使用进化多目标集成修剪解释单细胞RNA-seq

**摘要**

***动机***：近年来，单细胞RNA测序使我们能够发现细胞类型甚至亚型。 可用性的提高为从单细胞RNA序列数据识别细胞群体提供了机会。 计算方法已被用来揭示在多个细胞群体中基因表达的变异。 不幸的是，现有的可能会受到现实的限制例如实验噪声，数值不稳定性，高维数和计算可伸缩性。

***结果***：我们提出了一种进化的多目标集成修剪算法（EMEP），该算法可以解决这些现实的限制。我们的EMEP算法首先将无监督维数缩减应用于从原始高维到低维子空间的投影数据；

在这些新的子空间中应用了基本的聚类算法以生成不同的聚类结果形成簇集成。但是，大多数这些簇集成都是不必要的，体积大，但要花费额外的时间和内存。为了解决这个问题，

EMEP旨在从集合中动态选择合适的聚类结果。

此外，为了指导多目标集合的演化，包括三个聚类有效性指标：

总体集群偏差，集群内部紧密度和基本分区集群的数量被公式化为目标函数，利用进化的多目标优化来释放其细胞类型发现性能。我们将EMEP应用于55个模拟数据集和7个真实单细胞RNA-seq数据集，包括6个单细胞RNA-seq数据集和1个大规模数据集，具有3005个细胞和4412个基因。两个案例研究也被用来揭示EMEP的生物学相关性。我们发现，与其他聚类算法相比，EMEP可以获得更好的性能，证明了EMEP可以很好地识别细胞群。

***可用性和实现***：EMEP用Matlab编写，可从https://github.com/获得。

lixt314 / EMEP

联络人：kc.w@cityu.edu.hk

补充信息：补充数据可从Bioinformatics在线获得。

**1引言**

单细胞RNA-seq技术已被证明可有效用于通过基于转录组谱检测异种细胞群体中的亚群，发现新的细胞类型。 实际上，从单细胞RNA-seq数据中鉴定细胞类型被认为是无监督学习中的聚类问题。 因此，计算方法，包括k-均值，主成分分析和频谱聚类（SC）通常用于识别细胞类型。 RNA-seq的快速发展使我们对大量的单细胞RNA测序数据进行测序，带来计算上的挑战； 例如，抄本放大噪声，丢失事件，高维和数据稀疏（Kiselev et al。，2017; Wang et al。，2017）。 这些计算难题给在用于细胞群体解释的单细胞RNA-seq数据上开发有效的无监督聚类带来了困难。

过去，特定于应用程序的无监督聚类方法已经开发来解决这些计算难题；例如，Kiselev（2017）提出了一种称为单细胞共识聚类（SC3）的无监督聚类，该聚类通过共识方法整合了多个簇标记，并且可以从肿瘤细胞的转录组中改善细胞类型的识别。王（2017）通过从单细胞RNA序列数据中学习相似性度量，提出了通过多核学习（SIMLR）进行单细胞解释的方法。朱（2017）应用了经典的非负矩阵分解（NMF），并将其与其他非监督聚类方法进行了比较；结果表明，非NMF可以识别交互模块。张（2018a）利用了来自生物复制品（scVDMC）的多个单细胞群体用于单细胞RNA-seq解释。 scVDMC算法是一种具有嵌入式特征选择的多任务学习方法，可同时捕获差异表达的基因。张（2018b）提出了一个可解释的框架DendroSplit基于功能选择来发现多个级别的单细胞RNA-seq聚类问题。 Park（2018）提出了一部小说使用多个双重随机相似矩阵的SC框架形成用于对单元格类型进行聚类的相似矩阵。杨（2017）提出了通过迭代聚类的单细胞分析，以找到用于将细胞分为不同组的最佳签名基因集基于具有最佳参数的迭代聚类。然而，几乎不认为，这些无监督的聚类方法中的每一种都可以成为所有单细胞RNA-seq解释数据集的历史赢家。实际上，每种聚类算法都有其自身的优缺点。 不同的聚类算法在不同的单细胞RNA-seq数据集上提供不同的性能。因此，用户很难确定哪种聚类算法最适合单细胞RNA-seq数据。

聚类集成已成为一种有效的方法，可以将来自多个独立无监督聚类算法的解决方案集成到共识结果中。事实证明，聚类集成可以有效解决现实世界中的问题：医学诊断的集成聚类（Greene等，2004），模糊集成聚类（Avogadri和Valentini，2009），基于链接的聚类集成（LCE）方法（ Iam-On等人，2010a，b），基于图的共识聚类分析（Yu等人，2007），DNA微阵列数据的集成框架，用于蛋白质-蛋白质相互作用网络聚类的集成框架（Asuret等人，2007） ，集合非NMF方法（Greene等，2008）和基于知识的集群集合用于基于生物分子数据的癌症发现（Yu等，2011）。 集群合奏的详细列表可以参考过去的调查（Yang等，2010）。

不幸的是，大多数这些现有的集群集成方法可能会产生不必要的大型集成，但会花费额外的时间成本和内存消耗。

不幸的是，大多数这些现有的集群集成方法可能会产生不必要的大型集成，会花费额外的时间成本和内存消耗。为了解决这些限制，建议对集合进行修剪以从集合中选择合适的集群。

实际上，整体修剪的目标是减少群集数量，而不会牺牲准确性。直观地，整体修剪的目标既要使泛化性能最大化，又要使正则化的簇数最小化。不幸的是，这两个目标通常是矛盾的。在这两个目标之间进行权衡时，需要启用最佳决策。在这种情况下，考虑整体修剪是一个多目标问题，而不是一个单一目标问题。因此，提出了一种进化多目标集合修剪（EMEP）算法，以动态选择基本的聚类算法作为集合；它可以被认为是具有二进制权重的权重集合簇的特例。在55个模拟数据集，7个真正的单细胞RNA-seq数据集和2个案例研究中与其他方法进行的广泛比较表明，EMEP在几种最新的聚类方法上显示出竞争优势。

**2 材料和方法**

*2.1 EMEP方法论概述*

在本节中，我们提出用于单细胞RNA序列数据的EMEP算法。图1总结了EMEP的框架。

考虑具有n个细胞和m个基因的n×m个单细胞RNA-seq数据的矩阵X，我们提出的算法EMEP包括三个重要的组件（图1）。在第一部分中，可以采用NMF缩小基因空间的维数（即单细胞RNA-seq数据矩阵X）。代数上，NMF可以将X分解为非负n×r基矩阵W和非负r×m系数矩阵H的乘积。对于图1中不同数量的秩模型r，我们生成各种基向量W={W1; W2; ...;Wd}用于聚类，其中d是各种基本向量的数量。在这项工作中，我们将秩模型的数量设置为2到20。需要注意的是，可以选择任何基本的聚类算法来对集合W中的各个基本向量进行聚类，并获得多个聚类结果。例如，由于其简单性和高效的性能，我们可以在此步骤中选择K-means（KM）聚类算法。

|  |
| --- |
| 图1：EMEP管道的总体框架。 （a）第一部分是降低尺寸性； 可以采用NMF从单细胞RNA-seq数据矩阵X减少基因空间的维数。要注意的是，对于各种秩模型，该算法会生成各种基础向量W；（b）第二部分是基本分区； 例如，可以选择KM聚类算法作为基本划分算法。 对于集合W中的不同基向量，KM聚类算法可以获得多个聚类解π={π1，π2，……，πd}（c）第三部分是EMEP。 它从集合中删除不合适的聚类结果。 执行不同共识函数和聚类算法之间的自适应选择，以迭代方式生成聚类解 |

然后，EMEP删除了多个聚类结果π={π1，π2，……，πd}中的一些并进一步提高了聚类的泛化性能。给定一个聚类算法C和一组基本矢量W={W1; W2; ...;Wd} C：W→Y将每个基本向量W映射到标签空间Y，并且Cs表示带有所选向量si ∈{0,1}d的修剪的集合；其中si=1表示选择了Wi上的聚类结果。为了指导多目标集合修剪，选择了三个聚类有效性指标（即，总体聚类偏差，聚类内部紧实度和基本集合分区数）作为目标函数，以捕获在合奏修剪期间不断变化的聚类的多个特征。在进化过程中，对于不同的单细胞RNA-seq数据集，他们需要具有不同聚类算法的不同共识功能。因此，要对特定的单细胞RNA-seq数据进行聚类，在不同的进化阶段使用不同的聚类算法的不同共识功能将是有益的。因此，在整个进化过程中会保留具有不同聚类参数设置的不同共识函数池，从而导致不同聚类算法之间的进化选择竞争。在可用的共识函数中，为简洁明了，我们选择其中三个，包括基于三联的相似度（CTS）矩阵，基于SimRank的相似度（SRS）矩阵和基于SimRank的近似相似度（ASRS）矩阵。对于集群，选择并比较了KM聚类算法，SC和通过快速搜索和发现密度峰（CDP）进行聚类。

*2.2无监督降维*

为了解释高维单细胞RNA-seq数据集，采用NMF作为无监督的降维方法，将数据从原始高维空间投影到低维子空间（Gupta和Xiao，2011）。 NMF（Lee和Seung，2001年）是一种经过充分研究的无监督学习算法，它通过将以下目标（Frobenius范数）最小化，将atrix X分解为两个非负矩阵W ∈Rn×r和H ∈ Rr×m。 W和H的非负约束：



其中‖•‖F表示Frobenius范数。为了优化目标，迭代以下乘法更新规则，直到收敛为止，

优化这些目标后，可以通过NMF将高维单细胞RNA-seq数据集投影到低维子空间。 已经证明上述更新过程可以达到LNMF的局部最小值。 通过使用不同的秩模型r，NMF可以获得各种基本矩阵W。我们将这些基本矩阵W设置为集合W={W1; W2; ...;Wd}。 然后可以选择一个聚类算法作为基本划分算法，以对集合W中的每个不同基础矩阵进行聚类，并获得多个聚类结果π={π1，π2，……，πd}。

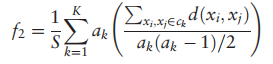
*2.3目标函数*

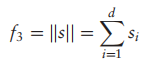
降维后，提出了EMEP算法以发展用于RNA-seq细胞类型发现的多个聚类结果。

为了指导发展，必须仔细设计目标函数。 我们注意到，整体修剪的目标是最大化泛化性能并最小化所选基本分区集群的数量。 因此，对于第一个目标，我们考虑两个目标函数：（i）聚类中心与其数据点之间的距离之和； （ii）同一集群内的数据一致性。 对于第二个目标，第三个目标函数旨在最大程度地减少所选基本分区集群的数量。

第一个目标函数与聚类偏差有关。 它计算分区的整体偏差（Mukhopadhyay等，2015）。 它被计算为数据点与其对应的聚类中心之间的距离的总和。d(ck.xi)是数据点xi及其对应的聚类中心ck之间的距离（例如，欧几里得距离）。 基于此定义，我们可以观察到它与KM具有相似的策略。

第二个目标函数是最小化聚类的紧密度（Iam-on等人，2010a，b）；这是另一种常用的测量方法。紧密度度量的是同一群集中每对数据点之间的平均距离；它可以表示为：

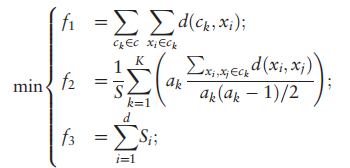
其中S是RNA-seq数据集中的细胞数。 K表示簇数。 ck是第k个聚类集。 ak是属于第k个簇的像元数。 从概念上讲，同一组中的元素应尽可能彼此靠近。 因此，应将f2值最小化。

最后一个目标函数是最大程度地减少为进行正则化而选择的基本分区集群的数量。给定聚类算法C和基向量W的集合，则构造每个基向量和标签空间之间的映射C：W→Y。 令Cs表示第二个经过修剪的集合，其二进制掩码向量si ∈{0,1}d。 基本分区集群的数量可以描述为：

*2.4帕累托最优方法*

在EMEP中，我们将整体聚类偏差，聚类内紧实度和基本分区聚类的数量设计为目标函数，并将整个多个聚类结果中的集合子集视为优化这三个目标函数的候选解决方案。 对于第一个和第二个目标，分数越低，每个基本向量W和标签空间Y之间的聚类越好。对于第三个目标，最小化所选基本分区聚类是合奏修剪的目标。 因此，针对高维单细胞RNA-seq数据集的整体簇修剪问题可被视为针对这些目标的多目标优化问题。

在这三个相互冲突的目标下解释单细胞RNA-seq数据集，困难在于存在可解释的数学解。 每个目标通常相互冲突。 换句话说，对一个目标有利的解决方案可能对另一个目标不利。 因此，很难找到一个满足所有目标函数的解决方案。一个以上的目标不能保证一个最优。这些目标之间的关系可以在本文中描述的：决策向量p1∈P被称为Pareto-dom inate决策向量p2∈P如果……，其中fe（）是先前定义的目标函数，E是最小化的目标函数的数量。

如果满足这些条件，则决策向量（也称为解）p1∈P决定了决策向量p2∈P。 以两个目标为例，设计空间（即解决方案空间）和目标空间之间的关系可以在图2中举例说明。可以对集合聚类修剪进行扩展，以找到非支配的解决方案集。 其多目标优化可总结如下：

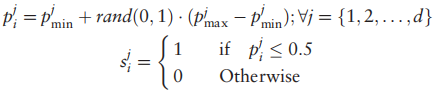
在这里，我们基于分解方法（tchebycheff方法）设计EMEP算法，以阐明所有非支配解。

*2.5进化多目标集合修剪*

在前面的部分之后，提出EMEP用于在本部分中对高维单细胞RNA-seq数据进行聚类和解释。

EMEP过程包括目标载体的定义，突变，交叉和基于tchebycheff的分解方法。 变异和交叉操作主要用于更新种群中的当前个体。 tchebycheff分解方法的重点是将多目标单细胞RNA-seq聚类问题分解为许多单目标单细胞RNA-seq聚类子问题。

•2.5.1目标向量定义

为了初始化，具有N个参数向量的总体对每个候选解pi={pi1,pi2,...pid}, i={1,2,...,N}进行编码；每个向量（或候选解）也与每个子问题相关联。 初始种群应通过在上下边界p内对个体进行随机化来尽可能覆盖整个搜索空间

其中si=1表示选择了基本群集Cj，而sij=0表示针对第i个解决方案pi删除了基本群集Cj。

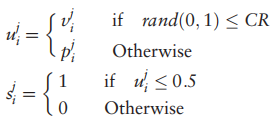
rand（0，1）是[0，1]范围内的随机变量。

•2.5.2突变

在初始化阶段之后，演化阶段主要用于突变和交叉操作。 变异和交叉操作的灵感来自差异进化（Das和Suganthan，2011年）。 诱变操作用于生成对应于溶液载体pi的突变载体vi，其描述如下：

其中r1，r2和r3是从总体中随机选择的三个指标。 F是称为比例因子的差分权重参数，可以缩放差异向量。

•2.5.3分频

在突变阶段之后，将交叉操作应用于目标向量pi和突变向量vi，以产生试验向量u，如下所示：其中CR∈[0,1]是交叉速率，它控制从突变载体复制的值的分数。 该交叉操作将突变向量vi的第j个参数复制到试验向量ui中的相应元素。 否则，它将从目标向量pi复制第j个参数。获得ui后，可以将其转换为二进制聚类选择空间si，该空间用于选择基本聚类以形成整体，从而为第i个个体生成最终聚类结果（p）。

•2.5.4 Tchebycheff分解方法

*2.6共识函数的池选择和聚类算法*

聚类算法对单细胞RNA-seq数据集的有效性取决于所选的共有函数及其相关的基本聚类算法。 但是，不同的单细胞RNA-seq数据集需要使用各种聚类算法的不同共识函数。 此外，要对单细胞RNA-seq数据集进行聚类，具有不同聚类算法的不同共识函数可以在进化的不同阶段相互竞争，并且比整体聚类算法中具有单个聚类算法的单个共识函数表现更好。

基于这样的观察，我们提出了共识函数和聚类算法的集合，作为进化多目标优化的自适应选择，其中共识函数池以及与每个相关基本聚类算法相对应的算法池竞争产生成功的下一代人口。如图3所示，共识函数和聚类算法的候选库被设计为展现出各种特征，以便它们可以在开发中实现强大的性能特征。从该图可以看出，pi和ui是两个修剪解决方案，它们表示从p中选择的聚类结果，以形成不同的集合。每个成员都分配有一个共识函数和相关的聚类算法，该算法取自各个池以产生最终的聚类结果。然后，如果所生成的试验向量ui优于目标向量pi，则共识函数和相关的聚类算法将与试验向量ui一起保留，后者将成为下一代的目标向量。对于共识功能库，考虑了三个共识功能，包括CTS矩阵（Klink等，2006），SRS矩阵（Calado等，2006）和ASRS矩阵（Iam-On等，2012）。 。对于池中的基本聚类算法，选择KM聚类算法，SC（Von Luxburg，2007）和CDP（在《科学》上发表（Rodriguez和Laio，2014））来进化解释那些用于细胞群体鉴定的单细胞RNA-seq数据集。

|  |
| --- |
| 图3.所提出的在共识函数和基本聚类算法之间进行迭代选择的自适应方法。 首先，在应用NMF和基本分区算法后获得多个聚类结果p。 pi和ui是两个修剪解决方案，它们表示从p中选择的聚类结果，以形成不同的集合。 然后，自适应选择可以从池中选择一个共识函数，然后找到相应的聚类算法以产生最终的聚类结果。 考虑了三个共识函数，包括CTS矩阵，SRS矩阵和ASRS矩阵。 对于池中的基本聚类算法，选择了KM聚类算法，SC和CDP。 最后，可以获得新的聚类结果π1\*和π2\*。 |

**4 讨论**

在这项研究中，基于观察到并非所有聚类结果都适合于所有单细胞RNA-seq数据分布的观察，提出了一种基于进化修剪（EMEP）的新颖多目标集成算法。在该算法中，采用降维方法将数据从原始的高维空间投影到低维子空间。提出了三个不同的聚类有效性指标，包括总体聚类偏差，聚类紧实度和所选基本分区聚类的数量，作为目标函数，以捕获不断变化的聚类的多个特征。之后，提出了EMEP来从集合中删除不合适的聚类，从而提高泛化性能。根据实验结果，与60多个单细胞RNA-seq数据集上的9种聚类方法相比，EMEP可以证明NMI和ARI的显着优势。进行了两个案例研究，包括胰岛单细胞和人类癌细胞，以证明EMEP可以从单细胞RNA-seq数据中清楚地区分不同的细胞类型。

尽管EMEP对于单细胞RNA-seq数据具有良好的性能，但是该算法存在一些局限性。由于EMEP是一种基于整体的方法，因此通常会很费时