大作业报告

1. 组内成员及分工

崔晏菲 2021210976，负责实现全连接模型和图卷积网络模型。

韩新宇 2021210967，负责实现树状分类器模型以及撰写报告。

2. 选题

我们选择 Project3: Hierarchical Cell type Classification 作为题目。

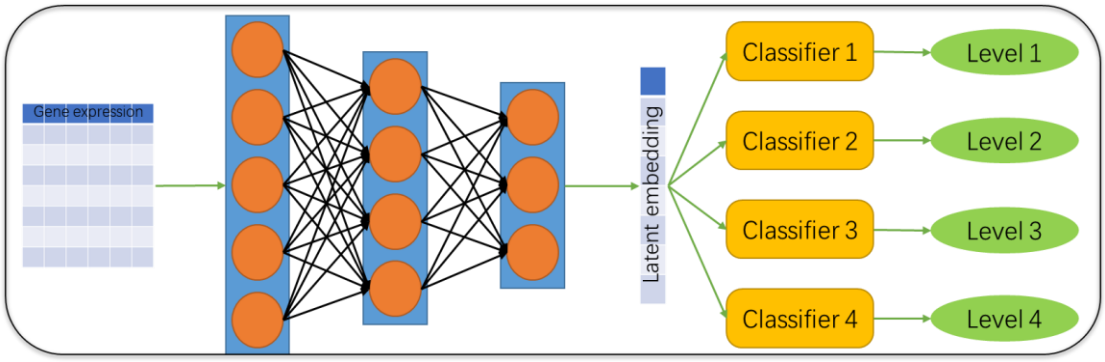
3. 方法与结果

我们设计了三套不同的模型：全连接模型、树状分类器模型和图卷积网络模型，方法和结果会按顺序一一说明。

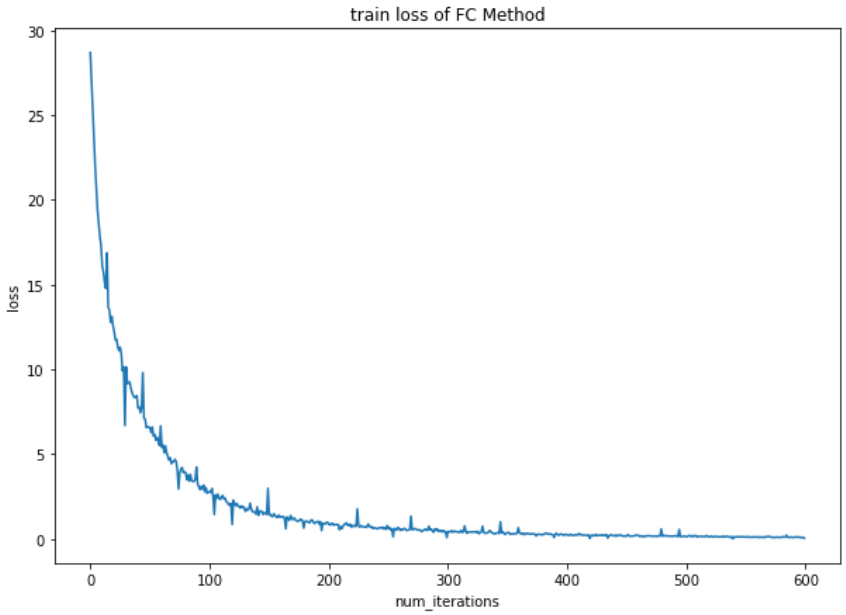
3.1 全连接模型

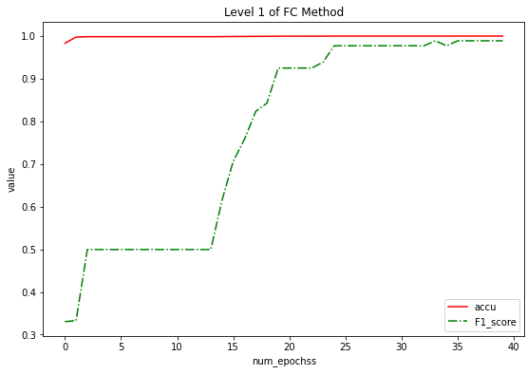
全连接模型也是最简单直接的分类器模型，因为样本有 4 个层级的标签，那么很自然地，我们可以采用分层分类的方法，在每一个层级，我们都可以将其视为一个单独的多分类任务，这样我们就有 4 个多分类任务。

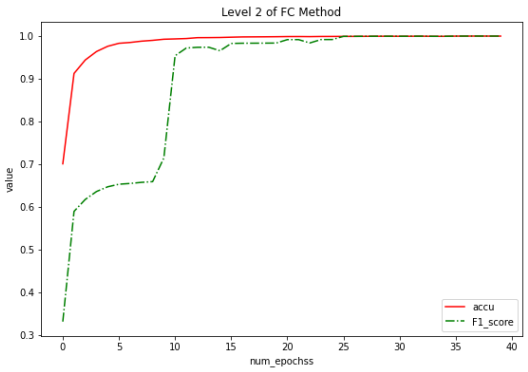
模型我们采用简单的全连接神经网络，我们使用了三层全连接，输出一个隐层 embedding，再根据这个 embedding，我们分别用 4 个分类器进行不同层级的分类任务。模型表示图如下：

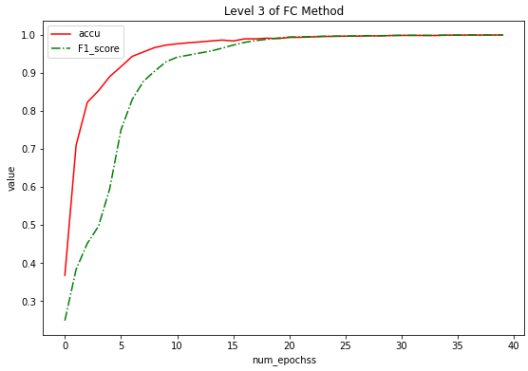


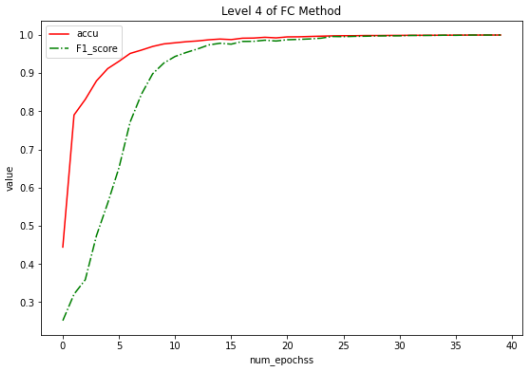
训练过程中，损失函数和各个层次的表现如下：











在验证集上的表现为:

level1, accu = 1.0000, f1_score = 1.0000

level2, accu = 0.9977, f1_score = 0.9725

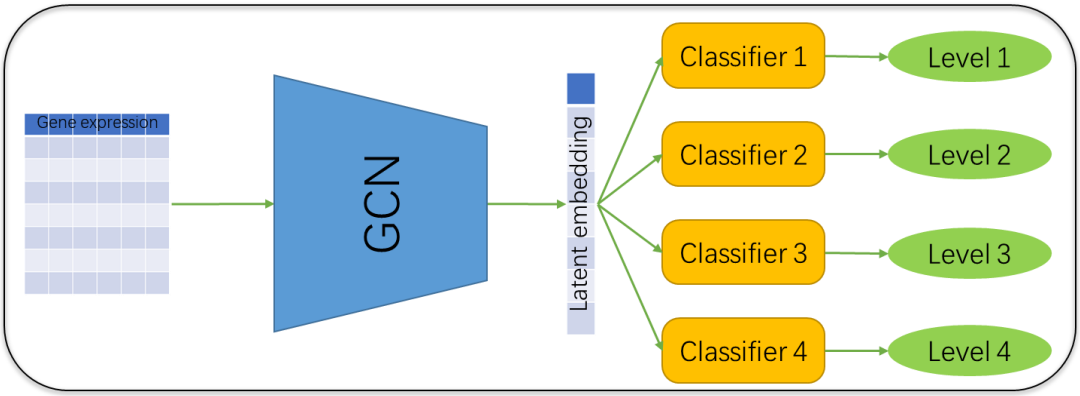
level3, accu = 0.9814, f1_score = 0.9669

level4, accu = 0.9810, f1_score = 0.9687

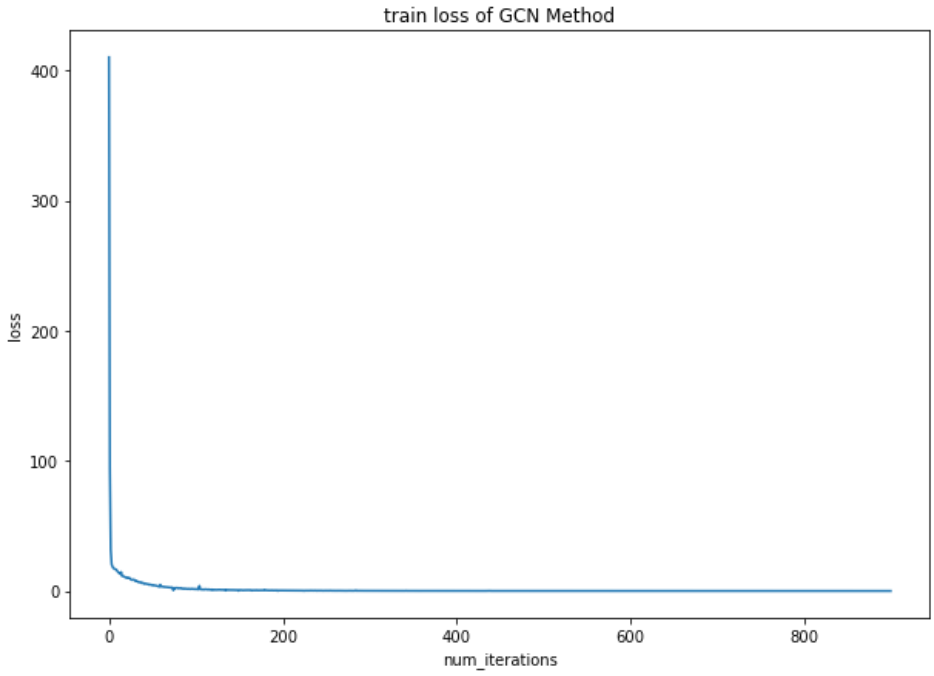
取得了相当好的效果。

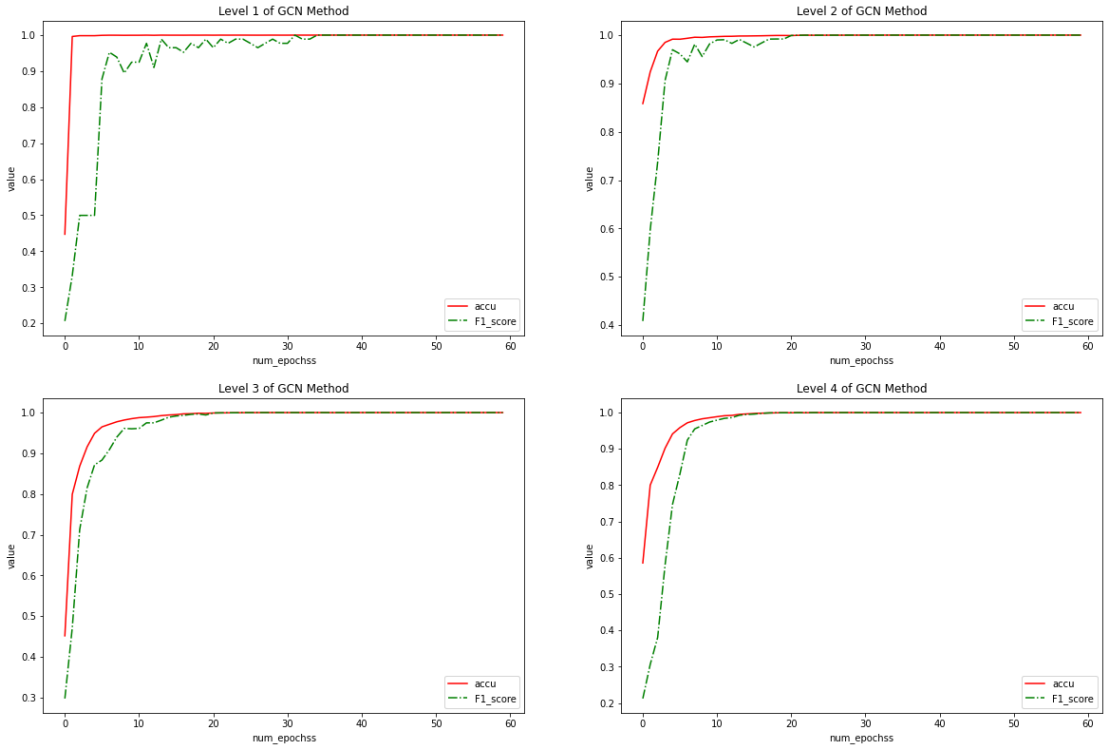
3.2 图卷积网络模型

鉴于全连接层很难给出可解释性，也很难提出生物上的意义，因此我使用图卷积网络代替全连接进行实验。图卷积网络的优势是可以更关注同一类别之间的关系，通过节点和边的关系，使得在生物上的含义更显著。下图是图卷积网络模型示意图：



训练过程中，损失函数和各个层次的表现如下：





可见，相比于全连接模型，图卷积网络收敛的速度更快，这说明边和节点的信息更有助于模型把握到基因表达矩阵中的关系，从而更好地为层次分类做服务。在验证集上的表现为:

level1, accu = 0.9998, f1_score = 0.9615

level2, accu = 0.9971, f1_score = 0.9975

level3, accu = 0.9833, f1_score = 0.9730

level4, accu = 0.9806, f1_score = 0.9697

效果比全连接层更好。

3.3 树状分类器模型

原始的特征提取-线性分类模型在层级分类问题上有一些缺陷。使用原始的模型时，当我们在做二级分类时，分类器并不能利用样本的一级分类结果。这意味着分类器需要适应更大的样本空间，并输出更多的预测结果。我们的线性分类器表达能力有限，在完成上述任务时

容易出现准确率的降低。

针对这一问题，我们提出如下改进。既然我们已知样本标签具有树状的层级结构，那么我们可以把原来每层设置一个分类器改为对每个分支节点设置一个分类器。同时，我们定义一个数据样本“相关”的分类器是那些位于样本标签到树根节点的路径上的分类器。换言之，对于树状的层级结构，我们在每个非叶子节点都设置一个分类器用于判断属于这个节点的样本对应的子类型。这种做法的好处是通过增加分类器数量，免去了分类器判断不相关样本的负担，提高了单个分类器的精确性。

我们的程序在实现这一功能时，采用了由统一的特征提取和树状层级结构的分类器组成的分类模型。每一个数据样本会产生多个标签，每个标签对应一个分类器。我们对数据的权重做了处理，确保对于一个数据样本，只有相关的分类器会在训练时受到影响。

根据此模型，得到的验证集上错误率为 2.0%。总体的 Macro-F1 分数为 0.9695。