计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(9):1−13 [doi: 10.15888/j.cnki.csa.009638]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

联合对比学习与图神经网络的自优化单细胞聚类①

蒋维康, 王劲贤

(复旦大学 计算机科学技术学院, 上海 200438)
通信作者: 王劲贤, E-mail: wangjx22@m.fudan.edu.cn

摘　要: 单细胞 RNA 测序技术 (single-cell RNA sequencing, scRNA-seq) 在单个细胞的水平上对转录组进行高通量测序分析, 其核心应用是识别具有不同功能的细胞亚群, 通常基于细胞聚类来完成. 然而, scRNA-seq 数据高维度、高噪声、高稀疏的特点使得聚类充满挑战. 常规的聚类方法表现不佳, 现有的单细胞聚类方法也大多只考虑基因的表达模式, 而忽略了细胞之间的关系. 针对这些问题, 提出了一个联合对比学习与图神经网络的自优化单细胞聚类方法 (self-optimizing single-cell clustering with contrastive learning and graph neural network, scCLG). 该方法采用自编码器来学习细胞的特征分布. 首先构建细胞-基因图, 使用图神经网络进行编码, 以有效利用细胞之间的关系信息. 通过子图采样和特征掩码获取增广视图用于对比学习, 进一步优化特征表示. 最后使用自优化的策略将聚类模块和特征模块联合训练, 不断优化特征表示和聚类中心, 实现更准确的聚类. 在 10 个真实的 scRNA-seq 数据集上的实验表明, scCLG 能够学习到细胞特征的良好表示, 在聚类精度上全面优于其他方法.

关键词: 单细胞 RNA 测序; 聚类; 对比学习; 图神经网络; 自编码器

引用格式: 蒋维康,王劲贤.联合对比学习与图神经网络的自优化单细胞聚类.计算机系统应用,2024,33(9):1–13. http://www.c-s-a.org.cn/1003-3254/9638.html

Self-optimizing Single-cell Clustering with Contrastive Learning and Graph Neural Network

JIANG Wei-Kang, WANG Jin-Xian

(School of Computer Science, Fudan University, Shanghai 200438, China)

Abstract: Single-cell RNA sequencing (scRNA-seq) performs high-throughput sequencing analysis of the transcriptomes at the level of individual cells. Its primary application is to identify cell subpopulations with distinct functions, usually based on cell clustering. However, the high dimensionality, noise, and sparsity of scRNA-seq data make clustering challenging. Traditional clustering methods are inadequate, and most existing single-cell clustering approaches only consider gene expression patterns while ignoring relationships between cells. To address these issues, a self-optimizing single-cell clustering method with contrastive learning and graph neural network (scCLG) is proposed. This method employs an autoencoder to learn cellular feature distribution. First, it begins by constructing a cell-gene graph, which is encoded using a graph neural network to effectively harness information on intercellular relationships. Subgraph sampling and feature masking create augmented views for contrastive learning, further optimizing feature representation. Finally, a self-optimizing strategy is utilized to jointly train the clustering and feature modules, continually refining feature representation and clustering centers for more accurate clustering. Experiments on 10 real scRNA-seq datasets demonstrate that scCLG can learn robust representations of cell features, significantly surpassing other methods in clustering accuracy.

Key words: single-cell RNA sequencing (scRNA-seq); clustering; contrastive learning; graph neural network (GNN); autoencoder

① 基金项目: 国家自然科学基金 (61972100)
收稿时间: 2024-03-28; 修改时间: 2024-04-23; 采用时间: 2024-05-09; csa 在线出版时间: 2024-07-26
CNKI 网络首发时间: 2024-07-29

作为生命体的基本结构和功能单位, 细胞储存着大量的生物信息. 在细胞的生长发育过程中, 诸多因素使得细胞转录信息趋向多样化, 激发了生物组织内细胞异质性的发生[1]. 近年来, 高通量测序技术[2]不断发展, 通过细胞的分子生物学特征识别细胞类型成为可能. 与传统批量测序 (bulk-seq) 技术[3]相比, 单细胞测序技术[4]允许在单一细胞层面下对基因组、转录组及表观组进行分析, 能准确呈现单细胞的基因构建及其表达状况, 从而揭示细胞间的异质性. 单细胞 RNA 测序技术 (scRNA-seq) 是单细胞测序技术的重要发展方向[5], 对单个细胞的转录组进行测序, 能够掌握每个细胞的变化, 探索基因调控机制, 发现新的细胞类型, 揭示复杂的生物学关联与过程.

scRNA-seq 技术的核心应用在于鉴定和表征细胞群体, 这通常需要通过细胞聚类来完成. 然而由于技术的不完善以及生物体本身的复杂多变性[6] (如测序深度浅, 捕获率低, 批次效应, 细胞状态变化等), 导致 scRNA-seq 数据整体呈现出高维度、高噪声、高变异性、高稀疏性和高度不平衡的特征, 这给传统的聚类方法带来了极大的挑战, 迫切需要研究新的方法.

针对 scRNA-seq 数据的聚类方法得到了广泛的研究, 目前已经有大量的相关工作. 早期的聚类方法主要使用传统的机器学习技术. pcaReduce[7]使用 K-means 聚类, 并结合 PCA 来进行降维, 识别簇相应的主成分并构建簇之间层次关系. Kiselev 等人提出了算法 SC3[8], 该算法对多种不同度量方法下的细胞距离矩阵数据应用 PCA 和拉普拉斯变换进行降维, 之后对所有主成分进行 K-means 聚类, 最后对所有的结果进行聚类集成. SIMLR[9]提出使用多个内核来学习样本相似性并在相似性矩阵上执行谱聚类. 为了发现和利用数据中的结构关系, 一些方法尝试用图来描述 scRNA-seq 数据, 细胞作为节点, 而边则代表细胞之间的关联性, 之后采用图聚类算法进行聚类. 2015 年, Xu 等人提出了一种基于准团 (clique) 的聚类方法 SNN-Cliq[10], 能够发现稀疏数据中的紧凑簇. SNN-Cliq 首先在细胞上构建共享最近邻居图, 之后在图上采用准团检测来识别细胞群体. Seurat[11]也使用共享最近邻居图来描述细胞之间的相似度, 并利用基于图的社区检测算法 Louvain[12]对共享最近邻网络进行分割并生成不相交的聚类.

尽管这些方法采用各种手段实现了对 scRNA-seq 数据的聚类, 但是其所使用模型的固有表达能力往往有限, 导致模型在各种不同组织, 不同平台和不同规模的数据[13]上的泛化能力不足.

近年来, 深度学习技术发展迅速, 其灵活强大的计算模型使其展现出相比于传统机器学习技术更优异的性能, 在包括计算生物学在内的各个学科领域中得到了广泛的应用[14,15]. 目前, 在 scRNA-seq 数据分析领域已经有了许多的基于深度学习的方法实践[16].

为了处理 scRNA-seq 数据中普遍存在的丢失事件 (dropout events), Eraslan 等人提出了一个深度计数自编码器模型 DCA[17]. 该模型把零膨胀的负二项分布模型 (ZINB)[18]与自编码器结构结合起来, 使用 ZINB 分布的负对数似然作为训练损失, 而不是传统的均方误差 (mean square error) 重构损失, 有效地实现了 scRNA-seq 数据的去噪和特征表示, 并在此基础上进行聚类. 深度嵌入聚类 (DEC)[19]是一个基于自编码器的聚类模型, 它使用学生 t 分布 (student's t-distribution) 对聚类分布进行建模, 将原分布和目标分布之间的 KL 散度 (Kullback-Leibler divergence) 作为聚类损失, 该损失能直接地为聚类任务优化特征表示, 并且由此实现了特征表示和聚类分配的同步优化. Tian 等人提出的模型 scDeepCluster[20]是对方法 DCA 和 DEC 的结合与优化, 一方面使用去噪的 ZINB 自编码器, 获得鲁棒的数据表示; 另一方面, 采用和 DEC 类似的聚类损失, 直接为聚类任务优化特征表示. scziDesk[21]在 scDeepCluster 的基础上做了进一步的优化, scziDesk 采用加权的软 K-means 算法进行聚类, 并在聚类过程中添加了成对的亲和力约束. 以上这些方法采用自编码器学习细胞的编码表示, 相比传统方法, 能得到更有意义的特征, 但是它们往往只针对基因表达信息 (细胞与基因的关系) 进行建模, 并未考虑细胞与细胞之间的结构关系信息, 这可能会影响类簇识别的准确性.

图神经网络 (graph neural network, GNN)[22]是深度学习领域目前热门的研究方向, 这主要得益于其在处理非结构化非欧氏空间的数据上的巨大优势. 此外, 传统的深度学习的一个核心假设是数据样本之间彼此独立, 然而实际的数据当中往往可能之间存在各种联系, 图神经网络能够充分利用图结构中所描述的节点与节点的关系信息, 这些结构关系信息将有助于模型学习到更深层次的数据特征. 就 scRNA-seq 数据而言, 由于测序的细胞往往来自同一个组织或器官, 细胞之间保持独立的可能性不大, 细胞间的关系是值得关注的重

要信息. 为了利用细胞与细胞之间的关系信息, 一些研究开始在 scRNA-seq 数据上构建相应的图结构, 并使用图神经网络进行建模. scGNN[23]首先使用一个特征自编码器获取细胞的特征表示, 并由此构建细胞-细胞图, 经过一个图自编码器以捕捉细胞之间的关系信息, 并使用特定于细胞类型的聚类自编码器帮助模型发现细胞类别相关的信息. GraphSCC[24]通过使用图卷积神经网络把细胞之间的结构关系整合到 scRNA-seq 聚类中, 它还利用双重自监督模块对细胞进行聚类并指导训练过程. Li 等人提出的 ScGSLC[25]是一个基于图相似性学习的单细胞聚类框架, 它将 scRNA-seq 数据和蛋白质-蛋白质相互作用网络有效整合成一个图, 使用图卷积网络对图进行嵌入, 并计算图之间的相似性来聚类细胞. scDSC[26]通过 K 近邻 (KNN) 算法构建细胞-细胞图, 利用图神经网络以更好地捕捉细胞之间的关系信息, 结合基于 ZINB 模型的自编码器模块学习基因的表达模式. scGAC[27]应用网络去噪改善构建的细胞图, 然后通过图注意力自编码器学习细胞表示, 该编码器以不同的权重传递细胞间信息, 捕捉细胞间潜在的关系, 最后以自优化方法确定细胞簇. Yu 等人提出了一个针对 scRNA-seq 数据的深度图嵌入聚类方法 scTAG[28], 该方法使用多核卷积拓扑图神经网络 (TAGCN) 利用图重构学习细胞的低维特征表示, 并在此基础上执行深度聚类. 这些方法的实践展现出引入细胞之间的关系信息对于聚类任务的重要性, 但是 scRNA-seq 数据具有高维稀疏, 高噪声的特点, 在此基础上计算得到的细胞之间的图结构可能质量较低, 存在较多噪声, 从而影响最终的聚类结果.

在自监督学习中, 对比学习是一种简明且有效的学习范式, 近年来发展迅速, 在自然语言处理, 计算机视觉等领域已经得到了广泛的研究与应用[29–32], 成为学习词句或图像表示的有效方法. 对比学习要义在于识别同类实例的共同属性和鉴别异类实例的差异性, 其核心目标是训练一个编码器, 该编码器能够为同类数据生成相似的编码, 而为不同类数据生成差异较大的编码. 在大量的实验中, 对比学习展现出优异的性能, 有时甚至优于有监督的方法[33]. 对比学习在处理 scRNA-seq 数据方面同样有着独特优势, 一方面其特性与聚类任务相适应, 另一方面, 与基于重构的特征学习方法相比, 对比学习只需在特征空间中对样本进行比较, 这在一定程度上部分规避了原始数据中的稀疏性和噪声的影响. Contrastive-sc[34]是一种新颖的针对 scRNA-seq 数据的对比学习聚类方法, 通过掩蔽一定比例的数据特征来获得增广数据. 与大多数对比学习的实践做法类似, 该方法将来自同一个原始样本的两个增广样本视为正样本, 而将所有其他样本对视为负样本. Wan 等人提出的 scNAME[35]方法改进了传统的对比损失, 提出了一种新的邻域对比损失, 将正样本的范围扩大到邻域, 同时结合一个辅助的掩码估计任务, 更好地描述了特征之间的相关性和细胞的成对相似性. 借鉴经典对比学习框架 MoCo[30]的做法, CLEAR[36]采用多种数据增强方法来模拟不同的噪声类型, 使用 InfoNCE[37]损失作为对比损失, 并结合编码器的动量更新策略为 scRNA-seq 数据生成良好的特征表示. scDCCA[38]是一种新型的深度对比聚类算法, 通过基于零膨胀负二项分布模型的去噪自编码器提取低维特征, 同时引入双重对比学习模块来捕捉细胞间的成对相似性. Xiong 等人提出了一个基于图对比学习的 scRNA-seq 数据插补方法 scGCL[39], 它通过对比学习总结全局和局部的语义信息, 并选取正样本来增强目标节点的表示.

由于 scRNA-seq 数据高维度、高稀疏、高噪声、高变异等特性, 细胞聚类一直充满挑战. 近年来, 深度学习技术特别是图神经网络和对比学习的广泛应用, 为聚类方法研究开辟了新的思路. 图神经网络能够有效利用细胞之间的关系信息, 这是很多模型所忽略的. 对比学习则能够在嘈杂的数据分布中学习到更加稳健的特征表示. 本文提出了一种联合对比学习和图神经网络的自优化单细胞聚类方法 scCLG. 该方法一方面利用图神经网络对构建的细胞-基因图进行建模, 不仅考虑了基因的表达模式, 也能够有效地挖掘细胞之间的潜在特征关系; 另一方面使用对比学习加强模型训练, 减轻了原始数据中噪声的干扰, 从而学习到对聚类更加友好的特征表示. 具体而言, scCLG 的主要工作如下.

(1) 对 scRNA-seq 数据构建细胞-基因图, 尽可能保留原始数据中的信息, 通过图神经网络对图结构进行编码, 综合得到细胞的特征表示, 并在解码器中引入 ZINB 模型提供对数据分布结构的约束.

(2) 通过子图采样和特征掩码的方式获取图的视图及其增广视图. 两个视图经过图编码器得到细胞的特征, 基于两个视图特征的差异在潜空间中进行对比学习, 这将指导编码器学习到更加清晰的特征表示.

(3) 使用一种自优化的策略, 将聚类中心作为参数, 同步优化特征表示和聚类中心, 进一步提升聚类效果.

scCLG 的创新主要在于图的构建方式和对比学习的策略. 首先, 选择构建细胞-基因图而非传统的细胞-细胞图, 能更准确地刻画原始数据. 其次, 采用了一种改进的对比学习策略, 使模型能学习到与聚类直接相关的特征. 为了验证该方法的有效性, 本文在多个真实的 scRNA-seq 数据集上进行了实验, 并与其他先进方法进行了比较, 实验结果表明, scCLG 能学习到良好的细胞特征表示, 在聚类任务上展现出超越其他方法的表现.

1 方法

1.1 概述

scCLG 整体将对比学习和图神经网络结合起来, 通过图神经网络学习细胞之间的结构关系信息, 对图结构进行数据增强, 使用对比学习进一步优化细胞特征表示. 模型的总体结构如图 1 所示.

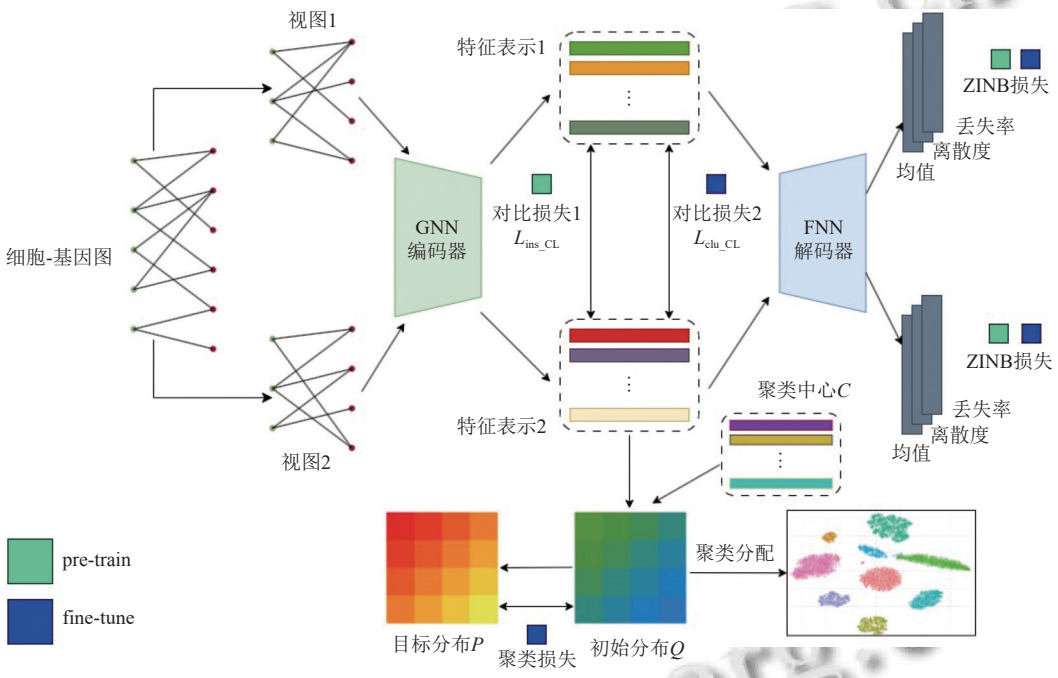

图 1 scCLG 整体结构

从图 1 可以看到, scCLG 采用基于 ZINB 模型的自编码器来学习细胞特征, 网络结构主要是由一个 GNN 编码器和一个 FNN (feedforward neural network, 前馈神经网络) 解码器构成, FNN 解码器后有 3 个独立的网络层用来计算 ZINB 分布的参数. 对图 1 所展现的模型工作流程, 进行简要介绍, 后续章节会详细说明. scCLG 的运行分为两个阶段, 第 1 阶段称为预训练 (pre-train) 阶段, 第 2 阶段称为微调 (fine-tune) 阶段. 在预训练阶段, 首先基于预处理后的表达矩阵构建细胞-基因二部图. 网络训练时, 通过子图采样和特征掩码的方式获取原始图结构的两个增广视图, 接着两个视图均通过 GNN 编码器编码得到细胞的特征表示. 最后, 将细胞的特征输入到 FNN 解码器中, 获得用于重建数据的 ZINB 分布参数. 在预训练阶段, scCLG 采用两种损失来优化模型. 第 1 种损失是基于重构的 ZINB 损失. 第 2 种损失是对比损失 (图 1 中的对比损失 1), 对于视图 1 中的细胞 i, 其正样本为视图 2 中的对应 (同一) 细胞 i', 其余均为负样本, 视图 2 中的细胞亦然. 损失的计算方式将在后文中具体说明. 在微调阶段, 在编码解码之外, 还增添了聚类部分, 参考文献[19,20]中的做法. 首先使用预训练后的 GNN 编码器得到细胞的特征, 基于此采用 K-means 算法进行初步聚类, 得到参数化的聚类中心. 在训练过程中, 通过特征表示 2 和聚类中心计算得到初始分布 Q, 再进一步得到目标分布 P, 将分布 P 和 Q 之间的 KL 散度作为聚类损失. P 是由 Q 计算得到的, 这是一个自优化的过程. 可以通过分布 Q 可以获得细胞的预测类别. 在这一阶段, scCLG 采用一种新的基于聚类的对比损失 (图 1 中的对比损失

2) 来进行对比学习. 对于视图 1 中的细胞i, 其正样本为其在视图 2 中对应细胞i'所在类簇的聚类中心, 其他簇的聚类中心则为负样本. 微调阶段的做法能够以一种端到端的方式同时优化特征表示和细胞聚类, 有利于提升聚类精度.

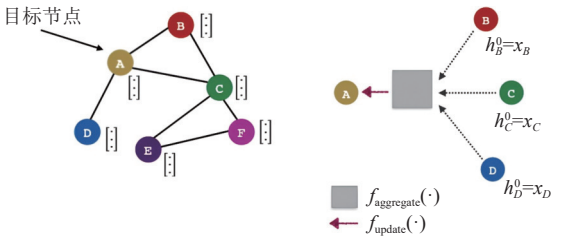

图 2 GraphSAGE 的节点聚合方式

1.2 数据预处理

对于原始表达数据矩阵 ($X_{m\times n}$, m 是细胞数, n 是基因数), 本文采用 Python 软件包 SCANPY[40]进行预处理, SCANPY 是一个常用的 scRNA-seq 数据处理工具. 预处理的策略本文参考了文献[20]中的做法. 具体而言, 首先过滤掉没有任何表达计数的基因和细胞. 然后计算每个细胞的总表达计数作为其文库大小 (library size), 用细胞的文库大小除以文库大小的中位数来获得每个细胞的大小因子 (size factor). 对所有的原始表达计数除以其对应细胞的大小因子, 这样就实现了数据的归一化. 归一化完成后, 进行对数变换 ($\log(x+1)$), 控制数据的范围. 接下来进行基因的选择, 根据标准化离散值, 只选取排在前一定比例的高变异基因. 最后进行标准分数 (Z-score) 标准化, 使数据的均值为 0, 标准差为 1.

1.3 图及增广视图的构建

模型中建立的是细胞和基因之间的二部图, 所谓二部图, 简单地说就是图的节点可以分成不相交的两部分, 而图的每一条边对应的两个节点分别属于这两部分. 对应地, 模型中图的这两部分节点就是细胞节点和基因节点. 若用$G=(V,E)$来表示图, 其中V是节点集合, E是边的集合, 则$V=\{V_c,V_g\}$, V_c,V_g分别代表细胞节点和基因节点. 若基因在某个细胞中有表达, 则该基因和该细胞之间存在一条边, 反之则没有. 集合E就是由这些边所构成的. 以上构建了图的基本骨架. 至于节点的表示, 细胞节点的表示直接采用细胞的基因表达, 即细胞的原始特征, 有$X_c\in R^{N_c\times N_g}$, N_c,N_g分别代表细胞数量和基因数量. 而基因节点的表示则是设定为随机初始化的可学习参数, 有$X_g\in R^{N_g\times F}$, F为基因初始特征向量的维度. 以上述方式构建细胞-基因二部图, 很大程度上保留了原始表达矩阵中的信息, 保持了细胞之间的结构关系, 而如果基于计算得到的细胞相似性来构建细胞-细胞图, 则可能由于细胞相似性的计算不准确而引入额外的噪音, 造成信息损失.

对于得到的细胞-基因图, 增广视图采用子图采样结合掩盖掉一部分特征的方式进行. 具体而言, 子图采样应用 Hu 等人在 2020 年提出的异构图采样技术 HGSampling[41], 该技术在 PyTorch Geometric[42]软件包中有相应的实现. 对于特征掩盖, 则是随机生成一定长度比例的 mask 向量, 将特征向量中对应位置上的值置为 0 即可, 细胞特征和基因特征都进行掩盖.

1.4 GNN 编码器

受文献[43]启发, scCLG 采用 GraphSAGE[44]这一图神经网络作为 GNN 编码器的选择. GraphSAGE 是一种归纳式 (inductive) 的网络, 其目的在于学习一种节点表示的方法, 也就是说如何通过一个节点的局部邻居采样并聚合以获取该节点的特征, 而非为每个节点单独训练其特征表示, 其节点聚合方式如图 2 所示. 相比于直推式 (transductive) 的学习, GraphSAGE 的优点在于训练时无需所有节点参与, 并且具有较强的泛化能力. 若用v,v'分别表示两个视图, 用f表示 GNN 编码器, 则视图通过 GNN 编码器的输出为相应细胞的特征表示, 即$H=f(v)\in R^{N_b\times d}, H'=f(v')\in R^{N_b\times d}$, 其中$N_b$是批量大小 (batch size), d是给定的细胞特征维度大小. 具体而言, 在本方法中, 由于所建立的图是二部图, 编码器网络的每一层的节点聚合方式可以表示如下:

$$\begin{cases} h_{c_i}^{(k+1)} = W'^{(k)}_c h_{c_i}^{(k)} + W_g^{(k)} mean_{g_j\in N(c_i)}(h_{g_j}^{(k)}) \\ h_{g_i}^{(k+1)} = W'^{(k)}_g h_{g_i}^{(k)} + W_c^{(k)} mean_{c_j\in N(g_i)}(h_{c_j}^{(k)}) \end{cases} \tag{1}$$

其中, $mean(\cdot)$表示输入向量的平均值, $N(c_i)$代表细胞i的所有邻居节点 (基因), 同样地, $N(g_i)$代表基因i的所有邻居节点 (细胞). $h_{c_i}^{(k)}\in R^{d_c^{(k)}}, h_{g_i}^{(k)}\in R^{d_g^{(k)}}$分别表示细胞$i$和基因$i$在$k$层的潜在表示, $d_c^{(k)}, d_g^{(k)}$代表细胞和基因在第k层的特征维度, 初始的特征输入为细胞和基因在图结构上的特征表示, $d_c^{(0)}=N_g, d_g^{(0)}=F$.

$mean_{g_j\in N(c_i)}(h_{g_j}^{(k)})$表示细胞$i$的所有邻居节点特征表示的均值, $mean_{c_j\in N(g_i)}(h_{c_j}^{(k)})$表示基因$i$的所有邻居节点特征表示的均值. 而 $W'^{(k)}_c, W_g^{(k)}, W'^{(k)}_g, W_c^{(k)}$ 都是可训

练的参数, $W_c^{(k)} \in R^{d_c^{(k+1)} \times d_c^{(k)}}, W'^{(k)}_c \in R^{d_c^{(k+1)} \times d_c^{(k)}}$ 分别表示第 k 层中用于转换邻居细胞节点和当前细胞节点的权重矩阵, $W_g^{(k)} \in R^{d_g^{(k+1)} \times d_g^{(k)}}, W'^{(k)}_g \in R^{d_g^{(k+1)} \times d_g^{(k)}}$ 分别表示第 k 层中用于转换邻居基因节点和当前基因节点的权重矩阵. 无论是细胞节点还是基因节点其特征都来自其本身和它的邻居节点的特征聚合.

1.5 ZINB 模块

相关方法中采用 ZINB (zero inflated negative binomial, 零膨胀的负二项分布) 模型来对 scRNA-seq 数据的分布进行建模[17,20,28,39,45], 本文同样假设 scRNA-seq 数据符合 ZINB 分布, 并进行相应建模. 具体而言, 负二项分布 NB 分布可以表示为:

$$\mathrm{NB}(x \mid \mu,\theta) = \frac{\Gamma(x+\theta)}{x!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x} \tag{2}$$

而 ZINB 分布则是在 NB 分布的基础上加了一个零膨胀因子, 具体表示如下:

$$\mathrm{ZINB}(x \mid \pi,\mu,\theta) = \pi\delta_0(x) + (1-\pi)\mathrm{NB}(x \mid \mu,\theta) \tag{3}$$

其中, δ_0 表示狄拉克 δ 函数. μ,θ,π 是 ZINB 分布的 3 个参数, 分别代表均值 (mean), 离散度 (dispersion) 和丢失率 (dropout). 为了估计这 3 个参数, 把 FNN 解码器的输出 D, 分别输入到 3 个独立的网络层中:

$$\begin{cases} M = S \times exp(DW_\mu) \\ \Theta = exp(DW_\theta) \\ \Pi = Sigmoid(DW_\pi) \end{cases} \tag{4}$$

其中, S 是一个对角矩阵, 其对角线上的值为对应位置细胞的大小因子 (size factor). M,Θ,Π 分别是参数均值, 离散度和丢失率的矩阵表示. W_μ, W_θ, W_π 是网络层中可训练的参数矩阵. exp, $Sigmoid$ 表示相应的激活函数. ZINB 损失定义为 ZINB 分布的负对数似然, 如下:

$$L_{\mathrm{ZINB}} = -\log(\mathrm{ZINB}(X^{\mathrm{count}} \mid \pi,\mu,\theta)) \tag{5}$$

其中, X^{count} 代表原始的表达矩阵.

1.6 训练过程

1.6.1 预训练阶段

首先按前文所述方法构建细胞-基因图. 在训练时, 通过子图采样和特征掩码的方式构建两个增广视图 v, v', 两个视图通过共享的 GNN 编码器 (f) 按式 (1) 得到对应细胞的特征表示, $Z = f(v), Z' = f(v')$. 之后对于编码得到的 Z, Z', scCLG 引入了对比学习用于指导编码器的训练. 具体地, 本文采用 InfoNCE 损失[37]作为对比损失. InfoNCE 损失是一种常用的样本之间的对比损失, 在这里, 正样本对仅为同一个细胞在两个视图中得到的不同表示, 而其他样本对都视为负样本. 对于细胞 i, 具体计算公式如下:

$$l(z_i) = -\log \frac{\mathrm{e}^{(sim(z_i, z'_i)/\tau)}}{\sum\limits_{j \neq i} \mathrm{e}^{(sim(z_i, z_j)/\tau)} + \sum\limits_{j} \mathrm{e}^{(sim(z_i, z'_j)/\tau)}} \tag{6}$$

其中, z_i, z'_i 分别表示细胞 i 在视图 1 和视图 2 中经过 GNN 编码器后得到的特征表示, e 代表自然对数的底数, τ 代表对比学习中的温度系数, 是一个超参数. $sim(\cdot,\cdot)$ 表示计算两个向量之间的相似度的函数, 此处使用的是余弦相似度, 即:

$$sim(a,b) = \frac{a \cdot b}{\|a\|\|b\|} \tag{7}$$

整体的对比损失为:

$$L_{\mathrm{ins\_CL}} = \frac{1}{2N_b}\sum_{i=1}^{N_b}[l(z_i) + l(z'_i)] \tag{8}$$

其中, N_b 是训练批次大小.

Z, Z' 继续通过 FNN 解码器以及 ZINB 模块得到 ZINB 分布的 3 个参数, 使用式 (5) 计算对应的损失. 预训练阶段最终的损失函数如下:

$$L_1 = \lambda_1 L_{\mathrm{ins\_CL}} + L_{\mathrm{ZINB}} \tag{9}$$

其中, λ_1 是用来平衡两部分损失的超参数. L_{ZINB} 是两个视图 ZINB 损失的平均值, 即:

$$L_{\mathrm{ZINB}} = \frac{1}{2}(L_{\mathrm{ZINB\_1}} + L_{\mathrm{ZINB\_2}}) \tag{10}$$

1.6.2 微调阶段

在这一阶段, 模型将基于预训练得到的编码器进行聚类并继续优化细胞的特征表示. 为了使聚类分配和特征学习能作为一个整体进行并相互受益, 受到文献[19]的启发, scCLG 采用一种自优化的策略来达成这一目标. 假设给定一个软聚类分配分布矩阵 Q, 其中 Q 的每一行代表一个细胞软聚类分配概率, 可以由矩阵 Q 计算得到目标分布矩阵 P, 通过最小化分布 P 和 Q 之间的 KL 散度, 来实现聚类分配的优化. 这个 KL 散度, 称为聚类损失.

$$L_{\mathrm{cluster}} = \mathrm{KL}(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{11}$$

其中, p_{ij}, q_{ij} 分别是矩阵 P 和 Q 中的元素. 在 scCLG

的设定下, 矩阵 Q 实际上代表细胞与类簇中心的相似度, 这里使用学生 t 分布来测算. 具体地, 细胞 i 的特征表示z_i和类别 j 的聚类中心c_j的关联程度可以用式 (12) 表示:

$$q_{ij}=\frac{\left(1+\left\|z_i-c_j\right\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'=1}^{K}\left(1+\left\|z_i-c_{j'}\right\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}} \tag{12}$$

其中, K 是类别数目, α是学生 t 分布的自由度, 固定设为 1. 聚类中心c_j是作为模型的一组参数存在, 在训练开始前通过 K-means 算法获取聚类中心对其初始化, 之后随着模型训练而更新. 用每个类别的分配频率对分布 Q 的二次项进行归一化, 得到目标分布矩阵 P, 具体如下:

$$p_{ij}=\frac{q_{ij}^2/f_j}{\sum_{j'=1}^{K}q_{ij'}^2/f_{j'}} \tag{13}$$

其中,

$$f_j=\sum_{i=1}^{N_b}q_{ij} \tag{14}$$

其中, f_j 代表软聚类频率, N_b是训练批次大小. 矩阵 P 被视为 Q 的目标, Q 和 P 应该尽可能接近, 按式 (11) 给出的损失优化模型, 可以得到置信度更高的聚类结果, 反过来也为聚类任务优化了细胞的特征表示, 聚类分配和特征学习是作为一个整体共同进行的. 由于矩阵 P 实际上完全是由矩阵 Q 计算得到的, 这是一个自我优化的过程.

首先通过前一个阶段训练的 GNN 编码器获取细胞的特征表示, 使用 K-means 算法进行初步聚类, 初始化聚类中心. 之后开始训练, 与前一阶段一样, 获取增广视图通过编码器和解码器计算 ZINB 损失. 不失一般性, 对视图 2 得到的中间特征进行上述的聚类分配, 得到相应的聚类损失. 前一个阶段的对比学习只将同一个细胞在不同视图中的特征视为正样本, 这可能会造成采样偏差[46], 产生不正确的负样本对 (将同一个类别的细胞视为负样本). 模型尝试改进对比学习的策略, 利用可能的类别信息来缓解这种情况. 具体地, 不再进行样本本身之间的对比, 而是考虑样本和类别 (聚类中心) 进行对比. 对于细胞 i 有如下的对比损失:

$$l'(z_i)=-\log\frac{e^{(sim(z_i,c_s)/\tau)}}{\sum_{j=1}^{K}e^{(sim(z_i,c_j)/\tau)}} \tag{15}$$

其中, z_i是细胞i在视图 1 中的特征, c_j代表视图 2 的聚类中心, c_s代表细胞 i 所在的那个类别的聚类中心, 通过矩阵 Q 来确定.

$$s=\arg\max_k(q_{ik}) \tag{16}$$

式 (15) 代表的对比损失将鼓励细胞与其所在的聚类中心靠拢, 而远离其他的聚类中心, 这是有利于聚类任务的. 进一步地, 整体的对比损失可以表示为:

$$L_{\text{clu\_CL}}=\frac{1}{N_b}\sum_{i=1}^{N_b}l'(z_i) \tag{17}$$

其中, N_b为训练批次大小.

经过以上分析, 微调阶段最终的损失函数如下:

$$L_2=\lambda_2L_{\text{clu\_CL}}+\lambda_3L_{\text{cluster}}+L_{\text{ZINB}} \tag{18}$$

其中, λ_2,λ_3是用来平衡各部分损失的超参数, 这里的 L_{ZINB}同样是两部分 ZINB 损失的平均.

2 实验与分析

2.1 数据集与性能评价指标

为了全面地评估本文提出的方法 scCLG, 从已发表的研究中收集了 10 个真实的 scRNA-seq 数据集. 这些数据集都带有专家注释的类别标签, 本文将这些数据集中自带的标签作为其中细胞的真实标签. 标签信息只在模型进行聚类性能评估的时候才会用到. 这些数据集涵盖了各种规模, 来自多个测序平台和多个不同组织器官的 scRNA-seq 数据. 数据集的详细信息在表 1 中列出, 包括数据集的名称和来源, 测序平台, 细胞数目, 基因数目以及真实的细胞类别数.

表 1 实验中用到的数据集 (共 10 个)

数据集	测序平台	细胞数目	基因数目	细胞类别数
Adam[47]	Drop-seq	3 660	23 797	8
Bladder[48]	10X	2 432	22 966	2
Klein[49]	inDrop	2 717	24 047	4
Mammary_Gland[48]	10X	3 282	22 966	4
Plasschaert[50]	inDrop	6 977	28 205	8
Pollen[51]	SMARTer	301	23 730	11
QS_Diaphragm[52]	Smart-seq2	870	23 341	5
QS_Limb_Muscle[52]	Smart-seq2	1 090	23 341	6
Romanov[53]	SMARTer	2 881	21 143	7
Tosches_turtle[54]	Drop-seq	18 664	23 500	15

本文采用两个广泛使用的聚类效果评价指标来对模型的聚类性能进行评估, 分别是标准化互信息 (normalized mutual information, *NMI*) 和调整兰德系数 (adjusted Rand index, *ARI*). *NMI* 和 *ARI* 都可以用来衡量聚类结果和真实类别分布之间的相似程度. 给定两个聚类分配 (标签向量) U 和 V, 它们之间的 *NMI* 的计算方式如下:

$$NMI = \frac{2I(U,V)}{H(U)+H(V)} \tag{19}$$

其中, $I(U,V)$代表 U 和 V 之间的互信息, $H(\cdot)$代表交叉熵. *NMI* 的取值范围为[0, 1], 数值越大 (越接近 1) 表示聚类效果越好. *ARI* 的计算方式如下:

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

其中, $\binom{\cdot}{\cdot}$表示组合数, n_{ij}表示同时在聚类结果 U 的第 i 类和聚类结果 V 的第 j 类中的样本数目, a_i代表 U 中第 i 类中的样本数量, b_j代表 V 中第 j 类中的样本数量, n 是样本的总数量. *ARI* 的取值范围是[−1, 1], 数值越大 (越接近 1) 表示聚类效果越好.

2.2 实验环境与设置

scCLG 是基于深度学习框架 PyTorch[55]实现的, Python 版本 3.10.9, PyTorch 版本 1.13.1, CUDA 版本 11.7. 硬件环境, CPU 为 Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50 GHz, 显卡为 NVIDIA GeForce RTX 3090, 操作系统为 Ubuntu 18.04.6 LTS. GNN 编码器是一个两层的 GraphSAGE, 其隐藏层的维度分别为 256, 64. 解码器是一个两层的全连接层网络, 其隐藏层的维度分别为 64, 256. 细胞-基因图中基因节点的特征维度为 64, 特征掩盖的比例为 0.2. 批量大小 (batch size) 为 1024. 学生 t 分布的自由度α设为 1, 对比损失中的温度系数τ设为 0.5.

2.3 比较的方法

scCLG 的聚类表现将和 9 个现有的先进 scRNA-seq 数据聚类方法进行比较, 分别是 Seurat[11]、CLEAR[36]、Contrastive-sc[34]、GraphSCC[24]、scTAG[28]、scDeepCluster[20]、scDHA[56]、scVI[57]、SIMLR[9]. Seurat 是一个广泛使用的 scRNA-seq 数据分析工具. 其聚类方法首先进行降维, 然后在共享的最近邻居图上使用 Louvain 方法. CLEAR 是一个基于自监督对比学习的综合性 scRNA-seq 数据分析工具. 它引入了一种新的数据增强方法, 并通过 InfoNCE 损失进行对比学习. Contrastive-sc 是一个自监督的对比学习聚类方法, 它通过随机掩盖细胞的基因表达来获取增广样本. GraphSCC 利用图卷积神经网络描述和利用细胞之间的结构关系, 通过双自监督模块优化模型学习到的表示. scTAG 是一个基于图神经网络的 scRNA-seq 数据聚类算法, 通过多核卷积拓扑图神经网络 (TAGCN) 对图结构数据进行编码. scDeepCluster 在去噪自编码器中引入了一个用来模拟 scRNA-seq 数据分布的 ZINB 模型. scDHA 首先利用非负核自编码器进行特征选择, 并使用基于变分自编码器 (VAE) 的自学习网络将数据进一步投影到低维空间. scVI 是一个用于分析 scRNA-seq 数据的综合工具. 它使用 ZINB 模型和变分自编码器以深度生成方式建模 scRNA-seq 数据. SIMLR 使用多核来学习样本之间的相似性并执行谱聚类.

2.4 聚类性能比较

为了对模型的聚类表现有更深的认识, 本文在 10 个真实的 scRNA-seq 数据集上将 scCLG 和上述 9 个聚类算法进行了比较. 图 3 (*ARI*) 和图 4 (*NMI*) 中展示了这些方法在不同数据集上的聚类结果, 其中每个数据集结果的第 1 列代表本文提出的方法.

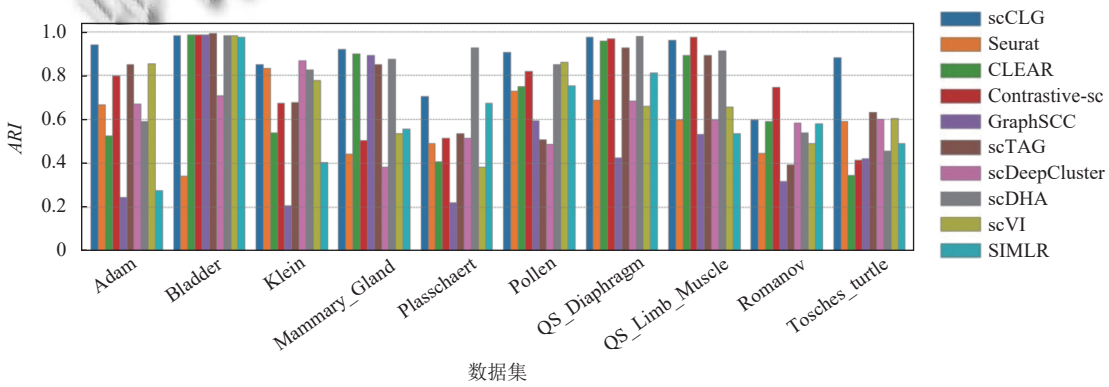

图 3 各个算法在不同数据集上的 *ARI* 柱状图

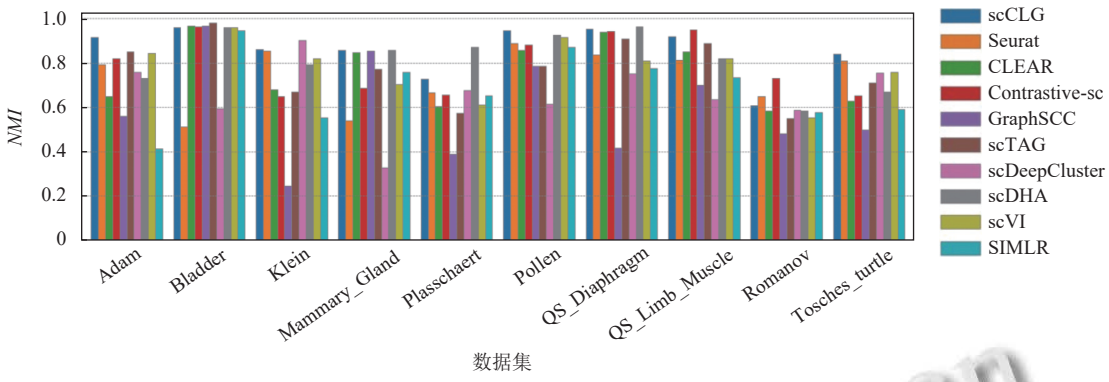

图 4 各个算法在不同数据集上的 *NMI* 柱状图

从图 3 和图 4 中的结果可以看到, scCLG 在 10 个数据集中的 4 个数据集 (Adam, Mammary_Gland, Pollen, Tosches_turtle) 上无论是 *ARI* 值还是 *NMI* 值都取得了最优的结果, 其他的最优结果分散在各个其他算法当中, 但其整体表现都不如本文提出的 scCLG. 即使是在 scCLG 没有达到最优的数据集上, scCLG 的性能也和最优的结果相当, 差距很小, 在每个数据集的结果当中, scCLG 的表现都处于领先水平. 从整体上看, scCLG 的聚类表现在每个数据集上都排在所有方法中的前列, 并且在不同数据集上的鲁棒性较好, 聚类性能没有发生较大的波动, 这表明 scCLG 泛化性较好, 对数据集不敏感, 并且可以有效处理不同规模的数据. 在细胞数最少的数据集 Pollen (301 个) 和细胞数最多的数据集 Tosches_turtle (18 664 个) 上, scCLG 都取得了最优的结果. 而一些其他的方法虽然偶尔也能取得最优结果, 但是其相对不稳定, 在不同的数据集上的聚类效果可能差异较大 (如: scDeepCluster, Contrastive-sc). 与同样基于图神经网络的方法 scTAG、GraphSCC 相比, scCLG 采用细胞-基因图这一图结构, 而不是细胞-细胞图, 这在一定程度上减少了信息损耗, 并且能避免引入额外的噪声, 此外 scCLG 还应用了对比学习, 能学习到更加利于聚类的特征表示. 和同样采用对比学习的 Contrastive-sc、CLEAR 相比, scCLG 对于图结构的建模更好地利用了细胞之间的结构关系信息, 同时改进了对比学习的策略, 有助于学习到更加有意义的特征表示.

图 5 中展示了不同的方法在 10 个数据集上的 *ARI*/*NMI* 指标平均值, 无论是 *ARI* 还是 *NMI*, scCLG 都取得了最高的数值, 这证明了其相对于其他方法的优越性.

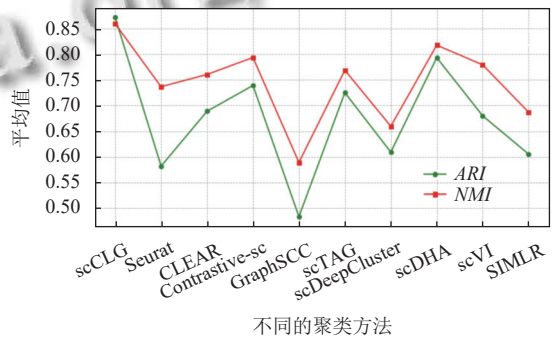

图 5 不同方法在 10 个数据集上的 *ARI*、*NMI* 均值

2.5 可视化分析

在 scRNA-seq 聚类的相关分析当中, 可视化是一种直观有效地展示不同类别的细胞的分布的手段. 本文采用 t-SNE[58]作为降维方法, 将各个方法得到的细胞的特征表示降到 2 维空间, 以便能在平面图形上把细胞的相对位置分布以散点图的形式绘制出来. 图 6、图 7 中分别展示了上述 10 个方法在 Klein 和 Mammary_Gland 这两个数据集上的 t-SNE 可视化结果. 图 6 和图 7 上的一个点代表一个细胞, 点的颜色代表其真实的类别标签.

在 Klein 数据集中, scCLG 将所有细胞分成了较为干净的 4 个部分, 不同类别之间区分明显, 同类别之间联系紧密, 这表明模型学习到了较为良好的特征表示. 其他方法中, scDHA 和 scVI 也有相对清楚的展示, 而 CLEAR 以及 GraphSCC 则表现较差, 整体呈线状分布并且难以区分各个类别. 在 Mammary_Gland 数据集上, scCLG 也相对清晰的显现出 4 个类别的存在 (蓝、绿、红、橙, 其中橙色细胞较少, 位于红色细胞的下方一些的位置), 相对于其他方法, scCLG 的可视化结果, 相同类簇聚集紧密, 不同类簇有区分, 没有大量的混杂,

并且较好地展现了橙色细胞 (endothelial cell, 内皮细胞) 这一类簇, 而其他方法容易将这一小的类别和其他类别混在一起.

为了进一步地验证 scCLG 的特征表示水平, 了解 scCLG 对于原始的数据特征的改善效果, 在 Romanov 数据集上进行了实验, 结果如图 8 所示.

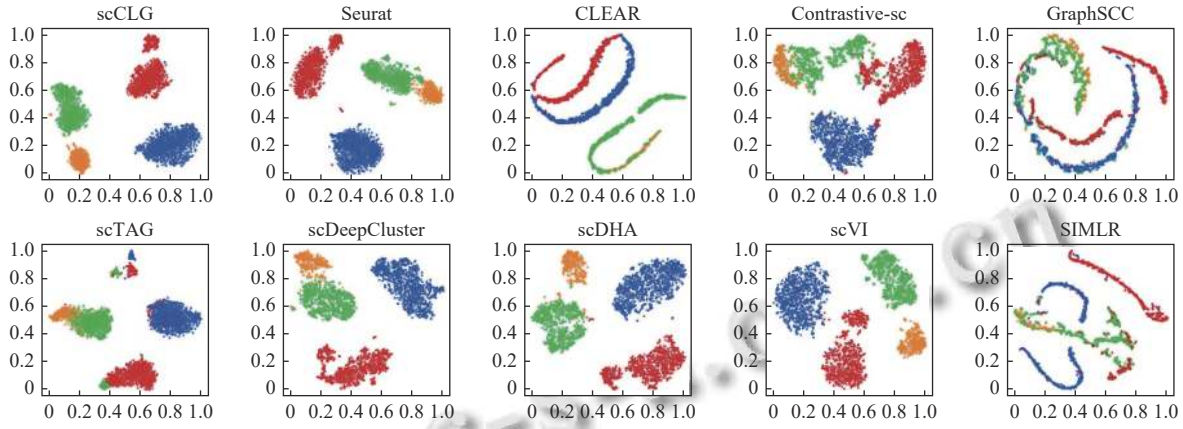

图 6 Klein 数据集上 10 个方法的 t-SNE 可视化结果

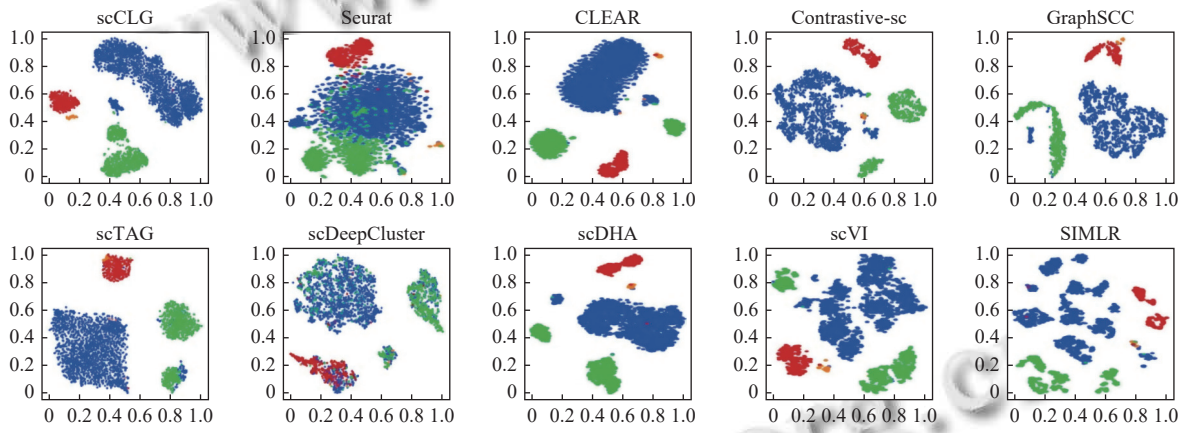

图 7 Mammary_Gland 数据集上 10 个方法的 t-SNE 可视化结果

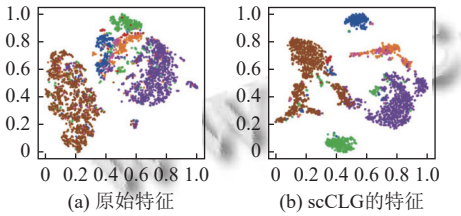

图 8 Romanov 数据集上原始特征和 scCLG 得到的特征的可视化结果对比

可以看到, scCLG 获得的特征表示与原始特征相比有较大的提升 (聚类结果 *ARI*=0.5965, *NMI*=0.6062), 这主要体现在对原始特征中混杂在一起的蓝色、绿色和橙色的细胞有了较为良好的区分, 同类型的细胞分布更紧密, 并且一定程度上减轻了棕色细胞和其他细胞的混杂.

2.6 消融实验

在微调阶段, scCLG 采用了自优化的聚类损失以及基于类簇中心的对比损失来训练模型, 从原理上看, 这朝着聚类任务的目标优化了细胞的特征表示, 一般来说对模型的性能是会有所提升的. 为了进一步验证微调阶段的有效性, 本文设计了相关的消融实验. 具体的做法是去掉微调阶段, 只进行训练阶段 1 (预训练阶段), 训练结束后对整个数据集的细胞进行编码得到其最终的特征表示, 并使用 K-means 算法来完成聚类. 在同样的 10 个数据集上进行了实验, 实验结果如图 9 所示. 图中的每一个点代表不同的数据集, 其横坐标代表消融实验 (仅 pre-train) 的结果, 纵坐标代表原始模型 scCLG 的结果, 图 9(a) 和图 9(b) 分别代表指标 *ARI* 和 *NMI* 的结果. 整体来看, 绝大部分数据点都落在了对角

线 $y=x$ 的上方, 也就是原始模型要优于消融模型, 这表明微调阶段确是对模型有帮助的, 并且在某些数据集 (如: Adam) 上有着比较明显的提升. 就 Romanov 数据集而言, 微调阶段反而让聚类精度下降, 这可能是因为聚类中心的初始化存在较多噪声, 经过训练引入了更大的偏差.

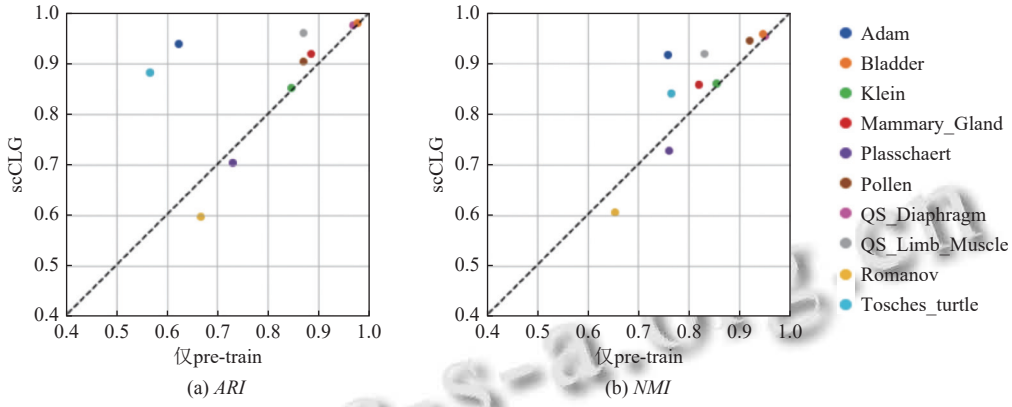

图 9 scCLG 和其变体的聚类表现的消融实验结果比较

3 结语

细胞聚类是 scRNA-seq 数据分析中的关键任务, 但是由于技术局限以及生物体的复杂性, scRNA-seq 数据整体呈现高维度、高稀疏、高噪声的特点, 这给聚类任务带来了挑战. 为此, 本文提出了一个联合对比学习和图神经网络的自优化单细胞数据聚类方法 scCLG. 对比学习是一种强大的特征学习范式, 图神经网络则能有效地刻画细胞间的关系. 本方法通过构建细胞-基因二部图, 而非常规的细胞-细胞图, 尽可能保留原始表达矩阵中的信息, 使用图神经网络进行编码, 描述和利用了细胞之间的结构关系信息, 这有助于模型学习到更加丰富的深层次特征表示. 使用子图采样和特征掩盖的方式进行数据增强, 生成不同的视图用于对比学习. 在第 1 阶段 (pre-train) 的训练中, 采用常规的样本间的对比学习结合用来保持数据的局部结构的 ZINB 损失, 初步地训练了图编码器. 在第 2 阶段 (fine-tune) 的训练中, 进行自优化的聚类, 逐步优化聚类结果, 并结合聚类分配改进了对比学习的策略, 使得模型能从中感知到聚类的相关信息, 由此继续提升模型的特征表示能力. 不同的图构建方式以及聚类层面的对比学习策略是本方法的主要创新之处.

在 10 个真实的 scRNA-seq 数据集上进行了实验, 并与其他 9 个方法进行了比较. 结果表明, scCLG 的聚类效果在近一半的数据集上能达到最优, 在所有的数据集上都处于领先水平, 并且就平均表现而言, scCLG 显著优于比较的其他方法. scCLG 展现出良好的鲁棒性, 可以有效处理来自不同组织、平台以及不同规模的各种 scRNA-seq 数据, 保持相对稳定的聚类性能, 这是其他方法所欠缺的. 同时, 可视化分析的结果表明, scCLG 可以学习到良好的细胞特征表示.

scCLG 的一个可能不足在于其运行成本, 模型的训练时间在细胞数量增多的时候有着相当的提升. 在对大规模的单细胞数据进行聚类的时候, 时间效率非常关键. 后续的研究应在保证聚类准确度的前提下, 加速模型学习的过程.

参考文献

1 Buettner F, Natarajan KN, Casale FP, *et al*. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology, 2015, 33(2): 155–160. [doi: 10.1038/nbt.3102]

2 Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Molecular Cell, 2015, 58(4): 586–597. [doi: 10.1016/j.molcel.2015.05.004]

3 Malikic S, Jahn K, Kuipers J, *et al*. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. Nature Communications, 2019, 10(1): 2750. [doi: 10.1038/s41467-019-10737-5]

4 Eberwine J, Sul JY, Bartfai T, *et al*. The promise of single-cell sequencing. Nature Methods, 2014, 11(1): 25–27. [doi: 10.1038/nmeth.2769]

5 文路, 汤富酬. 单细胞转录组分析研究进展. 生命科学, 2014, 26(3): 228–233.

6 Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature

Reviews Genetics, 2019, 20(5): 273–282. [doi: 10.1038/s41576-018-0088-9]
7 Žurauskienė J, Yau C. pcaReduce: Hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics, 2016, 17: 140. [doi: 10.1186/s12859-016-0984-y]
8 Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: Consensus clustering of single-cell RNA-seq data. Nature Methods, 2017, 14(5): 483–486. [doi: 10.1038/nmeth.4236]
9 Wang B, Ramazzotti D, de Sano L, *et al*. SIMLR: A tool for large-scale genomic analyses by multi-kernel learning. Proteomics, 2018, 18(2): 1700232. [doi: 10.1002/pmic.201700232]
10 Xu C, Su ZC. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics, 2015, 31(12): 1974–1980. [doi: 10.1093/bioinformatics/btv088]
11 Satija R, Farrell JA, Gennert D, *et al*. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology, 2015, 33(5): 495–502. [doi: 10.1038/nbt.3192]
12 Blondel VD, Guillaume JL, Lambiotte R, *et al*. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008. [doi: 10.1088/1742-5468/2008/10/p10008]
13 Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols, 2018, 13(4): 599–604. [doi: 10.1038/nprot.2017.149]
14 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]
15 Eraslan G, Avsec Ž, Gagneur J, *et al*. Deep learning: New computational modelling techniques for genomics. Nature Reviews Genetics, 2019, 20(7): 389–403. [doi: 10.1038/s41576-019-0122-6]
16 Bao SQ, Li K, Yan CC, *et al*. Deep learning-based advances and applications for single-cell RNA-sequencing data analysis. Briefings in Bioinformatics, 2022, 23(1): bbab473. [doi: 10.1093/bib/bbab473]
17 Eraslan G, Simon LM, Mircea M, *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications, 2019, 10(1): 390. [doi: 10.1038/s41467-018-07931-2]
18 Risso D, Perraudeau F, Gribkova S, *et al*. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications, 2018, 9(1): 284. [doi: 10.1038/s41467-017-02554-5]
19 Xie JY, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR, 2016. 478–487.
20 Tian T, Wan J, Song Q, *et al*. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nature Machine Intelligence, 2019, 1(4): 191–198. [doi: 10.1038/s42256-019-0037-0]
21 Chen L, Wang WN, Zhai YY, *et al*. Deep soft K-means clustering with self-training for single-cell RNA sequence data. NAR Genomics and Bioinformatics, 2020, 2(2): lqaa039. [doi: 10.1093/nargab/lqaa039]
22 Wu ZH, Pan SR, Chen FW, *et al*. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4–24. [doi: 10.1109/tnnls.2020.2978386]
23 Wang JX, Ma AJ, Chang YZ, *et al*. scGNN is a novel graph neural network framework for single-cell RNA-seq analyses. Nature Communications, 2021, 12(1): 1882. [doi: 10.1038/s41467-021-22197-x]
24 Zeng YS, Zhou X, Rao JH, *et al*. Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network. Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Seoul: IEEE, 2020. 519–522. [doi: 10.1109/BIBM49941.2020.9313569]
25 Li JY, Jiang W, Han H, *et al*. ScGSLC: An unsupervised graph similarity learning framework for single-cell RNA-seq data clustering. Computational Biology and Chemistry, 2021, 90: 107415. [doi: 10.1016/j.compbiolchem.2020.107415]
26 Gan YL, Huang XY, Zou GB, *et al*. Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network. Briefings in Bioinformatics, 2022, 23(2): bbac018. [doi: 10.1093/bib/bbac018]
27 Cheng Y, Ma XL. scGAC: A graph attentional architecture for clustering single-cell RNA-seq data. Bioinformatics, 2022, 38(8): 2187–2193. [doi: 10.1093/bioinformatics/btac099]
28 Yu ZH, Lu YF, Wang YH, *et al*. ZINB-based graph embedding autoencoder for single-cell RNA-seq interpretations. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 4671–4679. [doi: 10.1609/aaai.v36i4.20392]
29 Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 1597–1607.
30 He KM, Fan HQ, Wu YX, *et al*. Momentum contrast for unsupervised visual representation learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9726–9735. [doi: 10.1109/cvpr42600.2020.00975]
31 Grill JB, Strub F, Altché F, *et al*. Bootstrap your own latent a new approach to self-supervised learning. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1786.
32 Caron M, Misra I, Mairal J, *et al*. Unsupervised learning of visual features by contrasting cluster assignments.

Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 831.
33 Ericsson L, Gouk H, Hospedales TM. How well do self-supervised models transfer? Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5410–5419. [doi: 10.1109/cvpr46437.2021.00537]
34 Ciortan M, Defrance M. Contrastive self-supervised clustering of scRNA-seq data. BMC Bioinformatics, 2021, 22(1): 280. [doi: 10.1186/s12859-021-04210-8]
35 Wan H, Chen L, Deng MH. scNAME: Neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data. Bioinformatics, 2022, 38(6): 1575–1583. [doi: 10.1093/bioinformatics/btac011]
36 Han WK, Cheng YQ, Chen JY, *et al*. Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. Briefings in Bioinformatics, 2022, 23(5): bbac377. [doi: 10.1093/bib/bbac377]
37 van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.
38 Wang J, Xia JF, Wang HY, *et al*. scDCCA: Deep contrastive clustering for single-cell RNA-seq data based on auto-encoder network. Briefings in Bioinformatics, 2023, 24(1): bbac625. [doi: 10.1093/bib/bbac625]
39 Xiong ZH, Luo JW, Shi WW, *et al*. scGCL: An imputation method for scRNA-seq data based on graph contrastive learning. Bioinformatics, 2023, 39(3): btad098. [doi: 10.1093/bioinformatics/btad098]
40 Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. Genome Biology, 2018, 19(1): 15. [doi: 10.1186/s13059-017-1382-0]
41 Hu ZN, Dong YX, Wang KS, *et al*. Heterogeneous graph Transformer. Proceedings of the 2020 Web Conference. Taipei: ACM, 2020. 2704–2710. [doi: 10.1145/3366423.3380027]
42 Fey M, Lenssen JE. Fast graph representation learning with PyTorch geometric. arXiv:1903.02428, 2019.
43 Lee J, Kim S, Hyun D, *et al*. Deep single-cell RNA-seq data clustering with graph prototypical contrastive learning. Bioinformatics, 2023, 39(6): btad342. [doi: 10.1093/bioinformatics/btad342]
44 Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1025–1035.
45 Wei MQ, Liu RJ, Wang YJ, *et al*. Deep zero-inflated negative binomial model and its application in scRNA-seq data integration. SoutheastCon 2023. Orlando: IEEE, 2023. 901–905. [doi: 10.1109/SoutheastCon51012.2023.10115099]
46 Chuang CY, Robinson J, Lin YC, *et al*. Debiased contrastive learning. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 735.
47 Adam M, Potter AS, Potter SS. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: A molecular atlas of kidney development. Development, 2017, 144(19): 3625–3632. [doi: 10.1242/dev.151142]
48 The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature, 2020, 583(7817): 590–595. [doi: 10.1038/s41586-020-2496-1]
49 Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 2015, 161(5): 1187–1201. [doi: 10.1016/j.cell.2015.04.044]
50 Plasschaert LW, Žilionis R, Choo-Wing R, *et al*. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature, 2018, 560(7718): 377–381. [doi: 10.1038/s41586-018-0394-6]
51 Pollen AA, Nowakowski TJ, Shuga J, *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology, 2014, 32(10): 1053–1058. [doi: 10.1038/nbt.2967]
52 The Tabula Muris Consortium, Overall Coordination, Logistical Coordination, *et al*. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature, 2018, 562(7727): 367–372. [doi: 10.1038/s41586-018-0590-4]
53 Romanov RA, Zeisel A, Bakker J, *et al*. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nature Neuroscience, 2017, 20(2): 176–188. [doi: 10.1038/nn.4462]
54 Tosches MA, Yamawaki TM, Naumann RK, *et al*. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. Science, 2018, 360(6391): 881–888. [doi: 10.1126/science.aar4237]
55 Paszke A, Gross S, Massa F, *et al*. PyTorch: An imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 721.
56 Tran D, Nguyen H, Tran B, *et al*. Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature Communications, 2021, 12(1): 1029. [doi: 10.1038/s41467-021-21312-2]
57 Lopez R, Regier J, Cole MB, *et al*. Deep generative modeling for single-cell transcriptomics. Nature Methods, 2018, 15(12): 1053–1058. [doi: 10.1038/s41592-018-0229-2]
58 van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(86): 2579–2605.

(校对责编: 张重毅)