DOI:10.19850/j.cnki.2096-4706.2020.04.006

P-R 曲线与模型评估问题研究

张超

（南京信息工程大学，江苏 南京 210044）

摘　要：现阶段深度学习作为一种实现机器学习的技术，在分析模型问题和评估模型的方法上基本一致。文章从评估模型的角度，以混淆矩阵为基础，通过常用的 Accuracy，Precision 以及 Recall 等衡量模型的预测能力。研究结合深度学习近几年的竞赛任务分析样本均衡与非均衡下几种评估模型方法的差异，从几种评估指标之间的联系讨论 P-R 曲线评估模型之间的相关性，以及 P-R 曲线在目标检测任务中作为评估模型方法的合理性。

关键词：评估标准；Accuracy；P-R 曲线；mAP

中图分类号：TP181；TP311.1　　**文献标识码**：A　　**文章编号**：2096-4706（2020）04-0023-03

Research on P-R Curve and Model Evaluation

ZHANG Chao

（Nanjing University of Information Science and Technology，Nanjing 210044，China）

Abstract：At present，deep learning，as a technology to realize machine learning，is basically consistent in analyzing model problems and evaluating model methods. From the perspective of the evaluation model，based on the confusion matrix，this paper measures the prediction ability of the model from the commonly used Accuracy，Precision and Recall. This paper analyzes the differences of several evaluation models under the condition of sample equilibrium and non-equilibrium，discusses the correlation between the evaluation models of P-R curve from the relationship between several evaluation indexes，and discusses the rationality of P-R curve as the evaluation model method in the target detection task.

Keywords：model performance evaluation；Accuracy；P-R curve；mAP

0　引　言

在机器学习领域，模型评估的主流标准都以统计混淆矩阵下的 TP（true positive）、FP（false negative）、TN（true negative）、FN（false positive）去评价模型的优劣，例如图像分类竞赛的 ImageNet[1] 中使用的 Top-1、Top-5 的 Accuracy 评估标准。笔者在建模研究中，发现在使用以上几种评估指标评估模型时，单性能指标不能准确地评估模型，如何更优地评估模型以及如何迭代地对模型进行优化成为了研究热点之一。本文针对各种数据集样本分布不均衡问题进一步分析，在针对多种深度学习技术应用中的目标检测、图像分割以及图像分类的模型评估方法的研究中，我们从模型评估的最优化的角度，研究模型判定标准以及怎样的标准才能最优评估模型，以混淆矩阵为基础，探究评估模型最优方法。

本文从多类问题分析时可以理解为对主类和从类（其他）两种角度，研究以两类问题下的 P-R 曲线，并结合竞赛任务中的应用分析 P-R 曲线在实际场景中的评估方法的最优性。

1　混淆矩阵与几种评估指标的关系

我们以汽车分类问题为例，假设汽车类型有 A 和 B 两类，我们在均衡比例 A=5，B=5 的测试集去测试车型分类对两类样本的识别能力。假设分类器对在测试集中识别 A 类汽车共预测 7 张，预测 B 类汽车 3 张，其中识别成 A 类中包含 A 类 5 个，B 类 2 个，识别成 B 类中，实际包含 2 个 B 类和一个 A 类。根据以上模型预测输出能够得到对应的混淆矩阵如图 1 所示，本文从混淆矩阵进行分析几种评价指标之间的相互关系。

Confusion Matrix		True Value	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

（a）主从类混淆矩阵

Confusion Matrix		True Value	
		Car_A	Car_B
Predicted	Car_A	5	2
	Car_B	2	1

（b）车型分类混淆矩阵

图 1　车型分类的混淆矩阵

在图 1 汽车分类中，如果我们先以 Accuracy 评估模型为例，能够计算出 Accuracy=0·60，Precision=0·71，Recall=0·71。以上三种评估指标的计算是在样本均衡情形下得出的，

收稿日期：2020-01-16

假设样本中如果测试样本类是以 9：1 的方式呈现，预测结果 A=7（预测 A 集合中 A=6，B=1），在这种在数据不均衡的情况下，我们能够通过混淆矩阵计算 Accuracy=0·60，P=0·86，R=0·67。从样本不均衡和均衡条件的计算，可以看出在两种样本分布下 Accuracy 表现结果相同，而 Precision 和 Recall 随样本分布变化呈现反差，类别不平衡问题会导致正样本或负样本的比例过多，当仅使用 Accuracy 去评估分类器性预测能力就不能合理地评估，在 Everson R 中对分类问题样本不均衡情况下如何进行合理评估方法的讨论以及 Davis[2] 证明了 PR 曲线相比于文献 [2] 在样本不均衡下更关注正样本，更能反映分类器的好坏的推理。当针对其他分类问题时，我们将预测类主类和其他类从类归纳为二分类问题，能够对混淆矩阵预测总结计算 Accuracy 公式（1）所示。

$$Accuracy(p_+, r_+) = \frac{r_+ + r_-}{2} = \frac{1 - \frac{r_+}{p_+} + 2r_+}{2} = \frac{1}{2} - \frac{r_+}{2p_+} + r_+ = \frac{p_+ - r_+}{2p_+} + r_+ \quad (1)$$

从上分析，以及几种竞赛的评估分析中能够看到，Precision 和 Recall 是各种评估标准的基础，本文主要从 P-R 曲线在各种竞赛任务中的应用进行分析，论述 P-R 曲线的评估方法，并结合竞赛任务中的应用分析 P-R 曲线在实际场景中的评估方法。

2 P−R 曲线分析

Davis[2] 详尽地分析了通过 P、R 去关注分类器在预测主类的能力，能相对较好地忽略样本不均衡带来的问题（更关注于主类样本），能更有效地评估模型。本节主要从 P-R 曲线评估对象和在检测任务中通过两者之间的曲线去评估模型进行讨论，从坐标中的绘制方式到几种可能状态以及去分析如何获取理想条件下的最优 P-R 曲线，对比假设在 P-R 曲线在样本均衡时呈现。

2.1 Precision 和 Recall 之间的联系

在理想情况下，我们希望模型的精确率越高越好，同时召回率也越高越好，但实际情况下紧缺率和召回率总呈现反比状态。在各种竞赛的评估标准中，已有对 P-R 曲线权衡问题的讨论，在对两种指标进行平衡中，常用的方法是 F-Measure（加权调和平均），当 Measure=1 时即为 F1_Score 平衡两者之间的关系，如表 1 所示是几种文本分类算法在 CNN 和 DailyMail 数据集测试和验证集上 F1 的评估对比，在 NLP 类别的竞赛任务中，在样本无法均衡情形下，利用 F1 综合考虑 P 和 R 的评估结果，取 F1 较高指标时作为最优模型，能更准确地评估模型性能。

$$F1 = \frac{2 * P * R}{P + R} \quad (2)$$

表 1 文本分类算法在竞赛数据中 F1 上的性能对比

	CNN		DailyMail	
Attentive Reader	Val	Test	Val	Test
	61·6	63·0	70·5	69·0
MemNN	63·5	68	-	-
AS Reader	68·9	69·5	-	-
BIDAF	73·6	76·9	80·3	79·6

以上我们讨论 Precision 和 Recall 以及加权调和平均的应用和关系，如文中针对评估模型，很难仅依靠其中单一的指标去判断模型的优劣，因此如表 1 文本比赛中通过 Precision 和 Recall 之间的调和平均数、目标检测以及图像分割任务中 AP 的计算都是通过 P-R 曲线之间的联系进行模型分析和评估。

3 P−R 曲线在目标检测数据中的分析与评估

本节对目标检测竞赛任务中评估模型的方法使用 P-R 曲线进行讨论，本节主要从不同季赛情形下数据的分布情况进行分析，论证 P-R 曲线对模型评估的合理性并讨论通过 P-R 曲线获取 AP 的评估方式。

3.1 P−R 曲线评估合理性分析

在合理的模型的测试集中，我们期望理想情况下主类测试样本是无穷多，从类测试样本也是无穷多。以目标检测任务为例，在对评估模型性能过程中，我们期望测试集尽可能包含所有场景样例，去验证模型的鲁棒性。假设在测试样本中多类别 C_i，P-R 曲线是通过对象类别 i（主类）和其他类（从类），以分类角度去评估，从先验概率的角度来讲，Precision 最小值由 P（主类）和 P（从类）两个先验概率确定，为 P（主类）/（P（主类）+P（从类））。在主、从类样本均衡时，在少样本情况下每组 P-R 曲线呈现。需要考虑在多样本情形下 P-R 曲线如何呈现。

虽在很多算法中对分类样本不均衡问题从算法层面和数据增广方面进行处理样本不均衡问题。如检测算法 Faster RCNN 利用 RPN 抑制前背景以及一阶段检测算法 SSD 按一定比例取前背景样本预测样本。但样本不均衡问题仍不能很好地解决，对模型训练和评估存在极大影响。

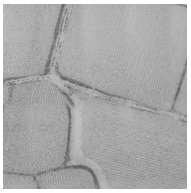
3.2 P−R 曲线在目标检测竞赛种的应用

从目标检测竞赛 VOC2007[3] 训练和测试数据以及 KITTI 数据，Person 类和其他类别相比明显出现类别不均衡问题。目标检测需要在大背景下同时对目标进行定位和识别，但由于前景和背景之间的不平衡，使得这一工作具有挑战性。基于深度学习的检测解决方案通常采用多任务分支网络体系，处理不同类别的分类任务和定位任务，其中分类任务的目标是识别给定框中的对象，而定位任务的目标是预测对象的精确边界框。无论是类别不均衡还是前景和背景不均衡的问题，对基于模型的训练和评估都有着极大的影响。

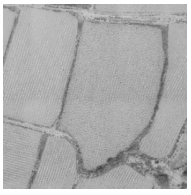
从在 VOC2007 检测数据的评估中的方法看，通过训练好的模型，对 VOCtest 数据（4 952 张）预测后，得到所有 Person 类的置信度得分，并对预测为 Person 类的样本进行排序，计算 Precision 和 Recall。在 VOC 竞赛评估中是对当前类评估时该类为主类，其他类和背景类为从类，得到两类问题的混淆矩阵，计算 Precision 和 Recall。根据计算一组 Range[0，0·1，1·0]（0 ~ 1·0 之间间隔为 0·1，11 个点）的 Recall，得到 Recall > Threshold 时对应最（下转 27 页）

0·75，通常情况下，阈值越大，识别越复杂，准确率越低。APs（area $< 32^2$），APm（$32^2 <$ area $< 96^2$），APl（area $> 96^2$），表示小目标、中等目标、大目标的 AP。由实验结果可知，该系统对于大目标的识别结果更优。

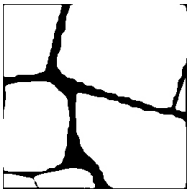
我们使用训练出来的权重文件进行识别效果的测试。测试所需要的图片必须不在训练集和验证集中，以避免测试结果的误差。其中两张图片和其识别结果如图 3 所示，我们可以观察到，我们所训练的模型可以将图片中的地块很好地识别出来。



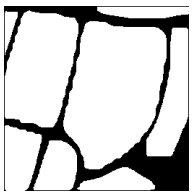
（a）地块原图 1



（b）地块原图 2



（c）识别结果 1



（d）识别结果 2

图 3　地块原图及识别结果

3　结　论

各式各样的卷积神经网络推动着深度学习不断向前发展。作为卷积神经网络中的一个主流实例分割算法，Mask R-CNN 在目标检测、人体姿态识别方面都有着很好的效果，灵活且易于掌握。本文中用 Mask R-CNN 进行网络训练和识别地块的效果较好，但在地块边缘仍存在识别不精准的情况。对此，可以扩大训练集的数量以及提高人工标注的准确度来取得更好的识别效果。

参考文献：

[1] 陈建廷，向阳 . 深度神经网络训练中梯度不稳定现象研究综述 [J]. 软件学报，2018，29（7）：2071−2091.

[2] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

作者简介：史宝坤（1998−），男，汉族，河北承德人，本科在读，研究方向：计算机科学；李欣（1999−），男，汉族，河北保定人，本科在读，研究方向：软件工程；魏春燕（1999−），女，汉族，河北石家庄人，本科在读，研究方向：电子信息科学与技术；安子湉（2000−），女，汉族，河北保定人，本科在读，研究方向：电子信息科学与技术；杜兵戈（2001−），女，汉族，河北石家庄人，本科在读，研究方向：数学与应用数学。

（上接 24 页） 大的 Precision，通过 11 点差值平均精度得到 Person 类的 AP，其计算如式（3）左公式所示。

$$\mathrm{AP}=\frac{1}{11}\sum_{r\in\{0,\,0.1,\,\cdots,1\}}\sum_{K=1}^{N}\max_{\tilde{k}}P(\tilde{k})\ \Delta(k)\Rightarrow$$

$$\mathrm{mAP}=\frac{\sum_{1}^{Q}\mathrm{AP}\ (q)}{Q}\qquad(3)$$

式（3）右公式是对 VOC2007 中差值平均精度计算类别对象的 AP 和多类 mAP 的方法，其中 Precision 是通过 11 点中根据 Recall 计算出的 Top-N 中取最大值。在 P-R 曲线中，虽然理论上说 PR 曲线呈现递减趋势，在 VOC2007 的评估中可能会出现某阶段上升情形，但总体上来说在多样本情形下整体是趋于下降趋势的，因此 P-R 曲线的评估更符合实际多样本情形下评估模型问题的标准。

4　结　论

考虑到几种模型精度评价标准都基于混淆矩阵对模型预测能力的统计进行评估，本文主要从 Precision 和 Recall 之间的关系进行讨论，分析了分类以及检测任务中实际场景中样本分布的样本均衡和前背景均衡问题。本文从 P-R 曲线的角度分析了其在面对以上问题时评估方法的合理性。近几年学术研究中通过 CNN 模型的拟合评估角度思考的方式，通过 P-R 角度思考的 AP-Loss[4] 为一种新思路，这也将是我们后续探索 P-R 曲线与模型结合的优化方向。

参考文献：

[1] DENG J，DONG W，SOCHER R，et al.ImageNet：a Large-Scale Hierarchical Image Database [C]//2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition （CVPR 2009），USA.IEEE，2009.

[2] DAVIS J，GOADRICH M.The Relationship Between Precision-Recall and ROC Curves [C]//ICML’06：Proceedings of the 23rd international conference on Machine learning，2006：233-240.

[3] EVERINGHAM M，GOOL L V，WILLIAMS C K I，et al.The Pascal Visual Object Classes （VOC） Challenge [J]. International Journal of Computer Vision，2010，88（2）：303-338.

[4] CHEN K，LI J G，LIN W Y，et al.Towards Accurate One-Stage Object Detection with AP-Loss [J].[2019-12-26].https://arxiv.org/abs/1904.06373?context=cs.CV.

作者简介：张超（1992.06−），男，汉族，河南固始人，硕士在读，研究方向：模式识别。