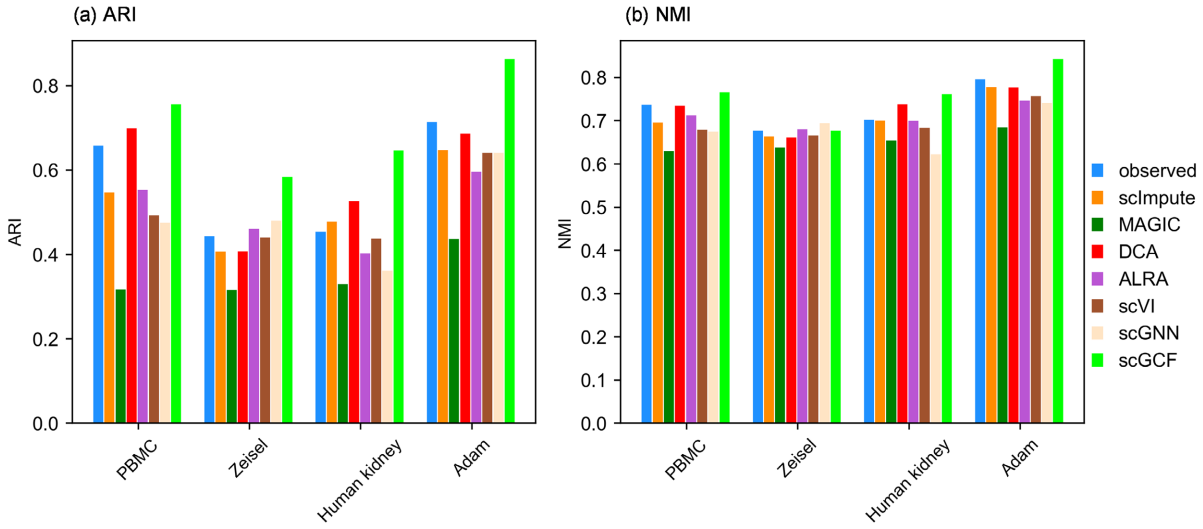

Figure 3. (a) Barplot of ARI values obtained by clustering using methods on complete datasets; (b) Barplot of NMI values obtained by clustering using methods on complete datasets

图 3. (a) 完整数据集上各算法聚类得到的 ARI 条形图；(b) 完整数据集上各算法聚类得到的 NMI 条形图

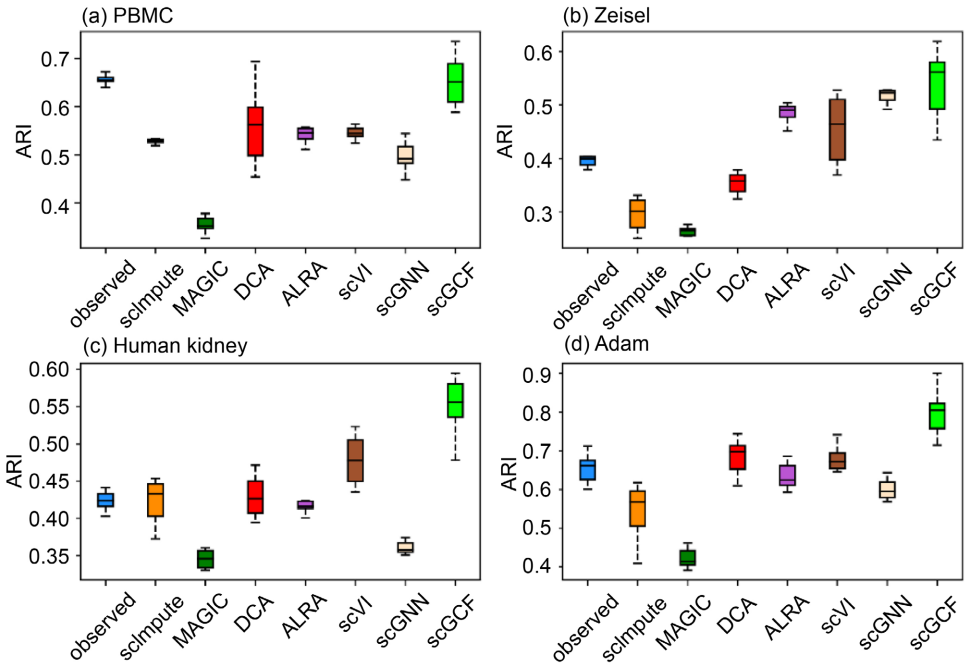

Figure 4. Boxplots of ARI values obtained by clustering using methods on 10 subsamples of the dataset. (a) PBMC dataset; (b) Zeisel dataset; (c) Human kidney dataset; (d) Adam dataset

图 4. 在数据集 10 个子样本上应用各算法聚类而获得的 ARI 箱线图。(a) PBMC 数据集；(b) Zeisel 数据集；(c) Human kidney 数据集；(d) Adam 数据集

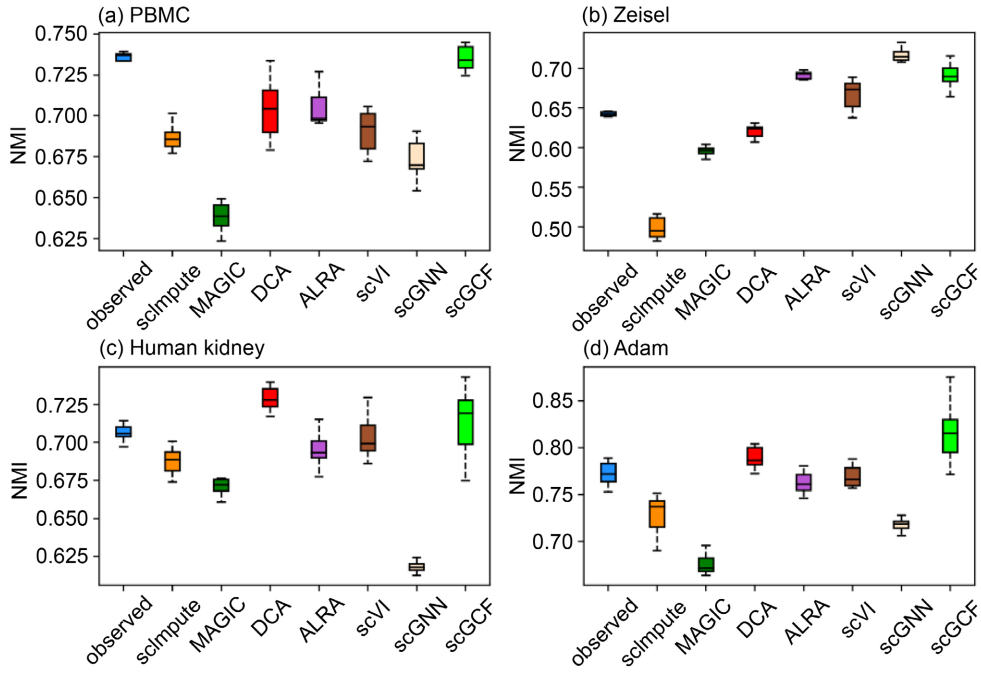

Figure 5. Boxplots of NMI values obtained by clustering using methods on 10 subsamples of the dataset. (a) PBMC dataset; (b) Zeisel dataset; (c) Human kidney dataset; (d) Adam dataset
图 5. 在数据集 10 个子样本上应用各算法聚类而获得的NMI箱线图。(a) PBMC数据集；(b) Zeisel数据集；(c) Human kidney数据集；(d) Adam数据集

集的 NMI 值仅低于 DCA。结合图 3~图 5 的实验结果，与其他对比算法相比 scGCF 具有较好的聚类精度和鲁棒性。

4. 结论

综上，本文提出了一种新的 scRNA-seq 数据填补算法 scGCF，它利用图协同过滤框架聚合得到细胞特征表示和基因特征表示，并使用基于 ZINB 分布的自编码器来填补 scRNA-seq 数据。在仿真数据集上验证了 scGCF 算法的有效性，并评估了其填补能力。同时在四个真实数据集上进行了聚类分析，表明 scGCF 良好的细胞聚类能力。scGCF 不仅在 scRNA-seq 下游聚类分析上增强了可靠性，还拓展了 scRNA-seq 数据的填补算法领域。

参考文献

[1] Luecken, M.D. and Theis, F.J. (2019) Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial. *Molecular Systems Biology*, **15**, e8746. https://doi.org/10.15252/msb.20188746

[2] Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science. *Nature Reviews Genetics*, **14**, 618-630. https://doi.org/10.1038/nrg3542

[3] Patel, A.P., Tirosh, I., Trombetta, J.J., *et al.* (2014) Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma. *Science*, **344**, 1396-1401. https://doi.org/10.1126/science.1254257

[4] Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., *et al.* (2015) Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-Seq. *Science*, **347**, 1138-1142. https://doi.org/10.1126/science.aaa1934

[5] Li, W.V. and Li, J.J. (2018) An Accurate and Robust Imputation Method sCimpute for Single-Cell RNA-Seq Data. *Nature Communications*, **9**, 997. https://doi.org/10.1038/s41467-018-03405-7

[6] Van Dijk, D., Sharma, R., Nainys, J., *et al.* (2018) Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, **174**, 716-729. https://doi.org/10.1016/j.cell.2018.05.061

[7] Linderman, G.C., Zhao, J. and Kluger, Y. (2022) Zero-Preserving Imputation of scRNA-seq Data Using Low-Rank Approximation. *Nature Communications*, **36**, 3139-3147.

[8] Eraslan, G., Simon, L.M., Mircea, M., *et al.* (2019) Single-Cell RNA-seq Denoising Using a Deep Count Autoencoder. *Nature Communications*, **10**, 390. https://doi.org/10.1038/s41467-018-07931-2

[9] Lopez, R., Regier, J., Cole, M.B., *et al*. (2018) Deep Generative Modeling for Single-Cell Transcriptomics. *Nature Methods*, **15**, 1053-1058. https://doi.org/10.1038/s41592-018-0229-2

[10] Wang, J., Ma, A., Chang, Y., *et al* .(2021) scGNN Is A Novel Graph Neural Network Framework for Single-Cell RNA-Seq Analyses. *Nature Communications*, **12**, 1882. https://doi.org/10.1038/s41467-021-22197-x

[11] Zappia, L., Phipson, B. and Oshlack, A. (2017) Splatter: Simulation of Single-Cell RNA Sequencing Data. *Genome Biology*, **18**, 174. https://doi.org/10.1186/s13059-017-1305-0

[12] Zheng, G.X., Terry, J.M., Belgrader, P., *et al.* (2017) Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nature Communications*, **8**, 14049. https://doi.org/10.1038/ncomms14049

[13] Young, M.D., Mitchell, T.J., Vieira Braga, F.A., *et al.* (2018) Single-Cell Transcriptomes from Human Kidneys Reveal the Cellular Identity of Renal Tumors. *Science*, **361**, 594-599. https://doi.org/10.1126/science.aat1699

[14] Adam, M., Potter, A.S. and Potter, S.S. (2017) Psychrophilic Proteases Dramatically Reduce Single-Cell RNA-Seq Artifacts: A Molecular Atlas of Kidney Development. *Development*, **144**, 3625-3632. https://doi.org/10.1242/dev.151142

[15] Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E. (2008) Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics*: *Theory and Experiment*, **2008**, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008