文章编号：2095-2163(2023)12-0191-05 中图分类号：TP399 文献标志码：A

大模型辅助的域适应算法在基因预测的应用

田雨竹，关佶红
（同济大学 电子与信息工程学院，上海 201804）

摘 要：域适应问题旨在解决由于源数据集和目标数据集存在域偏差，导致在源数据集上训练的模型在目标数据集上的泛化能力差的问题。当前域适应领域的工作通过强制特征空间中目标数据和源数据同分布，来对齐两个域的数据，从而提高模型在目标数据上的表现，这类方法在以下两种情况下表现不佳：一是两部分数据存在各自特有的类别；二是目标数据集原始特征质量不佳。针对这两个问题，本文提出使用预训练大模型增强目标数据集特征表示，且保留两个域数据的分布差异的域适应算法，并将其应用在生物信息中的空间数据缺失基因预测问题上。通过在多个数据集上的实验，本文提出的缺失基因预测方法在预测准确性上有所提升。

关键词：域适应问题；预训练大模型；缺失基因预测

Application of domain adaptation algorithm on gene prediction facilitated with foundational model

TIAN Yuzhu, GUAN Jihong
(College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: The focus of domain adaptation is to mitigate issues stemming from domain discrepancies between source and target datasets, which impairs the generalization of models trained on the source dataset when applied to the target dataset. Current approaches in this field strive to align the data distributions across domains in feature space, thereby enhancing the model′s performance on target data. These methods, however, may falter when distinct categories are present in each dataset or when the intrinsic features of the target dataset are of subpar quality. Addressing these challenges, this article proposes a domain adaptation algorithm that leverages pretrained large-scale models to enrich the feature representation of the target dataset while maintaining the distinct data distributions of both domains. Applied specifically to the prediction of missing genes in spatial transcriptomics data, the methodology outlined in this study has shown an increase in prediction accuracy across various datasets.

Key words: domain adaptation; large pretrained model; missing gene prediction

0 引 言

本文根据 RNA(Ribonucleic Acid)数据集解决对空间转录组数据缺失基因进行预测的问题，其本质是一个域适应领域的回归问题，其中空间转录组数据是目标数据集，参考 RNA 数据集是源数据集。

该问题形式化如下：令空间转录组测量的基因集合为 S，要从 S 预测未知基因集合 T，参考 RNA 数据集中的基因集合为 G，$T \subseteq G$，即所有待预测基因都包含在参考 RNA 数据集中。输入参考 RNA 数据集作为源数据集，空间转录组数据为目标数据集。源数据集为矩阵 $\boldsymbol{X}_{\mathrm{SRC}} \in R^{N_{\mathrm{SRC}} \times |G|}$（通常还包括细胞类型标注 Y_{SRC}），N_{SRC} 为源数据中的细胞个数，$\boldsymbol{X}_{\mathrm{SRC}\,ij}$ 是源数据集中第 i 个细胞的第 j 个基因的表达值，源数据集中通常还包括细胞类型标注 Y_{SRC}。目标数据集为矩阵 $\boldsymbol{X}_{\mathrm{TAR}} \in R^{N_{\mathrm{TAR}} \times |S|}$，其中 N_{TAR} 为目标数据中的细胞个数，$X_{\mathrm{TAR}\,ij}$ 是空间数据集第 i 个细胞的第 j 个基因的表达值。本文解决的域适应回归问题为：在源数据集 X_{SRC} 上训练一个映射函数，函数的输入为源数据集中细胞在 S 上的读数 x^{S}_{SRC}，输出为细胞在 T 上的读数预测 $\widehat{x^{T}_{\mathrm{SRC}}}$，并将该映射迁移至目标数据集 X_{TAR}，输入目标数据集中的细胞在 S 上的读数 x^{S}_{TAR}，输出细胞在 T 上的读数预测 $\widehat{X^{T}_{\mathrm{TAR}}}$，问题的输出应为 $\widehat{X^{T}_{\mathrm{TAR}}} \in R^{N_{\mathrm{TAR}} \times |T|}$，$\widehat{X^{T}_{\mathrm{TAR}}}_{ij}$ 为空间数据集第 i 个细胞

基金项目：国家自然科学基金(62172300)。
作者简介：田雨竹(1999-)，女，硕士研究生，主要研究方向：机器学习方法在单细胞数据上的应用。
通讯作者：关佶红(1969-)，女，博士，教授，主要研究方向：认知与智能信息处理。Email：jhguan@ tongji.edu.cn
收稿日期：2023-10-02

的第 j 个待预测基因的预测值。

当前通用的域适应方法应用在缺失基因预测的回归问题上的步骤：

(1)在源数据集 X_{SRC} 上训练一个多层感知机(Multi-Layer Perceptron, MLP)模型表示源数据集中的细胞在 S 上的读数 x^{S}_{SRC} 到细胞在 T 上的读数 $x^{\mathrm{T}}_{\mathrm{SRC}}$ 的映射；

(2)将 MLP 模型分为两部分,一部分为特征编码器,输入为源数据集中的细胞在 S 上的读数 x^{S}_{SRC},输出为对应的隐变量 z_{SRC},作为 x^{S}_{SRC} 的特征表示;另一部分为解码器,输入为 z_{SRC},输出为细胞在 T 上的预测读数 $\widehat{x^{T}_{\mathrm{SRC}}}$;

(3)训练目标数据的特征编码器,输入为目标数据集中的细胞在 S 上的读数 x^{S}_{TAR},输出为隐变量 z_{TAR};

目标数据的特征编码器的优化目标是使来自目标数据的 z_{TAR} 和来自源数据的 z_{SRC} 同分布。最优特征编码器参数 $\theta^{\mathrm{TAR}}_{\mathrm{ENC}}$ 可以通过最小化 KL(Kullback-Leibler divergence)散度求得,表达式为式(1)：

$$\theta^{\mathrm{TAR}}_{\mathrm{ENC}} = \arg\min_{\theta} D_{\mathrm{KL}}(P_{\mathrm{TAR}}(z) \parallel P_{\mathrm{SRC}}(z)) \quad (1)$$

(4)使用目标数据的特征编码器对目标数据进行编码,即输入 x^{S}_{TAR},输出 z_{TAR},并使用解码器对特征进行解码得到其基因预测值 $\widehat{x^{T}_{\mathrm{TAR}}}$。

使用一般的域适应方法的缺失基因预测存在两个问题：

(1)对分布不同的目标数据集和源数据集使用同分布假设会引入错误；

(2)目标数据的原始特征无法体现数据的真实分布。

产生问题(1)的原因是单细胞数据中经常出现目标数据集和源数据集的分布不同的情况,如特定类别只存在于目标数据集或源数据集中,并非两个数据集公共的类别,一个类别虽然是两个数据集公共的,但属于该类别的数据点个数存在较大差异。对于这样的数据,如果强行对源数据和目标数据进行同分布假设,最小化两个分布的距离,反而会引入错误。

产生问题(2)的原因是目标数据原始特征信息量小,表现为聚类质量差,即类和类之间没有很好地分隔开,而类内数据点弥散。这种情况增加了目标数据与源数据对齐的难度。在空间数据缺失基因预测问题中,聚类质量差的情况很突出。由于技术限制,空间数据集中包含的基因只有几十、几百个,而一个 RNA 数据集可包含数万个基因。因此,空间数据包含的信息很少,其细胞类别在原始特征(基因)空间的聚类效果很差。在空间数据聚类效果差的情况下,首先应该尽可能恢复目标数据集内部的真实类别结构,提高其特征表示质量,再与源数据集进行对齐。

本文提出一个能够处理数据集分布差异、增强数据特征表示的域适应回归算法,并将其应用在空间数据缺失基因预测问题上。

针对问题(1),本文提出舍弃特征空间同分布要求,而是对源数据和目标数据使用共同的自编码器,且在自编码器的隐空间接入两个分类器,分别对源数据进行由标注提供监督的分类任务训练以及对目标数据进行由伪标签提供监督的分类任务训练,伪标签的获取由预训练大模型辅助得到。通过在隐空间中对两部分数据分别进行分类任务训练,提升各自的聚类效果,同时通过使用同一个自编码器,约束了两部分数据公共部分的对齐。针对问题(2),本文提出使用预训练大模型为目标数据集补充信息,提升数据特征表示的质量,提升在补充信息后的特征空间中聚类效果,从而帮助与源数据的对齐。单细胞数据分析领域的预训练大模型利用数百万个细胞在数万个基因上的表达数据进行训练,有效习得基因-基因的相关性,因此本文提出利用大模型的基因-基因相关性,为目标数据集补充信息,增强其特征表示。

1 相关工作

1.1 域适应

具有代表性的域适应工作有基于反向传播的无监督域适应(Unsupervised Domain Adaptation by Backpropagation)[1]和对抗判别域适应(Adversarial Discriminative Domain Adaptation)[2]。

基于反向传播的无监督域适应主要解决域适应中的分类问题。首先,使用同一个特征提取器提取目标数据集和源数据集特征,再将来自源数据的特征 z_{SRC} 送入类别分类器 $\mathrm{CLF}_{\mathrm{cls}}$,而两部分数据的特征混合共同送入域分类器 $\mathrm{CLF}_{\mathrm{domain}}$。模型的损失为类别分类器预测得到的交叉熵 CE_{cls} 和域分类器预测得到的交叉熵的负数 $-CE_{\mathrm{domain}}$,公式(2),使得两部分数据的特征分布完全重合,以致域分类器无法分辨数据来自哪个域。

$$Loss = CE_{\mathrm{cls}} - CE_{\mathrm{domain}} \quad (2)$$

对抗判别域适应解决分类问题。首先,在源数据上训练特征提取器和分类器,接着为目标数据训练另一个特征提取器,目标数据、源数据经过其各自的特征提取器得到的特征经过一个判别器,来判别

其来自哪个域。对抗判别域适应本质上也是使源数据的特征和目标数据的特征分布完全一致。

反向传播的无监督域适应和对抗判别域适应都存在着无法处理源数据和目标数据分布不同,以及目标数据特征质量低的问题。

1.2　空间数据缺失基因预测

为空间数据进行缺失基因预测任务专门设计的算法有 stPlus[3]、SpaGE[4]。

stPlus 先将空间数据补零至与参考数据同维度,即对每个长度为 $|S|$ 的目标数据补零到长度为 $|S|+|T|$;利用补长的目标数据和源数据共同训练一个自编码器,并利用编码器的输出作为两部分数据的特征表示。在编码器输出构成的特征空间中,为每一个目标数据 z_{TAR} 找 K 个最近的源数据 z^1_{SRC}, $z^2_{\mathrm{SRC}},\cdots,z^K_{\mathrm{SRC}}$,并将这些源数据点中基因表达的加权和作为目标数据的预测基因值,距离目标数据点越近的源数据点的权重越大。stPlus 假设为源数据集中的每个数据点都可以表示为其 K 近邻的加权和,且这个局部近似关系可跨越域差异进行迁移。stPlus 中的线性假设对于基因表达来说过于简单,且在目标数据缺失大量信息时,自编码器的表示学习效果不佳。SpaGE 首先使用 PRECISE[5] 对数据进行对齐,再基于目标数据集中的数据点可近似于其在源数据集中的 K 近邻加权和的假设,对目标数据的基因进行预测。SpaGE 假设基因表达为线性关系,这一假设过于简单。

1.3　大型预训练模型的表示学习

近来,生物信息领域出现多个预训练大模型: cellLM, scBERT, scGPT, geneformer 等,其利用数百万细胞的基因表达学习基因-基因之间的相关性。通过仅保留数据本身的语义信息,大模型的特征可以帮助源数据和目标数据对齐。通过从大数据中学习得到的特征相关性补全数据中的缺失信息,大模型的特征可以增强数据的内在分布模式。

2　方法

2.1　方法概述

本文方法描述:

(1)将目标数据 X_{TAR} 输入 scGPT 预训练模型中,输出细胞在 K 个高方差基因上的表达预测值 X_{pred};

(2)将目标数据 X_{TAR} 和高方差基因预测值 X_{pred} 拼接得到扩展数据($X_{\mathrm{TAR}}\mid X_{\mathrm{pred}}$) $\in R^{N_{\mathrm{TAR}}\times(|S|+K)}$,并使用 Leiden 算法对扩展数据进行聚类,得到目标数据的伪标签 Y_{TAR};

(3)目标数据 X_{TAR}、源数据 X_{SRC}、目标数据的伪标签 Y_{TAR} 和源数据的类别标注 Y_{SRC} 作为输入,训练主模型;

(4)将目标数据 X_{TAR} 输入主模型,推断得到最终需要预测基因的预测值 $\widehat{X^T_{\mathrm{SRC}}}$。

2.2　使用 scGPT 对高方差基因进行预测

由于 scGPT 中预测的基因值为分箱值(binning),不是基因的真实值,且 scGPT 是由大量不同细胞条件的数据训练的,对于特定的预测任务需要进行微调才能使用。本文不直接使用 scGPT 来预测目标基因 T,而只是使用 scGPT 对高方差基因进行预测,起到补充信息的作用。scGPT 预测表现:当预留出部分已知基因,并使用 scGPT 基于剩余基因进行预测时,观察到各基因的预测值和真实值计算得到的皮尔森相关性(Pearson Correlation)不稳定,但 scGPT 预测带来的信息增益在于把多个基因的预测结果整合在一起,扩展目标数据时,扩展数据的聚类结果比只使用原始目标数据有所提升,体现在类内距离收紧,类间距离增大。

本文提出利用 scGPT 补充前 K 个高方差的额外基因的信息,用于扩展目标数据,并在扩展的目标数据上进行聚类以获得目标数据的伪标签,辅助其表示学习。首先,对源数据集的所有基因 $g\in G$ 计算方差,公式(3);其次,从 G 中选择 σ_g 最高的 K 个基因 G_K,作为需要 scGPT 补充的 K 的基因;最后,将目标数据集中已有的基因 S 输入,得到 scGPT 对 G_K 基因的箱值预测 $X_{\mathrm{pred}}\in R^{N_{\mathrm{TAR}}\times K}$,供后续步骤使用。scGPT 的模型架构如图 1 所示。

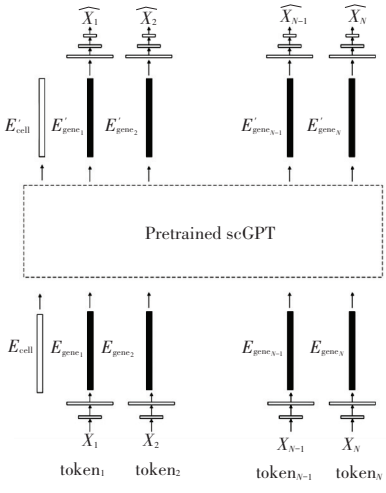


图 1　scGPT 模型架构

Fig. 1　The architecture of scGPT

$$\sigma_g^2=\frac{\sum_{i=1}^{N_{\mathrm{SRC}}}(X_{\mathrm{SRC}ig}-\mu_g)}{N_{\mathrm{SRC}}} \tag{3}$$

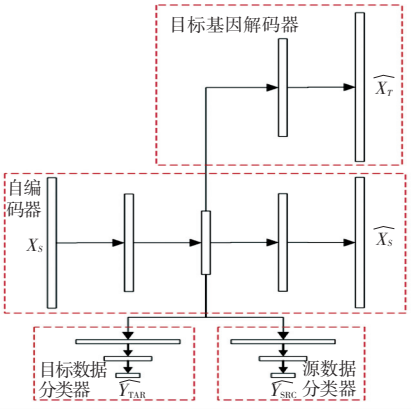
其中，$X_{\mathrm{SRC}\,ig}$ 为源数据中第 i 个细胞的基因 g 的表达值，μ_g 为源数据中基因 g 的平均值。

2.3 使用 Leiden 算法对扩展数据聚类

在 scGPT 的输出中，得到补充基因 G_K 的箱值预测 X_{pred}，将目标数据集中数据的特征从原来的 X_{TAR} 扩展为 $(X_{\mathrm{TAR}} \mid X_{\mathrm{pred}})$，即每个数据点的特征扩展为 $(x_{\mathrm{TAR}_1}, x_{\mathrm{TAR}_2}, \cdots x_{\mathrm{TAR}_{|S|}}, x_{\mathrm{pred}\,1}, \cdots, x_{\mathrm{pred}_K})$。本文在扩展数据 X_{extend} 上使用 Leiden 算法进行聚类，以获得目标数据集的伪标签 Y_{TAR} 供后续使用。

2.4 训练主模型

本文提出的主模型在隐空间中对齐源数据和目标数据，并为目标数据预测缺失基因。主模型的架构如图 2 所示。



①一个自编码器；②两个分类器；③一个从隐空间接出的目标基因解码器，以学习隐变量 z 到目标基因 T 的映射函数

图 2 主模型架构

Fig. 2 The architecture of main model

有两部分数据作为输入：第一部分数据为目标数据 $(X_{\mathrm{TAR}}, Y_{\mathrm{TAR}})$，其中 $X_{\mathrm{TAR}} \in R^{N_{\mathrm{TAR}} \times |S|}$，$Y_{\mathrm{TAR}}$ 为扩展数据聚类所得的伪标签；第二部分数据为源数据 $(X_{\mathrm{SRC}}, Y_{\mathrm{SRC}})$，$Y_{\mathrm{SRC}}$ 为源数据中的标注类别。本文将 X_{SRC} 拆成两部分：

(1) $X^S_{\mathrm{SRC}} \in R^{N_{\mathrm{SRC}} \times |S|}$，包含 S 中的基因，作为自编码器的输入；

(2) $X^T_{\mathrm{SRC}} \in R^{N_{\mathrm{SRC}} \times |T|}$ 包含 T 中的基因，作为目标基因解码器的真实值。

将这两部分数据输入自编码器。源数据 X^S_{SRC} 和目标数据 X^S_{TAR} 通过自编码器得到输出 $\widehat{X^S_{\mathrm{SRC}}}$ 和 $\widehat{X^S_{\mathrm{TAR}}}$，两部分数据都将预测值与真实值计算重构损失，公式(4)：

$$L_S = MSE(X^S, \widehat{X^S}) \tag{4}$$

同时，在自编码器的瓶颈(bottleneck)得到两部分数据的低维嵌入 Z，该嵌入接着输入源数据分类器、目标数据分类器和目标基因解码器。

若 z_i 为来自源数据集数据点的嵌入，则将其输入源数据分类器，并用该数据的标注作为真实标签，计算交叉熵，公式(5)：

$$L_{\mathrm{CE}_{\mathrm{SRC}}} = \mathrm{CE}(Y_{\mathrm{SRC}}, \widehat{Y_{\mathrm{SRC}}}) \tag{5}$$

从源数据集中学习由隐变量 z 预测目标基因 T 的映射函数。因此，对于来自源数据集的 z_i 还要将其输入目标基因解码器，得到 $\widehat{X^T_{\mathrm{SRC}}}$，并使用其中 T 的真实值，与预测值计算重构损失，公式(6)：

$$L_T = MSE(X^{\mathrm{T}}_{\mathrm{SRC}}, \widehat{X^T_{\mathrm{SRC}}}) \tag{6}$$

若 z_i 为来自目标数据集的数据点，则将其输入目标数据分类器，并使用在扩展基因空间中得到的聚类结果作为伪标签，计算交叉熵，公式(7)：

$$L_{\mathrm{CE}_{\mathrm{TAR}}} = \mathrm{CE}(Y_{\mathrm{TAR}}, \widehat{Y_{\mathrm{TAR}}}) \tag{7}$$

该模型的损失函数由 4 个部分组成：

(1) 自编码器对已知基因的重构损失 L_S；

(2) 目标基因解码器在源数据集上对待预测 T 基因的重构损失 L_T；

(3) 目标数据的分类交叉熵 $L_{\mathrm{CE}_{\mathrm{TAR}}}$；

(4) 源数据的分类交叉熵 $L_{\mathrm{CE}\,\mathrm{SRC}}$，最终损失为 4 部分的和，即公式(8)：

$$L = L_S + L_T + L_{CE_{\mathrm{TAR}}} + L_{CE_{\mathrm{SRC}}} \tag{8}$$

通过对最终损失 L 反向传播来更新模型参数。

2.5 预测缺失基因

模型训练完成后，只需要将目标数据 X^S_{TAR} 输入模型，得到目标基因解码器的输出 $\widehat{X^T_{\mathrm{TAR}}}$，即可得到对缺失基因的预测值。

3 实验

为了验证本文缺失基因预测方法的有效性，本文使用了 6 个数据集，组成了 4 组目标数据和源数据的组合，并与 stPlus，SpaGE，Seurat[6]，Liger[7]，gimVI[8] 5 个补全方法进行比较。数据集包含了不同空间测序技术和不同参考 RNA 数据的组合，以测试算法的鲁棒性。数据集的基本信息见表 1。

表 1 数据集基本信息

Table 1 Basic information on the datasets

数据集名称	细胞个数	基因个数
源数据集 Moffit	31 299	18 646
源数据集 AllenVISp	14 249	34 617
源数据集 AllenSSp	5 577	30 527
源数据集 Zeisel	1 691	15 075
目标数据集 MERFISH	13 819	155
目标数据集 osmFISH	3 405	33

首先，对数据进行基本的预处理，在每个细胞内