CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 43 卷第 11 期 2021 年 11 月
Vol. 43, No. 11, Nov. 2021

文章编号:1007-130X(2021)11-1910-10

一种基于混淆矩阵的多分类任务准确率评估新方法*

张开放,苏华友,窦　勇
(国防科技大学计算机学院,湖南 长沙 410073)

摘　要:多分类任务准确率评估对评判模型的分类效果具有重要的理论意义和应用价值。针对机器学习领域的多分类任务,在现有方法的基础上,通过拓展和迁移应用,给出一种新的评估方法。为了准确评估多分类任务模型的分类效果,将遥感图像分类效果评估方法引入多分类任务。针对多分类任务的实际特点,对该方法进行了改进与推广,以更好地评估分类器效能。基于 MNIST 手写字符集识别任务和 CIFAR-10 数据集分类任务的实验结果表明,同样是基于混淆矩阵进行计算,与现有的评估方法相比,该方法可以同时给出分类器整体的分类效果和单个类别的分类效果,对于改进训练过程有一定的指导意义。另一方面,该方法可以推广到任意的分类任务分类效果评估工作中,具有较好的应用前景。

关键词:多分类;准确率评估;混淆矩阵
中图分类号:TP181　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2021.11.002

A new multi-classification task accuracy evaluation method based on confusion matrix

ZHANG Kai-fang, SU Hua-you, DOU Yong
(College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: The accuracy evaluation of multi-classification tasks has important theoretical significance and application value to the classification effect of the evaluation model. Aiming at multi-classification tasks in the field of machine learning, based on the current existing methods, this paper proposes a new method by expanding and migrating applications. In order to accurately evaluate the classification effect of the multi-classification task model, this paper introduces the remote sensing image classification effectiveness evaluation method (R') into the multi-classification tasks. In view of the actual characteristics of the multi-classification tasks, the method improves and popularizes the R' method to better evaluate the performance of classifiers. The experimental results on the recognition task of MNIST handwritten character set and the classification task of the CIFAR-10 dataset show that, although the calculation is also based on the confusion matrix, compared with the existing evaluation indicators, the method can simultaneously give the overall classification performance of the classifiers and the classification efficiency of the individual categories, which can be beneficial to the training process. On the other hand, the method can be extended to the classification performance evaluation of any classification tasks, which has a good application prospect.

Key words: multi-classification; accuracy assessment; confusion matrix

* **收稿日期**:2020-08-16;**修回日期**:2020-10-30
基金项目:国家重点研发计划(2018YFB0204301)
通信作者:苏华友(shyou@nudt.edu.cn)
通信地址:410073 湖南省长沙市国防科技大学计算机学院
Address: College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, Hunan, P. R. China

1 引言

在机器学习领域,多分类任务[1-3]是指将样本实例分为3个及以上类别之一的问题(将样本实例分类为2个类别之一称为二分类)。由于分类算法和模型的局限性,对分类器的分类结果进行准确性评估是一个必须面对的问题[4,5]。另一方面,由于分类器过拟合现象的存在,恰当地选择准确率评价指标显得十分重要。现有的一些评价指标,诸如准确率[3]、Kappa系数[6]和$F1$值[3]等,都是基于混淆矩阵对总体分类效果进行的评估。它们很难给出单个类别的分类效果,这在某些实际应用中是不足以满足用户需求的(例如在MNIST(Mixed National Institute of Standards and Technology database)手写字符体识别任务中,数字0出现的概率和重要性往往会比其他数字大和高)。

本文将该方法引入多分类任务模型评估场景。该方法最初运用于地震预测领域[7],后被引入遥感图像目标识别效果评估领域[8,9],用于评估识别的效率。本文针对机器学习领域的多分类任务,对该方法进行拓展和迁移应用,并给出了理论推导过程。基于MNIST手写字符体识别和CIFAR-10(Canadian Institute For Advanced Research, 10 classes)数据集的多分类任务实验结果表明,与已有模型准确率评估方法相比,上述方法可以较好地评估模型分类准确率。值得一提的是,同样是基于混淆矩阵进行推理,该方法计算简单,并且可以同时给出分类器整体以及每一个类别的分类效果,对于评估和改进训练过程具有一定的指导意义,同时在特定的任务背景下应用前景广阔。

本文的主要工作如下所示:

(1)提出了一种新的多类别分类效果评价指标,该指标考虑真实标签和预测标签之间的数值差异,可以更好地反映分类模型的分类效果。

(2)从数学上给出了所提指标的理论推导及其性质证明。

(3)通过该指标可同时获得总体和单个类别分类效果,以改进分类模型训练过程。

(4)在不同的应用中评估了各指标在MNIST和CIFAR-10数据集上的分类效果,以验证其有效性和鲁棒性。

2 现有评估方法及其缺陷

本节主要介绍几种常见的模型准确率评价指标及其不足。不失一般性,考虑表1所示的三分类问题的混淆矩阵。表1中,l、m、n分别代表类别1、类别2、类别3的真实样本数,r、s、t分别代表结果中预测为3个类别的样本数;w是所有样本的总数;a、b、c代表被正确分类的样本数,d、f、g、e、i^*、h代表被错误分类的样本数。

Table 1 Confusion matrix of the three-category task

表1 三分类问题混淆矩阵

		分类结果			
		类别1	类别2	类别3	小计
真实标签	类别1	a	d	f	l
	类别2	g	b	e	m
	类别3	i^*	h	c	n
	小计	r	s	t	w

2.1 准确率

准确率作为分类问题最原始的评价指标,定义为正确预测的样本占总样本的百分比。对于表1所示的混淆矩阵,有:

$$accuracy=\frac{a+b+c}{w} \tag{1}$$

显然,这一指标没有考虑非对角线因素,也就是忽略了诸多的边界样本信息,尤其是在各个类别样本数量不均衡的情况下,它不能很好地评估分类效果的好坏。

2.2 *PR*曲线

*PR*曲线是描述精准率、召回率变化关系的曲线。其中P代表精准率(Precision),又叫查准率,是针对分类结果而言的,定义为所有被预测为正的样本中真实标签为正的样本的概率;R代表召回率(Recall),又叫查全率,是针对真实标签而言的,定义为所有实际为正的样本中被分类为正的样本的概率。曲线最初是针对二分类任务场景提出的,混淆矩阵如表2所示。其中,m、n分别代表类别1和类别2的真实样本数,s、t分别代表分类结果中预测为2个类别的样本数;w是所有样本的总数;a、b代表被正确分类的样本数,c、d代表被错误分类的样本数。表1和表2的a、b、c和d仅有局部意义,分别适用于三分类场景和二分类场景。

Table 2 Confusion matrix of the two-category task

表2 二分类问题混淆矩阵

		分类结果		
		类别1	类别2	小计
真实标签	类别1	a	c	m
	类别2	d	b	n
	小计	s	t	w

其 PR 值的计算如式(2)所示：

$$P=\frac{a}{s},R=\frac{a}{m} \tag{2}$$

对于多分类问题，实际上会获得多组混淆矩阵，也就会得到多组 PR 值，此时有 2 种处理方法：宏平均(macro-average)和微平均(micro-average)。宏平均是先计算每个混淆矩阵的 PR 值，然后再分别取平均；微平均则是计算出全局混淆矩阵的平均正负样本数，然后再计算整体的值。

这样，对于上述三分类问题，采用宏平均方式计算如式(3)所示：

$$P_{\text{macro}}=\frac{1}{3}\sum_{i=1}^{3}P_i,\quad R_{\text{macro}}=\frac{1}{3}\sum_{i=1}^{3}R_i \tag{3}$$

其中，P_i 和 R_i 分别代表类别 i 的精准率和召回率，具体计算方法为：$P_1=a/r,P_2=b/s,P_3=c/t;R_1=a/l,R_2=b/m,R_3=c/n$。

采用微平均方式(对于没有漏检的多分类任务而言，实际就是 2.1 节中的准确率)计算如式(4)所示：

$$P_{\text{micro}}=\frac{a+b+c}{w},\quad R_{\text{micro}}=\frac{a+b+c}{w} \tag{4}$$

可以看出，宏平均虽然加入了更多的非对角线元素，但是仍然只能给出所有类别整体的分类效果，而微平均则和 2.1 节的准确率等价。同时，PR 值是一对此消彼长的统计量，在实际应用中要做好两者的兼顾和取舍。

2.3 *F1* 值

为了解决 PR 值的上述问题，调和 PR 值，研究人员提出了 *F-measure*(或 *F-score*)方法，即：

$$F\text{-}score=(1+\beta^2)\frac{PR}{\beta^2P+R} \tag{5}$$

特别地，当 $\beta=1$ 时，认为 PR 值同等重要，称 $F1$ 值；有些情况下，如果认为 P 值更重要，就调整 β 值小于 1；反之，若认为 R 值比较重要，则调整 β 值大于 1。

虽然 *F-score* 给了更大的调节空间，一方面很难根据实际场景量化 β 值，另一方面仍然无法给出单个类别的分类评估结果。

2.4 *Kappa* 系数

Kappa 系数是统计学中的概念，一般用于一致性检验，也可以用来作为衡量分类精度的指标。其计算方法如式(6)所示：

$$Kappa=\frac{P_o-P_e}{1-P_e} \tag{6}$$

其中，P_o 代表总体分类精度(即 2.1 节中的准确率)，P_e 计算方法如式(7)所示：

$$P_e=\frac{\sum_{i=1}^{3}row_i col_i}{w^2} \tag{7}$$

其中，row_i 和 col_i 分别代表第 i 个类别的真实样本个数和分类预测的样本个数，具体为：$row_1=l$，$row_2=m$，$row_3=n$；$col_1=r$，$col_2=s$，$col_3=t$。一般情况下，根据 *Kappa* 系数大小进行如表 3 所示的一致性等级划分。

Table 3 Consistency level of *Kappa* coefficient

表 3 *Kappa* 系数一致性等级划分

系 数	一致性等级
0～0.2	极低
0.2～0.4	一般
0.4～0.6	中等
0.6～0.8	高度
0.8～1.0	几乎完全一致

同样，无法避免的是上述 *Kappa* 系数仍然不能给出单个类别分类结果的准确率评估。同时，这种等级划分的适用范围有限，等级划分缺乏一定的合理性，不能适应应用场景的变化迁移和满足用户特定的具体需求。

据作者所知，这方面的工作很少。然而，在一些特定的应用场景中，文献[10－13]进行了一些相关的工作。文献[14,15]研究了评估检索系统的问题，并定义了一些类似于 AP(Average Precision)的指标。文献[16－18]通过数学分析和一些特定实验比较了 AP 和其他一些指标。文献[19,20]提出了一些改进措施，以克服平均精度(mAP)的缺陷。文献[21－23]探究了在其他一些领域改变评价指标的可能性。但是，上述所有工作都只是试图调整或采用 AP 指数以在某些特定的应用场景中获得更好的性能[24-28]。他们很少关注怎样去克服 AP 及类似指标的固有缺点，且应用场景受限[29-32]。

3 *R′*方法介绍

R 方法是由许绍燮院士在 1973 年提出的，最初运用于地震预测的准确率评估，后来(1989 年)给出了更严格的理论推导和证明，并由王晓青研究员等人(1999 年，2002 年)进行了进一步的改进和推广[7]。Dou 等人[9](2004 年)将其引入遥感图像

分类效果评估,给出了理论推导,并进行了适当改进,称之为 R' 方法。基于上述原理,这里给出应用于多分类任务场景的评估方法,并仍称之为 R' 方法。

3.1 方法定义

不失一般性,仍以表 2 中的二分类问题为例,先给出 R' 方法的一般原理,然后进行多分类任务的拓展和推广。

以类别 1 为例,该类别的分类效率 $R(1)$ 定义如下:对该类别进行正确分类的概率与样本被预测为这个类别的概率之差,如式(8)所示:

$$R(m \mid s) = P(s \mid m) - P(s) \tag{8}$$

其中,$P(s|m)$ 代表该类别被正确分类的概率,计算方法如下:正确分类的样本数与该类别样本总数之比,如式(9)所示:

$$P(s \mid m) = \frac{a}{m} \tag{9}$$

$P(s)$ 代表样本被预测为该类别的概率,如式(10)所示:

$$P(s) = \frac{s}{w} \tag{10}$$

同样,$P(m)$ 代表这一类别在总样本中的出现概率,如式(11)所示:

$$P(m) = \frac{m}{w} \tag{11}$$

综上,可得:

$$R(m \mid s) = P(s \mid m) - P(s) = \frac{a}{m} - \frac{s}{w} \tag{12}$$

进而有:

$$R(m \mid s) + P(m) = P(s \mid m) - P(s) + P(m) = \frac{a}{m} - \frac{s}{w} + \frac{m}{w} \tag{13}$$

根据实际的分类结果,考虑以下 3 种可能出现的情况:

(1)该类别预测样本数小于该类别实际的样本数,即 $a \leqslant s < m$ 时:

$$R(m \mid s) + P(m) = \frac{a}{m} - \frac{s}{w} + \frac{m}{w} \leqslant \frac{s}{m} - \frac{s}{w} + \frac{m}{w} = \frac{s(w-m)}{mw} + \frac{m}{w} \leqslant \frac{m(w-m)}{mw} + \frac{m}{w} = 1 \tag{14}$$

(2)该类别预测样本数大于该类别实际的样本数,即 $a \leqslant m < s$ 时:

$$R(m \mid s) + P(m) = \frac{a}{m} - \frac{s}{w} + \frac{m}{w} \leqslant \frac{a}{m} - \frac{s}{w} + \frac{s}{w} = \frac{a}{m} \leqslant 1 \tag{15}$$

(3)分类结果完全正确,即 $a = m = s$ 时:

$$R(m \mid s) + P(m) = 1 \tag{16}$$

根据 $R(1)$ 值的定义,可得 $R(m|s) + P(s) = P(s|m) \geqslant 0$,即 $R(m|s) \geqslant -P(s)$。所以有:

$$-P(s) \leqslant R(m \mid s) \leqslant 1 - P(m) \tag{17}$$

也就是说,$R(1) \in [-P(s), 1-P(m)]$。它越接近于 $1-P(m)$,表明分类效果越好。为方便评估,本文进行以下改进,并定义为 $R'(1)$:

$$R'(m \mid s) = R(m \mid s) + P(m) = P(s \mid m) - P(s) + P(m) = P(s \mid m) - [P(s) - P(m)] \tag{18}$$

这样,$R'(1) \in [P(m)-P(s), 1]$。$R'(1)$ 值越接近于 1,分类效果越好。

3.2 R' 方法在多分类任务下的推广

对于多分类(假设类别数为 n)问题,显然不止一个类别需要预测。为此,对上述推理进行以下推广。

设 x 表示总样本中所有类别真实样本的总数,y 代表最终的分类预测结果,x_i 代表第 i 个类别的真实样本数量,y_i 代表第 i 个类别的预测样本数量,对于机器学习领域的多分类任务而言,每一个样本都会有一个预测标签,所以有:

$$\sum\nolimits_{i=1}^{n} x_i = x = w, \quad \sum\nolimits_{i=1}^{n} y_i = y = w \tag{19}$$

基于此,第 i 个类别分类正确的概率计算如式(20)所示:

$$P(y_i) = P(y_i \mid y)P(y) \tag{20}$$

其中,$P(y_i|y)$ 代表样本被分为第 i 个类别的条件概率,$P(y)$ 代表样本参与分类的概率(对于本文中的多分类任务场景,该概率实际为 1)。

进而,对所有类别而言,分类结果和真实标签一致的概率如式(21)所示:

$$P(y \mid x) = \sum_{i=1}^{n} P(y_i \mid x_i)P(x_i \mid x) \tag{21}$$

其中,$P(y_i|x_i)$ 代表第 i 个类别被正确分类的条件概率。

根据 3.1 节的结论,对于第 i 个类别有:

$$R'(x_i \mid y_i) = P(y_i \mid x_i) - [P(y_i) - P(x_i)] \tag{22}$$

进而对所有类别而言,有:

$$R'(x \mid y) = P(y \mid x) - [P(y) - P(x)] = \sum_{i=1}^{n} P(y_i \mid x_i)P(x_i \mid x) - P(y) + P(x) = \sum_{i=1}^{n} \frac{a_i}{x_i}\frac{x_i}{x} - \frac{y}{x}\frac{x}{w} + \frac{x}{w} = \sum_{i=1}^{n} \frac{a_i}{x} + \frac{x-y}{w} \tag{23}$$

其中,a_i 代表第 i 个类别的样本中被正确预测的样本数量。该值越接近 1,表明总体的分类效果越好。

这样,就可以通过这种方法同时获得分类器整体的分类效果评估值 $R'(x|y)$和单个样本分类效果的评估值 $R'(x_i|y_i)$。在某些应用场景下,用户如果特别关注某一类别的分类效果,可以在保证总体分类效果的前提下,通过调节 $R'(x_i|y_i)$来满足特殊分类需要。

上文给出了在多分类任务场景下的 R'方法。值得注意的是,该方法与 Dou 等人[9]的 R'方法有 2 点不同:(1)应用场景不同。如式(8)描述的那样,多分类任务场景下,该指标评估每个类别被正确分类的概率,并以样本数作为统计标准。与之不同的是,Dou 等人的方法以遥感图像像元的多少表征目标识别概率的高低。(2)适用条件不同。遥感图像识别往往包含像元的错漏现象,也就是某些像元不属于任何一个目标。而在一般的多分类任务场景下,正如式(20)中描述的那样,样本参与分类的概率 $P(y)=1$,也就是不存在样本不被归类的情况。

4 实验及结果分析

本文的实验基于 MNIST 手写字符体识别任务。这是一个 $n=10$ 的多分类问题。采用一种典型人工神经网络(LeNet-5)进行训练和测试,得到在测试样本精度最高的参数设置下的测试样本混淆矩阵,并计算出第2节描述的各评价指标,将在 4.1 节给出,以观察 R'方法的评价效果;同时,基于不同超参数设置,给出不同模型下 R'值对分类器的评估结果,将在 4.2 节给出,以评估 R'方法的鲁棒性;4.3 节通过改变某些样本的容量或者标签,对比在不改变上述容量或者标签的情况下,这些类别的值的变化,以此来进一步验证此方法对于单个类别的评估效果;4.4 节则将上述实验迁移到 CIFAR-10 数据集(对应的神经网络模型采用 VGG)并试图从另一个角度说明 R'方法的有效性。

4.1 不同评价指标的对比

实验中,测试样本最终在模型(最终测试准确率为 98.06%)下得出如表 4 所示的混淆矩阵(表中行表示实际标签,列表示预测标签;表中同时给出了每个类别的 R'值)。基于混淆矩阵,计算得出表 5 所示的各个评价指标取值(PR 值项分别给出 P 值和 R 值,用 P/R 表示)。

Table 5 Evaluation indices of test samples classification result

表 5 测试样本分类结果评价指标

评价指标	准确率	PR	$F1$	$Kappa$	R'
值	0.980 6	0.980 4/0.980 7	0.980 5	0.978 4	0.980 6

可以看出,在给定的参数设置下,R'值给出了与现有的评价指标相近的分类器评估取值。值得一提的是,表 4 说明了 R'值可以同时给出整体预测结果的评估指标以及单个类别的评估指标,这是其他指标无法做到的。为进一步说明 R'值的上述特性,图 1 给出了 10 个类别在不同指标体系下评估结果的雷达图(对于 R'值以外的评估指标,由于它们只给出了整体的分类效果评估值,这里对所有类别赋予同样的该评估值)。

Table 4 Confusion matrix of the test samples

表 4 测试样本分类结果混淆矩阵

	0	1	2	3	4	5	6	7	8	9
0	966	1	0	2	0	1	4	3	3	0
1	0	1 125	3	1	0	0	2	1	3	0
2	1	0	1 016	4	0	0	1	3	7	0
3	1	0	5	987	0	4	0	6	4	3
4	1	0	3	0	945	0	2	3	2	26
5	2	0	0	7	1	871	4	2	3	2
6	4	3	2	1	1	3	940	0	4	0
7	2	1	4	0	0	0	0	1 011	3	7
8	0	1	3	5	0	4	1	2	954	4
9	0	3	2	2	5	0	2	2	2	991
R'_i	0.986 0	0.991 3	0.983 9	0.977 3	0.965 3	0.977 4	0.981 4	0.983 0	0.978 4	0.979 8

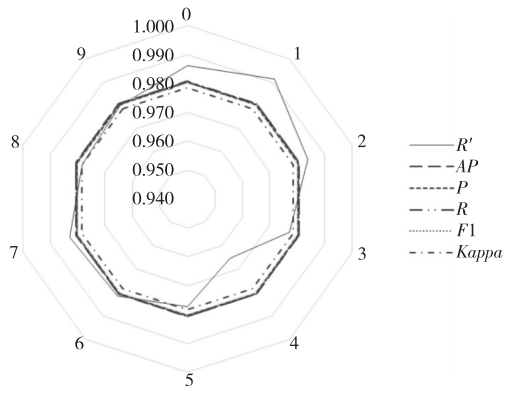

Figure 1 Appraised values for different categories of classification results under each indicator

图 1 各个指标下不同类别分类结果的评估值

同样可以看出,除了 R' 值以外,其他的评估指标雷达图均为正十边形(每个类别具有相同的全局评估值)。而对 R' 值而言,可以清楚地看出,实验结果对数字 0,1,2,7 识别率较高,对数字 4 识别率最差(数字 3,5,6,8,9 则介于两者之间)。这给某些场景下的特殊应用需求提供了直观、便利的评估结果和模型选择方法。

4.2 不同分类结果下 R' 值的对比

为进一步验证 R' 值的鲁棒性(在不同参数设置下,R' 值对不同模型的评价结果有无差异),本节进行了不同超参数设置(实际是不同学习率)下的 10 组实验,并对比其分类结果的评估值,如表 6(作为参考,同时给出了其他指标的评估 R' 值;或者更直观地将值绘制为图 2 的形式)所示。

可以看出,对于不同超参数设置下的分类结果,R' 值给出了不同的评估结果。R' 值根据不同模型的好坏,给出了其实际效果的评估结果,这说明了 R' 方法的鲁棒性。

4.3 R' 值对单个类别的评估效果

本节的实验采取改变训练样本标签的方法,以此来控制样本容量变化。具体而言,又分为以下 2 个步骤:

首先分别去除类别 0 和类别 6 的某些样本,减少类别 0 和类别 6 样本的容量,并通过 R' 方法来评估分类效果,称之为改变前;然后恢复这些训练样本的原始标签,同样通过 R' 方法来评估分类效果,称之为改变后。

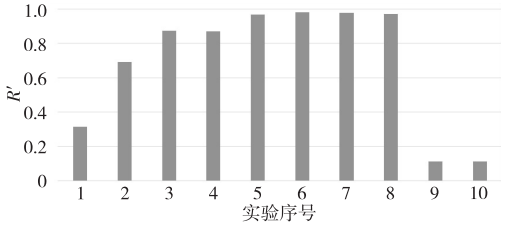

Figure 2 Classifier values under different hyper-parameter settings

图 2 不同超参数设置下分类器值

改变前后保持模型的其他参数不变。

表 7 给出了样本标签改变前后各个类别以及整体的 R' 值。

Table 7 R' value of each category before and after changing the sample label

表 7 改变样本标签前后各个类别的 R' 值

类别	改变前	改变后
0	**0.117 6**	**0.989 2**
1	0.991 2	0.989 0
2	0.977 1	0.972 1
3	0.982 3	0.983 0
4	**0.873 9**	**0.974 4**
5	0.964 7	0.969 0
6	**0.388 1**	**0.972 3**
7	0.981 7	0.970 4
8	**0.886 6**	**0.978 7**
9	0.972 6	0.980 7
整体	0.818 7	0.978 1

可以看出,在恢复类别 0 和类别 6 的样本容量之前(也就是改变前),他们的 R' 值很小(分别为 0.117 6和 0.388 1,如表 7 中加粗部分所示),对应

Table 6 Evaluation values of the classifier under different hyper-parameter settings

表 6 不同超参数设置下分类器评估值

评价指标	实验序号									
	1	2	3	4	5	6	7	8	9	10
R'	0.316 5	0.689 8	0.874 1	0.868 9	0.966 2	0.980 2	0.977 0	0.970 4	0.113 5	0.113 5
AP	0.302 7	0.683 2	0.873 0	0.868 0	0.965 9	0.980 0	0.977 0	0.970 2	0.100 0	0.100 0
P	0.273 1	0.800 7	0.886 0	0.897 0	0.965 8	0.980 2	0.976 8	0.970 7	0.011 4	0.011 4
R'	0.287 1	0.737 3	0.879 5	0.882 3	0.965 9	0.980 1	0.976 9	0.970 5	0.020 4	0.020 4
$Kappa$	0.238 4	0.655 1	0.860 1	0.854 3	0.962 4	0.9780	0.974 4	0.967 1	0.000 0	0.000 0

的类别4和类别8的R'值也得到一定的影响(分别为0.873 9和0.886 6,如表7中加粗部分所示)。恢复样本原始容量之后(也就是改变后),类别0和类别6对应的R'值得到显著提升(分别为0.989 2和0.972 3,如表7中加粗部分所示),对应的类别4和类别8的R'值也得到一定的提升(分别为0.974 4和0.978 7,如表中加粗部分所示)。值得说明的是,这对于优化和改进训练过程具有显著的指导意义,即可以通过观察单一类别或者某一些类别R'值的变化,采取必要的手段(如样本均衡)来改进训练过程。

回到3.2节的关于R'值方法推广。3.2节中给出了某一单个类别的R'值计算方法,如式(22)所示。

考察式(22),R'值方法在评估分类效果的时候,除了考虑在真实标签中样本被正确预测的概率$P(s_i|m_i)$之外,还进一步结合了样本被正确预测和错误预测的差异,即$P(s_i)-P(m_i)$。对于实验中因改变样本标签而导致样本不均衡的情形,这一差异被R'方法很好地提取了出来。

具体而言,考察表8和表9所示的训练样本容量改变前后的测试样本的混淆矩阵。表格中的行表示测试样本真实标签在2次实验中未发生变化,而表示预测标签的每一列则发生了一定的变化(尤其对类别0、类别4、类别6和类别8而言,如表7中加粗部分所示)。这解释了上述实验中这些类别值变化的原因。进一步说,R'方法可以很好地发现和指导解决训练过程中因样本不均衡等原因导致的分类效果评估的差异问题,进而指导和改进训练过程。

4.4 CIFAR-10数据集实验结果

为进一步说明R'方法的有效性和适用性,本节实验采用另一个多分类任务场景的经典数据集CIFAR-10进行验证。

CIFAR-10数据集是一个更接近普适物体的彩色图像数据集,一共包含10个类别的RGB彩色图像:飞机(airplane)、汽车(automobile)、鸟类(bird)、猫(cat)、鹿(deer)、狗(dog)、蛙类(frog)、

Table 8 Confusion matrix 1 before sample label changes

表8 改变样本标签前的混淆矩阵1

	0	1	2	3	4	5	6	7	8	9	sum
0	22	1	1	0	949	2	1	1	3	0	980
1	0	1 125	0	5	1	1	2	1	0	0	1 135
2	0	3	1 008	8	4	0	0	6	3	0	1 032
3	0	0	4	997	0	3	0	4	0	2	1 010
4	0	0	5	1	954	0	2	6	1	13	982
5	1	0	0	18	3	859	2	2	6	1	892
6	0	2	1	2	17	5	311	0	620	0	958
7	0	0	5	4	4	0	0	1 010	0	5	1 028
8	5	1	4	18	13	2	5	3	920	3	974
9	0	3	0	5	13	3	0	3	1	981	1 009
sum	28	1 135	1 028	1 058	1 958	875	323	1 036	1 554	1 005	10 000

Table 9 Confusion matrix 1 after sample label changes

表9 改变样本标签后的混淆矩阵1

	0	1	2	3	4	5	6	7	8	9	sum
0	971	1	0	0	1	0	3	1	3	0	980
1	0	1 122	3	1	0	1	3	1	4	0	1 135
2	6	1	1 003	3	3	0	2	7	7	0	1 032
3	0	0	4	994	1	4	0	3	2	2	1 010
4	0	0	8	0	956	0	1	1	1	15	982
5	4	0	0	11	0	863	3	1	9	1	892
6	7	2	2	1	5	7	930	0	4	0	958
7	2	1	8	4	2	0	0	996	4	11	1 028
8	3	1	2	4	1	1	1	2	955	4	974
9	3	2	0	4	4	1	0	1	3	991	1 009
sum	996	1 130	1 030	1 022	973	877	943	1 013	992	1 024	10 000

马(horse)、船(ship)和卡车(truck)。数据集中每幅图像的尺寸为 32 × 32,每个类别有 6 000 幅图像,数据集中一共有 50 000 幅训练图像和 10 000 幅测试图像。与 MNIST 的灰度图像不同,CIFAR-10 数据集由 3 通道 RGB 彩色图像组成,图像尺寸也比 MNIST 的 28 × 28 更大。此外,数据集是现实世界的真实物体,图像噪声更大,物体的比例、特征也都不尽相同,识别难度更大。但是,值得注意的是,CIFAR-10 数据集样本更加均衡,每个类别的样本数量都是 6 000,这对于进一步验证 4.3 节实验的设计思路更加方便和有效。

同样采用 4.3 节的实验设计方法,通过改变测试样本的标签来模拟样本不均衡的现象(这里将 cat 类别部分样本去除,将 deer 类别部分样本去除)。表10和表11分别给出了对应的混淆矩阵(表中同时给出了各个类别和整体上分类效果的评估 R'值,表中最后一列的 all 代表整体分类效果的 R'值)。

从表 10 和表 11 中可以看出,在恢复类别 cat 和类别 deer 的样本容量之前(也就是改变前),它们的 R'值很低(分别为 0.1 和 0.1,如表 10 中加粗部分所示),对应的类别 dog 和类别 horse 的值也受到一定的影响(分别为 0.808 3 和 0.811 7,如表 11 中加粗部分所示)。恢复原始标签之后(也就是改变后),类别 cat 和类别 deer 对应的 R'值得到显著提升(分别为 0.761 4 和 0.884 1,如表 11 中加粗部分所示),对应的类别 dog 和类别 horse 的 R'值也得到一定的提升(分别为 0.821 3 和 0.896 9,如表 11 中加粗部分所示),整体的分类效果评估指标也从 0.718 9 提高到 0.873 0。

Table 10 Confusion matrix 2 before sample label changes

表 10 改变样本标签前的混淆矩阵 2

	air-plane	auto-mobile	bird	cat	deer	dog	frog	horse	ship	truck	sum /all
air-plane	877	15	19	0	0	19	4	14	38	14	1 000
auto-mobile	11	939	1	0	0	6	5	4	8	26	1 000
bird	28	0	818	0	0	67	22	54	9	2	1 000
cat	9	2	24	0	0	871	24	53	8	9	1 000
deer	4	1	23	0	0	54	6	907	3	2	1 000
dog	7	1	15	0	0	911	6	58	1	1	1 000
frog	5	2	25	0	0	38	900	28	1	1	1 000
horse	5	1	12	0	0	55	5	917	2	3	1 000
ship	35	12	8	0	0	3	4	8	910	20	1 000
truck	14	43	1	0	0	3	1	10	11	917	1 000
sum	995	1 016	946	**0**	**0**	**2 027**	**977**	**2 053**	991	995	10 000
R'_i	0.877 5	0.937 4	0.823 4	**0.1**	**0.1**	**0.808 3**	0.902 3	**0.811 7**	0.910 9	0.917 5	0.718 9

Table 11 Confusion matrix 2 after sample label changes

表 11 改变样本标签后的混淆矩阵 2

	air-plane	auto-mobile	bird	cat	deer	dog	frog	horse	ship	truck	sum /all
air-plane	875	13	22	12	7	8	9	6	34	14	1 000
auto-mobile	7	937	2	4	2	1	2	2	8	35	1 000
bird	26	0	830	32	30	24	32	15	4	7	1 000
cat	9	4	38	758	34	113	13	16	3	12	1 000
deer	4	1	23	21	888	25	13	22	1	2	1 000
dog	3	2	16	89	30	826	5	27	0	2	1 000
frog	4	2	20	27	19	11	911	3	1	2	1 000
horse	6	1	15	12	27	34	3	896	2	4	1 000
ship	44	22	6	7	1	2	0	2	895	21	1 000
truck	19	40	1	4	1	3	3	2	13	914	1 000
sum	997	1 022	973	**966**	**1 039**	**1 047**	991	**991**	961	1013	10 000
R'_i	0.875 3	0.934 8	0.832 7	**0.761 4**	**0.884 1**	**0.821 3**	0.911 9	**0.896 9**	0.898 9	0.912 7	0.873 0

上述实验说明了 R' 方法对于 CIFAR-10 数据集的适用性和有效性,进一步说明了 R' 方法的可拓展性及其应用场景。

此外,结合 4.3 节和 4.4 节的实验结果,也就是样本容量发生变化前后评估指标的对比,可以看出该方法对不平衡数据集同样适用。也就是说,它不会因为样本数量的不均衡而影响对分类结果的评价,因为正如 3.2 节所强调的那样,该方法可以单独对每一个类别进行评估而不仅仅是对整体分类效果进行评估。在不平衡数据集上,即使整体的分类效果较好,对于样本数较少的类别而言,无论它的分类效果如何,它的评价指标都会被单独地呈现出来。这一点正是该方法的一个突出特点。

5 结束语

多分类任务模型准确率评估一直是一个值得讨论的问题,这不仅要涉及到模型选择问题,也对模型训练过程具有很好的指导意义。本文针对多分类任务场景下,尤其是用户关心特定类别分类效果的实际情况,现有的多分类任务准确率评价指标的不足,介绍和引入了用于评估模型分类准确率的 R' 方法。该方法具有严格的数学理论推导过程,不仅可以评估分类器整体的分类效果,而且还可以给出每一个类别的分类效果,不仅可以用于模型选择,而且对于更好地指导训练过程具有一定的意义。通过与已有评价方法的对比,基于 MNIST 的手写字符体识别任务和 CIFAR-10 数据集的多分类任务的实验验证,表明该方法具有很好的鲁棒性和有效性,可以用于多分类任务的分类准确率评估场景。同时值得一提的是,不仅对文中实验验证采用的 MNIST 手写字符体识别和 CIFAR-10 数据集分类这 2 个多分类任务,该方法还可以扩展到任意场景下的多分类任务问题,具有广泛的应用前景。

参考文献:

[1] Bishop C M. Pattern recognition and machine learning[M]. New York:Springer,2006.

[2] Mohri M,Rostamizadeh A,Talwalkar A. Foundations of machine learning[M]. Cambridge:The MIT Press,2012.

[3] Ben-David S, Shalev-Shwartz S. Understanding machine learning:From theory to algorithms[M]. New York:Cambridge University Press,2014.

[4] Candela J Q,Dagan I,Magnini B,et al. Machine learning challenges:Evaluating predictive uncertainty,visual object classification, and recognizing textual entailment[M]. Berlin:Springer Heidelberg,2006.

[5] Dou Yong,Qiao Peng,Jin Ruo-chun. Exploring the defects of the average precision and its influence[J]. SCIENTIA SINICA Informationis,2019,49(10):1369-1382. (in Chinese)

[6] Allouche O,Tsoar A,Kadmon R. Assessing the accuracy of species distribution models:Prevalence,kappa and the true skill statistic(tss)[J]. Journal of Applied Ecology,2006,43(6):1223-1232.

[7] Wang X Q. Problem and improvement of R-values applied to assessment of earthquake forecast[J]. China Earthquake Research,2001,16(3):75-83.

[8] Bruzzone L,Persello C. A novel protocol for accuracy assessment in classification of very high resolution multispectral and SAR images[C]// Proc of 2018 IEEE International Geoscience and Remote Sensing Symposium,2008:265-268.

[9] Dou A X,Wang X Q,Dou M W. A new approach to evaluate the accuracy of image classification result R'[C]// Proc of 2004 IEEE International Geoscience and Remote Sensing Symposium,2004:3033-3035.

[10] Chen S, Fern A,Todorovic S. Person count localization in videos from noisy foreground and detections[C]// Proc of 2015 IEEE Conference on Computer Vision and Pattern Recognition,2015:1364-1372.

[11] Gajda J,Sroka R,Stencel M,et al. Design and accuracy assessment of the multi-sensor weigh-in-motion system[C]// Proc of 2015 IEEE International Instrumentation and Measurement Technology Conference,2015:1036-1041.

[12] Li W K, Guo Q H. A new accuracy assessment method for one-class remote sensing classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2014,52(8):4621-4632.

[13] Jean-Marie F, Kenji O, Joachim E, et al. Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon:The ecological side of speciation[J]. BMC Ecology,2009,9:Article number 17.

[14] Yilmaz E, Aslam J A. Estimating average precision when judgments are incomplete[J]. Knowledge and Information Systems,2008,16:173-211.

[15] Piwowarski B,Dupret G,Lalmas M. Beyond cumulated gain and average precision:Including willingness and expectation in the user model[J]. arXiv:1209.4479,2012.

[16] Yan Y, Su W H,Zhu M. Threshold-free measures for assessing the performance of medical screening tests[J]. Frontiers in Public Health,2015,3:Article 57.

[17] Bestgen Y. Exact expected average precision of the random baseline for system evaluation[J]. Prague Bulletin of Mathematical Linguistics,2015,103:131-138.

[18] Henderson P,Ferrari V. End-to-end training of object class detectors for mean average precision[C]// Proc of Asian Conference on Computer Vision,2016:198-213.

[19] Ding P L K, Li Y K, Li B X. Mean local group average precision(mL-GAP):A new performance metric for hashing-

based retrieval[J]. arXiv:1811.09763,2018.

[20] He K, Lu Y, Sclaroff S. Local descriptors optimized for average precision[C]// Proc of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018:596-605.

[21] Jerome R, Jon A, de Rezende R S, et al. Learning with average precision: Training image retrieval with a listwise loss [C]// Proc of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:5106-5115.

[22] Andric K, Kalpic D, Bohacek Z. An insight into the effects of class imbalance and sampling on classification accuracy in credit risk assessment[J]. Computer Science and Information Systems,2019,16: Article 37.

[23] Mao H Z, Yang X D, Dally W J. A delay metric for video object detection: What average precision fails to tell[J]. arXiv:1908.06368,2019.

[24] Erener A. Classification method, spectral diversity, band combination and accuracy assessment evaluation for urban feature detection[J]. International Journal of Applied Earth Observation and Geoinformation,2013,21:397-408.

[25] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification[J]. Pattern Recognition Letters,2009,30(1):27-38.

[26] Mosley L. A balanced approach to the multi-class imbalance problem[D]. Iowa: Iowa State University,2013.

[27] Persello C, Bruzzone L. A novel protocol for accuracy assessment in classification of very high resolution images[J]. Geoscience and Remote Sensing,2010,48(3-1):1232-1244.

[28] Sharma R, Goyal A K, Dwivedi R K. A review of soft classification approaches on satellite image and accuracy assessment[C]// Proc of the 5th International Conference on Soft Computing for Problem Solving,2015:629-639.

[29] Qin Feng, Huang Jun, Cheng Ze-Kai, et al. A study on accuracy evaluation method for multi-label classifier[J]. Computer Technology and Development,2010,20(1):46-49. (in Chinese)

[30] Wu Ya-kun, Duan Fu, Yin Xue-mei. Research on accuracy evaluation of classifier[J]. Computer Development & Applications,2011,24(4):10-12. (in Chinese)

[31] Jiang Shuai. Researches on performance evalution of classifier based on AUC[D]. Changchun: Jilin University,2016. (in Chinese)

[32] Yang Bo, Qin Feng, Cheng Ze-kai. A new measure in classification learning system[C]// Proc of 2005 Digital Anhui Doctoral Science and Technology Forum,2015:525-530. (in Chinese)

附中文参考文献:

[5] 窦勇,乔鹏,靳若春. 探究平均准确度 AP 指标的缺陷及其影响[J]. 中国科学:信息科学,2019,49(10):1369-1382.

[29] 秦锋,黄俊,程泽凯,等. 多标签分类器准确性评估方法的研究[J]. 计算机技术与发展,2010,20(1):46-49.

[30] 武亚昆,段富,尹雪梅. 分类器准确率评估的研究[J]. 电脑开发与应用,2011,24(4):10-12.

[31] 蒋帅. 基于 AUC 的分类器性能评估问题研究[D]. 长春:吉林大学,2016.

[32] 杨波,秦锋,程泽凯. 一种新的分类学习系统评估度量[C]// 2005 年"数字安徽"博士科技论坛论文集,2005:525-530.

作者简介:

张开放(1993 -),男,河南开封人,硕士生,研究方向为机器学习。**E-mail**: zhangkaifang18@nudt.edu.cn

ZHANG Kai-fang, born in 1993, MS candidate, his research interest includes machine learning.

苏华友(1985 -),男,广西桂林人,博士,助理研究员,研究方向为人工智能。**E-mail**: shyou@nudt.edu.cn

SU Hua-you, born in 1985, PhD, assistant research fellow, his research interest includes artificial intelligence.

窦勇(1966 -),男,吉林吉林人,博士,研究员,研究方向为人工智能。**E-mail**: yongdou@nudt.edu.cn

DOU Yong, born in 1966, PhD, research fellow, his research interest includes artificial intelligence.