机器学习大作业

基于半监督学习的交互式相似细胞预测

2024.12.06

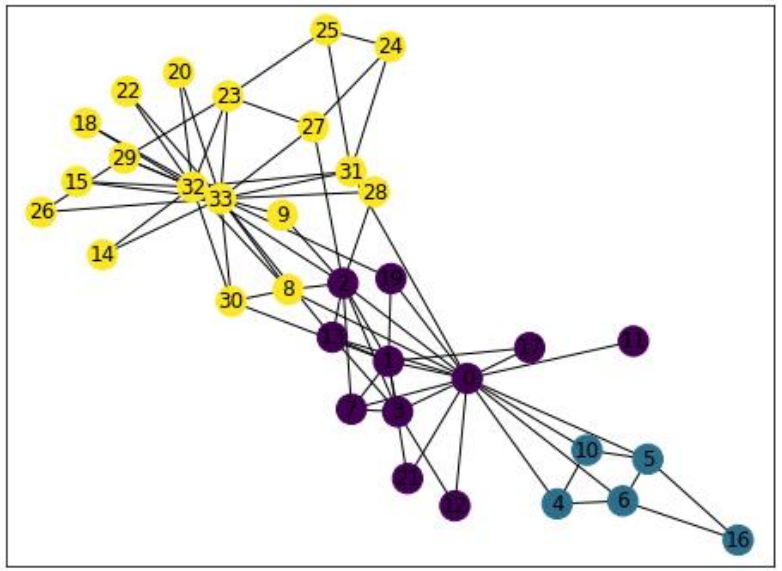
背景知识1——*单细胞测序数据分析*

- 单细胞测序数据分析是在单个细胞水平上，对基因组、转录组等进行高通量测序分析，旨在精细解释一个样本内细胞间的异质性。
- 简单来说，单细胞测序数据就是一个行是细胞，列是基因的高维稀疏矩阵。针对单细胞测序数据的工作流具体包括：质量控制、数据标准化、数据降维、细胞聚类、细胞类型注释和可视化。

- 然而，最关键的细胞聚类步骤由于分辨率的不同导致聚类结果无法反应真实的细胞类型，需要手动交互式的修正。而手动圈选必然会导致相似细胞遗漏或误选错误细胞，因此需要一种交互式的相似细胞预测方法弥补这一缺陷。

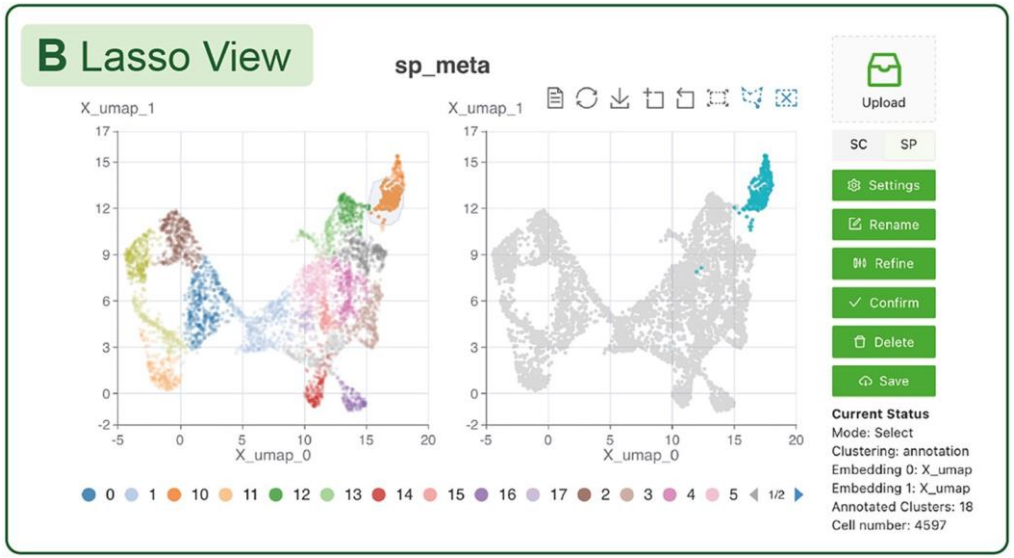
背景知识2——*标签传播算法LPA*

- Label Propagation Algorithm，也称作标签传播算法（LPA），是一种基于图的半监督学习的算法，常用于节点分类和图数据的聚类分析等。
- 该算法主要通过在各个节点之间传播标签并逐步确定，最终将图中的节点分为若干个簇或类别。

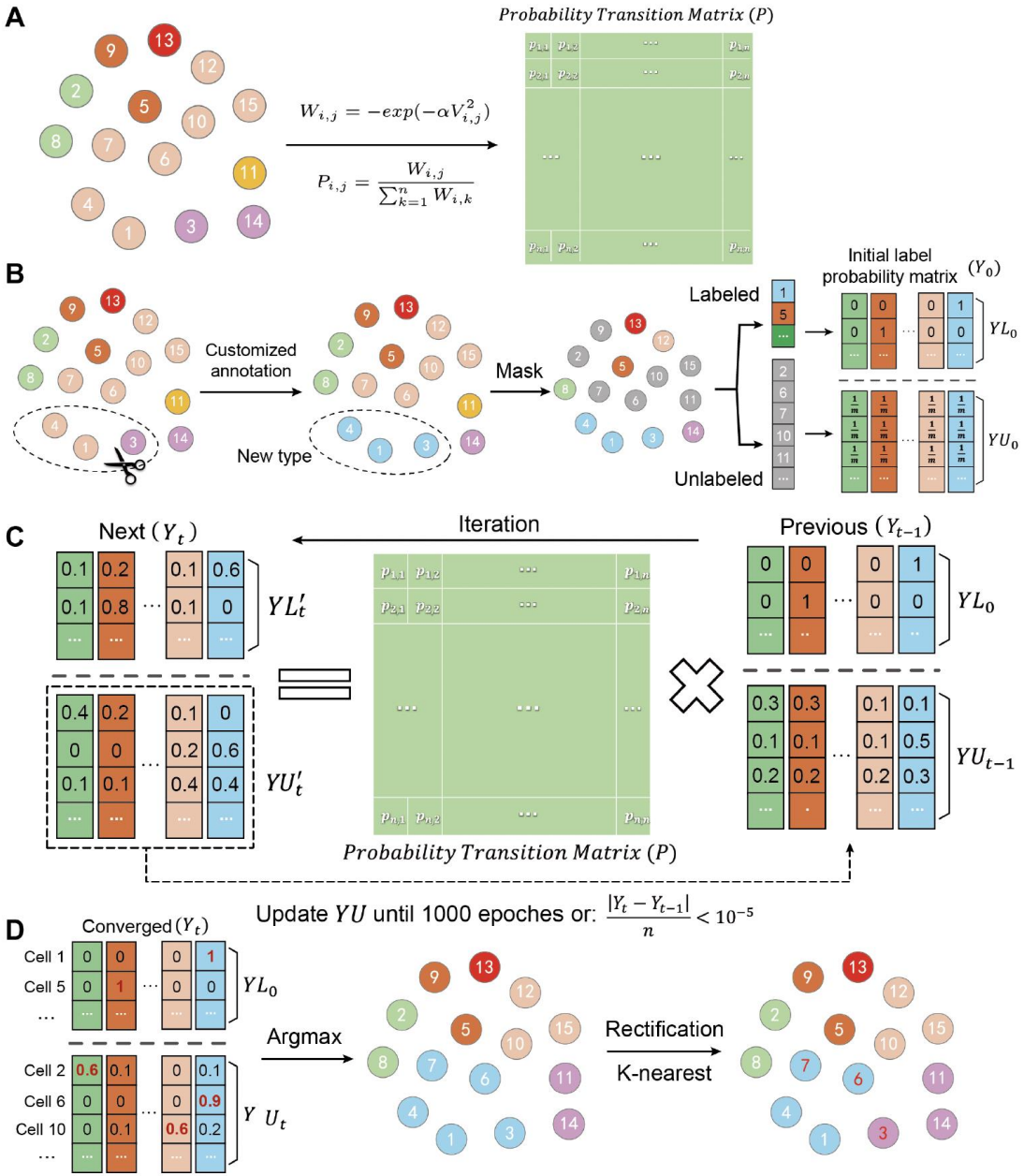


背景知识3——*Lasso-View*

- Lasso-View是PairPot数据库[1]中贡献的一种方法，前者是一种基于LPA算法来猜测用户感兴趣节点群的方法。
- 在PairPot对单细胞数据的分析中，每一个细胞作为图中的一个节点，当用户手动圈选一部分细胞后，Lasso-View能在毫秒级响应时间内识别该用户感兴趣的细胞亚群。



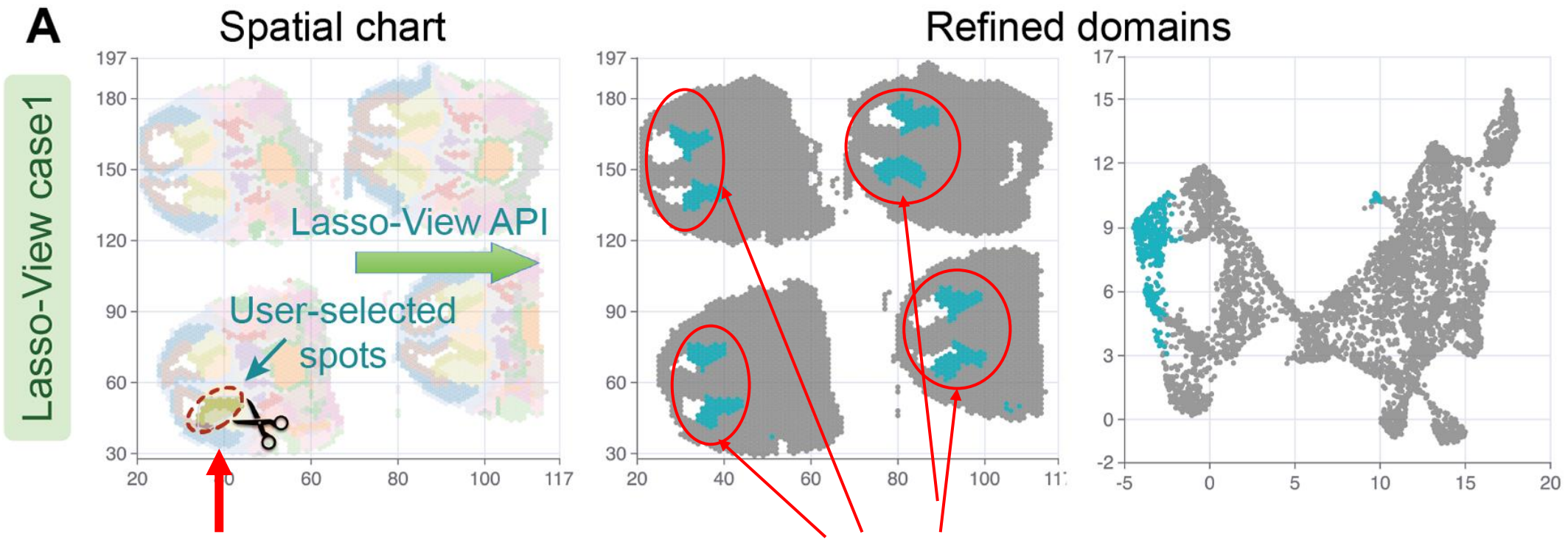
毫秒级启发式细胞注释 | Lasso-View启发式分析原理



- 在数据整合阶段，Pairpot为每个数据集生成概率转移矩阵。
- 在线分析阶段，用户圈选细胞并生成新的细胞类型U。
- 运用标签传播算法迭代生成所有细胞属于细胞类型U的概率。 该过程采用C++优化，可实现毫秒级响应。
- 运用K近邻调整生成结果。

Lasso-View启发式分析实例

选中Control 02切片上位于LGE区域的一块空间点位，调用Lasso-View的用户接口后，四个切片上的LGE区域都被突出显示。同时，被用户误选的空间点位也被剔除了。



模型评价方法

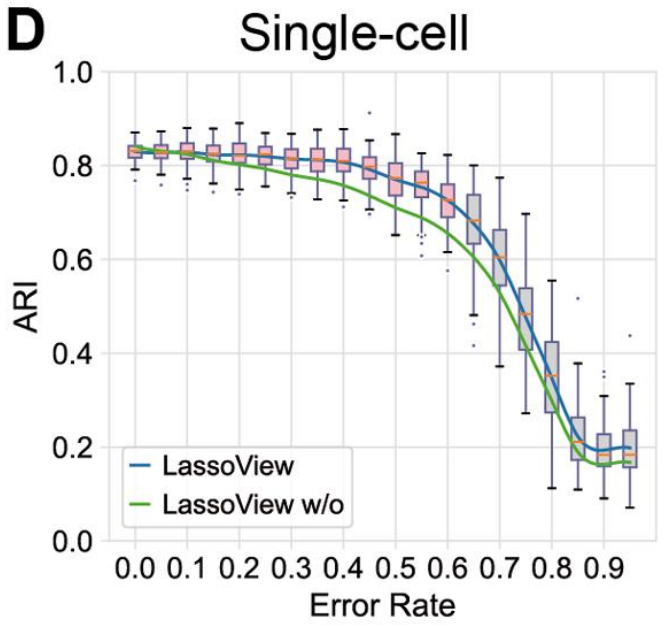
- 半监督学习的评价往往相对困难，因为没有办法生成足量的用户圈选细胞以及对应的真实值。
- 可以遮蔽大部分初始给定的注释，利用剩余的小部分注释信息，推断出所有细胞类型注释，再将推断出来的注释与初始注释进行比较，即可得出模型的性能。
- 为了贴合实际，可以在剩余的小部分注释信息中人为添加一些错误标签，并在这种情况下推断出所有细胞类型注释，并找出人为添加的错误标签
- Lasso-View中采用的评价方式是：遮蔽90%的初始注释，在10%剩余的标签中，添加错误标签（0–95%）；推断出所有细胞类型注释后，计算与初始注释的ARI值。

评价指标ARI

- ARI（Adjusted Rand Index，调整兰德指数）是一种用于衡量聚类结果与真实分类之间的相似度的评价方法。它通过比较聚类结果与真实分类之间的成对样本相似性来计算得分，匹配范围[-1,1]。
- 可以利用sklearn中的adjusted_rand_score函数来计算。
- 实验中需要计算两种ARI：
  1. 原始ARI_o（直接使用LPA的预测标签与真实标签的相似度）
  2. 修正ARI_r（修正后的标签传播结果与真实标签的相似度）
- 算法最终的评估：
  1. 准确性：错误率为0时的ARI值
  2. 鲁棒性：ARI>0.7时的错误率

评价指标ARI

- errPLA.py中提供了ARI的示例计算代码
- testPLA.py中提供了ARI变化的可视化代码
- mistake.txt中提供了参考的计算结果格式



ARI的可视化结果示意图。在每一个错误率下运行了100次。ARI值大于0.7的箱线图被标记为红色。（蓝线是ARI随错误率衰减的曲线，绿线是不处理错误标签的ARI衰减曲线）

作业具体要求

小组作业：1-3人

前提：题目给定聚类后的单细胞数据集（3000个细胞，2000个基因），以及用户所选择的细胞集合
目标：预测用户感兴趣的细胞集合，并对预测的结果进行评价

- **基础要求(10)：**阅读并复现PairPot论文[1]中基于随机游走的Lasso-View的方法，预测并可视化展示用户感兴趣的细胞群；使用论文中的方法对结果进行评价，屏蔽90%正常标签并模拟10%错误标签（错误率范围0-95%）。
- **进阶要求(10)：**使用其他算法实现，如决策树、支持向量机和神经网络等；也可以对论文中的标签传播算法进行优化。

Tips：我们将对所有人最终的ARI指标进行排名并赋分(这一部分占**10**分)，为了避免ARI过低，在实现的过程中请特别注意修改类内的错误标签。

如何进一步提升准确性和鲁棒性？

- Lasso-View实际上是一种基于图的随机游走模型，构建图的方法是采用欧式距离，优化图的构建方法可能是一种好的方案。
- 原问题实际上可以转化为分类问题。即用户选择的细胞可以当成一个新类型，然后用分类器判断所有的细胞中哪些细胞属于这种类型。分类器的构建可以采用支持向量机、决策树以及众多深度学习的方法。
- 可以对单细胞数据分析工作流进行优化。原始的表达矩阵是按照scanpy标准分析流得到的，可以通过对稀疏矩阵进行插补、降维和特征选择等方法得到更高质量的表达矩阵。
- 本次作业还提供了原始的表达矩阵（6000*20000），在进阶要求中可以使用这个矩阵进行训练（不强制）。

数据格式

- 单细胞数据集dataset.h5ad中存储了图中的所有细胞节点，每个细胞节点的维度数据主要是该细胞的注释（存储在annotation中）和它的基因表达情况（存储在稀疏矩阵中）。建议使用scanpy进行读取h5ad格式的数据。

```
import scanpy as sc
data = sc.read_h5ad("dataset.h5ad")
print(data.X) # Cell×Gene稀疏表达矩阵
```

[2] ✓ 0.0s

```
  (0, 23)       2.187500476837158
  (0, 29)       0.630852460861206
  (0, 37)       0.630852460861206
  (0, 87)       0.630852460861206
  (0, 88)       0.630852460861206
  (0, 104)      0.630852460861206
  (0, 113)      0.630852460861206
  (0, 117)      2.5200932025909424
  (0, 144)      0.630852460861206
  (0, 156)      1.8366167545318604
  (0, 167)      2.652409076690674
  (0, 179)      1.0146594047546387
```

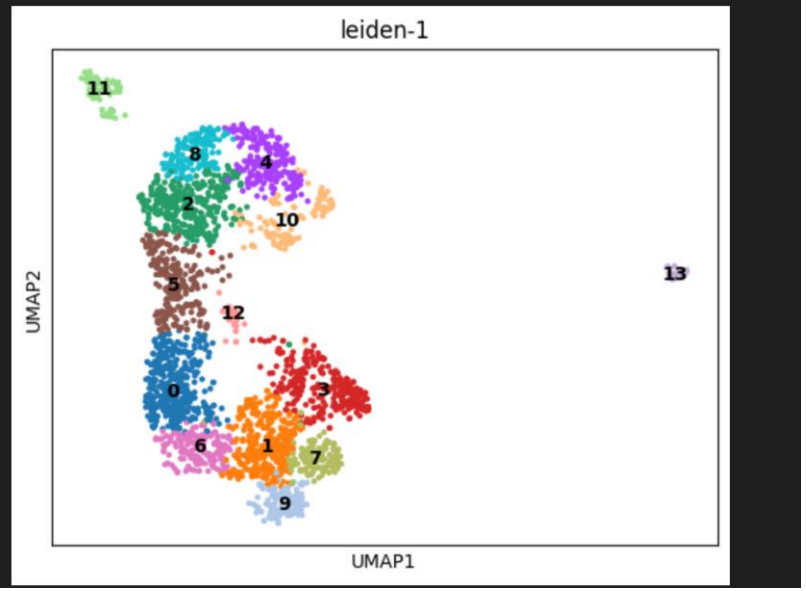
```
data # AnnData格式
```

[3] ✓ 0.0s

```
AnnData object with n_obs × n_vars = 3000 × 2000
    obs: 'leiden-1', 'annotation'
    uns: 'leiden-1_colors'
    obsm: 'X_umap'
```

```
# umap聚类的结果
sc.pl.umap(data, color=['leiden-1'], legend_loc="on data")
```

[4] ✓ 0.3s



- 共有10个测试数据。每个txt中的第一组数据是用户选定的细胞的索引，即输入；第二组数据是经过Lasso-View预测出的用户感兴趣的细胞集合的索引，即输出（供参考）。

2.txt

文件 编辑 查看

[13,42,46,66,70,112,166,194,290,332,342,405,444,513,518,537,710,720,723,950,957,963,1084,1085,1111,1132,1160,1207,1252,1308,1311,1337,1354,1443,1491,1536,1547,1585,1703,1712,1750,1800,1908,1928,1962,2065,2112,2131,2166,2292,2305,2324,2343,2346,2350,2382,2440,2715,2745,2747,2768,2932,2944,2969,2977,2992]

[13,42,46,66,70,112,153,166,194,290,332,342,343,382,405,423,444,513,518,537,689,702,710,716,720,723,734,859,863,950,957,963,1084,1085,1111,1132,1160,1207,1252,1308,1310,1311,1337,1354,1443,1491,1536,1547,1585,1703,1712,1750,1800,1908,1928,1962,1984,1991,2043,2065,2112,2131,2166,2178,2211,2241,2292,2305,2324,2343,2346,2350,2382,2427,2440,2533,2715,2745,2747,2768,2780,2830,2866,2932,2944,2969,2977,2987,2992]