Applied Intelligence
https://doi.org/10.1007/s10489-021-02223-7

Beyond AP: a new evaluation index for multiclass classification task accuracy

Kaifang Zhang[1] · **Huayou Su**[1] · **Yong Dou**[1]

Accepted: 20 January 2021

Abstract
Average precision *(AP)* and many other related evaluation indices have been employed ubiquitously in classification tasks for a long time. However, they have defects and can hardly provide both overall evaluations and individual evaluations. In practice, we have to strike a balance between whole and individual performances to satisfy diverse demands. To this end, we propose a new index for multiclass classification tasks, named R', which is an unbiased estimator of *AP*. Specifically, we improve the R index by taking the numerical differences between the real labels and predicted labels of each class into consideration. We evaluate its effectiveness and robustness on the *MNIST* and *CIFAR-10* datasets. Experimental results show that it is positively correlated with some related indices. More importantly, we can obtain both overall and individual evaluations, which can be beneficial for improving training processes and model selection. Furthermore, as an evaluation architecture, the index can be promoted to evaluate any classification task, thereby implying broad application prospects.

Keywords Machine learning · Multiclassification · Evaluation index · R' method

1 Introduction

It is widely known that the average precision *(AP)* [19] metric for classification tasks has defects [1]. Over the past few decades, many other evaluation indices have been proposed to improve or overcome the inherent drawbacks of *AP* [2–4, 6, 17, 18]. Some of these evaluation indices, such as the *PR* curve, *kappa* coefficient, and *F-1* score, are based on the confusion matrix to evaluate overall classification effects. However, they can hardly provide a classification effect evaluation for each category, so they are insufficient for meeting the demands of some practical applications (for example, in the *MNIST* handwritten character recognition task, the importance of the character *0* is usually higher than the importance of others).

✉ Huayou Su
shyou@nudt.edu.cn

Kaifang Zhang
zhangkaifang18@nudt.edu.cn

Yong Dou
yongdou@nudt.edu.cn

[1] College of Computer, National University of Defense Technology, Changsha, China

In the field of machine learning (*ML*), multiclass classification tasks [20, 21] imply that we have more than 2 classes to predict, and binary classification [22] tasks contain exactly 2 classes. Due to the limitations of classification algorithms and models, this is a problem that must be faced to evaluate the accuracy of classification results. On the other hand, due to classifier overfitting, it is vital to carefully and adequately choose accuracy evaluation indices.

To this end, this paper proposes the R' index, which is based on the R index [15, 16], as an alternative multiclass classification task evaluation scheme. We improve the R index and illustrate the theoretical derivation of R'. Experiments on multiclass classification tasks based on the *MNIST* and *CIFAR-10* datasets show that the proposed index can better evaluate the accuracy of classification results than some other existing indices. Moreover, the index is simple to calculate and can determine both overall and individual evaluations, which are meaningful for specific contexts where we emphasize the individual accuracies of certain classes. In other words, it has guiding significance for evaluating and improving the training process of a model.

Our main contributions can be stated as follows:

– We propose a new evaluation index for multiclass classification tasks that takes numerical differences between the real labels and predicted labels of each class into

Published online: 26 February 2021

consideration to effectively reflect the classification effects.
- We implement the theoretical derivation of the index as well as the mathematical proof of its property.
- We obtain overall and individual classification evaluations at the same time through the index based on the confusion matrix to improve the training process for a given model.
- We evaluate the index on the *MNIST* and *CIFAR-10* datasets in diverse application contexts to verify its effectiveness and robustness, as well as its broad application prospects.

2 Background and related work

This section first introduces several common model accuracy evaluation indices and then analyses their deficiencies in brief. We try to explain the motivation for the new index. Some related work is also provided.

2.1 Background

This subsection mainly introduces several common evaluation indices and their deficiencies. Without loss of generality, we consider the confusion matrix for the binary classification tasks shown in Table 1. The rows represent the real labels, while the columns represent the predicted labels.

Average precision Average precision [31, 32, 34–37] is the most primitive evaluation index for classification tasks, and it is defined as the percentage of correctly predicted samples out of the total samples. For the confusion matrix in Table 1, we have:

$$AP = (a+b)/w \tag{1}$$

PR curve The *PR* curve is a curve describing the relationship between P and R. P represents the accuracy rate (*Precision*), which is based on the classification results. R indicates the recall rate (*Recall*), which is related to the real labels. The *PR* is calculated as:

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

For multiclassification problems, we actually obtain multiple sets of confusion matrices and multiple sets of *PR* values.

Table 1 Confusion matrix for binary classification tasks

	Class 1	Class 2	Total
Class 1	a	c	m
Class 2	d	b	n
Total	s	t	w

At this time, there are two processing methods: macro averaging and micro averaging [33]. The macro averaging method first calculates the *PR* value of each confusion matrix and then averages them separately; micro averaging calculates the average numbers of positive and negative samples in the global confusion matrix and then calculates the overall *PR* value.

In this way, for ternary classification problems, the macro averaging method is formulated as:

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

where P_i and R_i represent the accuracy rate and recall rate of category i, respectively.

Using the micro averaging method (for multiclass tasks without omissions and missed detections, it is actually the accuracy metric in Section 2.1), we have:

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

Although macro averaging adds additional off-diagonal elements, it still only provides the overall classification effect of all categories altogether, while the micro average is equivalent to the accuracy in Section 2.1. At the same time, the PR value is a pair of statistics that eliminates each other's strengths. In practical applications, it is necessary to balance the two.

F-1 index The *F-measure* (or *F-score*) [31, 32, 36] is a coordinate between P and R, i.e.:

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

where $\beta \geq 0$ represents the balance between P and R; when $\beta = 1$, we call it the *F-1* score.

Kappa coefficient The *kappa* coefficient [34, 35] is a concept in statistics. It is generally used for consistency tests and can also be used as an index to measure classification effects. It is calculated as:

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

where P_o represents the overall classification accuracy (i.e., AP mentioned above), and the calculation formula for P_e is:

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

where row_i and col_i represent the number of real samples and the number of samples predicted by the classifier in the *i-th* category, respectively. In general, according to the values of the *kappa* coefficient, the consistency level is divided as shown in Table 2.

Table 2 Rank of *kappa* coefficient consistency

Coefficient	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1
Consistency	*slight*	*fair*	*moderate*	*substantial*	*perfect*

ROC and AUC The *ROC*(receiver operating characteristic) curve [29–31] is another curve commonly utilized to evaluate the performance of a classifier. The *ROC* curve is based on statistics and originates from the observation of electronic signals and the evaluation of medical diagnostic capabilities. With the continuous expansion of its applications and the deepening of research, the *ROC* curve can be applied in the fields of biology, engineering, medical imaging, parallel analysis, and machine intelligence.

The *AUC* (area under the curve) is the size of the area under the *ROC* curve. The larger the area under the *ROC* curve is, the better the model performance is. Generally, the *AUC* value is between 0.5 and 1.0, and a larger *AUC* represents a better performance. If the model is perfect, then its *AUC* equals 1, which proves that all positive examples are in front of negative examples. If the model is a simple two-class random model, then its *AUC* equals 0.5.

The *ROC* curve has a very good characteristic: when the distribution of the positive and negative samples in the test set changes, the *ROC* curve can remain unchanged. Class imbalance often occurs in actual datasets, i.e., negative samples are much more numerous than positive samples (or vice versa), and the distribution of positive and negative samples in the test dataset may also change over time. The *ROC* curve and *AUC* can eliminate the impact of sample category imbalance on the index results.

CLL The conditional log likelihood (CLL) [29, 30] is another index utilized to evaluate the quality of a classifier for class probability estimation. For a given classifier C and a set of examples $E = \{E_1, E_2, \cdots, E_t\}$, where $E_i = (a_{i_1}, a_{i_2}, \cdots, a_{i_n}, c_i)$, t is the total number of samples, n is the number of attributes, and c_i is the class label of E_i, the conditional log likelihood $CLL(C|E)$ of a classifier C on the samples E can be defined as:

$$CLL(E|D) = \sum_{i=1}^{t} log P_C(c_i|a_{i_1}, a_{i_2}, \cdots, a_{i_n}) \tag{8}$$

Classifiers with higher CLL values tend to have better class probability estimation performances. The CLL is especially useful for conditions when accurate rankings, not only accurate classifications, are considered [29]. However, it is generally employed in probability-based classifiers (e.g., Bayesian networks), which focus more on class probability estimation performance rather than on classification accuracy or the error rate [30].

To summarize, it can be seen that *AP* does not consider off-diagonal entries. When the number of samples in each category is not balanced, it cannot effectively evaluate classification effects [1, 9, 28].

As for the *PR*, it is frustrating to balance precision and recall in practical applications. The *F-score* and *kappa* coefficient provide us with much more space for adjustment; however, it is challenging to quantify β or the kappa coefficient according to actual scenes. In addition, these indices fail to determine the evaluation results for each category. Furthermore, the *ROC* curve is the same as the abovementioned *PR* curve, by which we mean it is an evaluation index that does not depend on the threshold used. In a classification model whose output is a probability distribution, if only the accuracy rate, precision rate, and recall rate are used when using rates as evaluation indices for model comparison purposes, they must always be based on a given threshold. For different thresholds, the evaluation results of each model will be different, so it is difficult to obtain a relatively reliable evaluation. For the CLL, this index is based on the prediction probability for a Bayesian network, which is slightly different from the confusion matrix.

To this end, we compare the AP, PR, F-1, and $kappa$ indices with our proposed R' in the following experiments to demonstrate the novelty and technical contributions of the R' index.

2.2 Related work

To our knowledge, little work has been done in this regard. However, in some specific application studies [26], some related work was conducted.

Benjamin et al. [10] and Emine et al. [7] studied the problem of evaluating retrieval systems and defined some indices similar to *AP*. Specifically, with incomplete relevance judgments, three related indices to *AP* were constructed accordingly. They defined a new metric family based on the concepts of several stopping criteria and the notion of satisfaction in the literature on information retrieval.

Paul et al. [8], Yan et al. [14], Bestgen et al. [12], and Ferri et al. [27] compared *AP* and other indices through a mathematical analysis as well as some specific experiments. Yan et al. [14] compared the *AP* and the *ROC* curve under the circumstances of disease classification through a mathematical analysis. Bestgen et al. [12] examined the efficient computation of the exact value of *AP* using a procedure they proposed. Their process was also conducted in the contexts of information retrieval and natural language processing research. Paul et al. [8] accomplished the end-to-end training of object class detectors for mean average precision calculation.

Pak et al. [5] and He et al. [6] proposed some improvements to overcome the defects of mean average precision

(*mAP*). Pak et al. [5] proposed a new performance metric for hashing-based retrieval called mean local group average precision (*mLGAP*) to overcome the defects of mean average precision (*mAP*) in the literature on hashing-based retrieval. He et al. [6] improved the extraction procedure for local descriptors based on average precision.

Revaud et al. [13], Kristina et al. [11], and Mao et al. [9] investigated the possibility of altering evaluation metrics in some other research areas. Revaud et al. [13] altered the loss function of image retrieval and, based on *mAP*, optimized the global *mAP* directly to obtain improved results. Kristina et al. [11] discussed the roles of the sample size and class distribution in the area of credit risk assessment based on real-life imbalanced datasets. They analyzed the sensitivities of diverse classification algorithms to class imbalances and sample size changes under the *AP* metric. Mao et al. [9] investigated a delay metric for latency-critical applications, such as video object detection, to capture the temporal natures that *AP* is insufficient for addressing.

However, all the aforementioned works only tried to adjust the *AP* index or adapt to it to obtain improved performances in specific application scenes. They seldom paid attention to overcoming the intrinsic drawbacks of *AP* or other similar indices, and the application contexts are thus limited [23–25].

3 Scheme of the R′ index

Initially applied to the evaluation of earthquake predictions, the *R* index was proposed by the academician Shaoxie Xu in 1973, and later (1989), he developed a more rigorous theoretical derivation and proof. Some further improvements were made by Wang et al. [15]. Dou et al. [16] introduced the index into the evaluation of remote sensing image classification efficiency. We improve it and apply it to multiclass classification tasks, and we name this improved version the R' index.

3.1 Definition of the R′ index

Without loss of generality, we still take the confusion matrix in Table 1 as an example. First, we give the general principles of the R' index, and then we generalize it for multiclass classification tasks. Taking class *1* as an example, the classification efficiency R is defined as the probability of correctly classifying a given category minus the probability that the sample is predicted to be in this category, i.e., :

$$R(m|s) = P(s|m) - P(s) \tag{9}$$

where $P(s|m)$ represents the probability that the category is correctly classified. The calculation involves determining the ratio of the number of correctly classified samples to the total number of samples in this specific category, i.e., :

$$P(s|m) = a/m \tag{10}$$

$P(s)$ represents the probability that the sample is predicted to be in that category, i.e., :

$$P(s) = s/w \tag{11}$$

In summary, we obtain:

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

Similarly, $P(m)$ is the probability that samples of this category appear among the total samples, i.e., :

$$P(m) = m/w \tag{13}$$

Then, we define the R' index of class *1* as:

$$R'(m|s) = R(m|s) + P(m) \tag{14}$$

According to the actual classification results, we consider the following three possible situations:

– the number of predicted samples is smaller than the actual number of samples in this category, i.e., $a \leq s < m$:

$$\begin{aligned} R'(m|s) &= a/m - s/w + m/w \\ &\leq s/m - s/w + m/w \\ &= s(w-m)/mw + m/w \\ &\leq m(w-m)/mw + m/w \\ &= 1; \end{aligned} \tag{15}$$

– the number of predicted samples is larger than the actual number of samples in this category, i.e., $a \leq m < s$:

$$\begin{aligned} R'(m|s) &= a/m - s/w + m/w \\ &\leq a/m - s/w + s/w \\ &= a/m \\ &\leq 1; \end{aligned} \tag{16}$$

– and the classification is completely correct, i.e. $a = s = m$:

$$R'(m|s) = a/m = 1 \tag{17}$$

In summary, we have $R'(m|s) \leq 1$. In addition, according to the definition of R (cf. (9)), we have $R(m|s) + P(s) = P(s|m) \geq 0$, which means $R(m|s) \geq -P(s)$. Then, we can obtain:

$$-P(s) \leq R(m|s) \leq 1 - P(m) \tag{18}$$

In this way, $R'(m|s) \in [P(m) - P(s), 1]$. The closer the R' index is to 1, the better the classification effect is.

3.2 Improvements in the R′ index

Consider a multiclass classification task where the number of categories is n. Suppose that the total number of real

samples in all categories is m, and s represents the total number of predicted samples in all the categories in the final predicted classification results; m_i is the number of real samples in the *i-th* category, and s_i is the number of predicted samples in the *i-th* category. For multiclass classification tasks that we consider in *ML*, every sample has a predicted label, i.e., :

$$\sum_{i=1}^{n} m_i = m = w, \sum_{i=1}^{n} s_i = s = w \tag{19}$$

Based on this, the probability that a sample is predicated to be in the *i-th* category is calculated as follows:

$$P(s_i) = P(s_i|s) \cdot P(s) \tag{20}$$

where $P(s_i|s)$ is the conditional probability that the sample is classified as being in the *i-th* category and $P(s)$ is the probability that the sample participates in the classification task (for classification tasks without omissions, this probability is always 1).

Furthermore, for all categories, the probability that the classification results are consistent with the real labels is:

$$P(s|m) = \sum_{i=1}^{n} [P(s_i|m_i) \cdot P(m_i|m)] \tag{21}$$

In the above formula, $P(s_i|m_i)$ represents the conditional probability that category i is correctly classified and $P(m_i|m)$ is the conditional probability that the samples of category i are in the whole samples. According to the conclusions in Section 3.1, for the *i-th* category:

$$R'(m_i|s_i) = P(s_i|m_i) - [P(s_i) - P(m_i)] \tag{22}$$

As a result, for all the categories, we have:

$$\begin{aligned} R'(m|s) &= P(s|m) - [P(s) - P(m)] \\ &= \sum_{i=1}^{n} [P(s_i|m_i) \cdot P(m_i|m)] - P(s) + P(m) \\ &= \sum_{i=1}^{n} [\frac{a_i}{m_i} \cdot \frac{m_i}{w}] - \frac{s}{m} \cdot \frac{m}{w} + \frac{m}{w} \\ &= \sum_{i=1}^{n} \frac{a_i}{w} + \frac{m-s}{w} \end{aligned} \tag{23}$$

where a_i is the number of correctly predicted samples among the samples belonging to category i.

In this way, we can use this index to simultaneously obtain the evaluation index $R'(m|s)$ for the overall classification effect of the classifier and the evaluation index $R'(m_i|s_i)$ for the classification effect of each class. In some application scenarios, if users pay special attention to the classification effect of a specific category, they can adjust $R'(m_i|s_i)$ to meet their specific classification requirements on the premise of ensuring the overall classification effect.

3.3 Further discussion

As mentioned above, we have defined and promoted R', and now we discuss its relationship with AP.

Property 1 R' is an unbiased estimator of AP.

Proof Let us still take the binary classification task as an example. Note that we have AP or $accuracy=(a+b)/w$ for a binary classification task (cf. (1)). The corresponding R' index is:

$$R'(m|s) = \sum_{i=1}^{n} \frac{a_i}{w} + \frac{m-s}{w} = \frac{a+b}{w} + \frac{m-s}{w} \tag{24}$$

Remember that for a case study with no omissions, we have (19), i.e., $m = s$, which implies that the total number of labels for real samples equals that for predicted samples. This means that each sample must be classified as one of the classes and marked with the corresponding label. Hence, we have:

$$\begin{aligned} E(R') &= E(\frac{a+b}{w} + \frac{m-s}{w}) \\ &= E(\frac{a+b}{w} + \frac{0}{w}) \\ &= E(\frac{a+b}{w}) \\ &= AP \end{aligned} \tag{25}$$

□

4 Experiments and results

The experiments we conduct are based on *MNIST* and *CIFAR-10*, which are both multiclass classification task datasets with *n=10*. In Section 4.1, based on *LeNet-5* and under the parameter settings that yield the highest accuracy for the test set, the confusion matrix of the test set is obtained. The evaluation indices described in Section 2.1 are calculated as comparisons to evaluate the effectiveness of the R' index. In Section 4.2, based on

Table 3 Indices of test set classification results

Index	*AP*	*PR*	*F-1*	*Kappa*	R'
Value	0.9806	0.9804/0.9807	0.9805	0.9784	0.9806

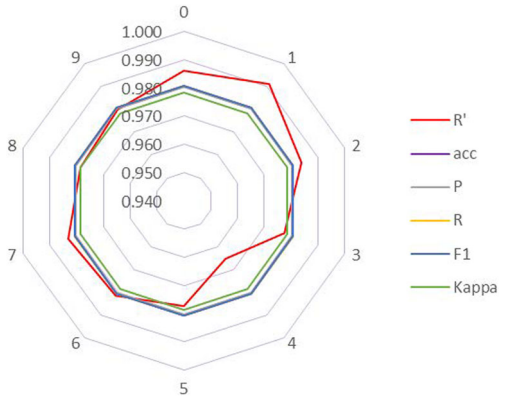

Fig. 1 Indices of the classification results for 10 classes

different hyperparameter settings, the R' indices of the classifiers are given to prove the robustness of R'. In Section 4.3, by changing the volumes of certain samples, we compare the R' indices of these categories with those obtained without changing the volumes to further verify the evaluation effects of the R' index for each category. In Section 4.4, we migrate the above experiments to *CIFAR-10* (the corresponding model is *VGG*) and attempt to illustrate the effectiveness of the R' index from another perspective.

4.1 Comparison between different indices

In the experiments, the test set finally yields the confusion matrix in Table 4 under the specific model utilized (the final test accuracy is *98.06%)* (the R' indices of all categories are also given in the table). Based on the confusion matrix, the value of each evaluation index shown in Table 3 is calculated.

It can be seen that under the given parameter settings, the R' index gives evaluations that are similar to those of the existing evaluation indices. To further illustrate the abovementioned characteristics of the R' index, Fig. 1 shows the radar chart of the evaluation results for the ten categories under different index architectures (for evaluation indices other than the R' index, the same evaluation indices are assigned to all categories since they only provide the overall classification effect evaluation index).

For the R' index, it can be seen that the experimental results have higher recognition rates for characters *0, 1, and 2* and the worst recognition rate for character *4* (the rates for characters *3, 5, 6, 7, 8, and 9* are in between). This provides intuitive and convenient evaluation results and model selection methods for special application requirements in certain scenarios (Table 4).

4.2 R′ indices for different classification results

To further verify the robustness of the R' index, ten sets of experiments with different hyperparameter settings are conducted. The R' indices of the results are shown in Table 5 (for reference, other indices are also given).

We can observe that for the classification results obtained under different hyperparameter settings, the R' index produces different evaluation results, thereby illustrating the robustness of the index. In fact, we can obtain the correlation coefficients (*CC*) of R' and other indices for the 10 models and the root mean square errors (*RMSEs*) in Table 6.

It can be seen that our proposed R' index behaves similarly to the existing indices and can also provide satisfactory estimations of the model performances.

More specifically, "indexes" to "indices" Fig. 2 shows the detailed behaviors of the 4 abovementioned evaluation indices, as well as that of our R' index. As discussed above, all 5 indices are scaled to [0, 1]. In all cases, we can observe that the R' index tends to behave similarly to *AP*. The box plots of the R' index for the 10 models are depicted in Fig. 3. We can tell that the R' index does express inconsistent evaluations for all the models of the 10 classes, thereby further verifying its robustness and effectiveness.

Table 4 Confusion matrix and R'_i for the test set classification results

	0	1	2	3	4	5	6	7	8	9
0	966	1	0	2	0	1	4	3	3	0
1	0	1125	3	1	0	0	2	1	3	0
2	1	0	1016	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	1011	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	0.9860	0.9913	0.9839	0.9773	0.9653	0.9774	0.9814	0.9830	0.9784	0.9798

Table 5 R' and other indices for 10 experiments

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R'	0.3165	0.6898	0.8741	0.8689	0.9662	0.9802	0.9770	0.9704	0.1135	0.1135
AP	0.0000	0.9020	0.9735	0.9847	0.9898	0.9929	0.9847	0.9878	0.0000	0.0000
P	0.3027	0.6832	0.8730	0.8680	0.9659	0.9800	0.9770	0.9702	0.1000	0.1000
R	0.2731	0.8007	0.8860	0.8970	0.9658	0.9802	0.9768	0.9707	0.0114	0.0114
F-1	0.2871	0.7373	0.8795	0.8823	0.9659	0.9801	0.9769	0.9705	0.0204	0.0204
$Kappa$	0.2384	0.6551	0.8601	0.8543	0.9624	0.9780	0.9744	0.9671	0.0000	0.0000

In addition, we can conclude that the box plots of classes *1* and *2* are more likely to have larger and compact value distributions, while classes *3, 5, 6, 7, 8,* and *9* behave in the opposite way; this also conforms to the observations in Section 4.1.

4.3 R′ indices for individual categories

The experiments in this subsection change the volumes of certain samples to simulate imbalanced samples. Specifically, the process is divided into the following 2 steps:

- Change the sample volumes of categories *0*, *4*, *6*, and *8* and evaluate the classification results through the R' index, which is called *before*.
- Restore the original volumes of these categories in the training set and evaluate the test set classification results by the R' index, which is called *after*.

Keep all the other parameters of the model the same before and after the above changes are made. Table 7 shows the R' indices of each category and the overall classification effects before and after the sample volume changes.

It can be seen that before the sample volumes of categories *0*, *4*, *6* and *8* are restored (i.e., *before*), their R' indices are relatively small (*0.8685*, *0.7519*, *0.8139*, and *0.7044*, respectively; bold in the table). After returning to the original volumes (i.e., *after*), the R' indices corresponding to categories *0*, *4*, *6* and *8* are significantly improved (to *0.9874*, *0.9745*, *0.9858*, and *0.9789*, respectively; bold in the table), and the overall classification effect evaluation index also increases from *0.8971* to *0.9789*. It is worth noting that this has guiding significance for optimizing and improving the training process of a given model, i.e., by observing the changes in the R' indices of each category or some categories , we can take necessary measures (such as ensuring sample equilibrium) to improve the training process.

Table 6 CCs between R' and other indices, as well as the RMSEs

	R'-AP	R'-P	R'-R	R'-F-1	R'-$Kappa$
CC	0.9703	1.0000	0.9939	0.9979	1.0000
RMSE	0.1398	0.0078	0.0600	0.0455	0.0579

Consider the generalization of the R' index in Section 3.2. The steps for calculating the R' index of each category are given in Section 3.2, i.e.:

$$R'(m_i|s_i) = P(s_i|m_i) - [P(s_i) - P(m_i)] \tag{26}$$

Examining this formula, when evaluating the classification effects, the R' index further combines the difference $P(s_i) - P(m_i)$ between the correct predictions and the incorrect predictions of the samples, in addition to considering the probability $P(s_i|m_i)$ that a sample is correctly predicted according to the real labels. For the sample imbalance caused by changing the sample volumes in the experiments, this difference is well extracted by the R' index.

More specifically, consider the confusion matrices of the test set before and after changing the training sample volumes, as shown in Table 7. The real labels of the test set represented by the horizontal rows in the table are not changed in the 2 experiments, and each column of the predicted labels is changed to a certain extent (especially for categories *0*, *4*, *6*, and *8*; bold in the table). This explains why the R' indices of these categories change in the above experiments. In other words, the R' index can fairly effectively observe the differences in classification evaluations caused by imbalanced samples, and it can guide and improve the training process as a result.

4.4 Results on the CIFAR-10 dataset

The same methodology as that in Section 4.3 is also utilized to conduct these experiments (the volumes of samples in categories *cat*, *deer*, *dog*, and *horse* are changed). Table 8 shows the corresponding confusion matrices that we finally obtain.

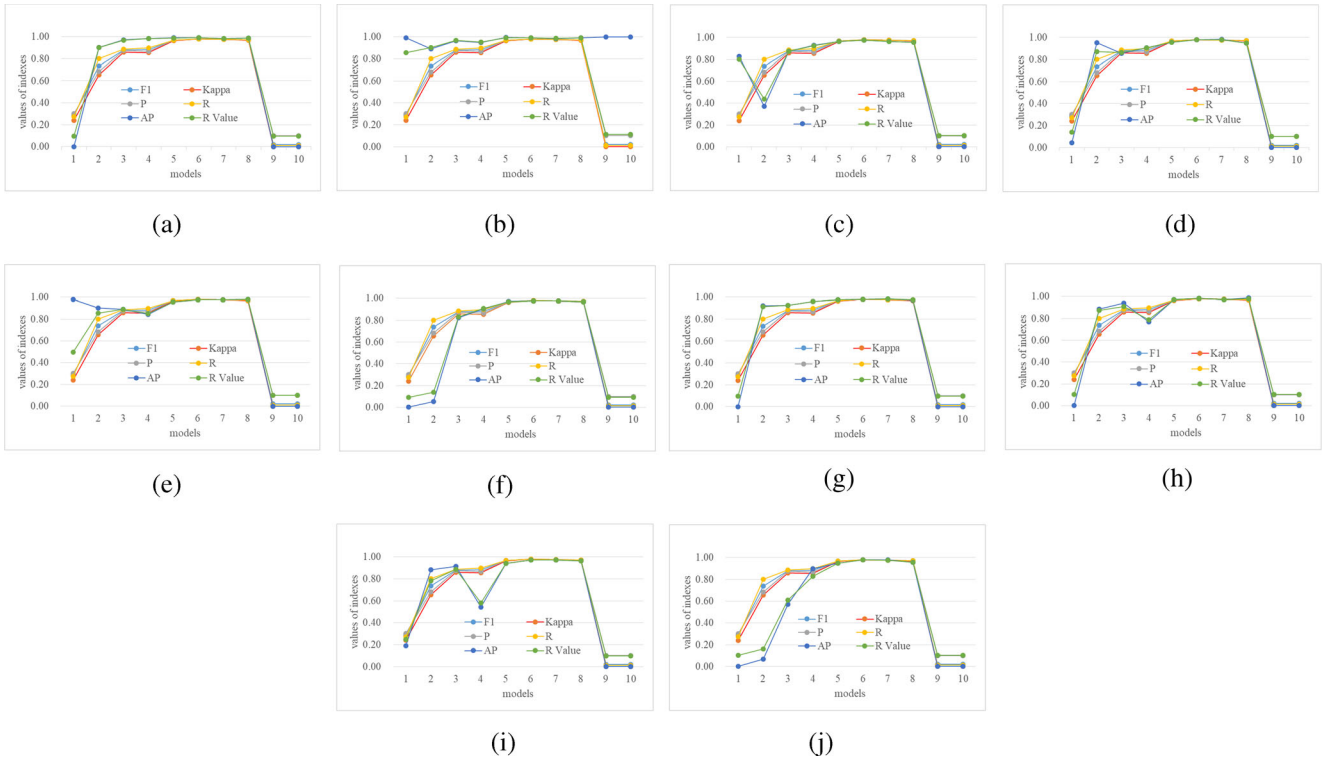

Fig. 2 Indices for 10 classes of different models. **a** class 0. **b** class 1. **c** class 2. **d** class 3. **e** class 4. **f** class 5. **g** class 6. **h** class7. **i** class 8. **j** class 9

It can be seen from the table that before recovering the sample volumes the *cat*, *deer*, *dog*, and *horse* categories (i.e., *before*), their R' indices are relatively small (*0.6410*, *0.7633*, *0.7247* and *0.8092*, respectively; bold in the table). After the original volumes are restored (i.e., *after*), the R' indices corresponding to the *cat*, *deer*, *dog*, and *horse* categories are significantly improved (to *0.7703*, *0.8941*, *0.8221* and *0.8980*, respectively; bold in table), and the overall classification effect evaluation index also increase from *0.8334* to *0.8756* (Table 8).

The above experiments demonstrate the applicability and effectiveness of the R' index on *CIFAR-10*, and this further illustrates the scalability of the R' index.

Table 9 shows the other indices obtained for the experiments in Sections 4.3 and 4.4. They tend to present slighter differences than those of the R' index because they combine the classification results of all categories together without distinguishing between the individuals, and this is a bit extensive. In other words, the R' index provides a more definite evaluation of each category than other indices, and based on this, we can perform many adjustments and elaborate management.

Fig. 3 Box plots of R'_i for 10 classes of different models

Table 7 Confusion matrix and R_i' before/after the sample volume changes (MNIST)

	0	1	2	3	4	5	6	7	8	9	*Sum/all*
0	839/968	7/1	40/2	14/2	4/1	40/1	6/3	10/1	3/1	17/0	980
1	0	1128/1126	1	3/1	0	1/0	0/4	1	1/2	0	1135
2	1/2	3	1007/1011	9/2	0/2	2/0	0/1	6/4	2/7	2/0	1032
3	0/1	0	3/8	989/982	0	10/4	1/0	3	0/8	4	1010
4	1/4	8/1	17/6	4/0	714/956	4/0	11/2	30/2	12/3	181/8	982
5	0/2	0/1	0	10/8	0/1	876/868	0/6	1	1/4	4/1	892
6	9/3	6/2	59/1	4/1	12/3	80/1	763/945	2/0	17/2	6/0	958
7	0	4/6	7/13	4/1	0	0	0	1005/998	1/4	7/6	1028
8	6/3	24/2	48/3	89/2	3/1	68/0	3/1	9/1	659/955	65/6	974
9	0/1	6/3	1	4/3	1/8	2/5	0/2	4/2	0/4	991/980	1009
Sum	**856/984**	1186/1145	1183/1046	1130/1002	**734/972**	1083/879	**784/964**	1071/1013	**696/990**	1277/1005	10000
$R_i'(before)$	**0.8685**	0.9887	0.9607	0.9672	**0.7519**	0.9630	**0.8139**	0.9733	**0.7044**	0.9554	**0.8971**
$R_i'(after)$	**0.9874**	0.9911	0.9783	0.9731	**0.9745**	0.9744	**0.9858**	0.9723	**0.9789**	0.9717	**0.9789**

Table 8 Confusion matrix and R_i' before/after the sample volume changes (CIFAR-10)

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Sum/all
Airplane	885/882	11	20/23	7/16	0/8	1/6	6/2	3/5	49/31	18/16	1000
Automobile	12/6	938/941	1/3	1/2	0/1	1/2	2/3	0/2	12/11	33/29	1000
Bird	40/32	5/1	862/812	10/46	19/36	8/19	39/27	5/15	7/9	5/3	1000
Cat	30/10	13/6	86/22	631/773	30/36	93/103	68/22	20/19	24/3	15/6	1000
Deer	18/6	3/1	79/13	26/25	749/900	19/20	43/10	40/20	14/3	9/2	1000
Dog	14/7	10/0	73/17	105/103	27/26	712/823	17/7	25/14	7/1	10/2	1000
Frog	7/5	2/1	22/16	9/29	1/14	1/10	948/918	5/5	1	4/1	1000
Horse	34/8	8/1	36/10	18/22	30/35	38/22	16/5	799/896	10/0	11/1	1000
Ship	33/42	15/16	3/5	0/7	0/1	0/1	1/2	0/3	924/909	24/14	1000
Truck	24/20	46/40	1/4	3/4	1/2	0/3	1/2	1	27/22	896/902	1000
Sum	1097/1018	1051/1018	1183/925	**800/1027**	**857/1059**	**873/1009**	1141/998	**898/980**	1075/990	1025/976	10000
$R_i'(before)$	0.8753	0.9329	0.8437	**0.6410**	**0.7633**	**0.7247**	0.9339	**0.8092**	0.9165	0.8935	**0.8334**
$R_i'(after)$	0.8802	0.9392	0.8195	**0.7703**	**0.8941**	**0.8221**	0.9192	**0.8980**	0.9100	0.9044	**0.8756**

Table 9 Other indices before/after sample volume changes

		AP	*P*	*R*	*F-1*	*Kappa*
MNIST	*before*	0.8971	0.8947	0.9074	0.9010	0.8856
	after	0.9789	0.9787	0.9790	0.9789	0.9765
CIFAR-10	*before*	0.8334	0.8334	0.8348	0.8341	0.8149
	after	0.8756	0.8756	0.8763	0.8760	0.8618

5 Conclusion

This paper proposes and illustrates the R' index for assessing classification effects. It has a rigorous mathematical theory derivation process and can supply us with both overall and individual evaluations. Experiments based on *MNIST* and *CIFAR-10* show that it exhibits great robustness and effectiveness and can be used for multiclass classification task evaluation. Furthermore, it is beneficial for improving the training process of a given model based on the individual evaluations that the index provides.

Acknowledgments The work is sponsored by National Key Research and Development Program of China (2018YFB0204301), and Open Fund of PDL (6142110190201).

References

1. Dou Y, Qiao P, Jin R (2019) Exploring the defects of the average precision and its influence. Scientia Sinica Informationis 49(10):1369–1382
2. Sharma R, Goyal AK, Dwivedi RK (2016) A review of soft classification approaches on satellite image and accuracy assessment. In: Pant M, Deep K, Bansal J, Nagar A, Das K (eds) Proceedings of fifth international conference on soft computing for problem solving. Advances in intelligent systems and computing, vol 437. Springer, Singapore. https://doi.org/10.1007/978-981-10-0451-3_56
3. Erener A (2013) Classification method, spectral diversity, band combination and accuracy assessment evaluation for urban feature detection. Int J Classification Applied Earth Observation and Geoinformation 21:397–408
4. Persello C, Bruzzone L (2010) A novel protocol for accuracy assessment in classification of very high resolution images. Geoscience and Remote Sensing 48(3-1):1232–1244
5. Ding PLK, Li Y, Li B (2018) Mean local group average precision (mLGAP): a new performance metric for hashing-based retrieval. arXiv:1811.09763
6. He K, Lu Y, Sclaroff S (2018) Local descriptors optimized for average precision. 596–605. IEEE Computer Society
7. Yilmaz E, Aslam JA (2008) Estimating average precision when judgments are incomplete. Knowl Inf Syst 16(2):173–211
8. Henderson P, Ferrari V (2017) End-to-end training of object class detectors for mean average precision. In: Lai SH, Lepetit V, Nishino K, Sato Y (eds) Computer vision - ACCV 2016. ACCV 2016. Lecture notes in computer science, vol 10115. Springer, Cham. https://doi.org/10.1007/978-3-319-54193-8_13
9. Mao H, Yang X, Dally WJ (2019) A delay metric for video object detection: what average precision fails to tell. arXiv:1908.06368
10. Piwowarski B, Dupret G, Lalmas M (2012) Beyond cumulated gain and average precision: including willingness and expectation in the user model. arXiv:1209.4479
11. Andric K, Kalpic D, Bohacek Z (2019) An insight into the effects of class imbalance and sampling on classification accuracy in credit risk assessment. Comput Sci Inf Syst 16(1):155–178
12. Bestgen Y (2015) Exact expected average precision of the random baseline for system evaluation. Prague Bull Math Linguistics 103:131–138
13. Revaud J, Almazan J, de Rezende RS, de Souza CR (2019) Learning with average precision: training image retrieval with a listwise loss [C]. The IEEE International Conference on Computer Vision (ICCV)
14. Yuan Y, Wanhua S, Mu Z (2015) Threshold-free measures for assessing the performance of medical screening tests[J]. Frontiers in Public Health. 3
15. Wang X (2001) Problem and improvement of R-values applied to assessment of earthquake forecast. China Earthquake Research 3(16):75–83
16. Dou AX, Wang XQ, Dou MW (2004) A new approach to evaluate the accuracy of image classification result R' [C]. IEEE International Geoscience & Remote Sensing Symposium. IEEE
17. Omri A, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43(6):1223–1232
18. Bruzzone L, Persello C (2008) A novel protocol for accuracy assessment in classification of very high resolution multispectral and SAR images. 265–268. IEEE
19. Candela JQ, Dagan I (2006) Machine learning challenges, evaluating predictive uncertainty, visual object classification and recognizing textual entailment. First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005: Revised Selected Papers[J]. Springer
20. Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. The MIT Press
21. Bishop CM (2006) Pattern recognition and machine learning (Information Science and Statistics) [M]. Springer-Verlag New York, Inc.
22. Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: from theory to algorithms[M]. Cambridge University Press, Cambridge
23. Chen S, Fern A, Todorovic S (2015) Person count localization in videos from noisy foreground and detections. 1364–1372. IEEE Computer Society
24. Gajda J, Sroka R (2015) Design and accuracy assessment of the multi-sensor weigh-in-motion system [C]. Instrumentation & Measurement Technology Conference. 1036–1041. IEEE
25. Li W, Guo Q (2014) A new accuracy assessment method for one-class remote sensing classification. Geoscience and Remote Sensing 52(8):4621–4632
26. Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, Ose K, Fotsing JM, Fontenille D, Besansky NJ, Costantini C (2009) Ecological niche partitioning between Anopheles gambiae molecular forms in cameroon: the ecological side of speciation. Bmc Ecology 9(1):17
27. Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern Recognition Letters 30(1):27–38
28. Mosley L (2013) A balanced approach to the multi-class imbalance problem[D]
29. Jiang L, Zhang H, Cai Z (2009) A novel bayes model: hidden naive bayes[J]. IEEE Transactions on Knowledge & Data Engineering 21(10):1361–1371
30. Jiang L, Cai Z, Wang D (2012) Improving tree augmented Naive Bayes for class probability estimation[J]. Knowl-Based Syst 26:239–245
31. Zhang C, Bi J, Xu S, Ramentol E, Fan G, Qiao B, Fujita H (2019) Multi-imbalance: an open-source software for multi-class imbalance learning[J]. Knowl-Based Syst 174(JUN.15):137–143
32. Sun J, Li H, Fujita H, Fu B, Ai W (2020) Class-imbalanced dynamic financial distress prediction based on adaboost-SVM ensemble combined with SMOTE and time weighting[J]. Information Fusion 54:128–144
33. Zhou L, Wang Q, Fujita H (2016) One versus one multi-class classification fusion using optimizing decision directed acyclic graph for predicting listing status of companies[J]. Information Fusion 36:80–89
34. Zhou L, Fujita H (2017) Posterior probability based ensemble strategy using optimizing decision directed acyclic graph for multi-class classification[M]. Elsevier Science Inc.
35. Zhou L, Tam KP, Fujita H (2016) Predicting the listing status of chinese listed companies with multi-class classification Models[J]. Inf Sci 328:222–236
36. Sun J, Lang J, Fujita H, Li H (2017) Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates[J]. Information Sciences. S0020025517310083
37. Deng T, Ye D, Ma R, Fujita H, Xiong L (2020) Low-rank local tangent space embedding for subspace clustering[J]. Inf Sci 508:1–21

Kaifang Zhang received B.S. degree in information engineering in 2014, and now is pursuing the degree in computer science and technology. His main interests include high performance computing, machine learning and reinforcement learning.

Huayou Su received his B.S., M.S., and Ph.D. degrees in computer science and technology in and 2008, 2010, and 2014 respectively. His main research fields involve high performance computing, and heterogeneous computing.

Yong Dou received the B.S., M.S., and Ph.D. degrees in computer science and technology in 1990, 1992, and 1995 respectively. His research interests include reconfigurable computing, high performance computing, and reinforcement learning.