武汉大学学报(信息科学版)
Geomatics and Information Science of Wuhan University
ISSN 1671-8860,CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：单细胞 RNA 测序数据的自监督低通滤波图聚类网络
作者：廖明辉，罗甫林，杜博
DOI：10.13203/j.whugis20220108
收稿日期：2022-12-02
网络首发日期：2023-01-11
引用格式：廖明辉，罗甫林，杜博．单细胞 RNA 测序数据的自监督低通滤波图聚类网络[J/OL]．武汉大学学报(信息科学版). https://doi.org/10.13203/j.whugis20220108

DOI:10.13203/j.whugis20220108
引用格式：
廖明辉，罗甫林，杜博. 单细胞RNA测序数据的自监督低通滤波图聚类网络[J]. 武汉大学学报（信息科学版），2023，DOI：10.13203/j.whugis20220108（LIAO Minghui, LUO Fulin, DU Bo. Self-supervised Low-pass Filted Graph Clustering Networks for Single Cell RNA Sequencing Data[J]. *Geomatics and Information Science of Wuhan University*, 2023, DOI: 10.13203/j. whugis20220108）

单细胞 RNA 测序数据的自监督低通滤波图聚类网络

廖明辉[1]　　罗甫林[2]　　杜博[3]

1 武汉大学计算机学院，湖北 武汉，430064
2 重庆大学计算机学院，重庆，400030
3 武汉大学计算机学院，湖北 武汉，430064

摘　要：近年兴起的单细胞 RNA 测序（single-cell RNA sequencing，scRNA-seq）技术可以测出每个单细胞的转录组表达量，利用单细胞 RNA 测序数据可以将具有相似生物学状态或相似功能的单细胞聚类成同一细胞群，从而指导下游生物学分析。针对单细胞 RNA 测序数据的复杂、高维、携带大量噪声的特点，提出了一种自监督低通滤波图聚类网络（Self-supervised Low-pass Filted Graph Clustering Network，SLFGCN）算法用于单细胞 RNA 测序数据的聚类研究。该方法首先构建了一个低通滤波的图卷积网络，以细胞为节点构建图网络结构，在谱域的图信息经过低通滤波图卷积操作后，获得更加平滑的图信号，即同一簇的细胞提取到更相似的节点特征，从而利于单细胞 RNA 测序数据聚类；然后，通过图自编码模型，建立自监督模块优化模型，进一步优化聚类效果。通过在单细胞 RNA 测序数据上与相关算法的对比实验结果表明，提出的方法能更好地获取单细胞 RNA 表达数据的内在特征，改善聚类效果。

关键词：单细胞 RNA 测序；图卷积网络；聚类；深度学习

收稿日期：2022-12-02
项目资助：国家杰出青年科学基金(62225113)；国家自然科学基金青年项目(62206202)；中国博士后面上项目(2022M712461)。
第一作者：廖明辉，硕士，研究方向为图卷积神经网络。minghui@whu.edu.cn
通讯作者：杜博，博士，教授。gunspace@163.com

复杂的生物组织和生命体是由形态各异、功能各异的细胞群组成。单细胞 RNA 测序（single-cell RNA sequencing，scRNA-seq）技术是对每一个细胞的 RNA 进行测序，得到所有基因在该细胞的表达量。与传统的批量测序不同，它具备分析单个细胞的生物学状态的能力，被广泛应用于肿瘤生物学[1]、胚胎发育学[2]、器官形成[3]等诸多生物学领域。细胞的每一个基因的表达量都可视为该细胞的一个特征，从 scRNA-seq 数据中挖掘生物信息的关键步骤是将生命状态、生物功能相似的单细胞聚类成一个集群。

传统的聚类方法不能很好地应用在大量、高维、复杂的 scRNA-seq 数据上，比如 K-mean 算法要求簇是凸形状的，谱聚类则要求簇密度是分布均匀。近年一些新的模型被提出用于 scRNA-seq 数据的聚类。文献[4]利用共享最近邻的思想来挖掘高维度数据信息。文献[5]提出多核学习的单细胞分析法（single-cell interpretation via multikernel learning， SIMLR），结合了多核机制从数据中学习到合适的距离度量以实现聚类。随着深度学习的发展，一些基于神经网络的聚类技术被提出来。文献[6]通过零膨胀负二项（zero-inflated negative binomial，ZINB）损失函数优化的自编码器来重构低维的 scRNA-seq 数据，在低维上利用简单的聚类方法，比如 K-mean 聚类将数据聚类。scDeepCluster [7]是一个面向 scRNA-seq 的深度嵌入聚类模型，它考虑了 scRNA-seq 数据的稀疏性和不均匀性。单细胞在生命体中的相对和绝对位置蕴含了丰富的生物学信息，生命体中相近的单细胞往往具有相似的基因表达特征。以上方法仅仅是对单细胞样本点的特征信息进行计算，而缺乏对单细胞样本点之间结构化信息的利用。

图卷积网络（Graph Convolutional Networks，GCN）[8]可以捕获样本之间的结构信息，GCN 和 GCN 的各种变体[9][10]已经被应用于各种需要考虑数据之间结构信息的场景中，比如预测交通流量 [11]和药物设计[12]。在聚类任务上，文献[13]提出了结构化的深度聚类网络（Structural Deep Clustering Network，SDCN）整合样本的结构信息。文献[14]采用图神经网络来表示基因之间的关系，获取单细胞之间的结构信息。文献[15]提出一种基于 GCN 的聚类模型对 scRNA-seq 数据进行聚类。scRNA-seq 数据由于测序手段的限制往往携带有噪声，以上方法虽然利用了单细胞样本点之间的结构化信息，但是未对测序数据的噪声问题提出解决办法。

为解决上述提到的问题，提高对诸如 scRNA-seq 数据等高维复杂和携带噪声的数据的聚类效果，本文提出了自监督低通滤波的图聚类网络模型（Self-supervised Low-pass Filted Graph Clustering Networks, SLFGCN）用于 scRNA-seq 数据聚类。首先引入低通滤波的图传播方式到图卷积网络中，然后通过图自编码（Graph Auto-Encoders, GAE）[16]模型建立自监督模型进行网络优化，进一步改善聚类效果。与现有的应用在 scRNA-seq 数据的聚类方法相比，SLFGCN 具有以下明显的优点：

1.图自编码器模块以细胞为节点构造图结构，向前传播的过程使用了邻近点的特征，挖掘到细胞之间的结构信息。

2. 构造了低通滤波的图卷积操作，在谱域中过滤掉高频的噪声信号，它使得图信号更加光滑，即同一簇的细胞具有更加相似的特征，更利于它们聚成同一个簇类。

1 模型架构

SLFGCN 模型框架如图 1 所示。SLFGCN 主要分为 3 个模块，分别是低通滤波 GCN (LFGCN)模块、GAE 模块和自监督模块。GAE 模块能在保证样本之间结构信息的同时获取样本点的低维表表示，LFGCN 模块用于获取聚类结果，自监督模块通过 GAE 获取的低维特征信息对 LFGCN 模块进行自主优化，进一步增强聚类的准确性。单细胞测序数据的基因表达矩阵 X 首先经过 KNN 构图之后传入图的编码解码模块，编码器输出的隐变量 X(h)经过 k-means 初步聚类之后由簇中心和样本点构造分布 Q 和 P；

另一方面，X 经过全连接层初步降维后构造 LFGCN 模块的输入 Z(0)，Z(0)经过 LFGCN 模块输出特征图 Z，后者经过 softmax 层得到样本点的聚类标签。

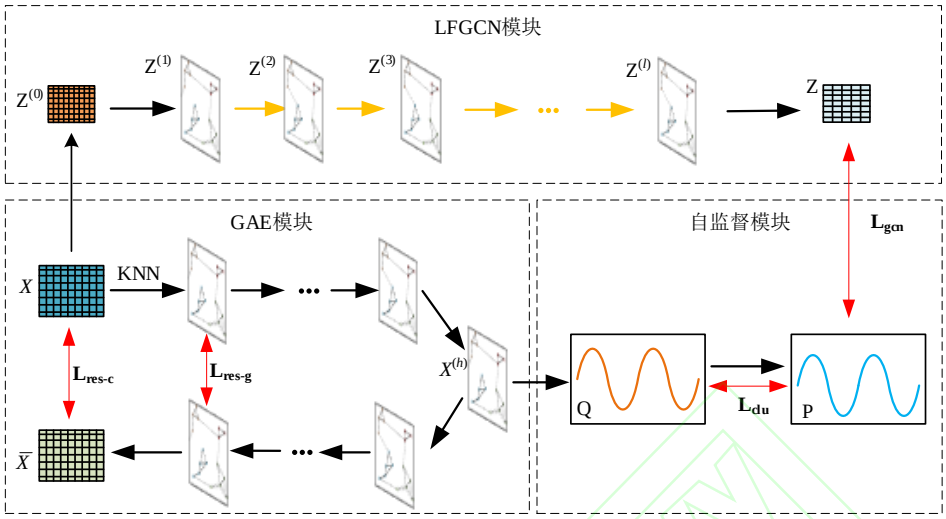


图 1 SLFGCN 模型框架

Fig. 1 Architecture of SLFCN

1.1 KNN 构图

在 scRNA-seq 样本中，细胞 i 的基因表达数据 x_i 可以表示为 d 维的向量，对于细胞 i 和细胞 j，它们的相似性可表示为：

$$s_{ij}=\frac{x_i\cdot x_j}{|x_i||x_j|} \tag{1}$$

(1)式中，$|x_i|$、$|x_j|$ 分别表示特征向量 x_i 和 x_j 的模，对于 N 个细胞的数据集，其相似性矩阵为 $S\in\mathbb{R}^{N\times N}$。对于细胞 i，按照 KNN 的思想选取距离细胞 i 最近即相似性最高的前 t 个细胞作为它的近邻点来构图，即构建图的连接矩阵 $A\in\mathbb{R}^{N\times N}$，细胞 i 与细胞 j 若为近邻则 a_{ij} 设置为 1，否则为 0。本文中 t 选取选为 $0.01\times N$ 和 20 中的最大值。

1.2 GAE 模块

对于单细胞 RNA 测序样本的基因表达数据 $X\in\mathbb{R}^{N\times d}$，其中 N 是细胞的个数，d 是数据的维度。在 GAE 中，采用主流的 GCN 传播方式[8]：

$$X^{(h)}=\mathrm{ReLU}(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(h-1)}W^{(h-1)}) \tag{2}$$

(2)式中，$\tilde{A}=A+I_N$，I_N 为单 N 阶单位矩阵；$\tilde{D}_{ij}=\sum_j\tilde{A}_{ij}$；$W^{(h-1)}$ 为可训练参数矩阵。经过 h 层传播后得到低维的特征矩阵 $X^{(h)}$ 被输入到下游的自监督模块，同时在 GAE 模块中继续传播得到重构特征矩阵 $\bar{X}$。参数矩阵 W 通过以下损失函数训练：

$$L_{res-c}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_{ij}-\bar{X}_{ij})^2 \tag{3}$$

此外，选择样本之间的内积运算来重构出样本之间的结构信息[17]：

$$\bar{A}=sigmoid(\bar{X}^{\mathrm{T}}\bar{X}) \tag{4}$$

(4)式中 $sigmoid(\cdot)$ 为激活函数，$\bar{A}$ 为重构的连接矩阵，图结构重构的损失函数定义如下：

$$L_{res-g}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(A_{ij}-\bar{A}_{ij})^2 \tag{5}$$

1.3 LFGCN 模块

由于原始特征矩阵 X 维度太大，在此使用全连接神经网络提取 X 的低维表示 $Z^{(0)}$ 作为 LFGCN 的原始输入：

$$Z^{(0)}=\mathrm{ReLU}(WX+b) \tag{6}$$

(6)式中，$\mathrm{ReLU}(\cdot)$ 为激活函数；W，b 为权重矩阵和偏置项；$Z^{(0)}\in\mathbb{R}^{N\times m}$ 作为 GCN 传播模块的初始输入，其中 m 远小于 d。$Z^{(0)}$ 被当作初始值输入 LFGCN。

对于 N 个节点的无向图 G(V,E)，V 是所有节点的集合，E 是所有边的集合。归一化的图的拉普拉斯矩阵 L 定义为：

$$L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad (7)$$

(7)式中，$A \in \mathbb{R}^{N\times N}$ 是图的连接矩阵，$D \in \mathbb{R}^{N\times N}$ 是图的度矩阵，$D_{ij} = \sum_j \mathrm{A}_{ij}$。$L$ 可特征值分解：$L = U\Lambda U^{\mathrm{T}}$，其中 Λ 是由特征值从小到大排列的对角矩阵，U 是特征值对应的特征向量组成的矩阵。

在图理论中，图的节点信号 $x \in \mathbb{R}^N$ 表示节点的特征向量，卷积操作就是使用频率响应函数 $\mathrm{g}_\theta = diag(\theta)$ 得到新的信号：

$$\mathrm{Y}=\mathrm{g}_\theta * x = U\mathrm{g}_\theta(\Lambda)U^T x \quad (8)$$

(8)式中，$U^T x$ 表示谱域上的信号，上述卷积操作也是在谱域上进行，左乘 U 之后又转换为节点域的信号表示。对于频率响应函数 g_θ 的选择，文献[18]提出 ChebNet 模型，对特征值的对角矩阵经过契比雪夫多项式的 K 阶式转化得到频率响应函数：

$$g_\theta(\Lambda) \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\Lambda}) \quad (9)$$

其中：

$$\tilde{\Lambda} = \frac{2}{\lambda_{\max}}\Lambda - I_N \quad (10)$$

θ_k 为可学习参数。契比雪夫多项式定义为：

$$T_0(x) = 1, T_1(x) = x$$
$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \quad (11)$$

因为 $\lambda_{\max}$ 为 2 [19]，所以 $\tilde{\Lambda}$ 的值落在[-1,1]，满足契比雪夫多项式中 x 取[-1,1]的要求。

对于图结构的特征矩阵 $X = [x_1, x_2, x_3, ..., x_n]^{\mathrm{T}}$，$x_i \in \mathbb{R}^d$ 表示第 i 个节点特征向量。在一个图中，如果邻近的节点有着更加相似的节点特征，那么对这样的图节点进行聚类任务也更加容易，同时该图的图信号更加光滑。基信号 u_q 的光滑程度可以用拉普拉斯－贝尔特拉米算子[19] $\Omega(\cdot)$ 衡量：

$$\Omega(u_q) = \frac{1}{2}\sum_{(v_i,v_j)\in E} a_{ij}\left\| \frac{u_q(i)}{\sqrt{d_i}} - \frac{u_q(j)}{\sqrt{d_j}} \right\|_2^2$$
$$= u_q{}^T L u_q = \lambda_q \quad (12)$$

其中 $u_q(i)$ 为 u_q 的第 i 个元素，（12）式表明低频（更小的特征值）的基信号更加光滑，这意味光滑的图信号含有更多低频的基信号。

在（9）式中，通常取 K=1，由此（8）（9）（11）式可推导为：

$$Y = \theta_0 x + \theta_1 U\tilde{\Lambda}U^T x \quad (13)$$
$$= \theta_0 x + \theta_1 g(L)x \quad (14)$$

（14）式中，$g(\cdot)$ 为频率响应函数；θ_0、θ_1 为可学习参数。如(12)式所示，在谱域中更多低频的信号表现为节点域更光滑的图信号，由此获得同一簇内更加相似的节点特征以利于聚类。受文献[20]的启发，取频率响应函数 $g(\cdot)$ 为：

$$g(L) = I_N - \frac{1}{2}L \quad (15)$$

（15）式与（10）式相比，前者为减函数，可以将谱域中的高频信号降低为低频。（14）（15）可得：

$$Y = \theta_0 x + \theta_1 (I_N - \frac{1}{2}L)x \quad (16)$$

将拉普拉斯矩阵 L 对称归一化：

$$L = I_N - D^{-1/2} A D^{-1/2} \quad (17)$$

（16）（17）式可得：

$$Y = \theta_0 x + \frac{1}{2}\theta_1 (I_N + D^{-1/2} A D^{-1/2})x \quad (18)$$

文献[8]提出再归一化，采用 $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ 代替 $I_N + D^{-1/2}AD^{-1/2}$ 以缓解梯度爆炸和梯度消失，其中 $\tilde{A} = A + I_N$，$D_{ij} = \sum_j \tilde{A}_{ij}$。鉴于此，（18）式可以得：

$$Y = \theta_0 x + \frac{1}{2}\theta_1 \tilde{D}^{-1/2} A \tilde{D}^{-1/2} x \quad (19)$$

为缓解 GCN 传播过程中的过平滑（Over-smoothing）问题[21]，引入恒等映射（Identity mapping）和初始残差连接（Initial residual connection）[18]，并使用 ReLU 函数作为激活函数，最后得低通滤波 GCN 传播公式 $Z^{(l+1)} = LFGCN(Z^{(l)})$ 为：

$$Z^{(l+1)} = \mathrm{ReLU}\left(\begin{bmatrix} (1-\alpha)Z^{(l)} \\ +\alpha_l Z^{(0)} \end{bmatrix} \begin{bmatrix} (1-\beta_l)I_N \\ +\omega_l W_1^{(k)} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} (1-\alpha)\tilde{D}^{-1/2}A\tilde{D}^{-1/2}Z^{(l)} \\ +\alpha_l Z^{(0)} \end{bmatrix} \begin{bmatrix} (1-\beta_l)I_N \\ +\omega_l W_2^{(l)} \end{bmatrix} \right) \quad (20)$$

其中 $Z^{(l)}$ 是第 l 层的特征矩阵；α_l、β_l、ω_l 是第 l 层超参数，在本研究中分别取 $\alpha_l = 0.3$，$\beta_l = \omega_l = 0.5\frac{1}{l}$；$Z^{(0)}$ 是网络的初始输入；I_N 是 N 阶单位矩阵；$W_1^{(l)}$、$W_2^{(l)}$ 是第 l 层可学习参数矩阵。

LFGCN 层的最后一层的输出 $Z^{(K)}$ 输入到全连接层，经过 softmax 层：

$$Z = softmax(W^{(K)}Z^{(K)} + b^{(K)}) \quad (21)$$

（21）式中 $Z \in \mathbb{R}^{N \times c}$ 是预测的样本点在各簇的分布概率，c 为簇的数量。$z_{ij} \in Z$ 表示细胞 i 在簇别 j 的概率，对于细胞 i，取最大概率对应的簇类别作为它的聚类结果。

1.4 自监督模块

GAE 模块经过（3）（5）式预训练后，取 GAE 模块的输出 $X^{(h)}$ 进行 K-means 聚类，得到 c 个类别和簇中心 u_j，j=1,2,...,c。这些簇类别随着迭代次数进行更新。

具体来说，细胞 i 从 GAE 模块学习到的表示 h_i 与第 j 个簇的簇中心 u_j 的相似性使用学生 t 分布作为核函数计算：

$$q_{ij} = \frac{(1 + \|h_i - u_j\|^2 / v)^{-\frac{v+1}{2}}}{\sum_{j'}(1 + \|h_i - u_{j'}\|^2 / v)^{-\frac{v+1}{2}}} \quad (22)$$

（22）式中，v 是学生 t 分布的自由度，在此设为 1，q_{ij} 是细胞 i 属于第 j 个类别的概率。基于概率分布 Q=[q_{ij}]，目标分布 P=[p_{ij}]:

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \quad (23)$$

（23）式中，$f_j = \sum_i q_{ij}$ 是簇别 j 的频率。使用 KL 散度来衡量 P 分布和 Q 分布之间的差异：

$$L_{clu} = KL(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (24)$$

最小化以上损失函数以获得更优的聚类效果。

为了整合细胞之间的整合结构信息，使用目标分布 P 监督 Z 的更新：

$$L_{gcn} = KL(P \| Z) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{z_{ij}} \quad (25)$$

通过最小化分布 P 和分布 Z 的差异，使得 LFGCN 模块也可以从 GAE 模块学习到信息，由此，LFGCN 模块可以兼顾样本的内容信息和结构信息。

最后，总的损失函数可以定义为：

$$L_{all} = L_{res-c} + aL_{res-g} + bL_{clu} + cL_{gcn} \quad (26)$$

（26）式中，a, b, c 均为超参数，在本研究中设为 0.0001、0.1、0.01。最小化总的损失函数 L_{all} 使 CGN 模块的输出 Z 能够结合细胞数据的内容信息和结构信息。最后，细胞 i 的聚类结果的标签为：

$$r_i = \arg\max_j z_{ij} \quad (27)$$

2 实验

2.1 实验数据

下载了 7 个来源于人源或鼠源的不同器官和组织的单细胞转录组测序数据集[22]，这些数据集的单细胞数量从几十到几千不等，如表 1 所示。

表 1 数据集列表

Tab.1 The list of datasets used in this study.

数据集	GSE/ID	细胞数	基因数	细胞种类
Biase	GSE57249	56	25733	4
Goolam	E-MTAB-3321	124	41427	5
Darmanis	GSE67835	466	22088	9
Deng	GSE45719	268	22431	6
Baron_mouse	GSE84133	1886	14878	13
Romanov	GSE74672	2881	24341	7
Zeisel	GSE60361	3005	19972	9

对上述的数据集采用文献[15]的方法提取出 2000 个基因作为特征，并缩放到[0,1]的范围。

2.2 评价指标

聚类结果采用三个经典的聚类评价指标进行评价，分别是聚类准确度（Clustering Accuracy, CA）[23]、标准互信息（Normalized Mu-tual Information, NMI）[7]和调整兰德系数（Adjusted Rand Index, ARI）[24]，以上指标得分越高则聚类效果越好。

2.3 参数设置

GAE模块各层的维度为d-512-256-64-10-64-256-512-d，d是输入数据的维度；GAE模块预训练时学习率设为0.00001；Batch size为32，每个数据集训练500个循环；采用Adam优化器。GCN模块共6层，其输入的初始特征表示 $Z^{(0)}$ 和隐藏层的维度均为256；训练时若超过400个循环效果仍未上升则终止训练。比较的方法的参数均用原始论文设置的参数，所有实验在NVIDIA GTX 2080Ti（12GB）中进行，使用的框架为Pytorch。模型实现代码将上传至：

https://github.com/WHUminghui/SLFGCN。

2.4 实验结果

2.4.1 聚类效果实验

对比方法选取了经典的基于深度学习的单细胞测序数据聚类方法scDeepCluster[7]、经典的基于图神经网络的GraphSCC[15]和传统的聚类方法 k-mean，实验结果如表 2 所示。从表 2 可以看到SLFGCN 除了在 Zeisel 数据集上是第二的效果，在其他的数据集上均取得第一的表现，SLFGCN 相较于 GraphSCC、scDeepcluster 和 K-mean，聚类效果分别提高了 7.53%、23.3%和 9.13%；而对于 Zeisel 这一数据集，该数据集是作者使用了新的测序手法 STRT/C1 对小鼠大脑皮层区域进行密集采样并测序处理得到，故可能该数据集中不同标签的单细胞也具有相似的基因表达特征，而使得 SLFGCN 的 LFGCN 模块作用效果不明显，聚类效果稍逊于 scDeepCluster；scDeepCluster 缺少对细胞之间结构关系的提取而在其他数据集上表现不佳；GraphSCC 虽然利用了 GCN 对结构信息进行提取而有了有较优的表现，但效果不及使用了低通过滤的图卷积方法的 SLFGCN，这是因为图信号经过低通过滤器的时候会变得更加平滑，而在节点域中的同一簇细胞提取出更相似的特征表示；K-mean 作为传统的机器学习方法，结构简单但是要求簇是凸的，这对于高维、复杂、大量的单细胞转录组测序数据往往不能满足。

表 2　各方法聚类效果对比

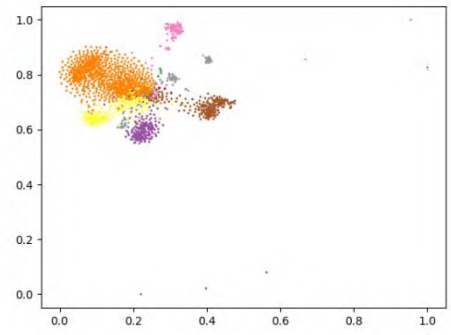
Tab.2　Clustering results on all seven datasets

数据集	指标	GraphSCC	scDeepcluster	kmean	SLFGCN
Biase	CA	**1**	0.9107	0.9821	**1**
	NMI	**1**	0.8209	0.9501	**1**
	ARI	**1**	0.7888	0.9556	**1**
Goolam	CA	**0.9516**	0.8306	0.9032	**0.9516**
	NMI	0.9513	0.8877	0.9184	**0.9516**
	ARI	**0.9808**	0.9097	0.9633	**0.9808**
Darmanis	CA	0.8219	0.7618	0.7918	**0.8348**
	NMI	0.7511	0.7469	0.7604	**0.7702**
	ARI	0.7448	0.6743	0.7544	**0.7639**
Deng	CA	0.6231	0.4963	0.8209	**0.8582**
	NMI	0.7484	0.6835	0.8880	**0.9084**
	ARI	0.5267	0.4127	0.8747	**0.8826**
Baron_mouse	CA	0.8383	0.5541	0.6506	**0.8956**
	NMI	0.8230	0.7429	0.8037	**0.9177**
	ARI	0.7915	0.4388	0.5921	**0.9464**
Romanov	CA	0.7796	0.7681	0.6762	**0.8133**
	NMI	0.5905	**0.6772**	0.5667	0.6566
	ARI	0.6007	0.6234	0.5082	**0.6954**
Zeisel	CA	0.9238	0.8136	0.8915	**0.9249**
	NMI	**0.8214**	0.7147	0.8062	0.8167
	ARI	**0.8778**	0.6890	0.8384	0.8698

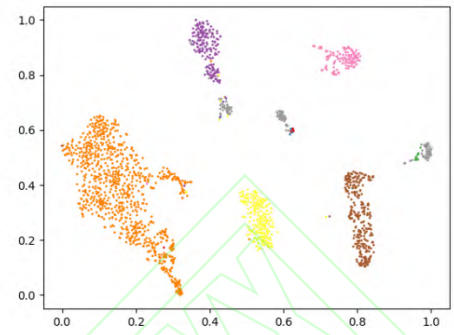
2.4.2 可视化实验

使用 SLFGCN 提取 Baron_mouse 和 Romanov 数据集的特征表示，即低通滤波 GCN 最后一层的输出特征 $Z^{(k)}$，以 t-SNE 算法[25]为降维工具对其降维到二维空间进行可视化，图中相同颜色的点即为同一类细胞。如图 2 所示，图 2（a）、图 2（c）分别

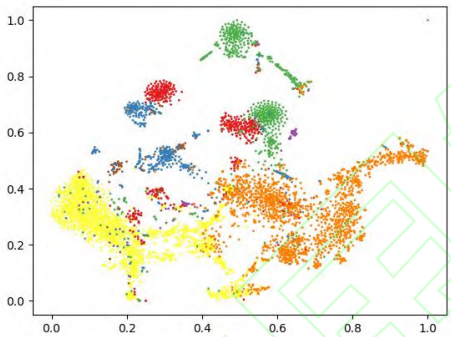
是 Baron_mouse、Romanov 原始数据的可视化结果；图 2（b）、图 2（d）则是经过 SLFGCN 提取到的特征表示的可视化结果。可以发现原始数据集中簇间样本有重叠且同一簇的样本点过于分散，但是 LFGCN 提取到的特征使得同一类别的样本点相距更近、簇间距离更加远，这是因为同一簇样本点的图信号变得更加光滑，在同一簇中获得更加相似的节点特征而利于聚类、分类等任务。



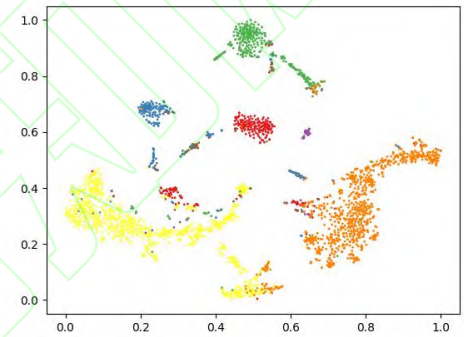
（a）Baron_mouses 数据集原始特征



（b）Baron_mouses 数据集经过 LFGCN 提取之后的特征



（c）Romanov 数据集原始特征



（d）Romanov 数据集经 LFGCN 提取之后的特征

图 2 LFGCN 提取的样本特征可视化图

Fig. 2 Visualization of features extracted by LFGCN

2.4.3 消融实验

为研究模块各组分的贡献，在 Baron_mouse 数据集上进行模块消融实验，实验结果如表 3 所示。在训练 GAE 模块时，不考虑样本点图的结构信息，即不使用 $L_{res\text{-}g}$ 时，聚类效果下降 1.32%。另外，SLFGCN 不使用低通过滤的图信号时，聚类效果下降了 9.59%，以上证明了低通滤波图卷积操作的优越性。

表 3　消融实验结果

Tab.3　The Results of ablation experiments on Baron_mouse

消融实验	CA	NMI	ARI
SLFGCN-$L_{res\text{-}g}$	0.8838	0.8985	0.9411
SLFGCN-LF	0.8396	0.8334	0.8221
SLFGCN	**0.8956**	**0.9177**	**0.9464**

2.4.4 超参数敏感性分析

对（26）式中 a, b, c 三个超参数在 Deng 数据集上分析其敏感性，固定 a-b、a-c、b-c 中的 a, b, c 为 0.0001、0.1、0.01 时，分别对 c、b、a 取不同值。实验结果如表 4 所示，因为自监督模块的修正作用，$L_{res\text{-}g}$ 对结果的影响不及 L_{clu} 和 L_{gcn}。

表 4　超参数分析实验结果

Tab.4　Results of experimental hyperparameter analysis

a	0.00001	0.00005	0.0001	0.0005	0.001
CA	0.8365	0.8465	**0.8582**	0.8523	0.8362

NMI	0.8362	0.8663	**0.9084**	0.8421	0.8741
ARI	0.8320	0.8757	**0.8826**	0.8701	0.8126
b	0.01	0.05	0.1	0.5	1
CA	0.6567	0.8059	**0.8582**	0.8246	0.6523
NMI	0.8045	0.8484	**0.9084**	0.8883	0.8041
ARI	0.5714	0.8125	**0.8826**	0.8749	0.5695
c	0.001	0.005	0.01	0.05	0.1
CA	0.6529	0.8134	**0.8582**	0.8246	0.8059
NMI	0.8012	0.8207	**0.9084**	0.8883	0.8493
ARI	0.5762	0.8462	**0.8826**	0.8748	0.8156

3 总结

针对多维、复杂、大量的单细胞转录组测序数据聚类效果差的问题，本文设计了一种低通过滤的图卷积网络传播方式，通过降低图信号的特征值而获得更加光滑的图信号表示，使得同一簇节点的特征更加相似而利于聚类任务进行。并且，被提取到的特征表示被结合图结构信息的 GAE 模块和自监督模块优化，实验表示该模型取得更优的聚类效果。

参 考 文 献

[1] Navin, N. et al. Tumor evolution inferred by single cell sequencing. Nature 472, 90–94 (2011).

[2] Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015).

[3] Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 360, 176–182 (2018).

[4] Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 31, 1974–1980 (2015)

[5] Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat. Methods 14, 414–416 (2017).

[6] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," Nat Commun, vol. 10, no. 1, p. 390, Jan 23 2019, doi: 10.1038/s41467-018-07931-2.

[7] T. Tian, J. Wan, Q. Song, and Z. Wei, "Clustering single-cell RNA-seq data with a model-based deep learning approach," Nature Machine Intelligence, vol. 1, no. 4, pp. 191-198, 2019, doi: 10.1038/s42256-019-0037-0.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[9] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," 2016.

[10] P. Velikovi, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," 2017.

[11] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, pp. 922-929.

[12] Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, and Y. Yang, "Communicative Representation Learning on Attributed Molecular Graphs," presented at the IJCAI, 2020.

[13] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, and P. Cui, "Structural Deep Clustering Network," presented at the Proceedings of The Web Conference 2020, 2020.

[14] J. Rao, X. Zhou, Y. Lu, H. Zhao, and Y. Yang, "Imputing Single-cell RNA-seq data by combining Graph Convolution and Autoencoder Neural Networks," biorxiv, 2020, doi:

10.1101/2020.02.05.935296.
[15] Zeng Y , Zhou X , Rao J , et al. Accurately Clustering Single-cell RNA-seq data by Capturing Structural Relations between Cells through Graph Convolutional Network[C]// 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020.
[16] Kipf T N , Welling M . Variational Graph Auto-Encoders[J]. 2016.
[17] T. Kipf and M. Welling, “Variational graph auto-encoders,” NIPS Workshop on Bayesian Deep Learning, 2016.
[18] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in Advances in Neural Information Processing Systems, 2016, pp. 3844–3852.
[19] Fan RK Chung and Fan Chung Graham. Spectral graph theory. Number 92. American Mathematical Society, 1997.
[20] Chen M , Wei Z , Huang Z , et al. Simple and Deep Graph Convolutional Networks[J]. 2020.
[21] Li Q , Han Z , Wu X M . Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning[J]. 2018.
[22] M. Krzak, Y. Raykov, A. Boukouvalas, L. Cutillo, and C. Angelini, "Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods," Front Genet, vol. 10, p. 1253, 2019, doi: 10.3389/fgene.2019.01253.
[23] J. M. Zhang, J. Fan, H. C. Fan, D. Rosenfeld, and D. N. Tse, "An interpretable framework for clustering single-cell RNA-Seq datasets," BMC Bioinformatics, vol. 19, no. 1, p. 93, Mar 9 2018, doi: 10.1186/s12859-018-2092-7.
[24] V. Y. Kiselev et al., "SC3: consensus clustering of single-cell RNA-seq data," Nat Methods, vol. 14, no. 5, pp. 483-486, May 2017, doi: 10.1038/nmeth.4236.
[25] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,”Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579–2605, 2008.

Self-supervised Low-pass Filted Graph Clustering Networks for Single Cell RNA Sequencing Data

LIAO Minghui[1] LUO Fulin[2] DU Bo[3]

1 School of Computer Science, Wuhan University, Wuhan 430064, China

2 School of Computer Science, Chongqing University, Chongqing 400030, China

3 School of Computer Science, Wuhan University, Wuhan 430064, China

Abstract: Single-cell RNA sequencing (scRNA-seq) provides high-resolution observation tools at the cell level for biological domains, such as embryonic development, cancer evolution and cell differentiation. A key step in using scRNA-seq data is to cluster cells with similar biological functions into one group. However, the current clustering methods are not able to perform the clustering task well in a large number of high-dimensional and complex scRNA-seq data, and don’t use the structural relationship information between samples. Here, we propose a GCN based deep clustering framework, named Self-supervised Low-pass Filted Graph Clustering Networks (SLFGCN). Firstly, a new propagation method of graph convolutional network is proposed. For

the proposed method, the graph information in the spectral domain passes through the frequency response function of the low-pass filter to obtain smoother node feature representation, which is more conducive to the clustering task. Secondly, we use the self-supervised module to optimize the network based on the representation learned from the low-pass filted GCN module and the representation learned from the graph auto-encoders module, which can obtain better clustering effect. Experiments indicate that our model outperforms the state-of-the-art methods in various evaluation metrics on real datasets. Further, the visualization results show that our model provides representations generating better intra-cluster compactness and inter-cluster separability.
Key words: Single-cell RNA sequencing; Graph convolutional network; Clustering;
Deep learning

First author: LIAO Minghui ,master, specializes in graph convolutional neural network. E-mail: minghui@whu.edu.cn

Corresponding author: DU Bo, PhD, professor. E-mail: gunspace@163.com

网络首发：

标题：单细胞RNA测序数据的自监督低通滤波图聚类网络

作者：廖明辉，罗甫林，杜博

DOI：10.13203/j.whugis20220108

收稿日期：2022-12-02

引用格式：

廖明辉，罗甫林，杜博. 单细胞RNA测序数据的自监督低通滤波图聚类网络[J]. 武汉大学学报（信息科学版），2023，DOI：10.13203/j.whugis20220108（LIAO Minghui, LUO Fulin, DU Bo. Self-supervised Low-pass Filted Graph Clustering Networks for Single Cell RNA Sequencing Data[J]. *Geomatics and Information Science of Wuhan University*, 2023, DOI: 10.13203/j. whugis20220108）