进度汇报 （2022/7/24）

本次汇报内容

- 阅读了细胞来源成分组成的相关论文
- 看了一下FS的wrapper方法
- 神经网络的办法
- 接下来要做的事

成分来源论文

部分论文对课题实际用处不是很大，然后有些论文是相互承接的，我找了一个最近出来的论文，感觉结果好像还行，郭老师之前说的是做了一个简单的向量平均，但是不知道后面有没有其他的操作，这篇文章用了一个中心极限定理来迭代了一下参数，效果我看上去还不错。

《Likelihood-based deconvolution of bulk gene expression data using single-cell references》，doi：https://www.genome.org/cgi/doi/10.1101/gr.272344.120.

论文的源码地址：https://github.com/songlab-cal/rna-sieve

方法的大致描述

- 输入：各种细胞的矩阵（抓取特征后的特征向量组成的矩阵A）、待测样本的向量b。然后就变成一个AX=b的一个解线性方程组的问题。X就是待解的百分比。
- 方法：
- - 首先是把特征矩阵中相同来源细胞组织的特征向量算平均值M和方差S。
- - 然后有一个比较有意思的过程就是，这个算法里面认为b是一个概率分布的结果，也就是每个不同细胞的基因表达g和所占权重a的积ag的求和。所以这个算法里面希望用中心极限定理来迭代整套的参数。

算法小结

不知道郭老师之前所说的效果不好是大概是什么程度，不过这个算法给出的实验结果里，占比预测结果和实际的结果差距最大是在百分之二十的样子，整体的趋势感觉还行。我觉得我们还有工作可以做——我们可以一开始通过别的方法先算出来一个来源分布的大致种类个数。

FS特征选取

特征选取上，我看了一个讲FS课的ppt，算是进一步理解了FS中wrapper方法和Filter方法的区别，之前我们使用的是PCA和fisherscore这一类的filter特征选取方法，简单的就比如SFS，或者是后面的SFFS，后面我们可以用wrapper方法来进一步选取较为重要的特征集合，这种方法应该是比简单的filter方法更优的，这个后面可以作为优化的内容。

神经网络的办法

首先要更正一个事情，上周我在报告里说神经网络组织分类模型效果不好，不能收敛。实际情况好像没有特别差，我测试的数据里面有四种组织，每个组织数量都大于50，最后的正确率是在81%，我想如果样本能再大一些，效果可能会更好。

然后我去看了特征融合，也问了李学长关于特征融合的基本情况，我的理解是——特征融合目前能进一步提升特征抽取的准确性，但是好像并不能解决成分来源占比这个问题，这个我后面可能会去尝试。

这周我在想一件事情，就是如果已经确定血液中存在的组织种类，这样再来计算组织成分占比是不是会更加准确。这样的话，假设我能判断组织的来源——如果组织来源只有十个不到，然后再根据FS抽取出来的特征来进行计算，那问题就变成了一个很简单的线性方程组。

我现在脑子里，课题的整体步骤大概是这样：

- 预处理
- ◦ 1、输入所有的基因数据集
- ◦ 2、首先做FS，得到n个的最佳特征向量，然后记为F
- ◦ 3、把基因数据连同标签一块放到神经网络里，做一下特征提取
- 样本处理
- ◦ 1、输入待测样本
- ◦ 2、把样本转成图片，然后做一个目标检测，得到一个组织种类的集合O
- ◦ 3、O和F组合成一个新的矩阵A，然后用X代表占比，B代表待测样本的指标含量
- ◦ 4、然后用某一种方法解AX=B这个方程（也许可以直接用《Likelihood-based deconvolution of bulk gene expression data using single-cell references》，因为这篇文章里本来也有给组织来源个数降维的步骤，不过也许混合起来用效果会更好）

疑问

我们怎么模拟混合的血液样本，简单的两列相加应该是可行的?

接下来要做的事

我现在想做的是看看能不能使用目标检测的方法，把一个抽血样本中的组织来源给识别出来，当然这还只是一个比较初步的设想，我下周可能要看一下目标检测的东西。因为之前没有接触过，所以可能有一些目标检测的基本的东西需要去了解一下。

可能的问题我列了出来

- 特征信息可能会在图片里面相互覆盖
- 特征信息比较分散
- 个别组织种类特征信息相似度过高
  不过这些问题只是我现在想出来的，可能后面具体去了解的过程中会新增或较少一些问题。