# 基于负二项分布的单细胞数据缺失值分治插补研究

熊珍珍[a, b]，张本龚[*b, c]

(武汉纺织大学 a.计算机与人工智能学院，b.应用数学与交叉科学研究中心，c.数理科学学院，湖北 武汉 430200)

**摘 要：**单细胞转录组测序（scRNA-seq，single cell RNA sequencing）技术为单个细胞高通量、高分辨率的深入研究提供了机会，为在单细胞层面研究细胞功能及其背后的基因调控机制提供了重要技术手段。然而这项技术也带来新的挑战，单细胞数据具有规模大、噪声高、异构性强等特点，特别是高比例的数据缺失（dropout）严重影响了下游分析的可靠性，甚至掩盖了基因与基因间的重要关系。这里提出一种基于负二项分布的分治插补策略 ND-Impute（Negative binomial distribution based Divide and conquer strategy for imputation）对 scRNA-seq 数据进行处理，该方法假设 scRNA-seq 数据符合负二项分布，利用包含特定损失函数的自动编码器获取数据的特异性参数，并使用分治策略估计潜在的基因表达值。通过聚类效果、相关性和误差分析等比较，表明该方法可以有效地恢复缺失数据，提高了后续研究分析的准确性。

**关键词：**单细胞转录组测序；数据缺失；插补策略；聚类分析

**中图分类号：**O211.9 **文献标识码：**A **文章编号：**2095－414X(2023)01－0014－07

## 0 引言

自然界中的生物是复杂且多样的，对生物学表征的研究，离不开对细胞的深入分析，无论是单细胞生物还是多细胞生物，单个细胞之间的差异都会产生不可估量的影响。随着细胞分离技术以及高通量测序技术的发展，bulk RNA-seq 技术逐渐被 scRNA-seq 技术所取代，scRNA-seq 技术的出现为单个细胞的高吞吐量和高分辨率转录组分析提供了新途径[1-2]。自 2009 年首个 scRNA-seq 技术发布以来，scRNA-seq 被越来越广泛地应用于基础科学研究中，这项技术尤其在肿瘤学[3]、遗传病学[4]、免疫学[5]等生物医学研究中发挥了重要作用。单细胞转录组测序的主要步骤是分离出单个细胞、将 mRNA 转为 cDNA、cDNA 扩增、高通量测序和数据分析等，测序所得数据可以使研究人员进一步了解异质性细胞类型、细胞发育谱系以及疾病发展状态，对于发现新的异质细胞类型和追踪细胞动态发育轨迹也有巨大帮助[6]。然而由于技术限制和生物因素，scRNA-seq 数据比 bulk RNA-seq 数据更复杂，噪声更大，目前 scRNA-seq 的通量已经由原来的几个细胞发展到成千上万个细胞，通量的增加提高了数据的复杂性。即使在现在的单细胞转录组测序研究中，数据噪声问题依然普遍，尤其是缺失（dropout）事件，dropout 的产生源于单细胞转录组的测序过程，因此难以避免[7]。

scRNA-seq 数据由于其生物学特性和测序技术因素产生噪声。首先在生物学方面，细胞的类型、大小和细胞的表达方式各不相同；其次对于测序技术方面，在使用细胞中表达量较低的 RNA 构建 cDNA 库的过程中，导致一些转录组的丢失是必然的，cDNA 库的扩增会引入更多的随机变异。同时，测序效率低会使每个细胞中的转录组只有很小一部分被成功测序。上述原因就导致每个细胞基因表达量中有很大比例的零值和低计数值，其中很多原本低表达的基因数据被忽略，测得结果为零，也就是数据缺失 dropout，然而其中有部分零值反映的是真正的零表达量，这极大混淆了低表达基因和不表达基因，从而影响下游分析的准确性和可信度[8]。

为了解决 scRNA-seq 数据中存在大量数据缺失的问题，研究人员开发了许多插补方法，这些方法大致可以分为三类：统计学方法、深度学习方法和集成学习方法。早期的一些方法从 scRNA-seq 数据的统计学分布特性出发，利用数学的方法和统计学思维解决数据中零值过多的问题。其中最经典的统计学方法是 2018 年的 SAVER[9]，该方法利用基因与基因的关系来恢复每

*通讯作者：张本龚（1979-），男，教授，博士，研究方向：计算机系统生物学和机器学习.

个细胞中每个基因的表达水平，通过具有 Poisson-LASSO 回归的经验贝叶斯方法估算先验参数，构建负二项分布模型（NB，Negative binomial distribution），最后输出后验分布的均值作为插补结果。但是这种方法会改变所有基因的表达水平，包括那些不受 dropout 影响的基因表达水平，因此需要考虑基因表达数据中零值为 dropout 的概率。

基于对 dropout 概率的分析，Wei[10]等人在 2018 年提出 scImpute，该方法基于混合模型计算每个表达值是 dropout 的概率，并利用相似细胞的相同基因的表达信息来估算 dropout 项的潜在表达值。但基于统计学的方法只能处理很小的数据量，所以目前越来越多出现和使用的插补算法都使用到了深度学习方法。基于深度学习的方法不仅能够处理大规模数据，还能利用深度学习中的神经网络自主学习 scRNA-seq 数据中的一些特征，达到比基于统计学的方法更优秀的插补效果[11]。比较有代表性的方法有 Eraslan G[12]等在 2019 年提出的基于深度自动编码器的数据插补去噪方法 DCA，该方法采用负二项噪声模型，考虑了数据的计数分布和稀疏性，通过改进的自动编码器学习给定数据集中的特殊参数，此方法与细胞的数量成线性比例，可以应用于数百万细胞的数据集，但该方法着重考虑的是基因表达数据的统计分布情况，而忽略了基因表达数据的特异性。Arisdakessian[13]等提出了 DeepImpute，一种基于深度神经网络的插补算法，该方法最大的特点就是利用了分治法的思想，把 scRNA-seq 数据分为多个子集后，使用 dropout 层和 loss 函数来学习数据中的模式，从而实现准确的插补，与 DCA 相反，该方法没有考虑到数据原本的统计分布，直接使用了深度学习方法，可能会产生更多的噪声。2020 年 Xu[14]等提出了把生成对抗网络 (GANs)用于 scRNA-seq 数据插补的方法 scIGANs，它使用生成的细胞而不是观察到的细胞来避免这些限制，并平衡主要和稀有细胞群之间的每个特性。该方法的主要思路首先是将基因表达矩阵转换成图片，具体规则为一张图片代表一个细胞，图片中的像素代表各个基因的基因表达水平，然后将转换的图片输入 GANs 模型中进行训练，最终获得插补结果。但将生成对抗网络运用到基因表达数据的图片形式，数据生成的图片与真实图片还是有所差距。深度学习方法相对于统计学方法，同样存在缺陷，例如模型选择不佳、网络训练过拟合等问题，可能会造成更多噪声的引入。随着插补方法的不断衍生，scTSSR[15]和 EnTSSR[16]这样的集成学习方法通过对比加权，实现了更全面的分析，然而要消耗大量的计算和内存资源。

在上述的工作中，各种方法都有切入面，面对基因表达数据，首先希望能找到一个合适的统计分布模型来规范数据，其次面对数据的大规模增长，需要引入深度学习方法处理巨大的计算量，而集成学习方法对资源的消耗十分巨大，因此，仍然需要一种高效的插补方法对 scRNA-seq 数据进行插补。自动编码器作为一种无监督的神经网络模型，它分为编码器和解码器两个模块，使用编码器将原数据进行压缩，学习数据中的特征，再使用解码器将数据还原。自动编码器是一个无监督的学习过程，它可以学习到输入数据的隐含特征。

本文旨在使用深度学习方法，利用自动编码器的特点来学习和处理 scRNA-seq 数据，并最终达到较好的缺失数据插补效果。这里提出一种基于负二项分布的分治插补策略（ND-Impute），以用于 scRNA-seq 数据的插补，该方法在两种基于深度神经网络的方法 DCA 和 DeepImpute 的分析思维上加以改进，既考虑了单细胞转录组测序数据的特殊统计学分布模型，也考虑到了基因表达数据的特异性。大量的实验结果表明，ND-Impute 在关键性能指标，即聚类效果、相关性和误差分析方面优于其他相似的方法。

## 1　工作介绍

### 1.1　数据集介绍

本文采用五个不同大小的 scRNA-seq 数据集（68KPBMC、293T Cells、Jurkat Cells、CD34+ cells 和 CD19+ B cells）来测试 ND-Impute 方法对 dropout 事件的插补能力。其中，68KPBMC[17]数据集是外周血细胞数据集，293T cells(10X Genomic)是人体上皮细胞，Jurkat cells (10X Genomic)是某种 T 淋巴细胞，CD34+ cells[18]是人类骨髓造血干细胞，CD19+ B cells 是免疫细胞的一种。这些公开数据集的大小如表 1 所示。

表 1　scRNA-seq 公开数据集大小

| Datasets | Number of cells | Number of genes |
|---|---|---|
| 68KPBMC | 68579 | 32738 |
| 293T Cells | 2885 | 3393 |
| Jurkat Cells | 3258 | 3204 |
| CD34+ Cells | 9232 | 1274 |
| CD19+ B Cells | 10085 | 478 |

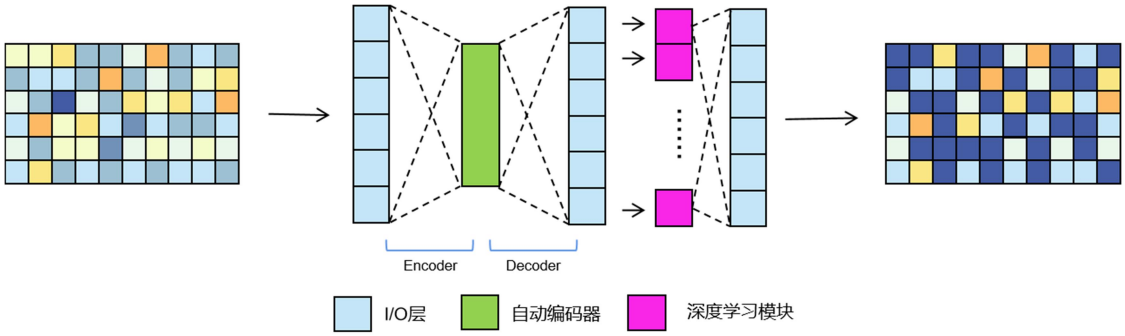


图 1 ND-Impute 算法的工作流程

### 1.2 负二项分布模型

首先，合理假设数据的分布情况有助于下游分析，scRNA-seq 数据的统计学分布情况比较特殊，其具有高稀疏性(high sparsity)，表现为数据中存在大量零值(zero inflation)，且数据也是高度离散的[18-19]。因此对于基于计数读取的 scRNA-seq 数据，描述连续型数据的正态分布不符合要求；其次泊松分布适合于描述单位时间（或空间）内随机事件发生的次数（事件发生的次数只能是离散的整数），但由于均值始终与方差相等，故无法体现 scRNA-seq 数据的高离散度特性。这里，从方差大于均值的角度出发，负二项分布是可行的。首先，二项分布描述的是在 $n$ 重伯努利试验中，事件 B 恰好发生 $x(0 \le x \le n)$ 次的概率，以二项分布 $X \sim B(n, p)$ 为例，$n$ 次试验中正好得到 $k$ 次成功的概率密度函数为：

$$P\{X = k\} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{1}$$

负二项分布描述的也是伯努利试验，这个分布描述的事件是当某个结果出现固定次数时，整个过程的数量分布。设试验持续到 $r$（$r$ 为整数）次失败，$p$ 表示一个事件在该伯努利试验中每次出现的概率，那么其负二项分布的概率密度函数为：

$$f(k; r, p) = \frac{(k+r-1)!}{k!(r-1)!} p^k (1-p)^r \tag{2}$$

该分布的均值和方差分别为：

$$\mu = \frac{pr}{1-p} \tag{3}$$

$$\sigma^2 = \frac{pr}{(1-p)^2} \tag{4}$$

进一步可以得到：

$$\sigma^2 = \mu + \alpha\mu^2 \tag{5}$$

从式(5)可以看出，方差随着均值增加呈现二次函数形式的递增，因此符合 scRNA-seq 数据的特点。同时，经过学者们的研究，负二项分布被证明是一种可以合理描述 scRNA-seq 数据的模型，并且该分布模型已被一些经典的插补方法所使用的，如 SAVER，scVI 等[20]。

## 2 方法

ND-Impute 算法的工作流程如图 1 所示，其主要包括两个模块：（1）数据集读取，获取数据集符合负二项分布模型的均值；（2）采用分治法构建训练网络，训练数据集，并得到插补结果。图中输入基因表达矩阵，矩阵中颜色越深说明表达量越高，颜色越浅说明表达量越低。

### 2.1 数据预处理

使用改进的自动编码器实现对数据中基于负二项分布参数的学习。自动编码器使用三个层来分析预测输入数据：输入层、隐藏层和输出层。这种改进自动编码器的特殊之处在于采用特定的噪声模型，这种噪声模型融合零膨胀负二项分布，构建零膨胀负二项分布损失函数，以对应 scRNA-seq 数据高稀疏性和高离散性的特点。零膨胀负二项分布是由负二项分量的均值和离散参数($\mu$ 和 $\theta$)以及表示点是 dropout 概率($\pi$)的混合系数来参数化的：

$$NB(x; \mu, \theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{x} \tag{6}$$

$$ZINB(x; \pi, \mu, \theta) = \pi\delta_0(x) + (1-\pi)NB(x; \mu, \theta) \tag{7}$$

对于基因表达矩阵中的每一个值，与传统的自动编码器不同，该自动编码器框架估算每个输入值的三个负二项分布参数：均值($\mu$)、dropout 概率（$\pi$）和离散度($\theta$)。因此，一个输入矩阵会得到三个输出层，三个输出层的大小都与输入一致，即具有相同数量的表达量。通过该自动编码器的处理，就得出了输入数据符合零膨胀负二项分布的均值。其中，这一模型框架默认有三个隐藏层，包含 64、32、64 个神经元，32 个神经元的隐藏层为瓶颈层。如图 2 所示，使 scRNA-seq 数据的复杂性和非线性能够被有效捕捉。

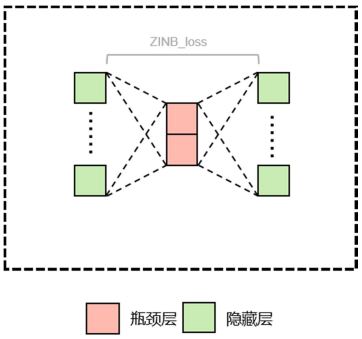


图 2　自动编码器结构

## 2.2　分治法构建并训练网络

为了保持 scRNA-seq 数据中的细胞特异性，并有效考虑到细胞与基因之间的关系，这里进一步使用深度神经网络来处理并训练由自动编码器获取的均值。采用分治法的思想，将数据集中所有的基因分割成子集，并构建子网络对数据进行训练。

网络构建首先要选择“目标基因”，将数据集中的所有基因分成具有相同基因数的子集作为目标基因，这些目标基因构成了子神经网络的输出层。对于子集 $n$ 中的每个目标基因 $g_i$，从子集 $n$ 以外的基因中选择 5 个相似度最高的预测基因，选择完成后删除在所有子神经网络中选择次数超过阈值的预测基因，剩余的预测基因为子神经网络的输入层。对于每个子集，训练一个四层神经网络，如图 1 和图 3 所示，预测基因组成的输入层、完全连接的隐藏层、dropout 层(与 scRNA-Seq 数据中的 dropout 不同，这里的 dropout 是按比例忽略一部分数据)和目标基因组成的输出层。其中使用线性整流函数（ReLU）作为激活函数，使用加权 MSE 作为损失函数，加入 dropout 层是为了防止过拟合，其中 dropout 参数设为 20%。

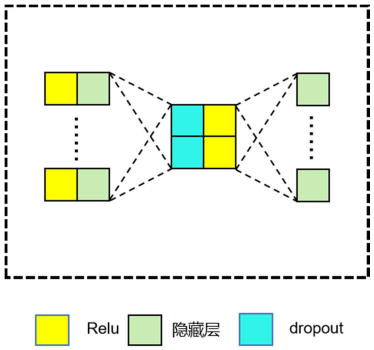


图 3　深度学习模块结构

将 95%的细胞作为训练集进行网络训练，剩下的 5%作为测试集，并行训练每个子模型，在每个网络训练完成后测试是否产生过拟合情况。由于每个子网络的结构较简单，可以观察到超参数调整导致的可变性较低。因此，将批次大小默认设置为 64，学习率设置为 0.0001。计算损失值时，使用加权均方误差（MSE）损失函数，加权操作可以赋予表达值较高的基因以更高的权重，从而加强基因的表达。针对一个细胞 $c$，其损失值 $loss\_mse$ 计算如下：

$$loss\_mse = \sum G_i (G_i - \hat{G}_i)^2 \tag{8}$$

其中 $G_i$ 是细胞 $c$ 的基因 $i$ 的值，$\hat{G}_i$ 为该基因训练过程中的给定值。

# 3　结果与分析

## 3.1　聚类分析

在 scRNA-seq 数据的分析实验中，聚类分析是必不可少的，聚类可以帮助确定数据中的细胞类型或亚型。本研究采取的数据聚类算法为 Scanpy，随机选取 2000 个细胞进行分析。

在聚类分析的评估中，使用本研究方法 ND-Impute 与其他两种方法深度学习 DCA 和 DeepImpute 进行对比。从图 4 中可以看出，ND-Impute 相比其他两种方法，聚类簇之间边界更加清晰，能相对显著区分出数据中的细胞类型。原始数据插补后，能在一定程度上恢复因测序技术造成的低表达量缺损，从而恢复或增强了细胞的特异性表达，使数据中的细胞类型或细胞亚群能够被识别出来。为了评估 ND-Impute 在细胞聚类上的有效性，使用两个聚类指标，归一化互信息(NMI)和调整兰德指数(ARI)。其中 NMI 用于度量两个聚类结果的相近程度，值域为 $[0,1]$，数值越高说明越接近，即划分越准确。ARI 能反映两种划分的重叠程度，取值范围为 $[-1,1]$，值越大说明效果越好。

设 $C = \{c_1, c_2, c_3, ..., c_k\}$ 表示聚类（cluster）划分，$R = \{r_1, r_2, r_3, ..., r_k\}$ 表示实际类别划分。

NMI 定义为：

$$NMI(C,R) = \frac{2MI(C,R)}{H(C)+H(R)} \tag{9}$$

其中 $MI(C,R)$ 表示 $C$ 和 $R$ 互信息：

$$MI(C,R) = \sum_{n=1}^{|C|}\sum_{m=1}^{|R|} P(n,m)\log\left(\frac{P(n,m)}{P(n)P(m)}\right) \tag{10}$$

其中，$P(n,m)$ 为 $(C,R)$ 的联合分布，$P(n), P(m)$ 分别为边缘分布，可以理解为样本中同时属于两个集合的概率，以及分别属于 $C, R$ 的概率。$H(C)$，$H(R)$ 分别表示 $C, R$ 的熵：

$$H(C) = -\sum_{n=1}^{|C|} P(n)\log(P(n)) \tag{11}$$

$$H(R) = -\sum_{m=1}^{|R|} P(m)\log(P(m)) \tag{12}$$

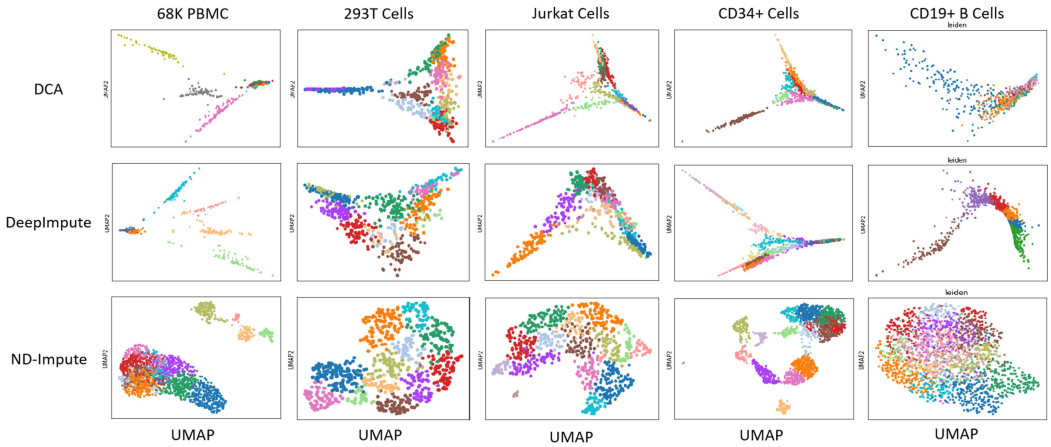


图 4 五个数据集经过三个算法插补后的聚类效果对比

表 2 五个数据集的聚类评估指标结果

| 评价指标 | 数据集 | 未插补 | DCA | DeepImpute | ND-Impute |
|---|---|---|---|---|---|
| NMI | 68KPBMC | 0.523 | 0.677 | 0.712 | 0.745 |
| | 293T Cell | 0.346 | 0.524 | 0.533 | 0.649 |
| | Jurkat Cell | 0.302 | 0.421 | 0.399 | 0.560 |
| | CD34+ Cells | 0.619 | 0.633 | 0.504 | 0.686 |
| | CD19+ B Cells | 0.427 | 0.592 | 0.632 | 0.597 |
| ARI | 68KPBMC | 0.501 | 0.683 | 0.779 | 0.791 |
| | 293T Cell | 0.324 | 0.336 | 0.511 | 0.539 |
| | Jurkat Cell | 0.255 | 0.288 | 0.395 | 0.504 |
| | CD34+ Cells | 0.563 | 0.703 | 0.654 | 0.728 |
| | CD19+ B Cells | 0.564 | 0.622 | 0.703 | 0.665 |

设 $a$ 为通过聚类方法正确分配到同一簇的细胞对数，$b$ 是错误地分配到同一簇中的单元对数，$c$ 为错误分配到不同簇的细胞对数，$d$ 为正确分配到不同簇的细胞对数。

ARI 定义为：

$$ARI = \frac{2(ad - bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{13}$$

三种方法在五个数据集上的 NMI 和 ARI 表现如表 2 所示。对比可得，ND-Impute 在两个指标的表现上要优于 DCA 和 DeepImpute。细胞数量大的数据集相比细胞数量小的数据集可以更好地训练网络，所以指标结果都相对表现更佳。

## 3.2 插补效果分析

对于获取的 scRNA-seq 数据集，由于无法确定真实的缺失值，这里采用了一种随机掩蔽部分 scRNA-seq 数据集基因表达矩阵中元素的方法，也就是用零代替表达矩阵中的部分非零值，比较插补后的数值是否能较好地还原被掩蔽的数值。选择随机掩码表达式矩阵中每个细胞的非零计数值的 5%。掩蔽遵循与之成比例的密度概率分布：

$$f(n) = e^{-\frac{n}{20}} \tag{14}$$

其中 $n$ 是原始计数值。原始值可作为评价插补方法性能的参考值，使用皮尔逊相关系数和均方误差(MSE)两种性能指标评估插补效果，以及恢复掩蔽值的准确性。图 5 和图 6 中分别显示了 ND-Impute、DCA 和 DeepImpute 的皮尔逊相关系数和均方误差对比情况，可以看出 ND-Impute 获取了更高的皮尔逊相关系数和更低的均方误差，这说明相比 DCA 和 DeepImpute，ND-Impute 更好地恢复了所有范围内的 dropout 值。

恢复掩蔽值的准确性如图 7 所示，插补值与真实值越接近，图中的点就会越接近斜线，并且分布更均匀。

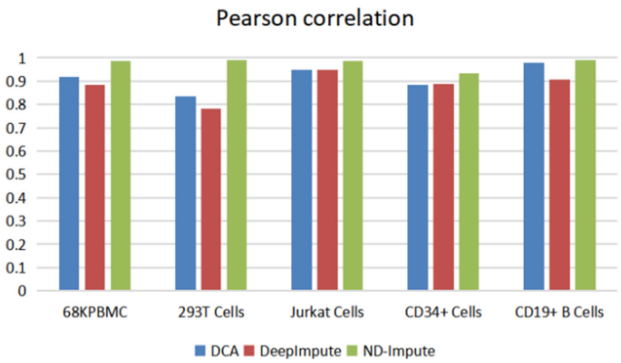


图 5 三个算法对掩蔽值插补后的皮尔逊相关系数

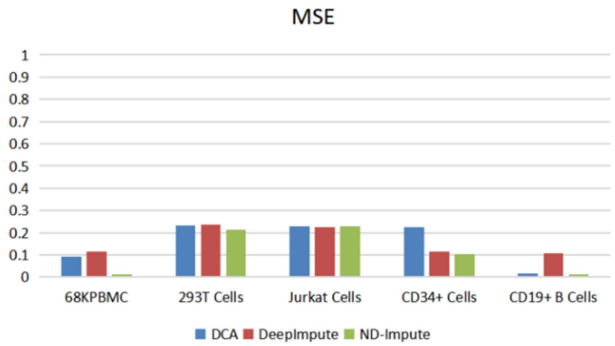


图 6 三个算法对掩蔽值插补后的均方误差

## 4 结论

scRNA-seq 数据中的 dropout 事件是生物信息学分析中的重要问题，其数据的庞大以及数据的嘈杂都不利于后续的生物学研究，传统的插补策略多使用单一思路解决这一问题。

本研究中提出了一种基于负二项分布的分治插补策略 ND-Impute 来处理 scRNA-seq 数据中的 dropout。ND-Impute 基于深度学习方法，在常规自动编码器基础上使用符合统计学分布的特异性损失函数，并将分治思想融入神经网络的训练中。这种思路和结构有效地还原了缺失的数值，既考虑了数据集的分布模型，也通过学习细胞与基因之间的关系和相似性保留了细胞的特异性表达。与另外两种深度学习方法不同，ND-Impute 不会只针对数据的统计学分布或细胞特异性，该方法同时考虑这两个方面的信息，以达到更好插补效果。通过实验比较，ND-Impute 显示出比另外两种深度学习方法更好的聚类效果，NMI 和 ARI 的平均值高于 DCA 和 DeepImpute，对于掩蔽值的恢复，不仅具有更高的皮尔逊相关系数，造成的误差值也更低，并且插补值与真实值也更加接近。综上所述，ND-Impute 在插补效果上相比一般的深度插补模型都有所优化，是一种表现良好的插补方法。

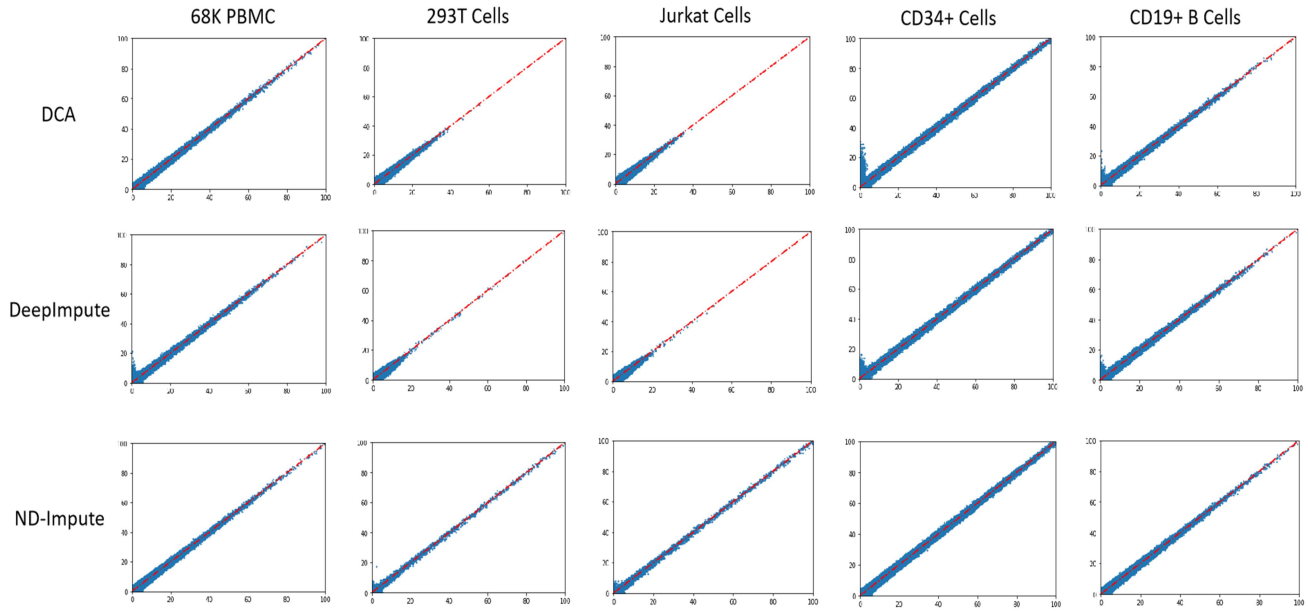


图 7　三个算法恢复的掩蔽值与原值对比

我们下一步研究将运用其他深度学习方法来学习 scRNA-seq 数据的特性，以达到更好的插补效果。同时将生成模型与统计学分布相结合，进一步提高数据的特异性。

**参考文献：**

[1] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science[J]. Nat Rev Genet, 2013, 14(09):618 - 630.

[2] Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell[J]. Nature Methods, 2009, 6(05):377-382.

[3] Zhu D, Zhao Z, Cui G, et al. Single-cell transcriptome analysis reveals estrogen signaling coordinately augments one-carbon, polyamine, and purine synthesis in breast cancer[J]. Cell Rep, 2018, 25 (08): 2285-2298.e4.

[4] Chen, Chongyi, Xing, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI)[J]. Science, 2017.14;356:189-194.

[5] Crinier A, Milpied P, Escali è re B, et al. High-dimensional single-cell analysis identifies organ-specific signatures and conserved NK cell subsets in humans and mice[J]. Immunity, 2018, 49 (05):971-986.e5.

[6] Kolodziejczyk A, Kim J K, Svensson V, et al. The technology and biology of single-cell RNA sequencing.[J]. Molecular Cell, 2015, 58(04):610-620.

[7] Andrews T S, Hemberg M. M3Drop: dropout-based feature

selection for scRNASeq[J]. Bioinformatics, 2019, 35(16):2865–2867.

[8] Zappia L, Phipson B, Oshlack A . Splatter: simulation of single–cell RNA sequencing data[J]. Genome Biology, 2017, 18(01):174.

[9] Mo H, Wang J, Torre E, et al. SAVER: gene expression recovery for single–cell RNA sequencing[J]. Nature Methods, 2018, 15(05):539–542.

[10] Wei V L, Li J J. An accurate and robust imputation method scImpute for single–cell RNA–seq data[J]. Nature Communications, 2018, 9(01):997.

[11] Ma Q, Xu D. Deep learning shapes single–cell data analysis[J]. Nature Reviews Molecular Cell Biology, 2022, 23(05): 303–304.

[12] Eraslan G, Simon L M, Mir Ce A M, et al. Single–cell RNA–seq denoising using a deep count autoencoder[J]. Nature Communications, 2019, 10:390.

[13] Arisdakessian C, Poirion O, Yunits B, et al. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single–cell RNA–seq data[J]. Genome Biology, 2019, 20(01):211.

[14] Xu Y, Zhang Z, You L, et al. scIGANs: single–cell RNA–seq imputation using generative adversarial networks[J]. Nucleic Acids Research, 2020, 48(04):e85.

[15] Jin K, Le O Y, Zhao X M, et al. scTSSR: gene expression recovery for single–cell RNA sequencing using two–side sparse self–representation[J]. Bioinformatics, 2020, 36(10).

[16] Lu F, Lin Y, Yuan C, et al. EnTSSR: a weighted ensemble learning method to impute single–cell RNA sequencing data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(06): 2781–2787.

[17] Zheng G X Y, Terry J M, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells[J]. Nature communications, 2017, 8(01): 1–12.

[18] Risso D, Perraudeau F, Gribkova S, et al. A general and flexible method for signal extraction from single–cell RNA–seq data[J]. Nature communications, 2018, 9(01): 1–17.

[19] Lopez R, Regier J, Cole M, et al. Bayesian inference for a generative model of transcriptome profiles from single–cell RNA sequencing[J]. bioRxiv, 2018: 292037.

[20] Hafemeister C, Satija R. Normalization and variance stabilization of single–cell RNA–seq data using regularized negative binomial regression[J]. Genome biology, 2019, 20(01): 1–15.

# Divide and Conquer Imputation for Dropouts in Single-Cell Data based on Negative Binomial Distribution

XIONG Zhen-zhen[a, b], ZHANG Ben-gong[b, c]

(a. School of Computer Science and Artificial Intelligence, b. Research Center of Nonlinear Science, c. School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan Hubei 430200, China)

**Abstract:** Single-cell RNA sequencing (scRNA-seq) technology provides opportunities for high-throughput, high-resolution in-depth research of single cells, and provides insights into cell functions and the underlying gene regulation mechanisms at the single-cell level important technical means. However, this technology also brings new challenges. ScRNA-seq data has the characteristics of large scale, high noise, and strong heterogeneity, especially the high proportion of data missing, which is called dropout. The problem of dropout seriously affects the reliability of the downstream analysis, and even covers up the important relationship between genes and genes. This paper proposed a divided and conquering imputation strategy based on negative binomial distribution ND-Impute to process scRNA-seq data. This method assumed that scRNA-seq data conform to the negative binomial distribution, utilized an autoencoder that incorporates a specific loss function to obtain data-specific parameters, and used a divide-and-conquer strategy to estimate potential gene expression values. The comparison of clustering effect, correlation, and error analysis showed that this method can effectively restore missing data and improve the accuracy of subsequent research and analysis.

**Key words:** single-cell RNA sequencing; dropout; imputation; clustering analysis

（责任编辑：周莉）