# Metric

我们将所提出方法的视网膜糖尿病分类的结果与其他参考方法的结果进行比较，使用了如下的两个指标，分别是*宏平均*（Macro-Average）和*加权平均*（Weight-Average）。下面是这两个指标的一些信息。

## 准确率、召回率、F-measure

从二分类入手，对于某 $Threshold = N$，我们可以通过 Label 集合与预测集合构建出*混淆矩阵*（Confusion Matrix）如下：

| | | True-Labels | True-Labels |
|---|---|---|---|
| Predict-Labels | 1 | TP | FP |
| Predict-Labels | 0 | FN | TN |

由上表可以定义*准确率*（Precision）函数：

$$P = \frac{TP}{TP + FP}$$

和*召回率*（Recall）函数：

$$R = \frac{TP}{TP + FN}$$

通过分析可知，Precision 描述的是 “预测为正确的可信度”；Recall 描述的是“查找正确的能力”。显然，只使用准确率或召回率其中的一个进行样本评估是不合理的，二者存在一定 track out 关系。为了均衡这一关系，可以使用 F-measure[[1] Yang Y . An Evaluation of Statistical Approaches to Text Categorization[J]. Proc Amia Annu Fall Symp, 1999, 1(1-2):358-362.]。其中 F1-score 的函数定义为：

$$F_1 = \frac{2 \times P \times R}{P + R}$$

## 宏平均（Macro-average）

二分类问题使用上述 F1-measure 并不会存在争议，但当我们在 n 个二分类问题上综合考察评价指标，上述方法失效。此时我们需要使用其他方法。在本文中我们使用了*宏平均*（Macro-average）和*加权平均*（Weight-Average）。

*宏平均*（Macro-average）是先对每一个分类求得统计指标，然后对所有指标求取算术平均值，最后对结果求 F1-score：

$$\mathrm{Macro\_}P = \frac{1}{n}\sum_{i=1}^{n} P_i$$

$$\mathrm{Macro\_}R = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$$\mathrm{Macro\_}F = \frac{2 \times \mathrm{Macro\_}P \times \mathrm{Macro\_}R}{\mathrm{Macro\_}P + \mathrm{Macro\_}R}$$

*加权平均*（Weight-Average）的目的是改善分类数据集不平衡的问题，其定义如下：

$$Weight\_avg = \frac{1}{n}\sum_{i=1}^{n} F_{1i} \times \theta_i$$

其中 $F_{1i}$ 是第 $i$ 个分类的 F1-score，$\theta_i$ 是第 $i$ 个分类的样本占比。

## 对比

相比于传统的 Accuracy 评价指标，宏平均和加权平均都更有优势。

- 宏平均权衡了分类问题中所有的类，并对他们进行调和，能够更好的反映分类任务的性能。
- 加权平均平衡了数据集各分类的占比，对本文所要解决的数据集严重不平衡任务尤其重要，是一个更加合理的评价指标。
- 传统 Accuracy 评价指标较差，尤其是当数据集不平衡时，Accuracy 将会收到更大的来自数据集的影响。