## 1.评价指标

**Loss**：categorical\_crossentropy分类交叉熵函数：

https://pic1.zhimg.com/80/v2-a925abffa7feb0dbd71be9e3d9e41990_hd.jpg

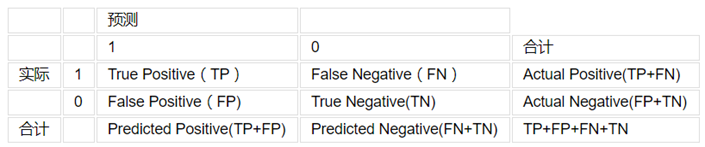
p are the predictions, t are the targets, i denotes the data point and j denotes the class.

某个蛋白质在所有分类中的损失=-六类求和 该蛋白质是否在所属类中\*log预测正确的概率

适用于多分类问题，并使用softmax作为输出层的激活函数的情况

**ROC**：　　考虑一个二分问题，即将实例分成正类（positive）或负类（negative）。对一个二分问题来说，会出现四种情况。如果一个实例是正类并且也被 预测成正类，即为真正类（True positive）,如果实例是负类被预测成正类，称之为假正类（False positive）。相应地，如果实例是负类被预测成负类，称之为真负类（True negative）,正类被预测成负类则为假负类（false negative）。

　　列联表如下表所示，1代表正类，0代表负类。

[](http://images2017.cnblogs.com/blog/785692/201709/785692-20170912171605219-830045428.png)

从列联表引入两个新名词。其一是真正类率(true positive rate ,TPR), 计算公式为TPR=TP/ (TP+ FN)，刻画的是分类器所识别出的 正实例占所有正实例的比例。另外一个是假正类率(false positive rate, FPR),计算公式为FPR= FP / (FP + TN)，计算的是分类器错认为正类的负实例占所有负实例的比例。还有一个真负类率（True Negative Rate，TNR），也称为specificity,计算公式为TNR=TN/ (FP+ TN) = 1-FPR。

其中，两列True matches和True non-match分别代表应该匹配上和不应该匹配上的

两行Pred matches和Pred non-match分别代表预测匹配上和预测不匹配上的

FPR = FP/(FP + TN) 负样本中的错判率（假警报率）

TPR = TP/(TP + FN) 判对样本中的正样本率（命中率）

ACC = (TP + TN) / (P+N) 判对准确率

　　在一个二分类模型中，对于所得到的连续结果，假设已确定一个阀值，比如说 0.6，大于这个值的实例划归为正类，小于这个值则划到负类中。如果减小阀值，减到0.5，固然能识别出更多的正类，也就是提高了识别出的正例占所有正例 的比类，即TPR,但同时也将更多的负实例当作了正实例，即提高了FPR。为了形象化这一变化，在此引入ROC，ROC曲线可以用于评价一个分类器。

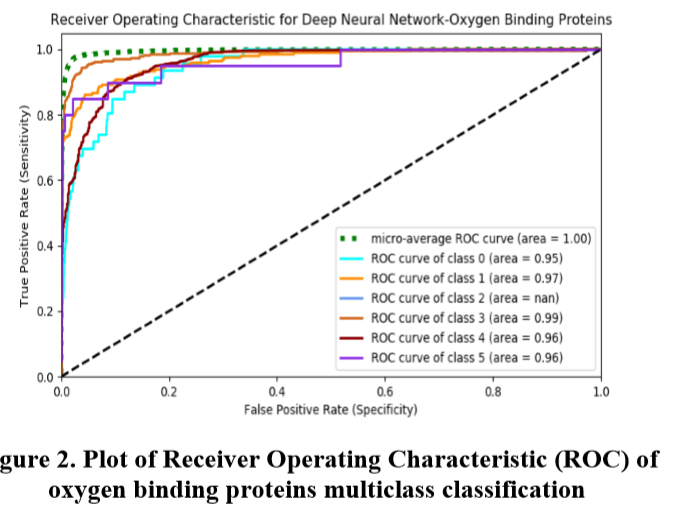
ROC曲线和它相关的比率

(a)理想情况下，TPR应该接近1，FPR应该接近0。

ROC曲线上的每一个点对应于一个threshold，对于一个分类器，每个threshold下会有一个TPR和FPR。

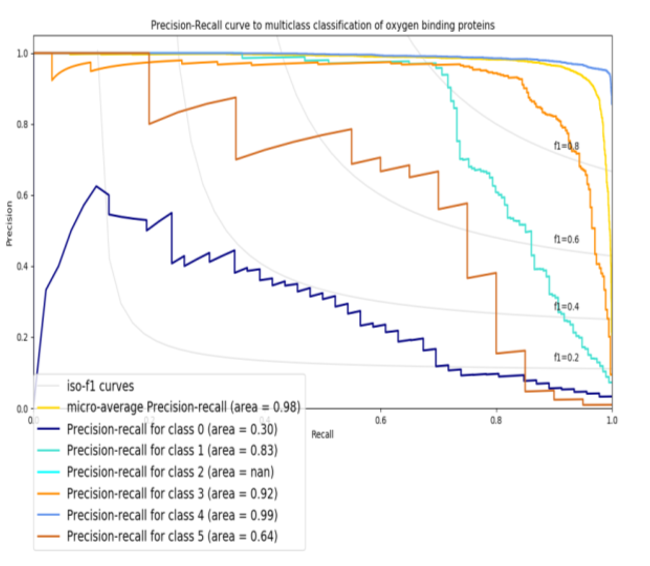
比如Threshold最大时，TP=FP=0，对应于原点；Threshold最小时，TN=FN=0，对应于右上角的点(1,1)

(b)随着阈值theta增加，TP和FP都减小，TPR和FPR也减小，ROC点向左下移动；



**AUC(Area under Curve)**：Roc曲线下的面积，介于0.1和1之间。Auc作为数值可以直观的评价分类器的好坏，值越大越好。

**Precision-recall：**横轴recall也就是TPR = TP/(TP + FN)，反映了分类器对正例的覆盖能力。而纵轴precision=TP/TP+FP。precision反映了分类器预测正例的准确程度。那么，Precision-recall曲线反映了分类器对正例的识别准确程度和对正例的覆盖能力之间的权衡。



## 2.分类

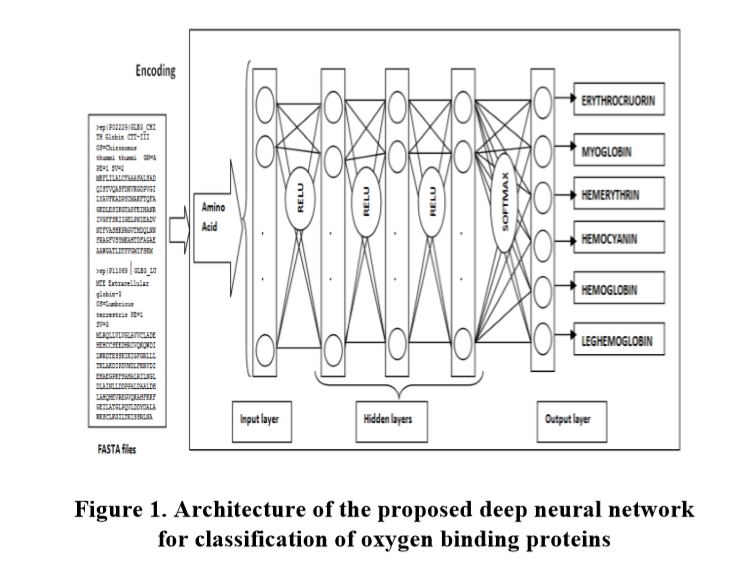
在uniprot上搜索分类数据

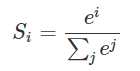
清除不符合的蛋白：name:hemerythrin NOT name:potential NOT annotation:(type:"alternative products" isoform) NOT name:probable fragment:no NOT subunit NOT name:like AND PE: keyword:"Oxygen transport [KW-0561]"

得到9大类蛋白：neuroglobin，myoglobin，leghemoglobin，hemoglobin，hemocyanin，hemerythrin，Flavohemoprotein，erythrocruorin，Cytoglobin

选取protein existence为evidence at protein level，evidence at transcript level，inferred from homology三个较高可信度的level

## 3.模型



5层神经网络，第一层为输入层，最后一层为输出层，中间三层隐藏层，使用relu函数，结点采用全连接，输出层用softmax函数，，得到每一个类别的可能性。

输入层输入20个特征，输出层给出标记。