

冯思凡, 王振友, 金应华. 基于自适应图正则化低秩表示的scRNA-seq数据分析方法[J]. 广东工业大学学报, 2025, 42(5): 129–136. doi: 10.12052/gdutxb.240018.
Feng Sifan, Wang Zhenyou, Jin Yinghua. ScRNA-seq data analysis based on adaptive graph regularization and low-rank representation[J]. Journal of Guangdong University of Technology, 2025, 42(5): 129–136. doi: 10.12052/gdutxb.240018.

基于自适应图正则化低秩表示的scRNA-seq 数据分析方法

冯思凡, 王振友, 金应华
(广东工业大学 数学与统计学院, 广东 广州 510520)

摘要: 单细胞RNA测序(scRNA-seq)技术可以用于研究单个细胞的基因表达情况, 生成大量的单细胞基因表达数据。这类数据通常具有高维度和复杂的结构, 需要进行降维和聚类分析来揭示细胞类型和状态之间的差异。本文提出了一种基于自适应图正则化低秩表示的scRNA-seq数据分析方法——scLRRAGR。该方法可以充分利用scRNA-seq数据的全局和局部信息进行图学习, 并通过自适应图正则化和引入秩约束来捕捉细胞之间的相似性和相互作用, 更好地反映细胞之间的聚类结构, 帮助揭示不同细胞类型和状态之间的差异。在应用该方法时, 可以将scRNA-seq数据转换为一个图结构, 其中每个节点表示一个单细胞样本, 边表示细胞之间的相似性或相互作用。然后使用该方法对图进行学习和优化, 得到最优的图表示。最后, 可以使用聚类算法将学习到的图表示应用于细胞类型和状态的识别。实验结果表明, 本文方法应用在scRNA-seq数据集上能够显著提高聚类性能。

关键词: scRNA-seq数据; 细胞聚类; 图正则化; 低秩表示; 秩约束

中图分类号: Q332

文献标志码: A

文章编号: 1007-7162(2025)05-0129-08

ScRNA-seq Data Analysis Based on Adaptive Graph Regularization and Low-rank Representation

Feng Sifan, Wang Zhenyou, Jin Yinghua

(School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou 510520, China)

Abstract: Single-cell RNA sequencing (scRNA-seq) can be used to study the gene expression of single cell and generate a large amount of single-cell gene expression data. This type of data generally has high-dimensional and complex structures, requiring dimension reduction and clustering analysis to reveal differences between cell types and states. A new scRNA-seq data analysis method (scLRRAGR) is proposed based on adaptive graph regularization low-rank representation. This method can fully utilize the global and local information of scRNA-seq data for graph learning, and capture the similarity and interaction between cells by adaptive graph regularization and the introduction of rank constraint. Its outcome can better reflect the clustering structure between cells and help to reveal differences between different cell types and states. When applying this method, scRNA-seq data can be transformed into a graph structure with each node representing a single-cell sample and edges representing similarities or interactions between cells. Then this method is used to learn and optimize this graph to obtain the optimal graph representation. Finally, typical clustering algorithms can use the optimal graph representation to recognize cell types and states. The experiment results show that the proposed method can significantly improve clustering performance on scRNA-seq datasets.

Key words: scRNA-seq data; cell clustering; graph regularization; low-rank representation; rank constraint

收稿日期: 2024-01-29 录用日期: 2024-04-08 网络首发日期: 2024-08-28

基金项目: 广东省自然科学基金资助项目(2023A1515012891)

作者简介: 冯思凡(1996-), 女, 硕士研究生, 主要研究方向为生物统计, E-mail: 649232944@qq.com

通信作者: 金应华(1982-), 男, 副教授, 博士, 主要研究方向为回归模型、聚类分析和深度学习等, E-mail: jyh@mail.ustc.edu.cn

单细胞RNA测序(scRNA-seq)是一种先进的技术,用于研究单个细胞的基因表达水平。随着技术的不断发展和成本的降低,scRNA-seq正变得越来越普遍。与批量RNA测序(RNA-seq)相比,scRNA-seq能够克服仅对数千或数百万细胞进行平均表达水平测量的限制,实现在单细胞级别上研究全基因组基因表达^[1-2]。通过scRNA-seq,研究人员能够更准确地解决关于组织的细胞组成、细胞类型的转录异质性以及结构变化等问题。

这项技术带来了许多令人兴奋的发现,例如新的细胞类型的发现,以及已知细胞类型中亚群和异质性的揭示^[3],这些在之前的研究中可能被忽视。scRNA-seq技术在神经学^[4]、免疫学^[5]、肿瘤学^[6]和发育生物学^[7]等领域被广泛运用,用于识别和分析复杂细胞群的异质性。与批量RNA-seq分析相比,scRNA-seq揭示了单个细胞之间的差异,从而能够回答传统方法无法解决的问题^[8-9]。通过对scRNA-seq数据进行无监督聚类和无偏分析,研究人员能够发现新的细胞类型和功能不同的细胞亚群,这种技术对于开发新的生物学见解和验证先验知识至关重要。

传统的计算和统计方法可以用于分析来自批量RNA-seq的数据,但它们并没有解决单细胞表达数据分析中的特殊挑战。如异常细胞群、转录本放大噪声和细胞周期等生物学效应,以及基因表达的固有随机性引起的生物异质性和技术噪音所导致的dropout事件。dropout事件指的是细胞处理、细胞裂解和文库制备过程中的技术噪音,导致实际表达的基因无法被检测到^[10]。目前的技术可能无法测量细胞中50%~95%的mRNA^[10-11]。虽然这些问题取决于所使用的特定技术,但在大多数scRNA-seq研究中存在一定的影响。因此,需要一种计算分析方法,考虑到scRNA-seq数据的特定性质,使研究人员能够准确可靠地识别、可视化和探索单个细胞的群体结构。

为了满足这一需求,许多专门针对scRNA-seq数据开发的聚类方法已被提出,这些方法隐式或显式地依赖于相似概念。如将Pearson相关性与稳健K-MEANS聚类相结合(Spearman Highly-expressed Genes-K-MEANS, Spearman-HiE-KM)^[12]的方法,该方法对初始中心点的选择较为敏感,容易陷入局部最优解。若使用共识聚类^[13]来获得稳定细胞分组(Single-cell Consensus Clustering, SC3)^[14],过程涉及多次运行聚类算法,计算成本较高。Pcareduce^[15]是将聚类与主成分分析(Principal Components Analysis, PCA)结合起来,在处理非线性数据时效果可能有

限。而Lin等^[16]探索使用神经网络(Neural Networks, NNs)^[17]进行聚类和降维,需要大量的数据来进行训练,对参数设置和模型结构的选择要求较高,同时过程缺乏解释性。SIMLR(Single-cell Interpretation via Multi-kernel Learning)^[18]是一种基于多核学习的方法^[19],可以直接学习单细胞之间的相似性,但需要预先确定核函数的选择,如果选择不当可能导致聚类结果不佳。还有RAFSIL^[20],一种基于随机森林的方法,从scRNA-seq数据中学习细胞之间的相似性,受随机森林中树的数量和深度等参数的影响,需要进行较多的参数调优工作。

本文受文献^[21]的启发,提出了一种基于自适应图正则化低秩表示的scRNA-seq数据分析方法——scLRRAGR。该方法可以充分利用scRNA-seq数据的全局和局部信息进行图学习^[22],并通过自适应图正则化^[23-25]和引入秩约束^[26]来捕捉细胞之间的相似性和相互作用,更好地反映细胞之间的聚类结构,帮助揭示不同细胞类型和状态之间的差异。此方法计算复杂度低,可以在不需要人工干预的情况下,自适应地选择最近邻进行图的构建,一定程度上提高了该方法在现实数据的聚类应用中(包括具有流形结构的数据)的适应性,并且得到的图具有良好的可解释性,图中每个元素直接揭示了对应两个细胞的相似关系。本文提出的scLRRAGR方法,对比现有的针对scRNA-seq数据开发的聚类方法有如下优势:(1)使用了图学习的方法来捕捉数据的非线性结构和相似性关系,更好地反映scRNA-seq数据的整体和局部结构,避免了其他方法可能存在的过度平滑或过度分散的问题;(2)采用自适应图正则化技术来平衡全局和局部信息,这种技术可以根据不同scRNA-seq数据集的特点自适应地调整图的权重,使图能准确地反映细胞之间的相似性关系;(3)使用了低秩表示方法进行降维和嵌入分析,在保留数据主要信息的同时,还可以减少噪声和异常值的影响,提高聚类结果的准确性和鲁棒性;(4)在模型中引入了一种新的秩约束,学习后的图具有非常清晰的聚类结构,即通过先验信息确定领域中已知的细胞类型或亚群数量 c ,可以将细胞按照其表达模式划分为不同的亚群,这有助于发现细胞类型或亚型、状态转换、细胞功能等信息,评估数据质量和异质性以及优化下游分析。

1 单细胞数据和分析方法

1.1 基因数据收集

表1列出了本文使用到的数据集及其属性,其中

Patel、Pollen和Treutlein来自<https://hemberg-lab.github.io/scRNA.seq.datasets/>, Usoskin和Zeisel来自<https://github.com/BatzoglouLabSU/SIMLR>。

表 1 scRNA-seq数据集及其属性
Table 1 ScRNA-seq datasets and and their attributes

Dataset	Number of cells	Number of genes	Number of populations
Patel	430	5 948	5
Usoskin	622	17 772	4
Treutlein	80	23 271	5
Pollen	301	9 966	11
Zeisel	3 005	30 000	9

1.2 数据预处理

1.2.1 标准化

在处理scRNA-seq数据时,“标准化”的预处理是非常重要的步骤之一^[2],旨在通过将可观测方差缩放到指定范围来调整数据集中的原始计数,以获得可变采样效果。标准化可以消除技术和实验差异,降低样本处理差异,消除基因表达量的偏差,提高不同实验之间的可比性。常用的标准化方法包括TPM (Transcripts Per Kilobase per Million Mapped Reads)、RPKM(Reads Per Kilobase per Million Mapped Reads)和DESeq等,它们都可以有效地消除样本之间的差异,提高数据的可靠性和可重复性。标准化在处理scRNA-seq数据中起着关键的作用,可以确保数据的准确性和可靠性,从而更好地进行后续的数据分析。

文献[27]比较了scRNA-seq数据的22种标准化方法。需要强调的是,该文献还比较了不同标准化对各种下游分析任务的影响,到目前为止还没有一个公认的“完美”标准化方法。所以要根据下游分析任务谨慎地选择合适的标准化方法。

本节采取了基于Scran的标准化方法。Scran被广泛测试并用于批量数据的校正任务,该方法是一类基于Delta方法^[28]的标准化方法,它使用了脱卷积方法来估计基于细胞池中基因线性回归的大小因子,能更好地利用数据集中所有细胞之间计数深度的差异,在使用时可通过R包Scran轻松调用。

1.2.2 特征选择

scRNA-seq数据经过标准化后,仍然保留生物异质性,但减少了基因表达的技术抽样效应^[2]。scRNA-seq数据集可能会包含多达30 000个基因,而通过标准化过程会删除一些在多个细胞中未检测到的基因。然而,在剩余的基因中仍有很多不能提供信息,

并且几乎没有计数。因此,标准的预处理流程还需要特征选择的步骤。特征选择是指仅选择相关特征的子集的过程,这些子集可以是信息量最大、变化最大或偏差最大的特征,其目的是排除可能不代表样本之间有意义的生物学变异的无信息基因。

如图1所示,scRNA-seq实验和由此产生的数据集集中在一个特定的组织上,因此,只有一小部分基因具有信息性和生物学可变性。传统的方法要么计算所有基因的变异系数(高度可变基因),要么计算平均表达水平(高度表达基因),以获得500~2 000个选定的基因,并将这些特征用于其下游分析步骤。然而,这些方法对以前使用的标准化技术非常敏感。一般地,scRNA-seq数据的预处理工作流程包括标准化和随后的对数变换。但是,由于不可能对精确的零进行对数变换,因此在对数据进行对数变换之前,分析人员通常会在所有归一化计数中添加一个小的伪计数。然而,伪计数的选择是任意的,可能会给变换后的数据带来偏差,从而对特征选择产生影响^[29]。

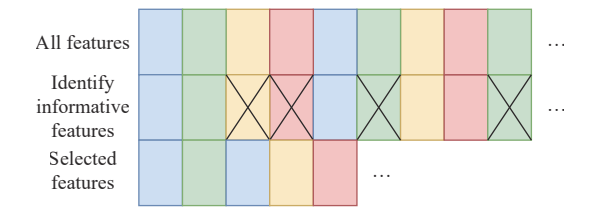


图 1 特征选择
Fig.1 Feature selection

对此,Germain等^[30]提出了一种基于偏差的保持对原始计数有效的特征选择方法。偏差可以以封闭的形式计算,并量化基因是否在细胞中表现出恒定的表达谱,此时这些表达谱没有信息。具有恒定表达的基因用多项式零模型来描述,它们用二项式偏差来近似。跨细胞的高信息基因将具有高偏差值,这表明零模型的拟合较差(即它们在跨细胞中没有显示出恒定的表达)。根据偏差值,该方法对所有基因进行排序,只保留高偏差的基因。相应的操作可通过R包scry实现。

1.3 具体算法

1.3.1 算法介绍

稀疏子空间聚类(Sparse subspace clustering, SSC)^[31]和低秩表示(Low Rank Representation, LRR)^[32]是两种基于图的聚类方法,适用于高维数据降维和聚类分析。SSC假定数据由多个低维子空间组成,每个子空间对应一个聚类。第一步通过字典学习或PCA等方法将原始数据降维,并利用稀疏表示的方

法对样本进行描述,第二步通过聚类系数矩阵的谱聚类或K-MEANS等方法将样本分成多个子空间,从而达到聚类的目的。LRR假定数据由多个低秩子空间组成,每个子空间对应一个聚类。第一步通过最小化低秩表示与原始数据之间的残差来求解低秩表示系数矩阵,第二步由谱聚类或K-MEANS等方法将样本分成多个子空间,从而达到聚类的目的。无论是SSC还是LRR都只致力于学习揭示样本表示关系的表示图,而忽略了数据的局部结构。同时因为这两种方法中很多表示值是负的,所以每个样本的表示系数不能清楚地显示样本之间的相似程度。因此,这两种传统方法得到的图都没有很好的可解释性;用它们来直接分析scRNA-seq数据,将会忽略该类数据中的局部结构。本文将介绍一个具有聚类性质的图学习方法——scLRRAGR,并用它来分析scRNA-seq类数据,该方法是基于自适应图正则化低秩表示,可捕获数据的局部和全局结构,确保所有元素都是非负的,能用来直接反映样本的相似程度,并且具有精确连通的结构。

scLRRAGR算法中目标函数为

$$\min_{Z,E} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 z_{ij} + \lambda_1 \|\mathbf{Z}\|_* + \lambda_2 \|\mathbf{E}\|_1$$

$$\text{s.t.} \begin{cases} \mathbf{X} = \mathbf{XZ} + \mathbf{E} \\ \text{diag}(\mathbf{Z}) = 0 \\ \mathbf{Z} \geq 0 \\ \text{rank}(\mathbf{L}_Z) = n - c \\ \sum_j z_{ij} = 1 \end{cases} \quad (1)$$

式中: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbf{R}^{m \times n}$ 为基因数为 m 、细胞数为 n 的scRNA-seq数据矩阵,每个细胞基因数据对应一列; c 为聚类数(一般可由先验信息得到); λ_1, λ_2 为正惩罚参数; $\|\cdot\|_1, \|\cdot\|_2$ 分别为 l_1, l_2 范数; $\|\cdot\|_*$ 为核范数(矩阵的奇异值之和)。式(1)中 $\mathbf{Z} \in \mathbf{R}^{n \times n}$ 是需要学习的细胞间关系表示图,其中元素 z_{ij} 表示细胞 $\mathbf{x}_j$ 在联合表示法中相对于细胞 $\mathbf{x}_i$ 的表示系数;等式 $\mathbf{X} = \mathbf{XZ} + \mathbf{E}$ (联合表示法)是为了实现LRR中的约束,它是SSC中的严格约束 $\mathbf{X} = \mathbf{XZ}$ 的弱化,引入了误差项 $\mathbf{E}$,此时的 $\mathbf{Z}$ 是数据矩阵 $\mathbf{X}$ 的稀疏表示图,即 $\mathbf{Z}$ 的第 j 列是细胞 $\mathbf{x}_j$ 在细胞间关系表示图中的表示系数; $\mathbf{L}_Z$ 为 $\mathbf{Z}$ 的拉普拉斯矩阵,定义为 $\mathbf{L}_Z = \mathbf{D} - (\mathbf{Z} + \mathbf{Z}^T)/2$, $\mathbf{D}$ 是一个对角矩阵,它的第 i 个对角元素 $D_{ii} = \sum_j (z_{ij} + z_{ji})/2$; 在表示图 $\mathbf{Z}$ 非负的条件下,有定理保证 $\mathbf{L}_Z$ 的特征根0的重数等于 $\mathbf{Z}$ 表示对象的连通分量的数量(即聚类数满足 $\text{rank}(\mathbf{L}_Z) = n - c$)^[33]; $\text{diag}(\mathbf{Z}) = 0$ 排除了单个细胞对自

身的表示影响, $\mathbf{Z} \geq 0$ 表示只考虑非负表示系数, $\sum_j z_{ij} = 1$ 是为了避免表示图 $\mathbf{Z}$ 的行元素都为0这种极端情况。式(1)目标函数中的第1项表示细胞间的距离越大,图中表示系数越小;第2项用细胞间关系表示图矩阵 $\mathbf{Z}$ 的核范数替换了 $\mathbf{Z}$ 的秩,能够保证优化结果的稀疏性,且方便计算,此时正惩罚参数 λ_1 越大,优化得到的 $\mathbf{Z}$ 的核范数越小;第3项表示正惩罚参数 λ_2 越大,约束 $\mathbf{X} = \mathbf{XZ} + \mathbf{E}$ 的误差项 $\mathbf{E}$ 范数越小。

一般地,式(1)中约束 $\text{rank}(\mathbf{L}_Z) = n - c$ 是不方便优化的;根据拉普拉斯矩阵 $\mathbf{L}_Z$ 是半正定的(即 $\mathbf{L}_Z$ 的所有特征值都大于等于0),以及Fan^[34]提出的定理,可以将优化问题(1)改写为式(2)的等价优化问题:

$$\min_{Z,E,F} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 z_{ij} + \lambda_1 \|\mathbf{Z}\|_* + \lambda_2 \|\mathbf{E}\|_1 + 2\lambda_3 \text{Tr}(\mathbf{F}^T \mathbf{L}_Z \mathbf{F})$$

$$\text{s.t.} \begin{cases} \mathbf{X} = \mathbf{XZ} \\ \text{diag}(\mathbf{Z}) = 0 \\ \mathbf{Z} \geq 0 \\ \sum_j z_{ij} = 1 \\ \mathbf{F}^T \mathbf{F} = \mathbf{I}_c \end{cases} \quad (2)$$

式中: $\mathbf{F} \equiv [\mathbf{f}_1^T, \mathbf{f}_2^T, \dots, \mathbf{f}_n^T]^T \in \mathbf{R}^{n \times c}$ 是一个使 $\mathbf{L}_Z$ 对角化(对角线为从大到小排序的特征值)的正交矩阵的前 c 列, λ_3 是一个正惩罚参数, $\text{Tr}(\cdot)$ 是迹运算符, $\mathbf{I}_c$ 表示单位矩阵;正惩罚参数 λ_3 越大, $\text{Tr}(\mathbf{F}^T \mathbf{L}_Z \mathbf{F})$ 越小。

利用增广拉格朗日方法^[35],将式(2)分离且改写成优化问题(3):

$$L(\mathbf{Z}, \mathbf{S}, \mathbf{F}, \mathbf{U}, \mathbf{E}) = \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \lambda_1 \|\mathbf{U}\|_* + \lambda_2 \|\mathbf{E}\|_1 + 2\lambda_3 \text{Tr}(\mathbf{F}^T \mathbf{L}_Z \mathbf{F}) + \langle \mathbf{C}_1, \mathbf{X} - \mathbf{XZ} - \mathbf{E} \rangle + \langle \mathbf{C}_2, \mathbf{Z} - \mathbf{S} \rangle + \langle \mathbf{C}_3, \mathbf{Z} - \mathbf{U} \rangle + \frac{\mu}{2} (\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{S}\|_F^2 + \|\mathbf{Z} - \mathbf{U}\|_F^2) \quad (3)$$

式中: $\mathbf{S}$ 和 $\mathbf{U}$ 为两个引入变量, $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ 为拉格朗日乘子, μ 为正惩罚参数, $\langle \cdot, \cdot \rangle$ 为矩阵内积, $\|\cdot\|_F$ 为矩阵的 F 范数。将式(3)的每个变量在其他变量固定的情况下交替求解,即可得到所有变量 $\mathbf{Z}, \mathbf{S}, \mathbf{F}, \mathbf{U}, \mathbf{E}$ 的解。

1.3.2 算法求解

Step1 更新 $\mathbf{Z}$: 固定变量 $\mathbf{S}, \mathbf{F}, \mathbf{U}, \mathbf{E}$, 可得

$$\mathbf{Z} = (\mathbf{X}^T \mathbf{X} + 2\mathbf{I})^{-1} (\mathbf{X}^T \mathbf{L}_1 + \mathbf{L}_2 + \mathbf{L}_3) \quad (4)$$

Step2 更新 $\mathbf{S}$: 固定变量 $\mathbf{Z}, \mathbf{F}, \mathbf{U}, \mathbf{E}$, 可得

$$\min_{s_i \geq 0, s_i, l=1, s_d=0} \|s_i - (h_i - g_i/\mu)\|_2^2 \quad (5)$$

式中: G 中的每一个元素由 $g_{ij} = \|x_i - x_j\|_2^2 + \lambda_3 \|f_i - f_j\|_2^2$ 计算得出, $H = Z + \frac{C_2}{\mu}$, s_i, h_i, g_i 分别表示 S, H, G 的第 i 行。式(5)具有显式解,可参考文献[26]。

Step3 更新 F : 固定变量 Z, S, U, E , 可得

$$\begin{aligned} F &= \arg \min_F \text{Tr}(F^T L_S F) \\ \text{s.t. } F^T F &= I, F \in \mathbb{R}^{n \times c} \end{aligned} \quad (6)$$

式(6)可通过特征值分解简单求解,其解为 L_S 的前 c 个最小特征值对应的 c 个特征向量的集合。

Step4 更新 U : 固定变量 Z, S, F, E , 然后利用奇异值阈值算子^[36]可得

$$U = \theta_{\lambda_1/\mu} \left(Z + \frac{C_3}{\mu} \right) \quad (7)$$

式中: θ 为SVT运算符。

Step5 更新 E : E 可通过求解最小化问题,得到

$$E = \Omega_{\lambda_2/\mu} \left(X - XZ + \frac{C_1}{\mu} \right) \quad (8)$$

式中: Ω 是收缩操作符^[37]。

Step6 更新 C_1, C_2, C_3, μ : 拉格朗日乘数 C_1, C_2, C_3 、惩罚参数 μ 分别用式(9)~(12)更新:

$$C_1 = C_1 + \mu(X - XZ - E) \quad (9)$$

$$C_2 = C_2 + \mu(Z - E) \quad (10)$$

$$C_3 = C_3 + \mu(Z - U) \quad (11)$$

$$\mu = \min(\rho\mu, \mu_{\max}) \quad (12)$$

式中: ρ 和 $\mu_{\max}$ 是常数。

可以认为,以上前5个步骤均有显式解。相关算法可在MATLAB中实现,具体如下:

算法1 scLRRARG算法

- 1) 输入: X ; 参数 $\lambda_1, \lambda_2, \lambda_3$; 聚类数 c
- 2) 初始化: 构造 c 个最近邻图作为 Z 的初始矩阵
 $U = Z, S = Z$; 计算 F 的初始矩阵; $E = 0$,
 $C_1 = C_2 = C_3 = 0, \rho = 1.2, \mu_{\max} = 10^8, \mu = 0.01$
 当未收敛时
 - (a) 利用(4)更新 Z ;
 - (b) 利用(5)更新 S ;
 - (c) 利用(6)更新 F ;
 - (d) 利用(7)更新 U ;
 - (e) 利用(8)更新 E ;

3) 直至收敛或超过最大迭代次数

4) 输出: Z, S, F, U, E

1.4 效果评估

本文将分析表1列出的5个带标签的scRNA-seq数据集,并将结果与当前的先进方法进行比较,评估scLRRARG算法在scRNA-seq类数据的聚类性能。

1.4.1 scLRRARG聚类方法

scLRRARG使用学习到的相似性作为向量嵌入每个细胞,总的来说,是通过谱聚类方法将习得的相似度矩阵进行聚类划分,并返回离散化后的聚类结果,用于确定每个细胞被划分到的聚类类别。

1.4.2 评估指标

为了评估聚类结果,本文计算了2个性能指标:调整Rand指数(Adjusted Rand index, ARI)和归一化互信息(Normalized Mutual Information, NMI)^[23, 38]。它们都是在scRNA-seq数据已知标记的背景下评估聚类结果的常用度量。

ARI的定义如下:假设将 n 个细胞聚成 k 个簇。设 $\{u_i\}_{i=1}^n$ 表示推断的聚类标签, $\{v_i\}_{i=1}^n$ 表示预先标注的标签,则有

$$\text{ARI} = \frac{\sum_{ls} \binom{n_{ls}}{2} - \left(\sum_l \binom{n_l}{2} \sum_s \binom{n_s}{2} \right) / \binom{n}{2}}{\left(\sum_l \binom{n_l}{2} + \sum_s \binom{n_s}{2} \right) / 2 - \left(\sum_l \binom{n_l}{2} \sum_s \binom{n_s}{2} \right) / \binom{n}{2}} \quad (13)$$

式中: l 和 s 枚举了 k 个簇, $n_l = \sum_i^n I(u_i = l)$, $n_s = \sum_i^n I(v_i = s)$, $n_{ls} = \sum_{i,j} I(u_i = l)I(v_j = s)$, $I(x = y)$ 为指示函数。

设 $p_l = n_l/n$, $q_s = n_s/n$ 和 $z_{ls} = n_{ls}/n$, 定义 $h(u) = -\sum_l p_l \log(p_l)$ 和 $h(v) = -\sum_s p_s \log(p_s)$ 分别为细胞的标签概率向量的熵, $i(u, v) = \sum_{l,s} z_{ls} \log(z_{ls}/p_l/q_s)$ 为互信息。然后将NMI定义为

$$\text{NMI} = i(u, v) / \sqrt{h(u)h(v)} \quad (14)$$

和ARI一样,NMI是一个完全重叠的聚类,它随着分歧的增加而减少。ARI和NMI的最大值和最小值分别为1和0。一般情况下,ARI或NMI值越大,聚类性能越好。

2 实验结果及讨论

2.1 scLRRARG的实验步骤

scLRRARG的具体步骤如图2所示。第一,对scRNA-seq数据进行预处理,其中行表示基因,列表

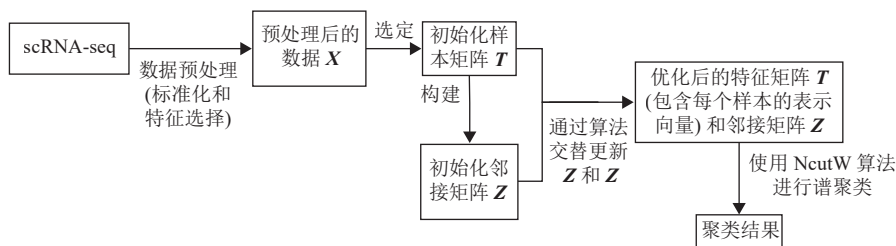


图2 scLRRAGR方法的处理流程图

Fig.2 Process flowchart of scLRRAGR method

示细胞。第二,对数据进行特征提取,得到初始化样本矩阵 T ,基于 T 构建初始化的邻接矩阵 Z ,结合低秩表示和自适应图正则化的技术,在保持数据局部结构的同时学习到最优的 T 和 Z 。第三,将邻接矩阵 Z 转换成样本之间的相似度,基于所得的相似度矩阵,使用NcutW函数对数据进行谱聚类,以此确定每个细胞被划分到的聚类类别。

2.2 scLRRAGR的实验结果对比

本节用到的6种对比聚类算法分别为Data-HiE-KM(Data Highly-expressed Genes-K-MEANS)^[39], Spearman-HiE-KM^[12](S-HiE-KM), SIMLR^[18], SC3^[14], SINCERA(Single Cell RNA-Seq Profiling Analysis)^[40], PcaReduce^[15],它们都是常见的scRNA-seq类数据的聚类方法。其中Data-HiE-KM是一种结合高亮特征提取和K-MEANS聚类的方法;S-HiE-KM是一种结合Spearman相关系数和K-MEANS聚类的方法;SIMLR是一种基于相似性传递的单细胞聚类方法;SC3是一种基于K-MEANS聚类和模块性优化的单细胞聚类方法;SINCERA是一种用于scRNA-seq图谱分析的计算管道,可以处理来自整个器官或分类细胞的scRNA-seq数据;PcaReduce是一种基于PCA的降维和聚类方法。

表2和表3分别列出了7种对比方法的NMI和ARI得分。针对Treutlein、Usoskin、Zeisel三个数据集,scLRRAGR方法聚类效果都是最好的;针对Patel、Pollen两个数据集,scLRRAGR方法只比SC3方法差,可能的原因是SC3受其算法设计、数据特征和参数配置等因素的影响,更适合处理这两个数据集,但实验结果同时也反映了SC3对参数和噪声更敏感,以及适用范围受限程度较高。以平均NMI和ARI得分为标准,scLRRAGR方法表现都是最好的。以上结果表明,scLRRAGR方法可提高scRNA-seq类数据的聚类性能。

2.3 稳定性

在本节中用实验证明了本文方法的收敛性。图3

表2 scRNA-seq数据集聚类方法的NMI值

Table 2 NMI values for clustering methods for scRNA-seq datasets

Method	Patel	Usoskin	Treutlein	Pollen	Zeisel	Average
scLRRAGE	0.95	0.98	0.80	0.90	0.91	0.91
Data-HiE-KM	0.76	0.37	0.33	0.84	0.51	0.56
S-HiE-KM	0.46	0.35	0.75	0.89	0.56	0.60
SIMLR	0.85	0.73	0.46	0.92	0.68	0.73
SC3	0.98	0.82	0.63	0.95	0.80	0.84
SINCERA	0.90	0.62	0.51	0.93	0.61	0.71
PcaReduce	0.60	0.60	0.39	0.93	0.57	0.62

表3 scRNA-seq数据集聚类方法的ARI值

Table 3 ARI values for clustering methods for scRNA-seq datasets

Method	Patel	Usoskin	Treutlein	Pollen	Zeisel	Average
scLRRAGE	0.93	0.93	0.71	0.91	0.80	0.86
Data-HiE-KM	0.78	0.29	0.19	0.71	0.46	0.49
S-HiE-KM	0.35	0.28	0.62	0.80	0.41	0.49
SIMLR	0.81	0.66	0.24	0.94	0.57	0.64
SC3	0.98	0.85	0.54	0.95	0.75	0.81
SINCERA	0.91	0.52	0.28	0.86	0.52	0.62
PcaReduce	0.48	0.55	0.38	0.89	0.51	0.56

显示了聚类准确性与迭代次数的关系(以Usoskin数据集为例),从图中明显可以看出聚类准确度逐渐增加,直至达到峰值后趋于平稳,经过几次迭代步骤后收敛到局部最优解。需要说明的是,由于K-MEANS对初始化非常敏感,因此图中呈现的聚类准确度会有轻微波动,这种情况是可以接受的。

scLRRAGR方法在进行低秩表示时,通过最小化残差矩阵(E)的低秩近似来重构原始数据,同时结合自适应图正则化技术,利用数据的邻近关系构建图结构(Z)来保持数据的局部结构,可以适应不同数据集的分布特征,这些过程对于存在异常值或噪声的数据具有一定的鲁棒性,提高了模型计算结果的稳定性。在实验中,本文在原始数据中引入0.1的噪声和10%的异常值处理后,实验结果如图4所示(以

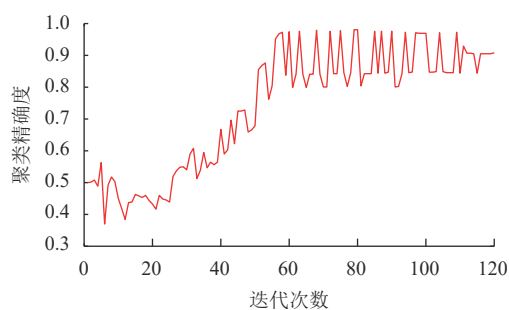


图3 Usoskin数据集聚类准确性与迭代次数的关系图

Fig.3 The relationship between clustering accuracy and the number of iterations in the Usoskin dataset.

Usoskin数据集为例),从图4中可以看出聚类准确度逐渐增加至峰值后也趋于平稳,准确度稍有降低,但算法仍然保持有效。说明即使面对干扰,scLRRAGR方法仍能够有效地处理类RNA-seq数据,并产生令人满意的结果。

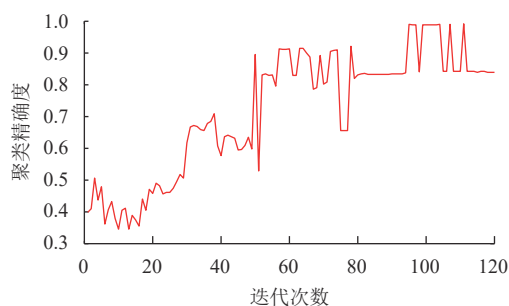


图4 引入0.1的噪声和10%的异常值处理后Usoskin数据集聚类准确性与迭代次数的关系图

Fig.4 The relationship between clustering accuracy and the number of iterations after introducing 0.1 noise and handling 10% outliers in the Usoskin dataset.

3 结论

本文提出了一种新的基于自适应图正则化低秩表示的scRNA-seq数据分析方法——scLRRAGR。准确对scRNA-seq数据进行分析是相关研究中的一个重要步骤,因为它们构成了识别、可视化和解释群体结构的基础。可靠和准确推断群体结构对于发现新的细胞(亚)类型、改善对现有细胞类型的表征和理解、解码健康(和异常)组织类型的细胞组成等都是必要的。

scLRRAGR方法使用了图学习来捕捉数据的非线性结构和相似性关系,运用自适应图正则化技术平衡了全局和局部信息,采用低秩表示方法进行降维和嵌入分析,并引入了秩约束来利用和学习先验信息。scLRRAGR从高维稀疏数据中学习相似性,能得到更好地反映scRNA-seq数据真实情况的聚类结

构,有助于揭示不同细胞类型和状态之间的差异。本文在真实数据集上的实验表明了scLRRAGR方法平均优于其他6种通用方法,展示了该方法在细胞聚类方面的良好性能。

参考文献:

- [1] BUETTNER F, NATARAJAN K N, CASALE F P, *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells [J]. *Nature Biotechnology*, 2015, 33(2): 155-160.
- [2] LUECKEN M D, THEIS F J. Current best practices in single-cell RNA-seq analysis: a tutorial [J]. *Molecular Systems Biology*, 2019, 15(6): e8746.
- [3] KUMAR P, TAN Y, CAHAN P. Understanding development and stem cells using single cell-based analyses of gene expression [J]. *Development*, 2017, 144(1): 17-32.
- [4] LLORENS-BOBADILLA E, ZHAO S, BASER A, *et al.* Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury [J]. *Cell Stem Cell*, 2015, 17(3): 329-340.
- [5] SHALEK A K, SATIJA R, ADICONIS X, *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells [J]. *Nature*, 2013, 498(7453): 236-240.
- [6] NAVIN N E. Delineating cancer evolution with single-cell sequencing [J]. *Science Translational Medicine*, 2015, 7(296): 296fs29.
- [7] KOWALCZYK M S, TIROSH I, HECKL D, *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells [J]. *Genome Research*, 2015, 25(12): 1860-1872.
- [8] SANDBERG R. Entering the era of single-cell transcriptomics in biology and medicine [J]. *Nature Methods*, 2014, 11(1): 22-24.
- [9] NAVIN N E. Cancer genomics: one cell at a time [J]. *Genome Biology*, 2014, 15(8): 1-13.
- [10] ADAM M, POTTER A S, POTTER S S. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development [J]. *Development*, 2017, 144(19): 3625-3632.
- [11] SVENSSON V, NATARAJAN K N, LY L H, *et al.* Power analysis of single-cell RNA-sequencing experiments [J]. *Nature Methods*, 2017, 14(4): 381-387.
- [12] GRÜN D, LYUBIMOVA A, KESTER L, *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types [J]. *Nature*, 2015, 525(7568): 251-255.
- [13] STREHL A, GHOSH J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions [J]. *Journal of Machine Learning Research*, 2002, 3: 583-617.

- [14] 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.
- [15] ŽURAUSKIENĖ J, YAU C. pcaReduce: hierarchical clustering of single cell transcriptional profiles [J]. *BMC Bioinformatics*, 2016, 17: 1-11.
- [16] LIN C, JAIN S, KIM H, *et al.* Using neural networks for reducing the dimensions of single-cell RNA-Seq data [J]. *Nucleic Acids Research*, 2017, 45(17): e156.
- [17] DEMUTH H B, BEALE M H, HAGAN M T, *et al.* Neural network design[M]. Boulder: University of Colorado, 2014.
- [18] 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.
- [19] LANCKRIET G R G, DE BIE T, CRISTIANINI N, *et al.* A statistical framework for genomic data fusion [J]. *Bioinformatics*, 2004, 20(16): 2626-2635.
- [20] POUYAN M B, KOSTKA D. Random forest based similarity learning for single cell RNA sequencing data [J]. *Bioinformatics*, 2018, 34(13): 79-88.
- [21] WEN J, FANG X, XU Y, *et al.* Low-rank representation with adaptive graph regularization [J]. *Neural Networks*, 2018, 108: 83-96.
- [22] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: Analysis and an algorithm. [C]//Advances in Neural Information Processing Systems (NeurIPS) . Vancouver: MIT Press, 2001.
- [23] YIN M, GAO J, LIN Z. Laplacian regularized low-rank representation and its applications [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 38(3): 504-517.
- [24] LIU J, CHEN Y, ZHANG J, *et al.* Enhancing low-rank subspace clustering by manifold regularization [J]. *IEEE Transactions on Image Processing*, 2014, 23(9): 4022-4030.
- [25] DU S, MA Y, MA Y. Graph regularized compact low rank representation for subspace clustering [J]. *Knowledge-Based Systems*, 2017, 118: 56-69.
- [26] NIE F, WANG X, JORDAN M, *et al.* The constrained laplacian rank algorithm for graph-based clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016.
- [27] AHLMANN-ELTZE C, HUBER W. Comparison of transformations for single-cell RNA-seq data [J]. *Nature Methods*, 2023, 20: 665-672.
- [28] DORFMAN R A. A note on the delta-method for finding variance formulae [J]. *Biometric Bulletin*, 1938, 1: 129-138.
- [29] TOWNES F W, HICKS S C, ARYEE M J, *et al.* Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model [J]. *Genome Biology*, 2019, 20: 1-16.
- [30] ROBINSON M D R, ROBINSON M D, SONREL A S, *et al.* PipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools [J]. *Genome Biology*, 2020, 21(1): 227.
- [31] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: Algorithm, theory, and applications [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(11): 2765-2781.
- [32] LIU G, LIN Z, YAN S, *et al.* Robust recovery of subspace structures by low-rank representation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 35(1): 171-184.
- [33] VON LUXBURG U. A tutorial on spectral clustering [J]. *Statistics and Computing*, 2007, 17: 395-416.
- [34] FAN K. On a theorem of Weyl concerning eigenvalues of linear transformations I [J]. *Proceedings of the National Academy of Sciences*, 1949, 35(11): 652-655.
- [35] WEN J, XU Y, LI Z, *et al.* Inter-class sparsity based discriminative least square regression [J]. *Neural Networks*, 2018, 102: 36-47.
- [36] LIU G, YAN S. Latent low-rank representation for subspace segmentation and feature extraction[C]// 2011 International Conference on Computer Vision. Barcelona: IEEE. 2011.
- [37] LIN Z, CHEN M, MA Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices[EB/OL]. arXiv: 1009.5055(2010-09-26)[2023-12-15]. <https://arxiv.org/abs/1009.5055>.
- [38] HUANG J, NIE F, HUANG H. A new simplex sparse learning model to measure data similarity for clustering[C]// Twenty-fourth International Joint Conference on Artificial Intelligence. Buenos Aires: AAAI, 2015.
- [39] JAIN A K, DUBES R C. Algorithms for clustering data[M]. Prentice Hall, 1988.
- [40] GUO M, WANG H, POTTER S S, *et al.* SINCERA: a pipeline for single-cell RNA-Seq profiling analysis [J]. *PLoS Computational Biology*, 2015, 11(11): e1004575.

(责任编辑: 张玮欣 英文审核: 熊荣斌)