进度汇报 （2022/7/10）

因为上周没有进行工作总结，所以这周我把这两周干的事情一块简单汇报一下。

汇报内容

- 对癌症类别以及单个组织来源的分类
- 接下来准备干的事情

1 对癌症类别以及单个组织来源的分类

目前我只用了两种方法，分别对癌症类别和癌症组织进行了分类。下面是大致的方法和情况。

1.1 使用FisherScore进行Feature Selection

我首先尝试了用FS的filter方法，直接用FisherScore对每个特征打分（分数越高说明信息量越多），然后选取了分数最高的若干的特征。最后使用了svm进行了分类。

FisherScore算的是一个类间方差K1和一个类内方差K2，然后用K2/K1作为分数。

下面是癌症分类的实验结果：

fs_score(above)	feature_number	accuracy
2.5	129	97%
3	49	95%
3.5	24	94%
5	5	83%

- SVM
- 训练集：测试集 4：1
- fs_score(above)：表示设定的分数阈值，超过该阈值则选取该特征

用现有的数据来做组织分类，因为又很多组织来源只有十个左右的样本，导致效果很差，我手工选了4个数量比较多（样本个数>70）的组织种类，然后做了一下，效果如下所示：

fs_score(above)	feature_number	accuracy
2.5	241	83%-72%
3	115	91%-83%
3.5	52	93%-81%
4	20	85%-82%

- SVM
- 训练集：测试集 4：1
- fs_score(above)：表示设定的分数阈值，超过该阈值则选取该特征
- 第三列运行了多次，准确率在这个范围波动

1.2 使用神经网络进行分类

上周老师最后发了一个用MLP网络做的一个预测癌症的文章。因为比较好奇效果，所以我用CNN又做了一次，感觉效果比想象的好。当然这只是对癌症种类的预测。

对于整个数据集的特征分布，其实我觉得卷积和池化也足够帮我们来进行特征抽取，只是收敛所需要的数据集会更大一些。

我直接使用了60483个特征作为输入，测试集和训练集比例为3：7，我用了三次卷积池化，这里不进行结构展示了。

癌症分类训练结果如下所示：

Epoch	loss	acc	val_loss	val_acc
1	68.7736	0.6635	2.6087	0.8362
2	5.6955	0.7686	4.3515	0.7155
3	2.9301	0.8158	2.6067	0.8362
4	1.1756	0.8891	6.0966	0.5862
5	1.2256	0.9084	0.5561	0.9138
6	0.8732	0.9122	0.4386	0.9914
7	1.0898	0.9325	4.2481	0.8017
8	0.7919	0.9373	1.3007	0.8707

Epoch	loss	acc	val_loss	val_acc
9	0.6382	0.9412	3.3948	0.7586
10	0.3972	0.9595	1.2354	0.9397

组织分类结果较差，这里不进行展示，最后的正确率大概是60%+。

1.3 小结

我现在感觉组织分类有一定样本数量的原因，目前我只知道样本选择对最后的分类结果有影响，但是样本数量和特征选择谁影响更大，这个我还没办法验证。

另外我个人有一个猜想，就是最后特征的个数应该在50个左右。

2 接下来干的事情

这周我找deconvolution的论文发现了我们现在研究的问题好像就是Cellular deconvolution。

Wikipedia上面有这个问题的一个介绍。
（https://en.wikipedia.org/wiki/Cellular_deconvolution#Current_strategies）

我现在在看这里面的论文，想先搞明白他们怎么做的。我会先看之前老师说过的reference based方法，然后再看reference-free方法。