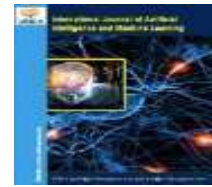




# International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

## Tempered Balanced Accuracy (tBA)

Hilal H. Nuha<sup>1\*</sup>, Hassan Rizky Putra Saillellah<sup>1</sup>

<sup>1</sup>School of Computing, Telkom University, Bandung, Indonesia

### Abstract

Class imbalance is present in most practical classification datasets, from intrusion detection to medical diagnosis, and it is well known that standard accuracy rewards a classifier that ignores the minority class. Balanced accuracy (BA) removes this bias by averaging the per-class recalls with equal weight. However, equal weighting is itself a fixed cost model: BA implicitly assumes that the importance of a class is exactly inversely proportional to its prevalence, which over-rewards a rare class that is not critical and under-rewards a rare class that is critical. This paper proposes the Tempered Balanced Accuracy (tBA), a weighted mean of per-class recalls in which the weight of class  $k$  is its prevalence raised to a tempering exponent  $t$ . The parameter  $t$  interpolates continuously between standard accuracy ( $t = 1$ ) and BA ( $t = 0$ ) and extends beyond BA ( $t < 0$ ) for rare-but-critical classes. We show that the chance level of tBA equals  $1/K$  for every  $t$ , that the score of a majority-class predictor has a closed form that decreases monotonically in  $t$ , and that  $t$  can be set directly from a desired misclassification cost ratio. Rules for choosing  $t$  are given for known cost ratios, for the absence of cost information ( $t = 0$  as the safe value against class collapse,  $t = 0.5$  as the symmetric compromise between accuracy and BA), and as a data-driven estimate from the effective number of classes. Through eight hand-calculated cases worked from labelled samples and nine further cases from confusion matrices, covering binary and multi-class problems, balanced and imbalanced classes at ratios from 4:1 to 999:1, class collapse, minority over-prediction, an absent class, and classifier ranking, we demonstrate that tBA exposes behaviors that both accuracy and BA hide, including a ranking reversal between two classifiers at  $t^* = 0.786$  and a slope of the tBA curve that identifies whether a classifier is weak on its large or its small classes. tBA requires no additional computation beyond the confusion matrix and reduces to existing metrics at its endpoints.

**Keywords:** balanced accuracy, class imbalance, evaluation metric, weighted recall, cost-sensitive evaluation, confusion matrix

## 1 Introduction

Imbalanced class distributions are the rule rather than the exception in applied classification. In network intrusion datasets, attack traffic typically constitutes less than 5% of records; in medical screening, disease prevalence is often below 1%; in manufacturing fault detection, defective items may be fewer than one per thousand. In such settings, standard accuracy, defined as the fraction of correctly classified samples, is dominated by the majority class. A classifier that always predicts the majority class attains an accuracy equal to the majority prevalence, which can exceed 99% while providing no useful information about the minority class [1].

Balanced accuracy (BA), defined as the arithmetic mean of per-class recalls, is the most widely adopted correction [2]. Under BA, the majority-class predictor scores exactly  $1/K$  for  $K$  classes, and every class contributes equally regardless of its size. Related alternatives include the geometric mean of recalls (G-mean) [3], the Matthews correlation coefficient (MCC) [4], Cohen's  $\kappa$  [5], macro-averaged  $F_1$ , and the area under the precision-recall curve [6]. Each of these removes the prevalence bias in a different way.

However, BA replaces one fixed cost model with another. Weighting every class by  $1/K$  is equivalent to assigning each class a misclassification cost inversely proportional to its prevalence. If the minority class is 0.1% of the data, BA treats a single minority error as one thousand times more costly than a single majority error. This is appropriate when the rare class is indeed critical, but it is arbitrary in general: it over-rewards a rare class that is merely uncommon, and it may still under-reward a rare class whose cost of omission is catastrophic. Cost-sensitive evaluation [7] allows arbitrary weights, but requires  $K$  free parameters that are rarely reported and hard to compare across studies. In the training literature, the same issue has been addressed by tempering the class prior, for example through logit adjustment with a temperature  $\tau$  [8] and through the class-balanced loss with an effective-number parameter  $\beta$  [9]. No equivalent single-parameter treatment exists on the evaluation side.

Motivated by the work of Brodersen et al. [2] and by the prior-tempering approach of Menon et al. [8], the main contribution of this paper is the Tempered Balanced Accuracy (tBA), a weighted mean of per-class recalls whose weights are the class prevalences raised to a single exponent  $t$ . The contributions are as follows: (i) tBA

is defined and shown to reduce to standard accuracy at  $t = 1$  and to BA at  $t = 0$ , so both existing metrics become special cases of one family; (ii) three properties are proven, namely  $t$ -invariant chance level, a closed-form monotone score for the majority-class predictor, and a direct rule for choosing  $t$  from a desired cost ratio; (iii) rules for choosing  $t$  are given, including a safe value, a symmetric compromise, and a data-driven estimate from the effective number of classes; (iv) eight cases worked from labelled samples and nine cases from confusion matrices are presented in full so that every number can be verified with a pocket calculator, covering balanced and imbalanced classes, binary and multi-class problems, imbalance ratios from 4:1 to 999:1, class collapse, minority over-prediction, an absent class, and the ranking of competing classifiers.

The remainder of this paper is organized as follows. Section 2 defines tBA, derives its properties, and gives rules for choosing  $t$ . Section 3 presents hand-calculated cases worked from labelled samples and Section 4 presents further cases from confusion matrices. Section 5 discusses the findings. Conclusions are given in Section 6.

## 2 Methodology

This section presents the proposed metric. The first subsection fixes the notation. The second subsection defines tBA. The third subsection derives its properties. The fourth subsection gives rules for selecting  $t$ .

### 2.1 Notation

Let  $K$  denote the number of classes and  $N$  the number of test samples. Let  $C$  be the  $K \times K$  confusion matrix with  $C_{ij}$  the number of samples of true class  $i$  predicted as class  $j$ . The support of class  $k$  is  $n_k = \sum_j C_{kj}$ , and its prevalence is

$$p_k = \frac{n_k}{N}, \quad \sum_{k=1}^K p_k = 1. \quad (1)$$

The recall of class  $k$  is

$$r_k = \frac{C_{kk}}{n_k}. \quad (2)$$

Standard accuracy and balanced accuracy are

$$\text{Acc} = \frac{1}{N} \sum_{k=1}^K C_{kk} = \sum_{k=1}^K p_k r_k, \quad (3)$$

$$\text{BA} = \frac{1}{K} \sum_{k=1}^K r_k. \quad (4)$$

Equation (3) makes explicit that accuracy is a prevalence-weighted mean of recalls, whereas (4) is the same mean with uniform weights. In the binary case, with the positive class as the minority,  $r_+ = \text{TP}/(\text{TP} + \text{FN})$  is the sensitivity and  $r_- = \text{TN}/(\text{TN} + \text{FP})$  is the specificity.

### 2.2 Tempered Balanced Accuracy

The proposed metric is defined as

$$\text{tBA}_t = \sum_{k=1}^K w_k(t) r_k, \quad w_k(t) = \frac{p_k^t}{\sum_{j=1}^K p_j^t}, \quad (5)$$

where  $t \in \mathbb{R}$  denotes the tempering exponent. The weight of each class is its prevalence raised to the power  $t$ , normalized so that the weights sum to one. The special cases are as follows:

- $t = 1$ :  $w_k = p_k$ , hence  $\text{tBA}_1 = \text{Acc}$ ;
- $t = 0$ :  $w_k = 1/K$ , hence  $\text{tBA}_0 = \text{BA}$ ;
- $0 < t < 1$ : minority classes are up-weighted, but not to parity;
- $t < 0$ : minority classes are weighted above majority classes;
- $t \rightarrow +\infty$ :  $\text{tBA}_t \rightarrow r_{\max}$ , the recall of the largest class;
- $t \rightarrow -\infty$ :  $\text{tBA}_t \rightarrow r_{\min}$ , the recall of the smallest class.

Therefore, accuracy and BA are the two anchor points of a single continuous family, and the evaluator states the cost model explicitly through one number  $t$  rather than inheriting the fixed model of either metric. Unless otherwise stated, we recommend computing  $p_k$  from the full dataset rather than from each test fold, so that the weights are identical across cross-validation folds.

### 2.3 Properties

Property 1 (Chance level is independent of  $t$ ). A classifier that predicts uniformly at random has  $r_k = 1/K$  for every class. Substituting into (5) gives

$$tBA_t = \frac{1}{K} \sum_{k=1}^K w_k(t) = \frac{1}{K} \quad (6)$$

for every  $t$ , since the weights sum to one. Consequently the chance-corrected form

$$tBA_t^{\text{adj}} = \frac{tBA_t - 1/K}{1 - 1/K} \quad (7)$$

is comparable across different values of  $t$  and across problems with different  $K$ .

Property 2 (Majority-class predictor). Let class 1 be the largest class. A classifier that always predicts class 1 has  $r_1 = 1$  and  $r_k = 0$  for  $k \neq 1$ , so

$$tBA_t^{\text{maj}} = w_1(t) = \frac{p_1^t}{\sum_{j=1}^K p_j^t}. \quad (8)$$

This is a closed-form curve that equals  $p_1$  at  $t = 1$ ,  $1/K$  at  $t = 0$ , and tends to 0 as  $t \rightarrow -\infty$ . Its derivative with respect to  $t$  is  $w_1(t)(\ln p_1 - \sum_j w_j(t) \ln p_j)$ , which is positive because  $p_1$  is the largest prevalence; hence (8) is strictly increasing in  $t$ , and lowering  $t$  monotonically reduces the reward for class collapse.

Property 3 (Choosing  $t$  from a cost ratio). Under (5), the ratio of the weight of class  $k$  to that of class  $j$  is  $(p_k/p_j)^t$ . If the evaluator wishes class  $k$  to count  $c$  times as much as class  $j$ , then

$$t = \frac{\ln c}{\ln(p_k/p_j)}. \quad (9)$$

For example, on a 95:5 dataset, requiring the minority class to count five times as much as the majority class gives  $t = \ln 5 / \ln(0.05/0.95) = 1.6094 / (-2.9444) = -0.5466$ . Requiring equal weight gives  $c = 1$  and  $t = 0$ , which recovers BA, and requiring  $c = p_k/p_j$  gives  $t = 1$ , which recovers accuracy.

Property 4 (Monotonicity in  $t$ ). Differentiating (5) yields

$$\frac{\partial tBA_t}{\partial t} = \sum_k w_k(t) r_k \ln p_k - \left( \sum_k w_k(t) r_k \right) \left( \sum_k w_k(t) \ln p_k \right), \quad (10)$$

which is the covariance of  $\ln p_k$  and  $r_k$  under the weights  $w(t)$ . Therefore,  $tBA_t$  increases with  $t$  exactly when large classes have higher recall than small classes, which is the typical situation, and decreases with  $t$  when the minority classes are recalled better than the majority. The sign of this derivative is itself a diagnostic: it tells the evaluator on which side of the class spectrum the classifier is weak.

## 2.4 Choosing $t$

The exponent  $t$  encodes how much the small classes are meant to count, so it should be fixed from the purpose of the evaluation before results are seen, reported with them, and held constant across folds and across the classifiers being compared. Four situations cover most cases.

Cost ratio known. If class  $k$  should count  $c$  times class  $j$ , (9) gives  $t$  directly; for a rare critical class this is typically between  $-0.3$  and  $-1$ .

Safe against class collapse:  $t = 0$ . By Property 2 the majority-class predictor scores above chance for every  $t > 0$  and exactly  $1/K$  at  $t = 0$ , so  $t = 0$  is the largest value at which a degenerate classifier cannot beat chance. When the only concern is not being misled by a classifier that ignores small classes, BA is the conservative default.

Symmetric compromise:  $t = 0.5$ . At  $t = 0.5$  the weights are  $w_k \propto \sqrt{p_k} = \sqrt{p_k \cdot (1/K)}$  up to normalisation, the geometric mean of the accuracy weights  $p_k$  and the BA weights  $1/K$ . It is the unique member of the family that sits symmetrically between the two standard metrics in weight space, and it is the appropriate choice when the majority class is itself of interest, when false alarms on it are costly, or when the minority evidence is too thin to justify parity.

Data-driven estimate. Let  $H(w) = -\sum_k w_k \ln w_k$  and  $K_{\text{eff}}(t) = e^{H(w(t))}$ , the effective number of equally weighted classes, which rises monotonically from  $e^{H(p)}$  at  $t = 1$  to  $K$  at  $t = 0$ . Solving  $K_{\text{eff}}(t) = 1/2(e^{H(p)} + K)$  by bisection gives a  $t$  that adapts to the imbalance: about 0.68 for 7:3, 0.51 for 95:5, 0.59 for 80:15:5, and 0.26 for 999:1. The target is stated in interpretable units, “as if there were  $K^*$  equally weighted classes”, rather than as an unmotivated exponent. Whichever rule is used,  $t$  must never be chosen to maximise a classifier’s own score.

Values below about  $-1$  make the score depend almost entirely on the smallest class and are rarely useful; in that regime reporting  $r_{\min}$  directly is clearer.

## 3 Hand-Calculated Cases from Labelled Samples

This section starts from the actual and predicted labels of small test sets, so that the entire computation, from labels to metric, can be checked by hand; Section 4 then works from larger confusion matrices directly. Eight

samples are given, four binary and four multi-class, covering balanced classes, mild and heavy imbalance, a majority-class predictor, minority over-prediction, and a fold in which one class is absent. In each sample table the index of a misclassified sample is marked with an asterisk. For every sample the same steps are applied, namely confusion matrix, per-class recall, prevalence, accuracy and BA, and tempered weights at  $t = \pm 0.5$ ; the five-value summary is collected in Table 9. The suggested  $t$  for each sample follows the rules of Section 2.4.

### 3.1 Sample 1: Binary, balanced (5:5)

The  $N = 10$  labels are listed in Table 1 (3 misclassified). The confusion matrix and supports are

**Table 1. Sample 1: actual and predicted labels (\* = misclassified).**

| no.       | 1 | 2 | 3 | 4* | 5 | 6 | 7* | 8 | 9 | 10* |
|-----------|---|---|---|----|---|---|----|---|---|-----|
| Actual    | 1 | 0 | 1 | 0  | 1 | 0 | 1  | 0 | 1 | 0   |
| Predicted | 1 | 0 | 1 | 1  | 1 | 0 | 0  | 0 | 1 | 1   |

$$C = \begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix}, \quad n = (5,5). \quad (11)$$

$$r = (3/5, 4/5) = (0.6000, 0.8000), \quad p = (0.5000, 0.5000).$$

so  $Acc = 7/10 = 0.7000$  and  $BA = (0.6000 + 0.8000)/2 = 0.7000$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.7071, 0.7071), \quad S = 1.4142, \quad w = (0.5000, 0.5000), \\ tBA_{0.5} = 0.5000(0.6000) + 0.5000(0.8000) = 0.7000.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.4142, 1.4142), \quad S = 2.8284, \quad w = (0.5000, 0.5000), \\ tBA_{-0.5} = 0.5000(0.6000) + 0.5000(0.8000) = 0.7000.$$

Observation. Every column equals 0.7000: with  $p_k = 1/K$  the weights are  $1/K$  for all  $t$ , so on balanced data  $tBA$  reduces to accuracy and BA even though the recalls differ ( $r_0 = 0.6000$ ,  $r_1 = 0.8000$ ). Suggested:  $t = 0$ ,  $tBA_0 = 0.7000$  (balanced, every  $t$  identical).

### 3.2 Sample 2: Binary, imbalanced (7:3)

The  $N = 10$  labels are listed in Table 2 (5 misclassified). The confusion matrix and supports are

**Table 2. Sample 2: actual and predicted labels (\* = misclassified).**

| no.       | 1 | 2 | 3* | 4* | 5 | 6* | 7* | 8 | 9 | 10* |
|-----------|---|---|----|----|---|----|----|---|---|-----|
| Actual    | 1 | 1 | 1  | 0  | 1 | 1  | 0  | 1 | 1 | 0   |
| Predicted | 1 | 1 | 0  | 1  | 1 | 0  | 1  | 1 | 1 | 1   |

$$C = \begin{pmatrix} 0 & 3 \\ 2 & 5 \end{pmatrix}, \quad n = (3,7). \quad (12)$$

$$r = (0/3, 5/7) = (0.0000, 0.7143), \quad p = (0.3000, 0.7000).$$

so  $Acc = 5/10 = 0.5000$  and  $BA = (0.0000 + 0.7143)/2 = 0.3571$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.5477, 0.8367), \quad S = 1.3844, \quad w = (0.3956, 0.6044), \\ tBA_{0.5} = 0.3956(0.0000) + 0.6044(0.7143) = 0.4317.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.8257, 1.1952), \quad S = 3.0210, \quad w = (0.6044, 0.3956), \\ tBA_{-0.5} = 0.6044(0.0000) + 0.3956(0.7143) = 0.2826.$$

Observation. All three class-0 samples are predicted as class 1, so  $r_0 = 0$  and the row falls from 0.5000 to 0.2143. The classifier is in effect a majority predictor with two extra errors; accuracy hides this and BA gives three samples half the weight.  $t = 0.5$  (0.4317) flags the collapse without letting three samples dominate. Suggested:  $t = 0.5$ ,  $tBA_{0.5} = 0.4317$  (mild imbalance, no cost information).

### 3.3 Sample 3: Binary, heavily imbalanced (18:2), majority-class predictor

The  $N = 20$  labels are listed in Table 3 (2 misclassified). The confusion matrix and supports are

**Table 3. Sample 3: actual and predicted labels (\* = misclassified).**

| no.       | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9   | 10  |
|-----------|----|----|----|----|----|----|----|----|-----|-----|
| Actual    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0   | 0   |
| Predicted | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0   | 0   |
| no.       | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19* | 20* |
| Actual    | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1   | 1   |
| Predicted | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0   | 0   |

$$C = \begin{pmatrix} 18 & 0 \\ 2 & 0 \end{pmatrix}, \quad n = (18,2). \quad (13)$$

$$r = (18/18, 0/2) = (1.0000, 0.0000), \quad p = (0.9000, 0.1000).$$

so  $Acc = 18/20 = 0.9000$  and  $BA = (1.0000 + 0.0000)/2 = 0.5000$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.9487, 0.3162), \quad S = 1.2649, \quad w = (0.7500, 0.2500), \\ tBA_{0.5} = 0.7500(1.0000) + 0.2500(0.0000) = 0.7500.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.0541, 3.1623), \quad S = 4.2164, \quad w = (0.2500, 0.7500), \\ tBA_{-0.5} = 0.2500(1.0000) + 0.7500(0.0000) = 0.2500.$$

Observation. The minority class is never predicted, so  $r = (1,0)$  and every column is the majority weight  $w_1(t)$  of (8): 0.90, 0.75, 0.50, 0.25, 0.10. If a missed minority sample costs five times a missed majority sample, (9) gives  $t = \ln 5 / \ln(0.1/0.9) = -0.7325$  and a score of 0.1667. Suggested:  $t = -0.7325$ ,  $tBA_{-0.7325} = 0.1667$  (minority critical, cost ratio  $c = 5$ ).

### 3.4 Sample 4: Binary, heavily imbalanced (17:3), minority over-prediction

The  $N = 20$  labels are listed in Table 4 (6 misclassified). The confusion matrix and supports are

**Table 4. Sample 4: actual and predicted labels (\* = misclassified).**

| no.       | 1  | 2   | 3   | 4   | 5   | 6   | 7   | 8  | 9  | 10 |
|-----------|----|-----|-----|-----|-----|-----|-----|----|----|----|
| Actual    | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0  | 0  |
| Predicted | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0  | 0  |
| no.       | 11 | 12* | 13* | 14* | 15* | 16* | 17* | 18 | 19 | 20 |
| Actual    | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 1  | 1  | 1  |
| Predicted | 0  | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1  | 1  |

$$C = \begin{pmatrix} 11 & 6 \\ 0 & 3 \end{pmatrix}, \quad n = (17,3). \quad (14)$$

$$r = (11/17, 3/3) = (0.6471, 1.0000), \quad p = (0.8500, 0.1500).$$

so  $Acc = 14/20 = 0.7000$  and  $BA = (0.6471 + 1.0000)/2 = 0.8235$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.9220, 0.3873), \quad S = 1.3093, \quad w = (0.7042, 0.2958), \\ tBA_{0.5} = 0.7042(0.6471) + 0.2958(1.0000) = 0.7515.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.0847, 2.5820), \quad S = 3.6666, \quad w = (0.2958, 0.7042), \\ tBA_{-0.5} = 0.2958(0.6471) + 0.7042(1.0000) = 0.8956.$$

Observation. All three positives are detected at the cost of six false alarms (a 35% false-positive rate), so the minority recall exceeds the majority recall and the row increases from 0.7000 to 0.9471. Any  $t \leq 0$  nearly ignores the false alarms;  $t = 0.5$  (0.7515) keeps 0.7042 of the weight on the majority class. Minority precision is only 3/9 and no  $t$  reflects it; MCC or  $F_1$  should be reported alongside. Suggested:  $t = 0.5$ ,  $tBA_{0.5} = 0.7515$  (false alarms on the majority matter).

### 3.5 Sample 5: Three classes, balanced (4:4:4)

The  $N = 12$  labels are listed in Table 5 (4 misclassified). The confusion matrix and supports are

**Table 5. Sample 5: actual and predicted labels (\* = misclassified).**

| no.       | 1  | 2   | 3* | 4 | 5 | 6 | 7* | 8* | 9 | 10 |
|-----------|----|-----|----|---|---|---|----|----|---|----|
| Actual    | 0  | 0   | 0  | 0 | 1 | 1 | 1  | 1  | 2 | 2  |
| Predicted | 0  | 0   | 1  | 0 | 1 | 1 | 0  | 2  | 2 | 2  |
| no.       | 11 | 12* |    |   |   |   |    |    |   |    |
| Actual    | 2  | 2   |    |   |   |   |    |    |   |    |
| Predicted | 2  | 1   |    |   |   |   |    |    |   |    |

$$C = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{pmatrix}, \quad n = (4,4,4). \quad (15)$$

$$r = (3/4, 2/4, 3/4) = (0.7500, 0.5000, 0.7500), \quad p = (0.3333, 0.3333, 0.3333).$$

so  $Acc = 8/12 = 0.6667$  and  $BA = (0.7500 + 0.5000 + 0.7500)/3 = 0.6667$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.5774, 0.5774, 0.5774), \quad S = 1.7321, \quad w = (0.3333, 0.3333, 0.3333), \\ tBA_{0.5} = 0.3333(0.7500) + 0.3333(0.5000) + 0.3333(0.7500) = 0.6667.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.7321, 1.7321, 1.7321), \quad S = 5.1962, \quad w = (0.3333, 0.3333, 0.3333), \\ tBA_{-0.5} = 0.3333(0.7500) + 0.3333(0.5000) + 0.3333(0.7500) = 0.6667.$$

Observation. As in Sample 1, all columns equal 0.6667; the balanced-data invariance holds for  $K > 2$ . Suggested:  $t = 0$ ,  $tBA_0 = 0.6667$  (balanced).

### 3.6 Sample 6: Three classes, imbalanced (5:3:2)

The  $N = 10$  labels are listed in Table 6 (4 misclassified). The confusion matrix and supports are

**Table 6. Sample 6: actual and predicted labels (\* = misclassified).**

| no.       | 1* | 2 | 3 | 4* | 5* | 6 | 7 | 8 | 9 | 10* |
|-----------|----|---|---|----|----|---|---|---|---|-----|
| Actual    | 0  | 1 | 2 | 0  | 0  | 1 | 2 | 0 | 1 | 0   |
| Predicted | 1  | 1 | 2 | 2  | 1  | 1 | 2 | 0 | 1 | 1   |

$$C = \begin{pmatrix} 1 & 3 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad n = (5, 3, 2). \quad (16)$$

$$r = (1/5, 3/3, 2/2) = (0.2000, 1.0000, 1.0000), \quad p = (0.5000, 0.3000, 0.2000).$$

so  $Acc = 6/10 = 0.6000$  and  $BA = (0.2000 + 1.0000 + 1.0000)/3 = 0.7333$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.7071, 0.5477, 0.4472), \quad S = 1.7020, \quad w = (0.4154, 0.3218, 0.2628), \\ tBA_{0.5} = 0.4154(0.2000) + 0.3218(1.0000) + 0.2628(1.0000) = 0.6676.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.4142, 1.8257, 2.2361), \quad S = 5.4760, \quad w = (0.2583, 0.3334, 0.4083), \\ tBA_{-0.5} = 0.2583(0.2000) + 0.3334(1.0000) + 0.4083(1.0000) = 0.7934.$$

Observation. The weak class is the largest one ( $r_0 = 1/5$ ,  $r_1 = r_2 = 1$ ), so accuracy (0.6000) is below BA (0.7333) and the row increases with falling  $t$ , to 0.8452 at  $t = -1$ . A negative  $t$  would reward a defect, namely predicting class 1 for four of five class-0 samples; the increasing slope is the diagnostic.  $t = 0.5$  (0.6676) keeps the failure visible. Suggested:  $t = 0.5$ ,  $tBA_{0.5} = 0.6676$  (the majority class is the weak one).

### 3.7 Sample 7: Four classes, imbalanced (10:5:3:2)

The  $N = 20$  labels are listed in Table 7 (4 misclassified). The confusion matrix and supports are

**Table 7. Sample 7: actual and predicted labels (\* = misclassified).**

| no.       | 1  | 2  | 3  | 4  | 5   | 6  | 7  | 8   | 9  | 10* |
|-----------|----|----|----|----|-----|----|----|-----|----|-----|
| Actual    | 0  | 0  | 0  | 0  | 0   | 0  | 0  | 0   | 0  | 0   |
| Predicted | 0  | 0  | 0  | 0  | 0   | 0  | 0  | 0   | 0  | 1   |
| no.       | 11 | 12 | 13 | 14 | 15* | 16 | 17 | 18* | 19 | 20* |
| Actual    | 1  | 1  | 1  | 1  | 1   | 2  | 2  | 2   | 3  | 3   |
| Predicted | 1  | 1  | 1  | 1  | 0   | 2  | 2  | 0   | 3  | 1   |

$$C = \begin{pmatrix} 9 & 1 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad n = (10, 5, 3, 2). \quad (17)$$

$$r = (9/10, 4/5, 2/3, 1/2) = (0.9000, 0.8000, 0.6667, 0.5000), \quad p = (0.5000, 0.2500, 0.1500, 0.1000).$$

so  $Acc = 16/20 = 0.8000$  and  $BA = (0.9000 + 0.8000 + 0.6667 + 0.5000)/4 = 0.7167$ .

For  $t = 0.5$ :

$$p^{0.5} = (0.7071, 0.5000, 0.3873, 0.3162), \quad S = 1.9106, \quad w = (0.3701, 0.2617, 0.2027, 0.1655), \\ tBA_{0.5} = 0.3701(0.9000) + 0.2617(0.8000) + 0.2027(0.6667) + 0.1655(0.5000) = 0.7603.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.4142, 2.0000, 2.5820, 3.1623), \quad S = 9.1585, \quad w = (0.1544, 0.2184, 0.2819, 0.3453), \\ tBA_{-0.5} = 0.1544(0.9000) + 0.2184(0.8000) + 0.2819(0.6667) + 0.3453(0.5000) = 0.6743.$$

Observation. Recall falls with class size, (0.9000, 0.8000, 0.6667, 0.5000), giving a smooth decreasing row from 0.8000 to 0.6373. At  $t = -1$  the two smallest classes (5 of 20 samples) carry 74% of the weight. With no class singled out,  $t = 0$  (0.7167) is the default; if the 2-sample class were critical with  $c = 3$ ,  $t = \ln 3 / \ln(0.1/0.5) = -0.6826$  gives 0.6600. Suggested:  $t = 0$ ,  $tBA_0 = 0.7167$  (no class singled out).

### 3.8 Sample 8: Three classes, one class absent from the test set (6:4:0)

The  $N = 10$  labels are listed in Table 8 (3 misclassified). The confusion matrix and supports are

**Table 8. Sample 8: actual and predicted labels (\* = misclassified).**

| no.       | 1 | 2 | 3 | 4 | 5* | 6* | 7 | 8 | 9 | 10* |
|-----------|---|---|---|---|----|----|---|---|---|-----|
| Actual    | 0 | 0 | 0 | 0 | 0  | 0  | 1 | 1 | 1 | 1   |
| Predicted | 0 | 0 | 0 | 0 | 1  | 2  | 1 | 1 | 1 | 0   |

$$C = \begin{pmatrix} 4 & 1 & 1 \\ 1 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad n = (6, 4, 0). \quad (18)$$

Class 2 has  $n_2 = 0$ , so  $r_2 = 0/0$  is undefined; with Laplace smoothing  $\tilde{r}_k = (C_{kk} + 1)/(n_k + 2)$ ,

$$\tilde{r} = (5/8, 4/6, 1/2) = (0.6250, 0.6667, 0.5000),$$

and the prevalence is taken from the full dataset,  $p = (0.60, 0.30, 0.10)$ , since the fold prevalence  $p_2 = 0$  would make  $p_2^t$  undefined for  $t < 0$ . Dropping the absent class instead would give  $BA = (0.6667 + 0.7500)/2 = 0.7083$  over  $K' = 2$  classes, not comparable with folds where  $K = 3$ . With smoothing,  $Acc = 0.7000$  and a-BA = 0.5972.

For  $t = 0.5$ :

$$p^{0.5} = (0.7746, 0.5477, 0.3162), \quad S = 1.6385, \quad w = (0.4727, 0.3343, 0.1930),$$

$$tBA_{0.5} = 0.4727(0.6250) + 0.3343(0.6667) + 0.1930(0.5000) = 0.6148.$$

For  $t = -0.5$ :

$$p^{-0.5} = (1.2910, 1.8257, 3.1623), \quad S = 6.2790, \quad w = (0.2056, 0.2908, 0.5036),$$

$$tBA_{-0.5} = 0.2056(0.6250) + 0.2908(0.6667) + 0.5036(0.5000) = 0.5742.$$

Observation. With smoothing and fixed prevalence the absent class contributes  $\tilde{r}_2 = 0.5$  at weight  $w_2(t)$ , so the value (0.5972 at  $t = 0$ ) is on the same scale as folds that contain all classes and  $K$  stays fixed. The suggested  $t$  is whatever the rest of the experiment uses. Suggested:  $t = 0$ ,  $tBA_0 = 0.5972$  (same  $t$  as other folds,  $a = 1$ ,  $p$  from full data).

**Table 9. Labelled samples: tBA at five values of  $t$ , suggested  $t$ , and tBA at that value. Sample 8 uses the full-dataset prevalence and  $a = 1$ .**

| S. | case            | $t = 1$ | $t = 0.5$ | $t = 0$ | $t = -0.5$ | $t = -1$ | sugg. $t$ | tBA    |
|----|-----------------|---------|-----------|---------|------------|----------|-----------|--------|
| 1  | 5:5             | 0.7000  | 0.7000    | 0.7000  | 0.7000     | 0.7000   | 0         | 0.7000 |
| 2  | 7:3             | 0.5000  | 0.4317    | 0.3571  | 0.2826     | 0.2143   | 0.5       | 0.4317 |
| 3  | 18:2 majority   | 0.9000  | 0.7500    | 0.5000  | 0.2500     | 0.1000   | -0.7325   | 0.1667 |
| 4  | 17:3 over-pred. | 0.7000  | 0.7515    | 0.8235  | 0.8956     | 0.9471   | 0.5       | 0.7515 |
| 5  | 4:4:4           | 0.6667  | 0.6667    | 0.6667  | 0.6667     | 0.6667   | 0         | 0.6667 |
| 6  | 5:3:2           | 0.6000  | 0.6676    | 0.7333  | 0.7934     | 0.8452   | 0.5       | 0.6676 |
| 7  | 10:5:3:2        | 0.8000  | 0.7603    | 0.7167  | 0.6743     | 0.6373   | 0         | 0.7167 |
| 8  | 6:4:0 absent    | 0.6250  | 0.6148    | 0.5972  | 0.5742     | 0.5509   | 0         | 0.5972 |

#### 4 Hand-Calculated Cases from Confusion Matrices

Having established the computation on labelled samples in Section 3, this section presents nine further cases that start directly from a confusion matrix, which allows larger  $N$  and more extreme imbalance. Each case gives the confusion matrix, the recalls, and the value of  $Acc$ ,  $BA$ , and  $tBA_t$  at  $t \in \{1, 0.5, 0, -0.5, -1\}$ . The first four cases are binary, the next four are multi-class, and the final case concerns the ranking of two competing classifiers. The confusion matrices are written with rows as true classes and columns as predicted classes. Intermediate quantities are rounded to four decimal places.

##### 4.1 Case 1: Binary, 95:5, Ordinary Classifier

Consider  $N = 100$  with 95 negatives and 5 positives,  $p = (0.95, 0.05)$ , and

$$C = \begin{pmatrix} 93 & 2 \\ 2 & 3 \end{pmatrix}. \quad (19)$$

The recalls are  $r_- = 93/95 = 0.9789$  and  $r_+ = 3/5 = 0.6000$ , so

$$Acc = (93 + 3)/100 = 0.9600,$$

$$BA = (0.9789 + 0.6000)/2 = 0.7895. \quad (20)$$

For  $t = 0.5$ :  $p^{0.5} = (0.9747, 0.2236)$  with sum 1.1983, giving  $w = (0.8134, 0.1866)$  and

$$tBA_{0.5} = 0.8134(0.9789) + 0.1866(0.6000) = 0.9082. \quad (21)$$

For  $t = -0.5$ :  $p^{-0.5} = (1.0260, 4.4721)$  with sum 5.4981, giving  $w = (0.1866, 0.8134)$ , which is the mirror image of the  $t = 0.5$  weights, and

$$tBA_{-0.5} = 0.1866(0.9789) + 0.8134(0.6000) = 0.6707. \quad (22)$$

For  $t = -1$ :  $p^{-1} = (1.0526, 20)$  with sum 21.0526, giving  $w = (0.05, 0.95)$ , exactly the swapped prevalences, and  $tBA_{-1} = 0.05(0.9789) + 0.95(0.6000) = 0.6189$ . It can be observed that  $tBA_{-1}$  is the accuracy the same

classifier would obtain if the class proportions were reversed, which gives  $t = -1$  a concrete interpretation in the binary case. The full row is given in Table 13.

#### 4.2Case 2: Binary, 95:5, Majority-Class Predictor

Same prevalences, but the classifier predicts negative for every sample:

$$C = \begin{pmatrix} 95 & 0 \\ 5 & 0 \end{pmatrix}, \quad r = (1,0). \quad (23)$$

Then  $Acc = 0.95$  and  $BA = 0.50$ . By Property 2,  $tBA_t = w_1(t)$ :

$$tBA_{0.5} = 0.8134, \quad tBA_{-0.5} = 0.1866, \quad tBA_{-1} = 0.0500. \quad (24)$$

Accuracy says the classifier is excellent, BA says it is at chance, and  $tBA_{-0.5}$  says it is far below chance. The evaluator can decide how harshly the collapse should be scored by choosing  $t$ ; the answer is not fixed by the metric.

#### 4.3Case 3: Binary, 95:5, Minority Over-Prediction

Now suppose the classifier predicts positive for 30 samples, which include all 5 true positives:

$$C = \begin{pmatrix} 70 & 25 \\ 0 & 5 \end{pmatrix}, \quad r = (0.7368,1). \quad (25)$$

Then  $Acc = 0.7500$  and  $BA = 0.8684$ . With the weights from Case 1,

$$\begin{aligned} tBA_{0.5} &= 0.8134(0.7368) + 0.1866(1) = 0.7859, \\ tBA_{-0.5} &= 0.1866(0.7368) + 0.8134(1) = 0.9509. \end{aligned} \quad (26)$$

This is the only binary case in which  $tBA_t$  decreases as  $t$  increases toward accuracy, because the minority recall exceeds the majority recall (Property 4). Comparing Cases 2 and 3 at  $t = 0$ , BA rates the over-predicting classifier (0.8684) far above the majority predictor (0.5000), whereas accuracy ranks them in the opposite order (0.75 versus 0.95).  $tBA$  at  $t = 0.5$  gives 0.7859 versus 0.8134, which is close to a tie, so an evaluator who considers 25 false alarms out of 95 negatives to be as bad as 5 missed positives would set  $t$  near 0.5; an evaluator who does not would set  $t \leq 0$ .

#### 4.4Case 4: Binary, 999:1, Detected Versus Missed Positive

Consider  $N = 1000$  with a single positive sample,  $p = (0.999,0.001)$ , and 9 false positives in both scenarios. In scenario (a) the positive is detected, in scenario (b) it is missed:

$$C^{(a)} = \begin{pmatrix} 990 & 9 \\ 0 & 1 \end{pmatrix}, \quad C^{(b)} = \begin{pmatrix} 990 & 9 \\ 1 & 0 \end{pmatrix}. \quad (27)$$

In both,  $r_- = 990/999 = 0.9910$ ;  $r_+ = 1$  in (a) and  $r_+ = 0$  in (b). The weights are:  $t = 0.5$ ,  $p^{0.5} = (0.9995,0.0316)$ ,  $w = (0.9693,0.0307)$ ;  $t = -0.5$ ,  $p^{-0.5} = (1.0005,31.6228)$ ,  $w = (0.0307,0.9693)$ . The results are given in Table 10.

**Table 10. Case 4: one positive in 1000 samples, detected (a) or missed (b).**

|              | $Acc (t = 1)$ | $t = 0.5$ | $BA (t = 0)$ | $t = -0.5$ | $t = -1$ |  |
|--------------|---------------|-----------|--------------|------------|----------|--|
| (a) detected | 0.9910        | 0.9913    | 0.9955       | 0.9997     | 1.0000   |  |
| (b) missed   | 0.9900        | 0.9606    | 0.4955       | 0.0304     | 0.0010   |  |
| difference   | 0.0010        | 0.0307    | 0.5000       | 0.9693     | 0.9990   |  |

Accuracy separates the two scenarios by 0.001, which is invisible in any reported table. BA separates them by exactly 0.5, which is the maximum any single class can contribute under equal weighting.  $tBA$  lets the evaluator place the separation anywhere in between or beyond: at  $t = 0.5$  the difference is 0.0307, at  $t = -0.5$  it is 0.9693. If a missed positive at this prevalence is worth ten missed negatives, (9) gives  $t = \ln 10 / \ln(0.001/0.999) = 2.3026/(-6.9068) = -0.3334$ .

#### 4.5Case 5: Three Classes, 80:15:5

Consider  $N = 100$ ,  $p = (0.80,0.15,0.05)$ , and

$$C = \begin{pmatrix} 76 & 3 & 1 \\ 4 & 10 & 1 \\ 2 & 2 & 1 \end{pmatrix}, \quad r = (0.9500,0.6667,0.2000). \quad (28)$$

Then  $Acc = 87/100 = 0.8700$  and  $BA = (0.95 + 0.6667 + 0.2)/3 = 0.6056$ . The weights and values are computed as follows.

For  $t = 0.5$ :  $p^{0.5} = (0.8944,0.3873,0.2236)$ , sum 1.5053,  $w = (0.5942,0.2573,0.1485)$ ,

$$\begin{aligned} tBA_{0.5} &= 0.5942(0.95) + 0.2573(0.6667) + 0.1485(0.2) \\ &= 0.7657. \end{aligned} \quad (29)$$

For  $t = -0.5$ :  $p^{-0.5} = (1.1180,2.5820,4.4721)$ , sum 8.1722,  $w = (0.1368,0.3159,0.5472)$ ,

$$\begin{aligned} tBA_{-0.5} &= 0.1368(0.95) + 0.3159(0.6667) + 0.5472(0.2) \\ &= 0.4501. \end{aligned} \quad (30)$$

For  $t = -1$ :  $p^{-1} = (1.25, 6.6667, 20)$ , sum 27.9167,  $w = (0.0448, 0.2388, 0.7164)$ ,  $tBA_{-1} = 0.3450$ .

At  $t = 0.5$  the 5% class carries 14.85% of the weight, roughly three times its prevalence; at  $t = -0.5$  it carries 54.72%, more than half. Note that in the multi-class case  $t = -1$  no longer corresponds to swapped prevalences: the weights  $(0.0448, 0.2388, 0.7164)$  are proportional to  $1/p_k$ , which is a harmonic reweighting rather than a permutation. The majority-class predictor on this dataset scores 0.8000, 0.5942, 0.3333, 0.1368, 0.0448 at  $t = 1, 0.5, 0, -0.5, -1$  respectively, by (8).

#### 4.6 Case 6: Five Classes, Collapsed Minority Versus Uniform Mediocrity

This case illustrates a structural weakness of BA: because it is an arithmetic mean, the total collapse of one class costs at most  $1/K$ , which becomes small when  $K$  is large. Let  $p = (0.50, 0.20, 0.15, 0.10, 0.05)$  and compare two classifiers:

$$\begin{aligned} r^{(a)} &= (0.90, 0.85, 0.80, 0.70, 0.00), \\ r^{(b)} &= (0.80, 0.75, 0.70, 0.65, 0.60). \end{aligned} \quad (31)$$

Classifier (a) is strong on four classes and never predicts the fifth; classifier (b) is mediocre everywhere but never collapses. The weights are, for  $t = 0.5$ ,

$$\begin{aligned} p^{0.5} &= (0.7071, 0.4472, 0.3873, 0.3162, 0.2236), \quad S = 2.0815, \\ w &= (0.3397, 0.2149, 0.1861, 0.1519, 0.1074), \end{aligned}$$

and for  $t = -0.5$ ,

$$\begin{aligned} p^{-0.5} &= (1.4142, 2.2361, 2.5820, 3.1623, 4.4721), \quad S = 13.8667, \\ w &= (0.1020, 0.1613, 0.1862, 0.2280, 0.3225). \end{aligned}$$

Table 11 gives the results.

**Table 11. Case 6: collapsed minority (a) versus uniform mediocrity (b),  $K = 5$ .**

|              | $Acc(t=1)$ | $t=0.5$ | $BA(t=0)$ | $t=-0.5$ | $t=-1$ |
|--------------|------------|---------|-----------|----------|--------|
| (a) collapse | 0.8100     | 0.7436  | 0.6500    | 0.5374   | 0.4210 |
| (b) uniform  | 0.7500     | 0.7264  | 0.7000    | 0.6746   | 0.6531 |

Accuracy prefers (a) by 6 points. BA prefers (b), but only by 5 points, because the collapsed class costs (a) just  $1/K = 0.2$  of the BA scale. At  $t = -0.5$  the gap widens to 13.7 points, and at  $t = -1$  to 23.2 points. An evaluator for whom a fully ignored class is disqualifying would report  $tBA_{-1}$  or lower; an evaluator who only requires that the collapse be noticed would report BA.

#### 4.7 Case 7: Four Classes, Intrusion-Detection Profile

Consider a profile typical of network intrusion datasets: one benign class and three attack classes of decreasing frequency,  $p = (0.85, 0.08, 0.05, 0.02)$ , with recalls  $r = (0.98, 0.90, 0.80, 0.50)$ . Then  $Acc = 0.9550$  and  $BA = 0.7950$ . The weights at  $t = -1$  and  $t = -2$  are

$$\begin{aligned} p^{-1} &= (1.1765, 12.5, 20, 50), \quad S = 83.6765, \quad w = (0.0141, 0.1494, 0.2390, 0.5975), \\ p^{-2} &= (1.3841, 156.25, 400, 2500), \quad S = 3057.63, \quad w = (0.0005, 0.0511, 0.1308, 0.8176). \end{aligned}$$

This gives

$$\begin{aligned} tBA_{0.5} &= 0.8967, \quad tBA_{-0.5} = 0.7027, \\ tBA_{-1} &= 0.6382, \quad tBA_{-2} = 0.5599. \end{aligned} \quad (32)$$

At  $t = -2$  the rarest attack class, which has 2% prevalence and 50% recall, carries 81.76% of the weight, and the score is close to that class's recall alone. Between  $t = -1$  and  $t = -2$  the benign class weight drops from 1.4% to 0.05%, so the score stops depending on benign recall entirely. This illustrates the practical range of  $t$ : values below about  $-1$  are useful only when the rarest class is the sole quantity of interest, in which case reporting  $r_{min}$  directly is clearer.

#### 4.8 Case 8: Ranking Reversal Between Two Classifiers

The most consequential use of a metric is to rank competing classifiers. Consider  $N = 100$ ,  $p = (0.90, 0.10)$ , and two classifiers:

$$C^{(1)} = \begin{pmatrix} 88 & 2 \\ 6 & 4 \end{pmatrix}, \quad C^{(2)} = \begin{pmatrix} 80 & 10 \\ 1 & 9 \end{pmatrix}. \quad (33)$$

Classifier 1 is conservative, with  $r^{(1)} = (0.9778, 0.4000)$  and Classifier 2 is aggressive, with  $r^{(2)} = (0.8889, 0.9000)$ . Since  $p^{0.5} = (0.9487, 0.3162)$  with sum 1.2649, the weights at  $t = 0.5$  are  $w = (0.75, 0.25)$  and at  $t = -0.5$  they are  $w = (0.25, 0.75)$ . Table 12 gives the results.

**Table 12. Case 8: ranking of a conservative (1) and an aggressive (2) classifier.**

|              | $Acc(t=1)$    | $t=0.5$       | $BA(t=0)$     | $t=-0.5$      | $t=-1$        |
|--------------|---------------|---------------|---------------|---------------|---------------|
| Classifier 1 | <b>0.9200</b> | 0.8333        | 0.6889        | 0.5444        | 0.4578        |
| Classifier 2 | 0.8900        | <b>0.8917</b> | <b>0.8944</b> | <b>0.8972</b> | <b>0.8989</b> |

Accuracy ranks Classifier 1 first; every other column ranks Classifier 2 first. Setting  $tBA_t^{(1)} = tBA_t^{(2)}$  and solving numerically gives the crossover at  $t^* = 0.786$ . Therefore, the two metrics in common use, accuracy and BA, disagree on this pair, and the disagreement is resolved only by stating where on the  $[0,1]$  interval the evaluator's cost model lies. tBA makes this statement explicit: any  $t < 0.786$  prefers Classifier 2. It can also be observed that  $tBA_t^{(2)}$  is nearly flat in  $t$  (from 0.8900 to 0.8989), because its two recalls are nearly equal, whereas  $tBA_t^{(1)}$  varies by 46 points; a flat tBA curve is itself evidence that a classifier's performance is uniform across classes. Fig. 1 plots  $tBA_t$  against  $t$  for Cases 1, 2, 3, 5, and 8.

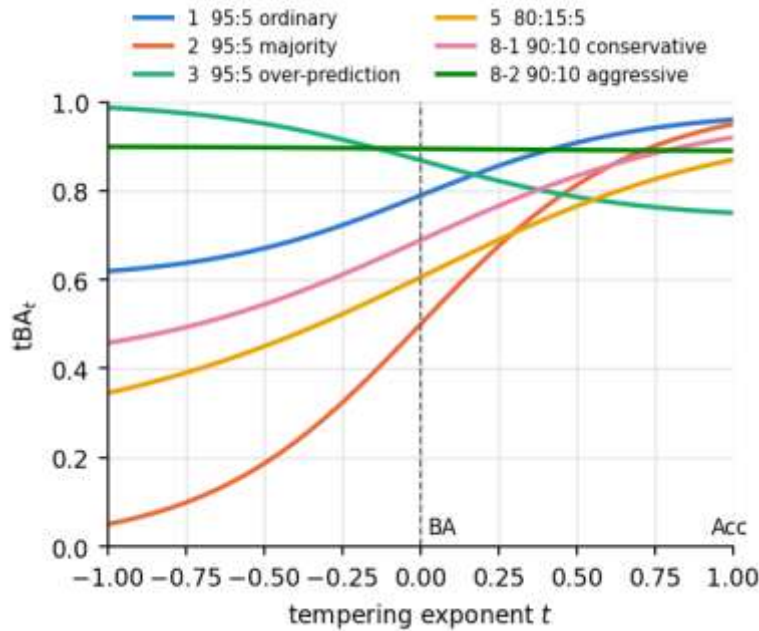


**Fig. 1.**  $tBA_t$  as a function of  $t$ . The dashed vertical line at  $t = 0$  is BA;  $t = 1$  is accuracy. Case 3 is the only decreasing curve, because its minority recall exceeds its majority recall. The two Case 8 curves cross at  $t^* = 0.786$ .

#### 4.9Case 9: Choosing $t$ from a Stated Cost Ratio

Property 3 is applied to Case 5. Suppose a domain expert states that an error on the 5% class should count three times as much as an error on the 80% class. Then

$$t = \frac{\ln 3}{\ln(0.05/0.80)} = \frac{1.0986}{-2.7726} = -0.3962. \quad (34)$$

The weights at  $t = -0.3962$  are

$$p^t = (1.0924, 2.1206, 3.2773), \quad S = 6.4904, \quad w = (0.1683, 0.3267, 0.5049),$$

and  $w_3/w_1 = 3.0000$  as required. The middle class receives  $w_2/w_1 = 1.94$ , which follows automatically from the power law. Substituting the recalls of Case 5 gives

$$\begin{aligned} tBA_{-0.3962} &= 0.1683(0.95) + 0.3267(0.6667) + 0.5049(0.2) \\ &= 0.4787. \end{aligned} \quad (35)$$

By contrast, BA would have imposed a cost ratio of exactly 1 between all classes, and accuracy a ratio of 0.0625, neither of which was what the expert asked for.

#### 4.10 Summary

Table 13 collects all cases. Three observations follow. First, accuracy and BA can rank the same pair of classifiers in opposite orders (Case 8), and tBA locates the crossover. Second, BA's penalty for ignoring a class is capped at  $1/K$  (Case 6), whereas tBA's penalty is adjustable and, for  $t < 0$ , larger. Third, for a rare and critical class (Case 4), BA's fixed weight of 0.5 may itself be too small;  $t < 0$  raises it, with the required value computable in closed form from the cost ratio.

**Table 13. Summary of all cases:  $tBA_t$  at five values of  $t$ .**

| Case | $p$  | Acc    | $t = 0.5$ | BA     | $t = -0.5$ | $t = -1$ |
|------|------|--------|-----------|--------|------------|----------|
| 1    | 95:5 | 0.9600 | 0.9082    | 0.7895 | 0.6707     | 0.6189   |

|     |               |        |        |        |        |        |
|-----|---------------|--------|--------|--------|--------|--------|
| 2   | 95:5          | 0.9500 | 0.8134 | 0.5000 | 0.1866 | 0.0500 |
| 3   | 95:5          | 0.7500 | 0.7859 | 0.8684 | 0.9509 | 0.9868 |
| 4a  | 999:1         | 0.9910 | 0.9913 | 0.9955 | 0.9997 | 1.0000 |
| 4b  | 999:1         | 0.9900 | 0.9606 | 0.4955 | 0.0304 | 0.0010 |
| 5   | 80:15:5       | 0.8700 | 0.7657 | 0.6056 | 0.4501 | 0.3450 |
| 6a  | 50:20:15:10:5 | 0.8100 | 0.7436 | 0.6500 | 0.5374 | 0.4210 |
| 6b  | 50:20:15:10:5 | 0.7500 | 0.7264 | 0.7000 | 0.6746 | 0.6531 |
| 7   | 85:8:5:2      | 0.9550 | 0.8967 | 0.7950 | 0.7027 | 0.6382 |
| 8-1 | 90:10         | 0.9200 | 0.8333 | 0.6889 | 0.5444 | 0.4578 |
| 8-2 | 90:10         | 0.8900 | 0.8917 | 0.8944 | 0.8972 | 0.8989 |

## 5 Discussion

Recommended default. When no cost information is available, we suggest reporting  $tBA_t$  at  $t \in \{1, 0.5, 0, -0.5\}$ , i.e. accuracy, an intermediate value, BA, and one value beyond BA, in a single row, and choosing the headline value by the rules of Section 2.4:  $t = 0$  when classes are balanced or equally important,  $t = 0.5$  when the majority class itself matters or the minority evidence is thin, and the cost-ratio value when a rare class is critical. This costs nothing beyond the confusion matrix and communicates the full shape of the curve in Fig. 1. When cost information is available,  $t$  should be fixed in advance by (9) and stated in the paper.

The slope of the row is diagnostic. By Property 4 the sign of  $\partial tBA_t / \partial t$  is the sign of the covariance between  $\ln p_k$  and  $r_k$ . Across the labelled samples, a decreasing row (Samples 2, 3, 7) means small classes are under-recalled, the usual failure on imbalanced data; an increasing row (Samples 4, 6) means the classifier is over-committing to small classes at the expense of the large one, as produced by aggressive resampling or a threshold shifted too far; a flat row (Samples 1, 5) means uniform performance. Samples 2 and 6 have similar imbalance ratios and opposite slopes, and only the slope tells the two situations apart. None of this is visible from accuracy or BA alone.

Small-sample caution. With supports as small as  $n_1 = 2$  (Sample 3),  $r_1$  can only be 0, 0.5 or 1, and one sample moves  $tBA_{-1}$  by 0.45. BA has the same sensitivity since it also gives the small class weight  $1/K$ . On real test sets any  $t \leq 0$  value should be accompanied by a confidence interval or repeated cross-validation.

Relation to other metrics. tBA shares with BA the limitation that it uses only the per-class recalls, i.e. the row-normalized confusion matrix, and therefore ignores precision. Case 3 shows that tBA at  $t \leq 0$  can still reward heavy over-prediction of the minority class; MCC,  $\kappa$ , and the precision-recall AUC are complementary for that failure mode. tBA is not a proper scoring rule and is computed at a fixed operating point, like accuracy and BA. A reference implementation is provided (see Program Availability). The parameter  $t$  is analogous to the temperature  $\tau$  of logit adjustment [8]: both rescale the influence of  $\ln p_k$ , one at training time and one at evaluation time, and a natural practice is to evaluate with the same  $t$  that was used to train.

Missing classes. When a class is absent from a test fold,  $r_k$  is undefined and BA implementations silently drop the class, changing  $K$  between folds. Because tBA computes  $p_k$  from the full dataset, the weights are unaffected; the undefined  $r_k$  can be replaced by the Laplace-smoothed estimate  $\tilde{r}_k = (C_{kk} + a) / (n_k + 2a)$  with  $a = 1$ , which equals 0.5 for an absent class and keeps the denominator of (5) constant across folds.

## 6 Conclusions

This paper proposed the Tempered Balanced Accuracy (tBA), a weighted mean of per-class recalls with weights proportional to the class prevalence raised to an exponent  $t$ . Standard accuracy and balanced accuracy are recovered at  $t = 1$  and  $t = 0$  respectively, and values of  $t < 0$  weight rare classes above parity. The chance level of tBA is  $1/K$  for every  $t$ , the score of a majority-class predictor is a closed-form increasing function of  $t$ , and  $t$  can be set from a desired cost ratio as  $t = \ln c / \ln(p_k/p_j)$ . Nine hand-calculated cases showed that accuracy and BA can disagree on the ranking of two classifiers, with the crossover at  $t^* = 0.786$  in the example given; that BA's penalty for ignoring one of  $K$  classes is capped at  $1/K$  while tBA's is adjustable; and that at 999:1 imbalance the difference between detecting and missing the single positive is 0.001 under accuracy, 0.5 under BA, and 0.9693 under  $tBA_{-0.5}$ . Future work will evaluate tBA on public imbalanced benchmarks and study the choice of  $t$  as a function of the imbalance ratio.

### Program Availability

A Python implementation of tBA, together with the script that reproduces every hand-calculated case in Section 4 and the synthetic simulation, is available at <https://github.com/hilalnuha/tBA>.

## References

- [1] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [2] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 3121–3124.
- [3] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th Int. Conf. Machine Learning (ICML)*, 1997, pp. 179–186.
- [4] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.
- [5] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
- [6] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.
- [7] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2001, pp. 973–978.
- [8] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
- [9] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277.