计算机科学
COMPUTER SCIENCE

联合ZINB模型与图注意力自编码器的自优化单细胞聚类

孔凤玲, 吴昊, 董庆庆

引用本文

孔凤玲, 吴昊, 董庆庆. 联合ZINB模型与图注意力自编码器的自优化单细胞聚类[J]. 计算机科学, 2023, 50(12): 104-112.

KONG Fengling, WU Hao, DONG Qingqing. Self-optimized Single Cell Clustering Using ZINB Model and Graph Attention Autoencoder [J]. Computer Science, 2023, 50(12): 104-112.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000167

联合 ZINB 模型与图注意力自编码器的自优化单细胞聚类

孔凤玲　吴　昊　董庆庆
云南大学信息学院　昆明 650500
(2378015890@qq.com)

摘　要　单细胞数据聚类在生物信息分析中具有重要作用,但受测序原理和测序平台的限制,单细胞数据集普遍存在高维稀疏性、高方差噪声和基因数据缺失的问题,导致单细胞数据在聚类分析和应用方面仍面临诸多挑战。现有的单细胞聚类方法主要针对细胞和基因表达间的关系进行建模,忽略了对细胞间潜在特征关系的充分挖掘以及对噪声的去除,导致聚类结果不理想,从而阻碍了后期对数据的分析。针对上述问题,提出了一种联合零膨胀负二项(Zero Inflated Negative Binomial,ZINB)模型与图注意力自编码器的自优化单细胞聚类算法(Self-optimized Single Cell Clustering Using ZINB Model and Graph Attention Autoencoder,scZDGAC)。该算法首先使用 ZINB 模型并结合可扩展的 DCA 去噪算法,通过 ZINB 分布更好地拟合数据特征分布,提升自编码器的去噪性能,并减小噪声和数据丢失对 KNN 算法输出的影响;然后通过图注意力自编码器在不同权重的细胞之间传播信息,更好地捕获细胞间的潜在特征进行聚类;最后 scZDGAC 采用自优化的方法使原本两个独立的聚类模块和特征模块相互受益,不断迭代更新聚类中心,进一步提升聚类性能。为了对聚类结果进行评价,文中使用调整兰德指数(ARI)和标准化互信息(NMI)两个通用评价指标。在 6 个不同规模的单细胞数据集上与其他算法进行对比实验,结果表明,所提聚类算法在聚类性能上较其他方法有很大提高,很好地展现了该算法的鲁棒性。

关键词:深度聚类;scRNA-Seq;ZINB 模型;自优化;DCA;图注意力自编码器
中图法分类号　Q811.4

Self-optimized Single Cell Clustering Using ZINB Model and Graph Attention Autoencoder

KONG Fengling,WU Hao and DONG Qingqing
School of Information Science and Engineering,Yunnan University,Kunming 650500,China

Abstract　One of the most important aspects of single-cell data analysis is the clustering of individual cells into clusters of subpopulations. However,due to the limitation of sequencing principle and sequencing platform,the obtained single cell dataset generally has high-dimensional sparsity,high variance noise and a large amount of data loss,which lead to many challenges in cluster analysis and application of single cell data. Single-cell clustering methods proposed in recent years mainly model the relationship between cell and gene expression,ignoring the full mining of the potential characteristic relationship between cells and the removal of noise,resulting in unsatisfactory clustering results,which hinders the later analysis of data. In view of the above problems,a self-optimized single-cell clustering algorithm(scZDGAC) combining zero expansion negative binomial(ZINB) model with graph attention autoencoder is proposed. The algorithm firstly uses ZINB model combined with extensible DCA denoising algorithm,better fit data feature distribution through ZINB distribution,to improve the denoising performance of autoencoder,and reduce the impact of noise and data loss on the output of KNN algorithm. And then using the graph attention autoencoder to spread the information between cells of different weights,which can better capture the potential features between cells for clustering. Finally,scZDGAC uses the self-optimization method to make the originally two independent clustering modules and feature modules benefit from each other,and constantly update the clustering center iteratively to further improve the clustering performance. In order to evaluate the clustering results,this paper uses adjusted RAND index(ARI) and standardized mutual information(NMI) as two general evaluation indicators. Compared with six single cell datasets of different scales,experimental results show that the clustering performance of the proposed clustering algorithm has greatly improved.

Keywords　Deep clustering,scRNA-seq,ZINB model,Self-optimization,DCA,Graph attention autoencoder

到稿日期:2022-10-21　返修日期:2023-03-14
基金项目:国家自然科学基金(62061049);云南省基础研究项目(2018FB100)
This work was supported by the National Natural Science Foundation of China(62061049) and Yunnan Fundamental Research Projects(2018FB100).
通信作者:吴昊(haowu1982@vip.163.com)

1 引言

单细胞 RNA 测序技术的快速发展使得在细胞水平上测量转录组基因表达成为可能，研究人员能够进一步增加对细胞类型的了解，并揭示跨组织、不同发育阶段和生物体等细胞亚群之间的异质性和多样复杂性[1-2]。与大量的 RNA-seq 数据相比，scRNA-seq 数据面临的问题是数据高维稀疏，大部分的测量值为零（由于 RNA 捕获率低）[3]。测序平台的最新发展也极大地提高了细胞吞吐量，使得细胞数量呈指数式增加，达到数千万个[4]。由于每个细胞的测序深度相对较浅，目前的技术特别容易导致数据丢失[5]。此外，即使在同一组细胞中，scRNA-seq 数据中的基因表达水平也存在很大差异，或者在短暂的细胞状态中相当平稳，这使得细胞间的关系很难捕获。上述技术和生物因素的结合带来了许多变化和噪声[3]，使得细胞与基因表达间的特征和细胞间潜在的特征关系很难被学习到，进一步阻碍了单细胞的聚类。

为了解决这些问题，获得更好的聚类性能，研究人员最近研发了大量的单细胞转录表达聚类方法。在单细胞转录表达聚类方面，Wang 等[6]提出了一种单细胞解释的学习方法 SIMLR，该方法通过结合多核的光谱聚类方法来学习数据结构之间的相似性。Satija 等[7]提出了一种基于共享邻近图（Seurat）的聚类方法，采用鲁汶算法来检测细胞群落。Lin 等[8]通过插补单细胞 RNA-seq 数据进行快速和准确的聚类，采用简单的隐式推断过程来减少 scRNA-seq 数据中丢失率对下游聚类的影响。Mei 等[9]提出了带秩约束的相似性学习来完成对单细胞数据的聚类，通过度量细胞之间的全局和局部关系，构建了一个相似矩阵，然后推导出一个相似块对角矩阵，得到最终的聚类结果。Kiselev 等[10]提出了一种单细胞表达谱的无监督聚类方法（SC3），该方法首先使用主成分分析（PCA）和拉普拉斯变换来降低数据的维度，然后测量 3 对细胞距离以获得 6 个投影，并将 K-means 算法应用于上述投影，建立了一个共识矩阵，最终作为层次聚类的输入。Yang 等[11]提出了基于集成学习的单细胞聚类方法（来自 Ensemble）SAFE，该方法结合了 T-SNE＋K-means，CIDR，Seurat 和 SC3 的聚类结果。在此基础上，SAME[12]进一步集成了 5 种聚类方法，从而学习到单细胞数据更好的特征。尽管上述方法采用简单的线性降维技术，直接学习基于原始噪声数据矩阵的相似度或距离表示获得简单的聚类结果，但由于数据的高维稀疏性和噪声的存在，距离测量不准确，线性降维方法无法捕获隐藏在 scRNA-seq 数据中的非线性结构，从而导致这些方法在高维稀疏性和噪声数据集上的聚类效果不佳。此外，将降维方法和聚类方法分离，容易导致更大的聚类结果偏差。

近年来，深度学习技术已经被广泛应用于计算生物学，并且表现出比以往的传统机器学习算法[13-14]更好的性能。自动编码器是一种常见的神经网络类型，它能够通过编码器和解码器来学习有效的数据压缩，而且不需要任何的监督过程[15]。它可以在重构去噪数据的同时，实现潜在空间中高维数据的非线性降维。scDeepCluster[16]使用基于零膨胀负二项模型的自动编码器对单细胞数据的统计特征进行断层建模，同时通过深度嵌入聚类方法（DEC）进行聚类。受 DEC[17]的启发，Li 等[18]提出了一种基于无监督的单细胞深度嵌入聚类算法（DESC），该算法通过迭代式地优化目标函数来对 scRNA-seq 数据聚类，并且在聚类精度和稳定性之间取得了适当的平衡，但是该算法只采用深度自编码器来对数据结构进行预训练，将聚类过程从数据去噪中分离出来，并没有学到更合适的聚类潜在空间。Chen 等[19]设计了一种基于自训练的 soft K-means 聚类算法（scZiDesk）。他们选择前 500 个高变基因作为默认输入，大大减少了运行时间和内存消耗。但 soft K-means 可能会得到仅与部分数据相匹配的结果，从而出现过拟合问题。上述方法将表示学习与聚类过程结合到一个统一的框架中来提高聚类性能。然而，此类方法在特征学习过程中只考虑基因表达信息，而没有明确地描述不同细胞之间的关系。

为了将表达信息和关系信息同时嵌入到潜在特征空间中，scDSC[20]利用基于 K-近邻分类算法的图神经网络来捕获细胞之间的关系，将细胞间的结构信息与自编码器学习到的特征信息进行集成并逐层传播，在两种结构中获得了丰富的特征学习信息，取得了不错的聚类结果。scGAC[21]基于图注意力网络对单细胞数据聚类，通过构造单元图和基于图注意力的自动编码器来学习细胞的潜在特征，以利用基因表达和细胞-细胞关系间的信息进行特征学习和聚类。GCN 通过相邻节点的信息传播来学习图中节点的特征，考虑节点特征和图形结构，证明了通过 GCN 学习到的特征可以改善聚类结果[22]。scGNN[23]将新兴的图卷积网络（GCN[24]）结合到了单细胞聚类中。scGNN 将 GCN 集成到其多自动编码器框架中，首先利用特征自编码器构造 GCN 的细胞图；然后通过基于 GCN 的自编码器学习图中细胞的特征，使用 K-means 运算得到聚类结果。然而，由于所构造的图可能包含一些连接不同类型细胞的噪声边，而基于 GCN 的自动编码器直接在相邻细胞之间传播信息并不对它们进行区分，因此会混淆不同类型的细胞，从而导致出现错误的聚类结果。

上述聚类方法都只针对单细胞数据普遍存在的噪声问题或高维稀疏性问题进行研究，而且大多数只考虑到细胞与基因表达间的特征关系，忽略了细胞之间的潜在特征关系，以至于最终的聚类结果并不是最优的。因此，单细胞数据集中的噪声、测序数据自身的高维稀疏性、单细胞数据的高丢失率（基因的零表达值），以及对细胞间的特征关系挖掘不充分等问题的存在，使得对单细胞数据聚类成为一项特别具有挑战性的任务。

针对单细胞数据存在的高维稀疏性、高噪声、大量的数据丢失，以及未充分挖掘细胞间的特征关系等问题，本文提出了一种单细胞聚类算法 scZDGAC（联合 ZINB 模型与图注意力自编码器的自优化单细胞聚类算法）。

1）用 ZINB 模型对单细胞的基因数据分布进行建模，并结合自编码器与可拓展 DCA[25]去噪算法对原始的表达矩阵进行去噪处理，去噪后的数据对后续使用 KNN 构造细胞相似性邻接矩阵提供了准确性和稳定性。

2）通过图注意力自编码器学习细胞间的潜在关系，在不同权重的细胞之间进行信息传播，以充分地利用细胞与基因

表达间以及细胞与细胞之间的潜在特征关系。

3)为了优化细胞间特征学习和聚类,采用自优化聚类方法,使原本两个独立的聚类模块和特征模块相互受益,不断地更新和迭代优化聚类中心,改善单细胞数据的聚类性能。

为验证 scZDGAC 算法的有效性,本文在 6 个真实单细胞数据集上将其与其他聚类算法进行了对比实验,结果表明相比其他算法,scZDGAC 算法在聚类性能上有较大的提高。

2 实验方法

2.1 scZDGAC 基本架构

scZDGAC 首先将包含有 M 个基因和 N 个细胞的原始矩阵 $\boldsymbol{D}\in\mathbb{R}^{M\times N}$ 进行数据预处理,即进行质量控制、过滤线粒体、低质量细胞和高变基因的选择[18];其次对筛选后的数据进行标准化处理 $\boldsymbol{H}\in\mathbb{R}^{M'\times N'}(M'<M, N'<N)$;最后通过对 $\boldsymbol{H}$ 进行对数转换和 Z 分数归一化得到表达式矩阵 $\boldsymbol{X}\in\mathbb{R}^{m\times n}(m<M', n<N')$,并将其作为 ZINB 模型的输入数据。为了对预处理后的数据 $\boldsymbol{X}$ 去噪,捕捉 scRNA-seq 数据的特征,本文使用基于自动编码器 ZINB 模型并结合可拓展 DCA[25]去噪算法,从而得到更好的去噪数据;其次使用 K 邻近算法获得图邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times n}$,将特征矩阵和邻接矩阵一同输入到图注意力自编码器网络中进行细胞间特征的学习和初始聚类;最后采用自优化聚类对初始聚类结果不断更新优化。图 1 给出了 scZDGAC 的架构,它主要包括 3 部分:细胞特征去噪模块(ZINB)、图注意力自编码器和自优化聚类模块。

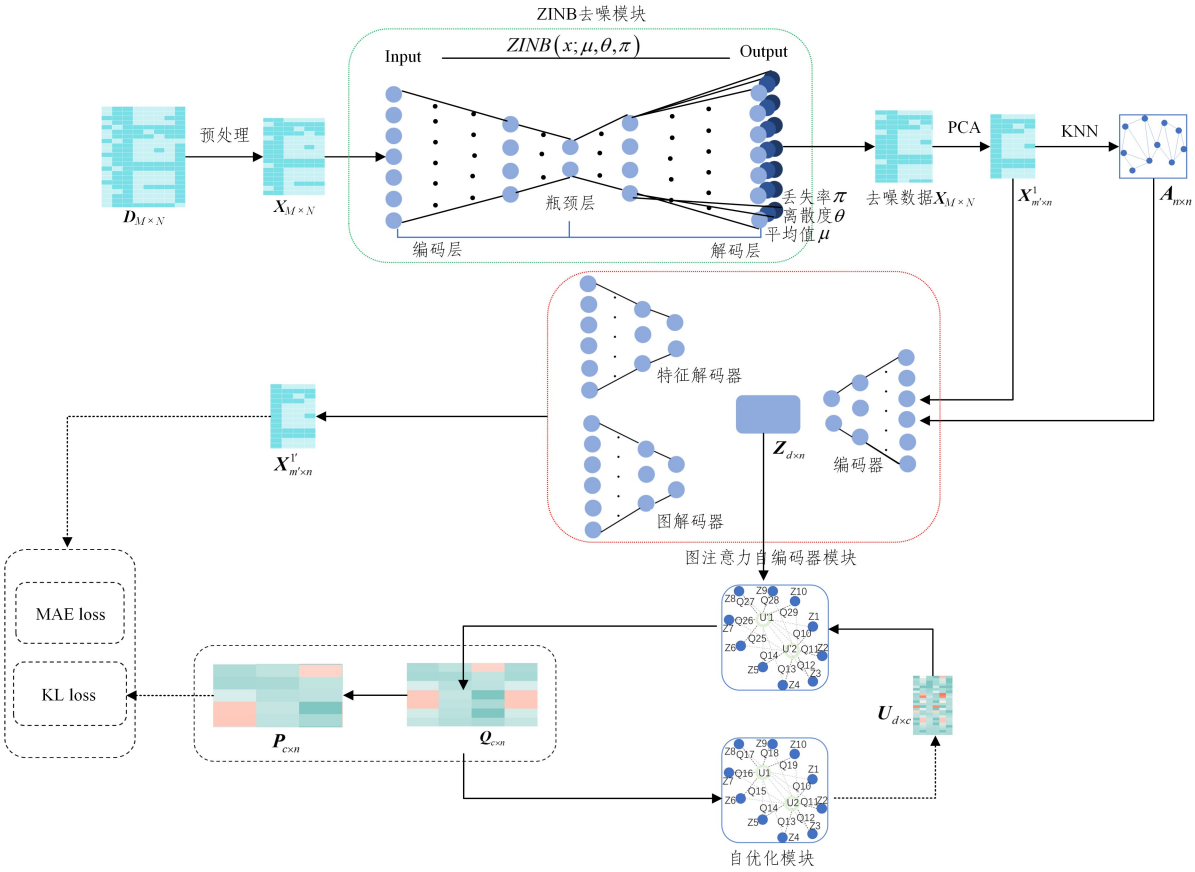


图 1 scZDGAC 网络架构图

Fig. 1 scZDGAC network architecture diagram

2.2 细胞特征去噪模块 ZINB

为了对预处理后的数据进行去噪,捕捉 scRNA-seq 数据的特征,本文采用基于自动编码器的 ZINB 模型并结合可拓展 DCA[25]去噪算法来描述数据的特征性能。ZINB 分布可用于对高度稀疏和过度分散的基因表达数据进行处理,并且 DCA 去噪算法能很好地捕获细胞与细胞间及细胞与基因间的结构信息,从而得到更好的去噪数据,且该过程有利于后续 KNN 算法的稳定性和准确性。自编码器的损失函数是 ZINB 分布的似然性。

首先,ZINB 分布结合 DCA 算法对矩阵 $\boldsymbol{X}$ 进行降噪处理得到 DCA 去噪后的重构特征矩阵 $\boldsymbol{X}$,由于预处理的数据矩阵 $\boldsymbol{X}$ 与 DCA 重构数据具有相同的维数,因此本文利用主成分分析(PCA)对重构数据 $\boldsymbol{X}$ 进行初始降维,得到一个新的特征矩阵 $\boldsymbol{X}^1\in\mathbb{R}^{m'\times n}(m'<m)$。

ZINB 分布是由两个分量(零点质量和负二项分量)组成的一个混合模型。零点质量代表数据中多余的零值,负二项分量(NB)代表计数分布。

$$NB=(x;\mu,\theta)=\frac{\Gamma(x+\theta)}{\Gamma(\theta)\Gamma(x+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{x} \tag{1}$$

$$ZINB(x;\pi,\mu,\theta)=\pi\delta(x)+(1-\pi)NB(x;\mu,\theta) \tag{2}$$

其中,π,μ 和 θ 是 ZINB 分布的参数[26],分别代表数据丢失的概率、平均值和离散度;Γ 表示 *Gamma* 分布;$\delta(x)$ 为 ZINB 的概率质量函数,可以视为真实的基因表达值被观测为 0 的概率。基于 ZINB 模型的自编码器将 3 个独立的全连接层与解码器的最后一层连接起来,以评估 ZINB 的 3 个参数 π,μ,θ。

实际上，本文使用这些参数的矩阵形式进行损失函数的计算，并将 π,μ,θ 的矩阵计算形式定义为：

$$\begin{cases} E=RELU(\boldsymbol{X}\boldsymbol{W}_E) \\ B=RELU(E\boldsymbol{W}_B) \\ D=RELU(B\boldsymbol{W}_D) \\ \boldsymbol{\Pi}=sigmoid(D\boldsymbol{W}_\pi) \\ \bar{\boldsymbol{M}}=\exp(DW_\mu) \\ \boldsymbol{M}=\mathrm{diag}(S_j)\bar{\boldsymbol{M}} \\ \boldsymbol{\Theta}=\exp(D\boldsymbol{W}_\theta) \end{cases} \tag{3}$$

其中，$\boldsymbol{X}$ 是通过对数转换和 $\boldsymbol{Z}$ 分数归一化得到的表达式矩阵；E,B,D 分别代表编码层、瓶颈层和解码层；S_j 表示每个细胞因子的大小，它是总细胞数与中位数 S 的比值；$\boldsymbol{W}$ 代表编码器和解码器的权重矩阵；$\boldsymbol{\Pi},\boldsymbol{M}$ 和 $\boldsymbol{\Theta}$ 分别为 π,μ 和 θ 的矩阵形式。用输出层预测的负二项分量（式(3)中的 $\bar{\boldsymbol{M}}$）的均值替换原始计数值生成去噪矩阵 $\boldsymbol{X}$，即该方法的最终输出。最后 DCA 的损失函数为 ZINB 的似然分布，表示为：

$$NLL_{\mathrm{ZINB}}=-\log(ZINB(\boldsymbol{X}|\mu,\theta)) \tag{4}$$

$$\hat{\boldsymbol{\Pi}},\hat{\boldsymbol{M}},\hat{\boldsymbol{\Theta}}=\arg\min_{\Pi,M,\Theta} NLL_{\mathrm{ZINB}}(\boldsymbol{X};\boldsymbol{\Pi},\boldsymbol{M},\boldsymbol{\Theta})+\lambda\|\boldsymbol{\Pi}\|_{\mathrm{F}}^2 \tag{5}$$

其中，NLL_{ZINB} 函数表示 $ZINB$ 分布的负对数似然，λ 表示参数。

基于 ZINB 模型的自动编码器可看作是对 scRNA-seq 数据的低通滤波，在这个过程中，相比其他低通滤波的方式，ZINB 能够更好地保证计数矩阵数据中的非零值数据信息受损率更低，进行细胞间特征拟合时更准确，获得的聚类结果更好。因为 ZINB 分布在对数据进行处理时把数据分为零值和非零值，使得计数矩阵中大量的零值不会影响到非零值的拟合过程。DCA 去噪方法具有很高的鲁棒性并能快速去噪，以及消除数据中的技术差异和捕获真实数据集中的细胞群体结构，改善下游分析，获得更好的聚类结果。

2.3 图注意力自编码器

为了更好地学习细胞间的潜在特征关系，本文设计了一个图注意力自编码器，该编码器是由两个堆叠的图注意层和一个结构对称的解码器组成，它将高维数据投影到低维潜在空间，也将拓扑信息嵌入到细胞的潜在特征空间中，更好地进行细胞间潜在特征关系的学习。

给定一个细胞图，图注意层通过聚合其具有不同权重的相邻细胞的特征来学习细胞间的特征。由于权重是根据细胞及其相邻的特征自动分配的，因此可以自然地捕获细胞间潜在的特征关系，从而使学习到的特征可以更好地进行聚类。

具体地，图注意层的功能描述如下：

$$h'_i=\sigma(\sum_{j\in N_i}\alpha_{ij}\boldsymbol{W}h_j) \tag{6}$$

其中，h_i' 是细胞 i 的新特征，N_i 是 KNN 图中细胞 i 的邻接细胞集合，h_j 是细胞 j 的输入特征，$\boldsymbol{W}$ 是一个可学习的转换矩阵，σ 是一个非线性激活函数，α_{ij} 为权重系数，表示细胞 j 对细胞 i 的重要性。

为了测量一个细胞对另一个细胞的重要性，本文使用最原始的 GAT[27] 网络将它们的特征连接起来获得注意力系数 e_{ij}，可以表示为：

$$e_{ij}=LeakyRELU(\boldsymbol{\sigma}^{\mathrm{T}}[\boldsymbol{W}h_i\|\boldsymbol{W}h_j]) \tag{7}$$

其中，$LeakyRELU$ 是一个非线性激活函数，$\boldsymbol{\sigma}$ 是一个可学习的权重向量，$\|$ 是连接运算符。

与 GAT 不同的是，本文将相似性信息整合到注意力系数中，注意力系数则通过高斯核变换两个细胞之间的距离来进行计算，这样既考虑了细胞间的重要性和对特征的充分学习，也考虑到了数据与聚类间的关系以及训练数据的分布对聚类的影响。计算方式如下：

$$e_{ij}=\exp(-|\boldsymbol{\sigma}_1^{\mathrm{T}}\boldsymbol{W}h_i-\boldsymbol{\sigma}_2^{\mathrm{T}}\boldsymbol{W}h_j|)^2 \tag{8}$$

其中，$\boldsymbol{\sigma}_1$ 和 $\boldsymbol{\sigma}_2$ 是细胞 i 及其邻近细胞 j 的可学习权重向量。

通常，注意力系数通过 softmax 函数进行归一化，以便在与其他不同的细胞之间进行比较时具有可比性，可以表示为：

$$\alpha_{ij}=\mathrm{softmax}_j(e_{ij})=\frac{\exp(e_{ij})}{\sum_{j\in N_i}\exp(e_{ik})} \tag{9}$$

与 GAT 类似，本文采用多头注意力来稳定学习过程，G 个独立的注意模块共同学习特征，可以表述为：

$$h_i'=\bigoplus_{g=1}^{G}\sigma(\sum_{j\in N_i}\boldsymbol{\alpha}_{ij}^g\boldsymbol{W}^g h_j) \tag{10}$$

其中，$\oplus$ 表示聚合操作。本文对自动编码器的前三层和最后一层分别使用拼接和平均函数进行聚合操作。

在网络训练过程中，我们将特征矩阵 $\boldsymbol{X}^1$ 和邻接矩阵 $\boldsymbol{A}$ 这两个矩阵作为输入，一起输入图注意力自编码器中，两者分别提供细胞与基因表达信息以及细胞间的拓扑信息。为了更好地约束学习过程，我们使用 MAE 来计算重构的特征矩阵 $\boldsymbol{X}^{1'}\in\mathbb{R}^{m'\times n}$ 与输入的特征矩阵 $\boldsymbol{X}^1$ 之间的重构损失。

$$L_r=\sum_{i=1}^{m'}\sum_{j=1}^{n}|\boldsymbol{X}_{ij}^1-\boldsymbol{X}_{ij}^{1'}| \tag{11}$$

2.4 自优化聚类

经过图注意力自编码器预训练后，特征矩阵 $\boldsymbol{Z}\in\mathbb{R}^{d\times n}$ 可以从细胞与细胞的关系以及细胞与基因表达的关系两个方面描述细胞，d 表示自编码器的瓶颈层大小。通过在 $\boldsymbol{Z}$ 上执行 K-means[28] 聚类算法，可以得到一个简单的聚类结果。但是，由于聚类模块与特征学习模块为两个互相独立的模块，因此本文采用了一种不断迭代自优化的聚类方法，使两个模块能够相互受益，以此来改进最终的聚类结果。

首先，本文采用学生 t-分布来测量细胞和聚类簇中心（由 K-means 初始化）之间的相似性，

如 DEC[29] 一样。在细胞水平上进行归一化操作后，隶属度矩阵 $\boldsymbol{Q}$ 可以表示为：

$$\boldsymbol{q}_{ij}=\frac{(1+\|Z_i-U_j\|)}{\sum_{k=1}^{c}(1+\|Z_i-U_k\|^2)^{-1}} \tag{12}$$

其中 Z_i 为细胞 i 的嵌入，U_j 为聚类簇中心 j 的嵌入，c 为细胞类型的数量。

然后，基于 $\boldsymbol{Q}$ 矩阵构造一个更优化的隶属度矩阵 $\boldsymbol{P}$，定义为：

$$\boldsymbol{P}_{ij}=\frac{\boldsymbol{q}_{ij}^2/\sum_{i=1}^{n}\boldsymbol{q}_{ij}}{\sum_{k=1}^{c}(\boldsymbol{q}_{ij}^2/\sum_{i=1}^{n}q_{ik})} \tag{13}$$

隶属矩阵 $\boldsymbol{P}_{ij}$ 被作为 $\boldsymbol{Q}$ 的目标，用于优化和重新分配细胞之间的关系拓扑图。在重新分配之后，平方项会使细胞间的隶属度分布更加明确。除此之外，总成员数较少的簇将会

获得更高的成员特征关系。

根据 $\boldsymbol{Q}$ 和 $\boldsymbol{P}$ 来更新嵌入的聚类中心和细胞嵌入，从而获得更好的聚类结果。一方面，为了更好地描述聚类，聚类中心的嵌入通过细胞的加权平均值进行更新，$\boldsymbol{Q}$ 作为权重，计算式如下：

$$U_j=\frac{\sum_{l_i=j}\boldsymbol{q}_{ij}Z_i}{\sum_{l_i=j}\boldsymbol{q}_{ij}} \tag{14}$$

其中，l_i 是迭代聚类过程中细胞 i 的聚类标签，$j\in(1,n)$，定义如下：

$$l_i=\arg\max_j(\boldsymbol{q}_{ij}) \tag{15}$$

另一方面，为了监督细胞间潜在特征的学习，增强底层的聚类结构，我们将 $\boldsymbol{Q}$ 和 $\boldsymbol{P}$ 之间的 KL 散度作为聚类损失，计算式如下：

$$L_c=\sum_{i=1}^{c}\sum_{j=1}^{n}\boldsymbol{P}_{ij}\log\frac{\boldsymbol{P}_{ij}}{\boldsymbol{q}_{ij}} \tag{16}$$

在训练过程中，本文基于潜在的特征学习来计算轮廓系数[29]，进而监测聚类的性能，细胞 i 的轮廓系数 s_i 为：

$$s_i=\frac{b_i-a_i}{\max(b_i-a_i)} \tag{17}$$

其中，a_i 为细胞 i 到所有包含它的簇中其他点的距离，b_i 表示细胞 i 到某一不包含它的簇内的所有点的平均距离。聚类结果的轮廓系数是将所有细胞的轮廓系数求平均，即该聚类结果总的轮廓系数，轮廓系数的值介于[−1,1]，越趋近于 1 代表内聚度和分离度都相对较优。细胞和集群中心的潜在特征将进行迭代的微调，直到轮廓系数收敛。

综上所述，总的损失函数计算式如下：

$$L=L_r+\gamma L_c \tag{18}$$

其中，L_r 是重构损失函数，L_c 是聚类损失，γ 为平衡两个损失函数的超参数。损失函数将潜在的特征学习和聚类集成到一个统一的框架，从而获得更好的最终聚类结果。

3 实验结果与分析

本文在 6 个真实数据集上将 scZDGAC 与 10 种聚类方法进行了对比，结果如表 1 和表 2 所列。scZDGAC 在 ARI 和 NMI 评分方面表现更好，在 5 个数据集上的两个聚类指标都高于其他聚类方法。实验环境：内存 32 GB，处理器 Intel Corei7-10700F@2090 GHz 八核，Windows 10 操作系统。

表 1 不同聚类方法在 6 个数据集上的 ARI 指标

Table 1 ARI values of different clustering methods on six datasets

methods	Zeisel	Alts1	Altas1se	GSE130114	Macaque	Mouse
DESC	0.6307	0.5691	0.5563	0.6355	0.5529	0.3617
scDRHA	0.6270	0.6040	0.5183	0.6916	0.5406	0.6863
SC3	0.5894	0.5048	0.4355	0.6958	0.4437	0.5864
scGAE	0.6331	0.5812	0.6958	0.6459	0.4128	0.6342
scGNN	0.6973	0.4080	0.6141	0.6686	0.5657	0.6511
scGAC	0.7237	0.4626	0.6018	0.7375	0.6331	0.7130
Seurat	0.5901	0.4674	0.3885	0.6626	0.4423	0.4128
scDeepCluster	0.5391	0.3316	0.4752	0.5880	0.5671	0.6342
scZiDESK	0.6545	0.4752	0.4913	0.6507	0.5400	0.5941
SCVI	0.4211	0.4474	0.5108	0.5730	0.4366	0.5160
scZDGAC	**0.8541**	**0.6302**	**0.7115**	**0.7905**	0.5786	**0.8337**

表 2 不同聚类方法在 6 个数据集上的 NMI 指标

Table 2 NMI values of different clustering methods on six datasets

methods	Zeisel	Alts1	Altas1se	GSE130114	Macaque	Mouse
DESC	0.7042	0.6363	0.6298	0.6754	0.6863	0.5365
scDRHA	0.7270	0.6874	0.6021	0.7126	0.6269	0.6755
SC3	0.6712	0.5941	0.6234	0.7088	0.6413	0.6841
scGAE	0.6615	0.6462	0.7008	0.6636	0.5489	0.6592
scGNN	0.7063	0.5800	0.6870	0.7342	0.6249	0.6940
scGAC	0.7359	0.5382	0.6371	0.7144	0.6615	0.7147
Seurat	0.6576	0.6173	0.5259	0.7032	0.5015	0.5489
scDeepCluster	0.6256	0.4840	0.6184	0.5930	0.6889	0.6592
scZiDESK	0.6792	0.6184	0.5681	0.6753	0.6584	0.6513
SCVI	0.5654	0.5974	0.5692	0.5752	0.5331	0.6303
scZDGAC	**0.8112**	**0.7184**	**0.7177**	**0.7615**	**0.7131**	**0.7550**

3.1 细胞嵌入分析

对于基于深度学习的 scZDGAC，SCVI[30]，scDeepCluster，DESC，scGAE，scGNN，scDRHA[26] 和 scGAC 等方法，在自编码器的瓶颈层获得细胞间潜在特征作为（scGNN 的特征自编码器）细胞嵌入。对于 SC3 和 Seurate 两种聚类方法，由于没有合适的细胞嵌入，因此不能进行嵌入分析。为了量化细胞嵌入的簇内纯度与簇间分化，本文根据标注的细胞类型标签计算轮廓系数（如式(17)中定义）。为了将聚类结果进行可视化，在细胞嵌入过程中，如果获得的嵌入维度超过 256，则使用 PCA 将数据降维到 256 维，然后通过 t-分布随机领域嵌入（T-SNE）[31] 进一步将数据维数减少到 2 维进行展示，二维可视化平面上的点将根据其标注的细胞类型进行着色。

3.2 聚类数目与 K 参数设置

在聚类过程中首先对数据进行初始化聚类，该过程借鉴 Adaptive K-means[32] 聚类思想来确定最佳的聚类数目。在聚类过程中引入 Davies-BouldinIndex（DBI）指标来确定聚类个数。DBI 值越小，代表各聚类内的数据对象关系越紧密且聚类间的差异大，表明此聚类数目下的聚类结果最佳，则选择该聚类结果下的聚类数目作为最终确定的聚类数。

对于 scZDGAC，用于构造图邻接矩阵的默认邻接数 K 的计算式如下：

$$K=round\left(\frac{N_{\text{cells}}}{10\times N_{\text{clusters}}}\right) \tag{19}$$

其中，N_{cells} 表示细胞数量；N_{clusters} 表示聚类数目，取值在 6～20 之间。

3.3 评价指标

该实验使用调整兰德指数 ARI[33] 和标准化互信息 NMI[34] 这两个通用的评价指标来评估 scZDGAC 和对比方法的聚类性能。ARI 的取值范围为[−1,1]，值越大表示聚类性能越好；NMI 的取值范围为[0,1]，值越接近于 1 表示聚类结果越好。

ARI 用来度量聚类结果与真实标签值之间的相似性，计算式如下：

$$ARI=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}} \tag{20}$$

其中，n_{ij} 表示聚类 i 初始真实值和聚类结果 j 之间共享的细胞数，n 表示总的细胞个数，$a_i=\sum_j n_{ij}$，$b_j=\sum_i n_{ij}$。

NMI 用于测量聚类结果和真实值之间的归一化互信息，

其计算式为：

$$NMI=\frac{2MI(U,V)}{H(U)+H(V)} \tag{21}$$

其中，U 表示聚类结果，V 表示真实值，$MI(U,V)$ 表示 U 和 V 之间的互信息，H 表示交叉熵。

为了验证所提方法的有效性，本文将 scZDGAC 方法和最近的 10 种聚类方法进行了对比，并取 ARI 和 NMI 的平均值进行估计。

3.4 数据集和预处理

本文实验在 6 个真实数据集上进行了评估，如表 3 所列，这些数据集的大小不同，并且来自于不同的组织与批次。例如，数据集 GSE130114 包含 1453 个细胞、30792 个基因和 12 种细胞类型，这些数据细胞来自小鼠皮质组织；Zeisel 数据集包含 3005 个细胞、19972 个基因和 9 类细胞类型，这些数据细胞来自于小鼠的皮质和海马两个不同组织。

表 3 6 个真实数据集的基本信息

Table 3 Basic information of six real datasets

数据信息	Zeisel	Alts1	Altas1se	GSE130114	Macaque	Mouse
Cell	3005	3500	8500	1453	2157	2187
Gene	19972	29452	139832	30792	39608	39671
Cell type	9	17	19	12	10	16

在进行聚类操作之前，首先，将计数矩阵 $\boldsymbol{H}\in\mathbb{R}^{M'\times N'}$ 进行预处理，过滤掉在任何基因中都没有表达的基因，然后对计数数据进行对数转换和 Z 分数归一化处理，则有一个标准化输出的 X，它由下面的计算式计算得到：

$$Y'=\log(1+diag(S_{j\in(1,n)})^{-1}\boldsymbol{H}) \tag{22}$$

$$X=zscore(Y') \tag{23}$$

其中，$diag$ 表示对角化；S_j 表示每个细胞因子的大小，它是总细胞数与中位数 S 的比值。该数据处理的优点是可以保留数据集大小的差异，并将离散值转换为连续值，从而为后续模型的训练提供更大的灵活性。

3.5 对比实验

为了进一步验证 scZDGAC 算法的聚类性能，本文将所提方法与其他 10 种聚类方法进行了比较分析。本文将这些方法分为传统的聚类方法、基于深度神经网络的聚类方法和基于 GNN 的聚类方法。

在比较分析中，使用了两个通用的聚类评价性能指标（ARI 和 NMI）来评估每种聚类方法的性能，图 2 给出了这些聚类方法在 6 个真实的单细胞数据集上的聚类性能；并且由图 3 所示的折线图可以明显看出，scZDGAC 在所有数据集中的平均得分最高。

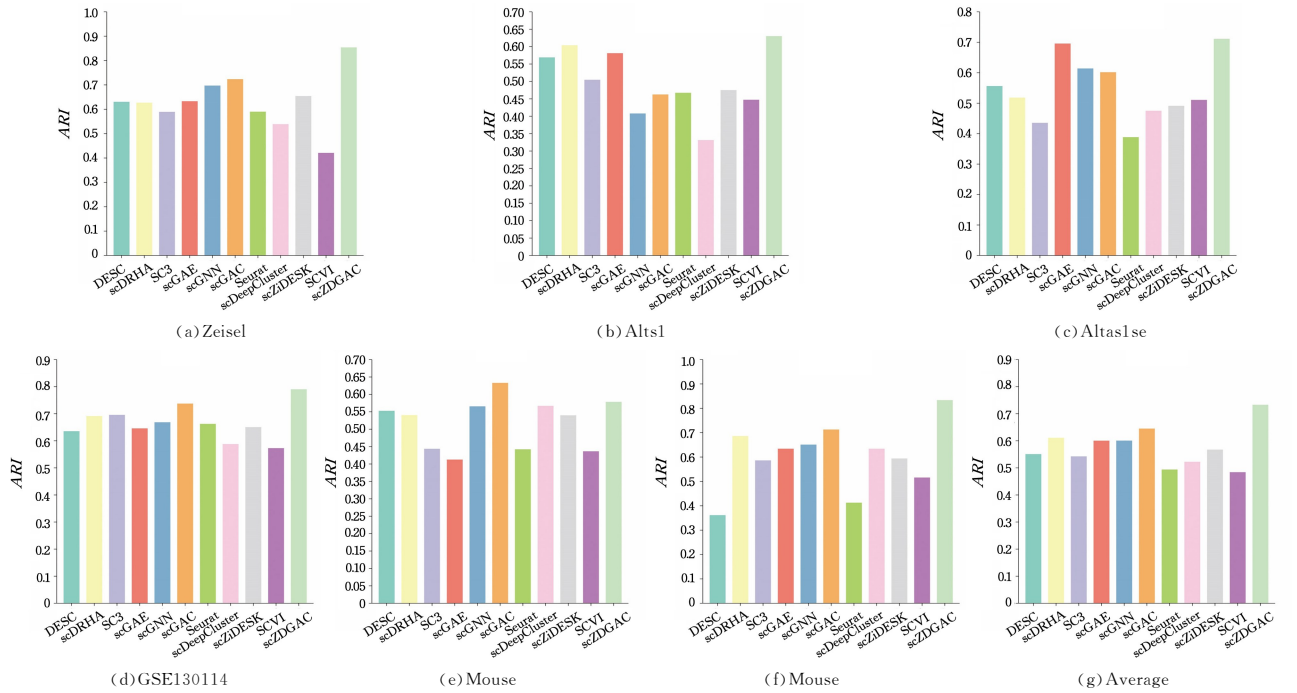


图 2 scZDGAC 与其他 10 种聚类算法的对比

Fig. 2 Comparison between scZDGAC and other 10 clustering algorithms

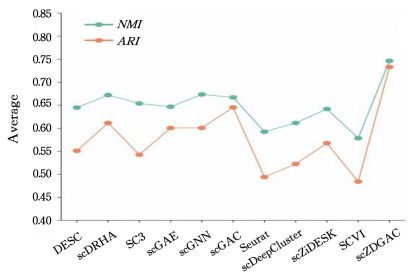


图 3 11 种聚类方法在所有数据集上的平均指标值

Fig. 3 Average index value of 11 clustering methods on all datasets

由图 2 可以观察到，scZDGAC 比其他 10 种对比方法表现得更好，具体来说，其在 Zeisel，Alts1，Altas1se，GSE130114 和 Mouse 这 5 个数据集上的 ARI 和 NMI 的结果值明显优于其他几种聚类方法。特别是在 Zeisel 数据集上，与 scGAC 相比，scZDGAC 在 ARI 上提升了 13%，NMI 提升了 7.5%。在 GSE130114 数据集上与次优方法相比，scZDGAC 在 ARI 上提升了 5.3%，在 NMI 上提升了 4.71%。在 Mouse 数据集上与 scGAC 相比，scZDGAC 在 ARI 上提升了 12%，在 NMI 上提升了 4.03%。为了直观地展示聚类效果，验证该模型在

去噪和提取高维数据低维特征方面及聚类方面的有效性，本文利用 t-SNE 将自优化聚类后的结果投影到二维空间中进行可视化。图 4 给出了 scZDGAC 已识别的细胞类型和 10 种对比方法的可视化图，本文选择了数据集 Zeisel 具有代表性的聚类结果可视化图，每个点代表一个细胞类型，不同的颜色表示预测的不同的细胞。

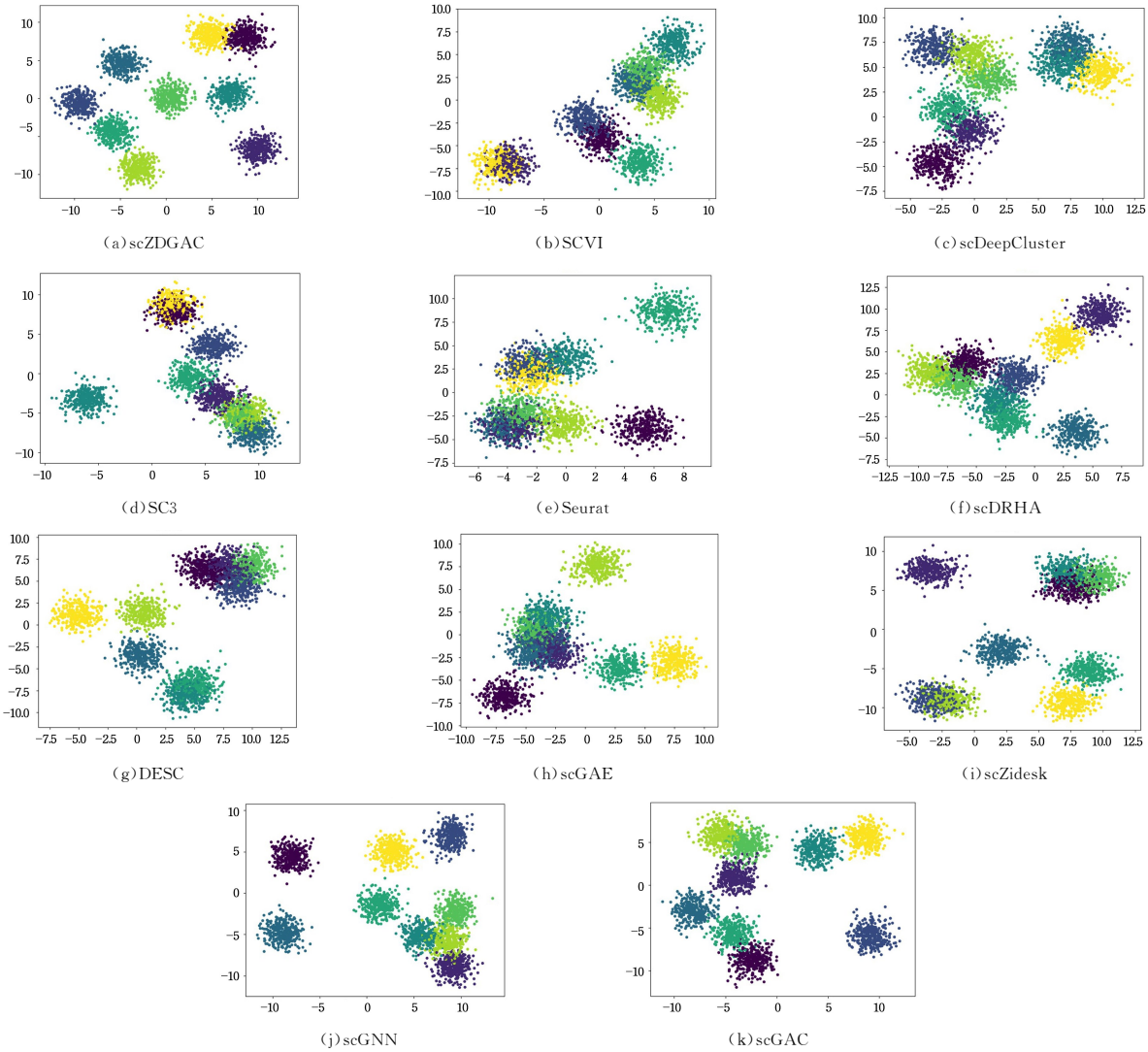


图 4 Zeisel 数据集的可视化图

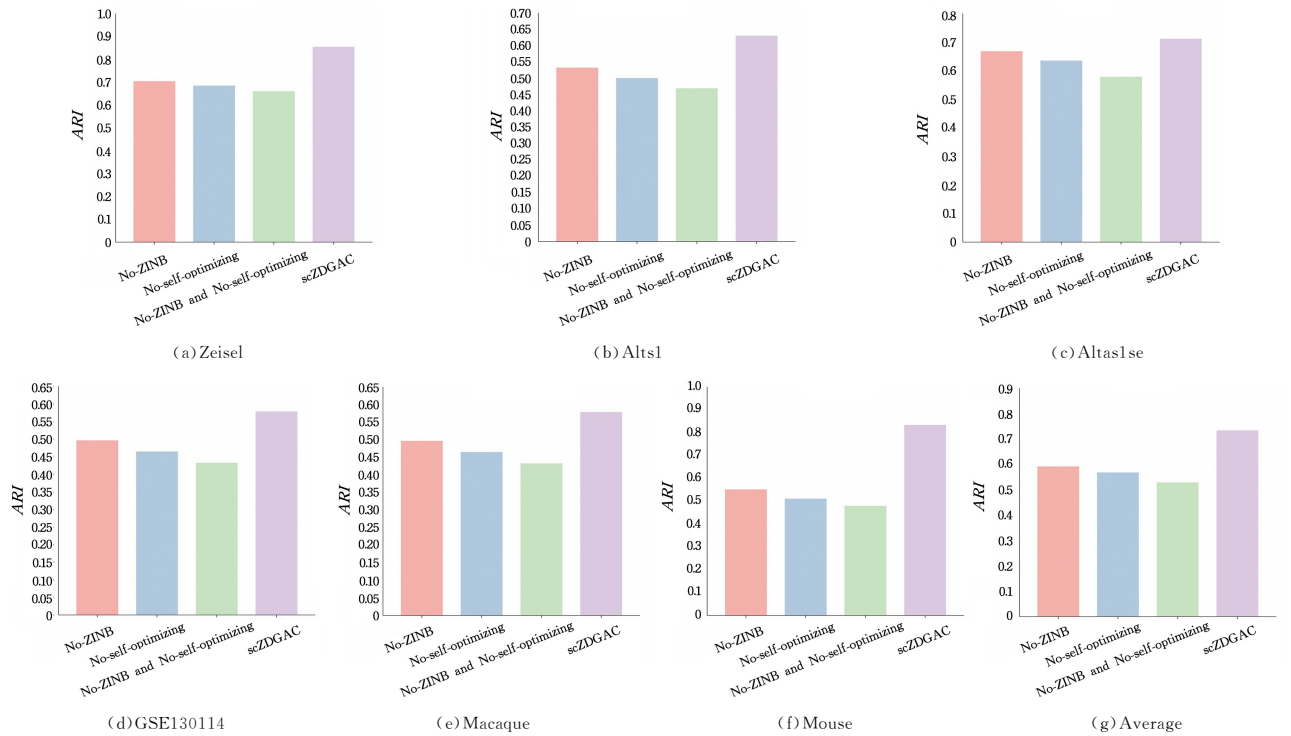
Fig. 4 Visualization of Zeisel dataset

3.6 消融实验

本文提出的 scZDGAC 模型主要由细胞特征去噪模块 ZINB、图注意力自编码器和自优化聚类模块构成。因此，本文设计消融实验来探讨细胞特征去噪模块 ZINB 和自优化模块对 scZDGAC 聚类性能的贡献，使用了 scZDGAC 的 3 种变体：1）只有图注意力自编码器和自优化聚类两个模块组成的网络架构；2）只有细胞特征去噪模块 ZINB 和图注意力自编码器两模块组成的网络；3）只有图注意力自编码器一个模块的网络。通过这 3 个消融实验来分别验证细胞特征去噪模块 ZINB 和自优化聚类模块对聚类性能的重要性。

如图 5 所示，scZDGAC 与其他 3 种变体相比取得了较好的聚类性能，结果表明细胞特征去噪模块 ZINB 和自优化聚类模块对整个模型的性能发挥了重要作用。在没有细胞特征去噪模块 ZINB 时，其在 6 个真实数据集上的聚类性能较差，原因可能是噪声未去除而导致所构造的 KNN 图错误率相对较高且不稳定，该 KNN 图在进一步使用图注意力自编码器在潜在特征空间中进行细胞种群结构信息与特征学习时包含了更多的错误信息，导致聚类性能较差。因此，为图注意力自编码器模块提供高质量的 KNN 图是非常重要的。

为了进一步验证自优化聚类模块对聚类性能的重要性，本文采用只有细胞特征去噪模块 ZINB 和图注意力自编码器构成的网络结构在 6 个真实数据集上进行实验验证。由图 5 可以看出，其在 6 个真实数据集上的聚类性能较差，原因是在图注意力自编码器中，聚类模块与特征学习模块相互独立，采用迭代自优化聚类方法，可以使两个模块相互受益，不断优化聚类中心并重新分配成员关系，并帮助图注意力自编码器更新学习到的嵌入，使其更具确定性，最终的聚类效果更好。最后，将细胞特征去噪模块 ZINB 和自优化聚类模块全部去除，只保留图注意力自编码器模块，在 6 个真实数据集上进行验证，由图 5 可以看出，与上面的两种变体网络结构相比，此变体网络的聚类性能较差。通过实验表明，细胞特征去噪模块 ZINB 和自优化聚类模块对聚类结果十分重要。



(a) Zeisel (b) Alts1 (c) Altas1se

(d) GSE130114 (e) Macaque (f) Mouse (g) Average

注：(g)为 scZDGAC 及 3 种变体在 6 个数据集上的平均 ARI 得分。

图 5 scZDGAC 及其他 3 种不同变体的性能比较(ARI)

Fig. 5 Performance comparison(ARI) of scZDGAC and three different variants

结束语 单细胞 Seq-RNA 数据的聚类分析是后续进行单细胞序列分析的一个关键步骤。但是，单细胞数据在进行聚类分析时受到高维稀疏性、高度可变性和高噪声等的影响，使得单细胞聚类在下游分析中变得复杂化，并导致了对结果的错误解释。本文提出了一种单细胞聚类的新方法 scZDGAC。scZDGAC 首先在细胞特征去噪模块 ZINB 部分对单细胞的基因数据分布进行建模，并结合自编码器与可拓展 DCA 去噪算法对原始的表达矩阵进行去噪处理，通过 ZINB 分布提高自编码器的去噪性能；然后利用图注意力机制在相似的细胞间进行信息共享，以充分利用细胞与基因表达间以及细胞与细胞之间的潜在特征关系，学习聚类效果友好的细胞嵌入；最后利用自优化聚类模块使聚类模块与特征学习模块互相交互，并更新学习到的嵌入，优化聚类中心并重新分配成员关系，使其更具确定性，聚类性能更好。

实验结果表明，scZDGAC 在 6 个真实数据集上取得了更好的聚类性能，优于其他 10 种专门针对单细胞数据集设计的聚类方法。此外，scZDGAC 在真实单细胞数据集聚类问题上展现出良好的鲁棒性，无论是在数据集细胞数量大小、细胞类型和不同组织间提取的细胞数据以及时间批次之间的差异上，该方法相较于其他的聚类方法都能取得更好的聚类结果。

ScZDGAC 的一个局限性在于它的运行效率。scZDGAC 的运行时间会随着数据集中细胞数量的增加而显著增加，值得注意的是，其他考虑成对细胞间关系的方法，如 scGNN，效率也低。因此，进一步优化 scZDGAC 在大规模单细胞数据上的聚类效率对于算法的实际应用十分关键，接下来的研究将考虑在利用细胞之间关系的同时加速学习过程。

参 考 文 献

[1] HWANG B, LEE J H, BANG D. Single-cell RNA sequencing technologies and bioinformatics pipelines[J]. Experimental & Molecular Medicine, 2018, 50(8): 1-14.

[2] GUO M, DU Y, GOKEY J J, et al. Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth[J]. Nature Communications, 2019, 10(1): 1-16.

[3] HU H, LI Z, LI X, et al. ScCAEs: deep clustering of single-cell RNA-seq via convolutional autoencoder embedding and soft K-means[J]. Briefings in Bioinformatics, 2022, 23(1): bbab321.

[4] MACOSKO E Z, BASU A, SATIJA R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets[J]. Cell, 2015, 161(5): 1202-1214.

[5] ANGERER P, SIMON L, TRITSCHLER S, et al. Single cells make big data: New challenges and opportunities in transcriptomics[J]. Current Opinion in Systems Biology, 2017, 4: 85-91.

[6] WANG B, ZHU J, PIERSON E, et al. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning [J]. Nature Methods, 2017, 14(4): 414-416.

[7] SATIJA R, FARRELL J A, GENNERT D, et al. Spatial reconstruction of single-cell gene expression data[J]. Nature Biotechnology, 2015, 33(5): 495-502.

[8] LIN P, TROUP M, HO J W K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data[J]. Genome Biology, 2017, 18(1): 1-11.

[9] MEI Q, LI G, SU Z. Clustering single-cell RNA-seq data by rank constrained similarity learning[J]. Bioinformatics(Oxford, England), 2021, 37(19): 3235-3242.

[10] KISELEV V Y, KIRSCHNER K, SCHAUB M T, et al. SC3: consensus clustering of single-cell RNA-seq data[J]. Nature Methods, 2017, 14(5): 483-486.
[11] YANG Y, HUH R, CULPEPPER H W, et al. SAFE-clustering: single-cell aggregated(from ensemble) clustering for single-cell RNA-seq data[J]. Bioinformatics (Oxford, England), 2019, 35(8): 1269-1277.
[12] HU H R, YANG Y, JIANG Y, et al. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble[J]. Nucleic Acids Research, 2020, 48(1): 86-95.
[13] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[14] ERASLAN G, AVSEC Ž, GAGNEUR J, et al. Deep learning: new computational modelling techniques for genomics[J]. Nature Reviews Genetics, 2019, 20(7): 389-403.
[15] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science(New York), 2006, 313(5786): 504-507.
[16] TIAN T, WAN J, SONG Q, et al. Clustering single-cell RNA-seq data with a model-based deep learning approach[J]. Nature Machine Intelligence, 2019, 1(4): 191-198.
[17] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]// International Conference on Machine Learning. PMLR, 2016: 478-487.
[18] LI X, WANG K, LYU Y, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis[J]. Nature Communications, 2020, 11(1): 1-14.
[19] CHEN L, WANG W, ZHAI Y, et al. Deep soft K-means clustering with self-training for single-cell RNA sequence data[J]. NAR Genomics and Bioinformatics, 2020, 2(2): lqaa039.
[20] GAN Y, HUANG X, ZOU G, et al. Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network[J]. Briefings in Bioinformatics, 2022, 23(2): bbac018.
[21] CHENG Y, MA X. scGAC: a graph attentional architecture for clustering single-cell RNA-seq data[J]. Bioinformatics(Oxford, England), 2022, 38(8): 2187-2193.
[22] BO D, WANG X, SHI C, et al. Structural deep clustering network[C]// Proceedings of the Web Conference 2020. 2020: 1400-1410.
[23] WANG J, MA A, CHANG Y, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses[J]. Nature Communications, 2021, 12(1): 1-11.
[24] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[J]. arXiv: 1609.02907, 2016.
[25] ERASLAN G, SIMON L M, MIRCEA M, et al. Single-cell RNA-seq denoising using a deep count autoencoder[J]. Nature Communications, 2019, 10(1): 1-14.
[26] ZHAO J, WANG N, WANG H, et al. SCDRHA: A scRNA-Seq Data Dimensionality Reduction Algorithm Based on Hierarchical Autoencoder[J]. Frontiers in Genetics, 2021, 12: 733906.
[27] VELICKOVIC P, CUCURULL G, Casanova A, et al. Graph attention networks[J]. arXiv: 1710.10903, 2017.
[28] HARTIGAN J A, WONG M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the Royal Statistical Society, Series c(Applied Statistics), 1979, 28(1): 100-108.
[29] ROUSSEEUW P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.
[30] LOPEZ R, REGIER J, COLE M B, et al. Deep generative modeling for single-cell transcriptomics[J]. Nature Methods, 2018, 15(12): 1053-1058.
[31] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 2579-2605.
[32] TANG Y W. Research on an adaptive clustering Algorithm based on K-Means [J]. Science and Technology Wealth Guide, 2012(2): 143-143.
[33] HUBERT L, ARABIE P. Comparing partitions [J]. Journal of Classification, 1985, 2(1): 193-218.
[34] STREHL A, GHOSH J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions[J]. Journal of Machine Learning Research, 2002, 3(Dec): 583-617.

KONG Fengling, born in 1997, postgraduate, is a student member of China Computer Federation. Her main research interests include biological information technology, image processing, etc.

WU Hao, born in 1982, Ph.D, lecturer, is a senior member of China Computer Federation. His main research interests include image processing, computer vision and bioinformatics analysis, etc.

(责任编辑:何杨)