Knowledge-Based Systems 210 (2020) 106490

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Imbalance accuracy metric for model selection in multi-class imbalance classification problems

Ebrahim Mortaz *

Lubin school of business, Pace University, One Pace Plaza, New York, NY 10038, USA

ARTICLE INFO

Article history:
Received 28 May 2020
Received in revised form 17 July 2020
Accepted 25 September 2020
Available online 28 September 2020

Keywords:
Machine learning
Classification accuracy
Multi-class problems
Imbalance datasets
Knowledge discovery

ABSTRACT

The overall accuracy, macro precision, macro recall, F-score and class balance accuracy, due to their simplicity and easy interpretation, have been among the most popular metrics to measure the performance of classifiers on multi-class problems. However, on imbalance datasets, some of these metrics can be unfairly influenced by heavier classes. Therefore, it is recommended that they are used as a group and not individually. This strategy can unnecessarily complicate the model selection and evaluation in imbalance datasets. In this paper, we introduce a new metric, imbalance accuracy metric (IAM), that can be used as a solo measure for model evaluation and selection. The IAM is built up on top of the existing metrics, is simple to use, and easy to interpret. This metric is meant to be used as a bottom-line measure aiming to eliminate the need for group metric computation and simplify the model selection.

1. Introduction

Most of the real-world datasets are either heavily or moderately imbalance [1]. Numerous techniques [2,3] and open-source libraries [4,5] have been developed to help the analysts train and test classifiers for such imbalance binary, multi-class and multi-label datasets. Among the learning techniques for imbalance datasets, some of the techniques use under- or over-sampling methods to balance out the data [6,7], some modify the loss functions in the learning algorithms to accommodate the imbalance conditions [8], and some deploy an ensemble of algorithms [9]. Nevertheless, in evaluating the performance of classifiers for multi-class datasets, defining an ultimate go-to accuracy metric, which is simple to compute and easy to interpret, remains a challenge. Some of the existing metrics can misleadingly distort the results, such as overall accuracy, and some do not offer a complete account of the performance, such as macro precision or macro recall when used individually. Hence, the analysts are advised to consider multiple metrics in their evaluations to steer clear of potential controversies and obtain a complete account of the true performance of the classifiers. With multiple metric computation scheme, the gap among the metrics is studied and the smallest metric is selected to report as the final accuracy value for model selection [10–12]. This advice enables evaluations from different perspectives and is immune to the controversies manifested in imbalance datasets. However, some issues could still emerge when implementing this scheme. For instance, in model selection, different classifiers could have different lowest metrics and no single classifier could dominate all others on all the metrics in the group. As an example, one classifier presents the lowest overall accuracy, and another presents the lowest class balance accuracy. Which classifier should be selected to be used in the final classifying product? Without the domain knowledge, should we select the one with the majority vote on all metrics? Or it would be better to consider computing more advanced metrics? Multi-class ROC graphs or multiple precision recall curves are recommended in [13]. Another metric is proposed by [14] which is called the *M* measure, a generalization approach that aggregates all pairs of classes based on the characteristics of the AUC (area under ROC curve). Using the extended G-mean (geometric means of recall values of every class) in a cost sensitive boosting algorithm is also proposed in [15]. However, by employing more advanced metrics, the issue of interpretation emerges. After all, the popularity of the commonly used metrics such as overall accuracy is owed mostly to their simplicity. Their simple computation and easy interpretation can leverage against the occasional shortcomings. Therefore, it should be noted that the interpretability of a metric plays an important role to make adoption of a metric appealing.

We provide three examples after the definitions in the methodology section to explain the issues we brought up in model selection. We discuss five most popular and most intuitive accuracy metrics, overall accuracy, macro average precision, macro average recall, F-score and class balance accuracy proposed by [12]. These metrics have been used in most of open-source libraries

* Correspondence to: One Pace Plaza, W490, New York, NY 10038, USA.
E-mail address: emortaz@pace.edu.

https://doi.org/10.1016/j.knosys.2020.106490
0950-7051/

for evaluating the classifiers in imbalance datasets such as [4,5]. There are other measures such as the M measure [13], Matthew's correlation coefficient proposed by [16], relative classifier information proposed by [17], confusion entropy developed by [17] and the ROC curve for multi-class datasets explained in [18] to evaluate the classifiers' performance as well. However, those metrics might not be as popular as the five we study in this paper possibly due to their somewhat counter intuitive structure or limited applications - for instance, applicable to binary classification problems only. The impact of class imbalance in classification performance metrics based on the binary confusion matrix has been extensively studies in [19].

In this paper, we introduce a new metric for evaluating the classifiers on multi-class datasets. This metric is mathematically designed to be lower than the five popular metrics at all times aiming to eliminate the need for multiple metrics computation. This does not mean that the computation of multiple metrics is inaccurate. The new metric is meant to simplify and accelerate the evaluation process. It is built up on top of the five popular metrics, is simple to compute and easy to interpret and consistent with other measures. The metric is primarily designed for multi-class imbalance datasets but it can very well be used in balanced and binary classifications as well. With that, we define two research questions:

1. With so many metrics adopted by the research community, would it be useful to introduce a new metric for evaluating a classifier's performance in multi-class imbalance datasets?
2. What additional benefits this new metric could offer that are not already offered?

In the methodology section, we define the metrics that we study, introduce the imbalance accuracy metric along with its properties, and provide multiple numerical examples to compare the new metric against the existing ones. In Section 3, we study the new metric offerings and use multiple examples to show how to implement the new metric for model selection in practice. Finally in Section 4, we draw the research conclusions and discuss future research opportunities.

2. Methodology

We first define the five metrics that are studied in the paper. We make several examples to explain the issues that might be encountered when we evaluate the classifier performance on multi-class datasets. We then introduce the IAM, and its properties with four Theorems and proofs.

2.1. Definitions

Let us assume that we train a classifier for a k-class dataset. Then, we assume that C^k is the k by k confusion matrix of the classifier and c_{ij} are the elements of C^k where $i, j = 1, 2, \ldots, k$. The rows and columns show the true and predicted values at each class, respectively. Let us define $c_{i.}$, and $c_{.i}$ as (1) and (2):

$$c_{i.} = \sum_{j=1}^{k} c_{ij} \quad \forall i \in \{1, \ldots, k\} \tag{1}$$

$$c_{.i} = \sum_{j=1}^{k} c_{ji} \quad \forall i \in \{1, \ldots, k\} \tag{2}$$

Then, we can define accuracy, macro average precision, macro average recall, F-score and class balance accuracy in (3)–(7) as below:

Table 1
Distribution of y_i for example 1.

Class	1	2	3	4
Frequency	5000	500	100	25

Overall accuracy:

$$\text{ACC} = \frac{\sum_i c_{ii}}{\sum_{ij} c_{ij}} \tag{3}$$

Macro average precision:

$$\text{MAP} = 1/k \sum_{i=1}^{k} \frac{c_{ii}}{c_{.i}} \tag{4}$$

Macro average recall:

$$\text{MAR} = 1/k \sum_{i=1}^{k} \frac{c_{ii}}{c_{i.}} \tag{5}$$

F-score:

$$\text{F} = 1/k \sum_{i=1}^{k} \frac{2\frac{c_{ii}}{c_{.i}}\frac{c_{ii}}{c_{i.}}}{\frac{c_{ii}}{c_{.i}} + \frac{c_{ii}}{c_{i.}}} \tag{6}$$

Class balance accuracy :

$$\text{CBA} = 1/k \sum_{i=1}^{k} \frac{c_{ii}}{\max(c_{i.}, c_{.i})} \tag{7}$$

ACC measures the overall accuracy, which is the percentage of correctly classified instances (sum of diagonal values in confusion matrix) against all instances. MAP and MAR are the average precision and recall for each class. The precision value is the number of correctly classified instances divided by the number of instances labeled by the algorithm as correct, and the recall value is the number of correctly classified instances divided by the number of correct instances in the data. The F-score is the harmonic mean of the precision and recall. Finally, the CBA assesses the classifier's predictive power by dividing the diagonal value (correct instances) by max value of sum of on/off diagonal values.

2.2. Examples

We provide three examples below to discuss the issues that could emerge using the five popular metrics. The first example shows when the multiple metric scheme functions well and examples 2 and 3 reveal a few deficiencies.

Example 1.

Consider $DS_1 = \{(x_i, y_i)\}_{i=1}^{n}$ as a multi-class dataset. We split the dataset into train, validation, and test sets and ask two analysts to train a classifier and report the performance of their classifiers on DS_1. The distribution of y_i for the test set is given in Table 1.

The first analyst using ML_1 produces the confusion matrix, CM_1, on the test set where the rows and columns show the true and predicted values at each class, respectively.

$$CM_1 = \begin{bmatrix} 4900 & 90 & 10 & 0 \\ 255 & 245 & 0 & 0 \\ 45 & 5 & 45 & 5 \\ 11 & 3 & 1 & 10 \end{bmatrix}$$

The accuracy metrics are given in Table 2.

Table 2
Accuracy results for example 1 analyst a.

ACC	MAP	MAR	F-score	CBA
0.92	0.78	0.58	0.65	0.57

Table 3
Accuracy results for example 1 analyst b.

ACC	MAP	MAR	F-score	CBA
0.92	0.76	0.56	0.63	0.55

Table 4
Distribution of y_i for example 2.

Class	1	2	3
Frequency	301	215	202

Table 5
Accuracy results for example 2 analyst a.

ACC	MAP	MAR	F-score	CBA
0.40	0.42	0.41	0.41	0.41

Table 6
Accuracy results for example 2 analyst b.

ACC	MAP	MAR	F-score	CBA
0.41	0.42	0.42	0.42	0.40

With multiple metrics approach, the CBA metric offers the lowest accuracy value and should be used as the metric to report.

b. The second analyst using ML_2 has produced CM_2.

$$CM_2 = \begin{bmatrix} 4900 & 90 & 10 & 0 \\ 250 & 250 & 0 & 0 \\ 50 & 10 & 35 & 5 \\ 9 & 4 & 2 & 10 \end{bmatrix}$$

The accuracy metrics are given in Table 3.

With multiple metrics approach, the CBA metric offers the lowest accuracy value on the test set for both classifiers and should be used as the metric for model selection. The CBA, which offers the lowest value for both classifiers, is higher for ML_1 and thereby ML_1 is selected for production.

Example 2.

a. Consider $DS_2 = \{(x_i, y_i)\}_{i=1}^{n}$ as a multi-class imbalance dataset where the distribution of y_i is given in Table 4. We ask two analysts again to report the performance of classifiers on DS_2. The first analyst produces CM_3 using ML_3 on the test set.

$$CM_3 = \begin{bmatrix} 100 & 102 & 99 \\ 105 & 100 & 10 \\ 102 & 10 & 90 \end{bmatrix}$$

Multiple metrics are computed and shown in Table 5. The ACC metric produces the lowest value.

b. The second analyst produces CM_4 using ML_4 on the test set.

$$CM_4 = \begin{bmatrix} 114 & 86 & 101 \\ 100 & 100 & 15 \\ 110 & 10 & 82 \end{bmatrix}$$

Based on Table 6, this time the CBA offers the lowest accuracy value and should be the performance metric to report.

Example 2 shows that the model selection could become puzzling at times. Problems similar to example 2 validate our first question whether it would be useful to have a metric that offers the smallest value at all times. Example 3 is similar to example 2 with more classifiers.

Table 7
Distribution of y_i for example 3.

Class	1	2	3	4	5
Frequency	2200	2150	1080	650	505

Table 8
Accuracy results for example 3.

Model	ACC	MAP	MAR	F-score	CBA
ML_5	0.37	0.34	0.33	0.33	0.31
ML_6	0.36	0.37	0.33	0.35	0.31
ML_7	0.35	0.35	0.35	0.35	0.30

Example 3.

Consider $DS_3 = \{(x_i, y_i)\}_{i=1}^{n}$ as a multi-class imbalance dataset where the distribution of y_i is given in Table 7. Three classifiers, ML_5 to ML_7, are trained and tested. The accuracy results are given in Table 8.

The results in Table 8 again show that it is difficult to find out what model is outperforming the others.

2.3. Imbalance accuracy metric

Imbalance accuracy metric (IAM) for a k-class dataset is mathematically defined in (8).

$$\text{IAM} = 1/k \sum_{i=1}^{k} \frac{c_{ii} - \max(\sum_{j\neq i}^{k} c_{ij}, \sum_{j\neq i}^{k} c_{ji})}{\max(c_{.i}, c_{i.})} \tag{8}$$

In computing the IAM, the max value of total off-diagonal items ($\sum_{j\neq i}^{k} c_{ji}$ or $\sum_{j\neq i}^{k} c_{ji}$) are subtracted from the diagonal values (c_{ii}), divided by the max sum in the corresponding row or column ($\max(c_{.i}, c_{i.})$), and finally averaged ($/k$) to obtain the expectation.

The IAM is built up on top of the ACC and CBA metrics. However, the interpretation of IAM is different. The ACC is the expectation that a random instance is correctly classified and the CBA, measures the classifier's power in correctly classifying a random instance [12]. The IAM, on the other hand, shows how well a classifier is expected not to classify a random instance in the incorrect classes. The higher the value of IAM, the better the classifier functions in not labeling the instances incorrectly. This interpretation aligns with the conservative nature of the IAM's mathematical presentation. According to the IAM's definition, any class (heavy or not) can have only $\frac{1}{m}\%$ (m number of classes) impact on the metric, and that protects the IAM from being unfairly influenced by the heavier classes.

2.4. Properties of IAM:

We present four Theorems discussing the properties of IAM. It is by Theorem 1 that IAM always varies between -1 and 1. This shows that the IAM is on a different scale from ACC and CBA. It is by Theorems 2–4 that the IAM is always smaller than CBA, MAP, MAR, F-score and ACC metrics and thereby can be used as a bottom-line accuracy metric in model selection.

Theorem 1. $-1 \leq IAM \leq 1$.

Proof. We consider the extreme cases to find the lower and upper bounds for *IAM*. For simplicity, we denote $c_{..}$, $c'_{i.}$ and $c'_{.i}$ by (9)–(11):

$$c_{..} = \sum_{i=1}^{k} \sum_{j=1}^{k} c_{ij} \tag{9}$$

$$c'_{i.} = \sum_{j=1}^{k} c_{ij} \qquad \forall i = 1, \ldots, k \text{ and } j \neq i \tag{10}$$

$$c'_{.i} = \sum_{j=1}^{k} c_{ji} \qquad \forall i = 1, \ldots, k \text{ and } j \neq i \tag{11}$$

Let us first assume $c_{ii} \leq \max(c'_{i.}, c'_{.i})$. With that assumption,

$$c_{ii} \leq \max(c'_{i.}, c'_{.i}) \rightarrow \min(c_{ii} - c'_{i.}, c_{ii} - c'_{.i}) \leq 0 \tag{12}$$

On the other hand by definition:

$c'_{.i} \leq \max(c_{i.}, c_{.i})$, $c'_{i.} \leq \max(c_{i.}, c_{.i})$

Therefore,

$c'_{.i} - c_{ii} \leq \max(c_{i.}, c_{.i})$, $c'_{i.} - c_{ii} \leq \max(c_{i.}, c_{.i})$

Then,

$c_{ii} - c'_{.i} \geq -\max(c_{i.}, c_{.i})$, $c_{ii} - c'_{i.} \geq -\max(c_{i.}, c_{.i}) \rightarrow$

$$\rightarrow \min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.}) \geq -\max(c_{i.}, c_{.i}) \tag{13}$$

From (13):

$$\frac{\min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.})}{\max(c_{i.}, c_{.i})} \geq -1 \tag{14}$$

From both (12) and (14),

$$-1 \leq 1/k \sum_{i=1}^{k} \frac{\min(c_{ii} - c'_{i.}, c_{ii} - c'_{.i})}{\max(c_{.i}, c_{i.})} \leq 0$$

$$\rightarrow -1 \leq 1/k \sum_{i=1}^{k} \frac{c_{ii} - \max(c'_{i.}, c'_{.i})}{\max(c_{.i}, c_{i.})} \leq 0 \tag{15}$$

Now, let us assume $c_{ii} \geq \max(c'_{i.}, c'_{.i})$. With that assumption,

$$c_{ii} \geq \max(c'_{i.}, c'_{.i}) \rightarrow \min(c_{ii} - c'_{i.}, c_{ii} - c'_{.i}) \geq 0 \tag{16}$$

Also, by definition:

$c_{ii} \leq \max(c_{i.}, c_{.i})$, $c'_{.i} \leq \max(c_{i.}, c_{.i})$, $c'_{i.} \leq \max(c_{i.}, c_{.i})$

Then,

$c_{ii} - c'_{.i} \leq \max(c_{i.}, c_{.i})$, $c_{ii} - c'_{i.} \leq \max(c_{i.}, c_{.i})$

And,

$\min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.}) \leq \max(c_{i.}, c_{.i})$

Therefore:

$$\frac{\min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.})}{\max(c_{i.}, c_{.i})} \leq 1 \tag{17}$$

From (16) and (17)

$$0 \leq 1/k \sum_{i=1}^{k} \frac{\min(c_{ii} - c'_{i.}, c_{ii} - c'_{.i})}{\max(c_{.i}, c_{i.})} \leq 1$$

$$\rightarrow 0 \leq 1/k \sum_{i=1}^{k} \frac{c_{ii} - \max(c'_{i.}, c'_{.i})}{\max(c_{.i}, c_{i.})} \leq 1 \tag{18}$$

From (15) and (18), $-1 \leq IAM \leq 1$.

Theorem 2. *IAM ≤ CBA*

Proof. By definition:

$\min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.}) \leq c_{ii}$

Because $\max(c_{i.}, c_{.i}) > 0$,

$$\frac{\min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.})}{\max(c_{i.}, c_{.i})} \leq \frac{c_{ii}}{\max(c_{i.}, c_{.i})}$$

The sum across all classes and averaging the output would yield:

$$1/k \sum_{i=1}^{k} \frac{\min(c_{ii} - c'_{.i}, c_{ii} - c'_{i.})}{\max(c_{i.}, c_{.i})} \leq 1/k \sum_{i=1}^{k} \frac{c_{ii}}{\max(c_{i.}, c_{.i})}$$

Which means IAM ≤ CBA, and that completes the proof.

Theorem 3. *IAM ≤ MAP, IAM ≤ MAR, IAM ≤ F-score*

Proof. By transitive law: As proved in [12], by definition:

CBA ≤ MAP and CBA ≤ MAR

Therefore, by transitive law:

IAM ≤ MAP and IAM ≤ MAR

Also because F-score is the harmonic mean of macro precision and macro recall, and by transitive law:

IAM ≤ min(MAR, MAP) ≤ F-score

Which completes the proof.

Proposition 1. *If $a \leq b$ and $b > 0$ and $c \geq 0$, then $a/b \leq (a+c)/(b+c)$*

Proof. By contradiction: To compare the two fractions in the conclusion statement, we multiply both the denominator and the numerator of each fraction by the denominator of the opposite fraction.
As b, b+c > 0,

$$\text{If } \frac{a}{b} > \frac{a+c}{b+c} \rightarrow \frac{a(b+c)}{b(b+c)} > \frac{b(a+c)}{b(b+c)}$$
$$\rightarrow ab + ac > ab + bc \rightarrow a > b$$

which contradicts the conditional hypothesis. Therefore,

$$\frac{a}{b} \leq \frac{a+c}{b+c}$$

Theorem 4. *IAM ≤ ACC*

To simplify understanding the proof, we present the proof for a binary classification first. We assume that C^2 is the 2 by 2 confusion matrix of the algorithm and c_{ij} are the elements of C^2. We also assume $c_{12} \geq c_{21}$. Note that the last assumption does not have any impact on the validity of the proof and is for simplicity.

$$\text{ACC} = \frac{c_{11} + c_{22}}{c_{11} + c_{12} + c_{21} + c_{22}}$$

$$\text{IAM} = 1/2[\frac{c_{11} - c_{12}}{c_{11} + c_{12}} + \frac{c_{22} - c_{12}}{c_{22} + c_{12}}]$$

Based on Proposition 1:

$$\frac{c_{11} - c_{12}}{c_{11} + c_{12}} \leq \frac{c_{11} - c_{12} + c_{21} + c_{22}}{c_{11} + c_{12} + c_{21} + c_{22}}$$
$$\frac{c_{22} - c_{12}}{c_{22} + c_{12}} \leq \frac{c_{22} - c_{12} + c_{11} + c_{21}}{c_{22} + c_{12} + c_{11} + c_{21}}$$

If we sum the two inequalities then,

$$1/2[\frac{c_{11} - c_{12}}{c_{11} + c_{12}} + \frac{c_{22} - c_{12}}{c_{22} + c_{12}}]$$
$$\leq 1/2[\frac{c_{11} - c_{12} + c_{21} + c_{22} + c_{22} - c_{12} + c_{11} + c_{21}}{c_{11} + c_{21} + c_{12} + c_{22}}]$$

$$\rightarrow IAM \leq 1/2[\frac{2c_{11}-2c_{12}+2c_{21}+2c_{22}}{c_{11}+c_{21}+c_{12}+c_{22}}]$$
$$\rightarrow IAM \leq \frac{c_{11}+c_{22}-(c_{12}-c_{21})}{c_{11}+c_{21}+c_{12}+c_{22}}$$
$$\xrightarrow{c_{12}-c_{21}\geq 0} IAM \leq \frac{c_{11}+c_{22}}{c_{11}+c_{21}+c_{12}+c_{22}}$$

Which shows that IAM $\leq$ ACC and completes the proof for the binary target problems.

Proof. Now, let us assume that we train a classifier to classify a k-class variable ($k > 2$). We denote C^k as the k by k confusion matrix of the classifier and c_{ij} as the elements of C^k. We present the proof under three cases, first when $c'_{i.} \geq c'_{.i} \quad \forall i = 1, \ldots, k$, and second when $c'_{i.} \leq c'_{.i} \quad \forall i = 1, \ldots, k$, and third when neither case holds true.

First case: $c'_{i.} \geq c'_{.i} \quad \forall i = 1, \ldots, k$

By definition:

$$\text{ACC} = \frac{\sum_i c_{ii}}{c_{..}}$$
$$\text{IAM} = 1/k \sum_{i=1}^{k} (\frac{c_{ii}-c'_{i.}}{c_{ii}+c'_{i.}})$$

Based on Proposition 1:

$$\frac{c_{11}-c'_{1.}}{c_{11}+c'_{1.}} \leq \frac{c_{11}-c'_{1.}+(c_{..}-(c_{11}+c'_{1.}))}{c_{11}+c'_{1.}+(c_{..}-(c_{11}+c'_{1.}))}$$
$$\vdots \qquad\qquad \vdots$$
$$\frac{c_{kk}-c'_{k.}}{c_{kk}+c'_{k.}} \leq \frac{c_{kk}-c'_{k.}+(c_{..}-(c_{kk}+c'_{k.}))}{c_{kk}+c'_{k.}+(c_{..}-(c_{kk}+c'_{k.}))}$$

By summing both sides of all fractions and dividing them by k:

$$1/k[\frac{c_{11}-c'_{1.}}{c_{11}+c'_{1.}}+\cdots+\frac{c_{kk}-c'_{k.}}{c_{kk}+c'_{k.}}] \leq 1/k[\frac{c_{11}-c'_{1.}+c_{..}-(c_{11}+c'_{1.})+\cdots+c_{kk}+c'_{k.}+c_{..}-(c_{kk}+c'_{k.})}{c_{..}}]$$

$$\rightarrow \text{IAM} \leq 1/k[\frac{kc_{..}-2(c'_{1.}+\cdots+c'_{k.})}{c_{..}}] \tag{19}$$

We then: $c_{..} \leftarrow \sum_i c_{ii} + \sum_i c'_{i.}$ in the numerator of the right hand side of (19), and obtain:

$$\rightarrow \text{IAM} \leq 1/k[\frac{k(\sum_i c_{ii}+\sum_i c'_{i.})-2\sum_i c'_{i.}}{c_{..}}]$$
$$\rightarrow \text{IAM} \leq \frac{\sum_i c_{ii}+(1-2/k)\sum_i c'_{i.}}{c_{..}}$$
$$\xrightarrow[k>2]{1-2/k>0} \text{IAM} \leq \frac{\sum_i c_{ii}}{c_{..}} \rightarrow \text{IAM} \leq \text{ACC}$$

Which shows that IAM $\leq$ ACC and completes the proof for k-class problems when $c'_{i.} \geq c'_{.i} \quad \forall i$.
Similarly, when $c'_{i.} \leq c'_{.i} \quad \forall i$:

$$\rightarrow \text{IAM} \leq 1/k[\frac{k(\sum_i c_{ii}+\sum_i c'_{.i})-2\sum_i c'_{.i}}{c_{..}}]$$
$$\rightarrow \text{IAM} \leq \frac{\sum_i c_{ii}+(1-2/k)\sum_i c'_{.i}}{c_{..}}$$
$$\xrightarrow[k>2]{1-2/k>0} \text{IAM} \leq \frac{\sum_i c_{ii}}{c_{..}} \rightarrow \text{IAM} \leq \text{ACC}$$

Lastly, for the third case, let us assume for a $z \in \{1, \ldots, k\}$, $c'_{z.} \leq c'_{.z}$ and $c'_{i.} \geq c'_{.i} \quad \forall i \in \{1, \ldots, k\},\ i \neq z$. Note that selecting multiple z does not impact the proof. By definition and because $c'_{z.} \leq c'_{.z}$:

$$\frac{c_{11}-c'_{1.}}{c_{11}+c'_{1.}} \leq \frac{c_{11}}{c_{11}+c'_{1.}}$$
$$\vdots$$
$$\frac{c_{zz}-c'_{.z}}{c_{zz}+c'_{.z}} \leq \frac{c_{zz}}{c_{zz}+c'_{.z}} \leq \frac{c_{zz}}{c_{zz}+c'_{z.}}$$
$$\vdots$$
$$\frac{c_{kk}-c'_{k.}}{c_{kk}+c'_{k.}} \leq \frac{c_{kk}}{c_{kk}+c'_{k.}}$$

By Proposition 1:

$$\frac{c_{11}-c'_{1.}}{c_{11}+c'_{1.}} \leq \frac{c_{11}+(c_{..}-(c_{11}+c'_{1.}))}{c_{11}+c'_{1.}+(c_{..}-(c_{11}+c'_{1.}))}$$
$$\vdots \qquad\qquad \vdots$$
$$\frac{c_{zz}-c'_{.z}}{c_{zz}+c'_{.z}} \leq \frac{c_{zz}+(c_{..}-(c_{zz}+c'_{z.}))}{c_{zz}+c'_{z.}+(c_{..}-(c_{zz}+c'_{z.}))}$$
$$\vdots \qquad\qquad \vdots$$
$$\frac{c_{kk}-c'_{k.}}{c_{kk}+c'_{k.}} \leq \frac{c_{kk}+(c_{..}-(c_{kk}+c'_{k.}))}{c_{kk}+c'_{k.}+(c_{..}-(c_{kk}+c'_{k.}))}$$

By summing both sides of all fractions and dividing them by k:

$$\text{IAM} \leq 1/k[\frac{kc_{..}-(c'_{1.}+\ldots+c'_{z.}+\ldots+c'_{k.})}{c_{..}}]$$
$$\text{IAM} \leq 1/k[\frac{k(\sum_i c_{ii}+\sum_i c'_{i.})-(c'_{1.}+\ldots+c'_{z.}+\ldots+c'_{k.})}{c_{..}}]$$
$$\text{IAM} \leq \frac{\sum_i c_{ii}+(1-1/k)\sum_i c'_{i.})}{c_{..}}]$$
$$\xrightarrow{1-1/k>0} \text{IAM} \leq \frac{\sum_i c_{ii}}{c_{..}} \rightarrow \text{IAM} \leq \text{ACC}$$

That completes the proof for the third case.

2.4.1. Interpretation of IAM

How to interpret different signs in IAM is summarized below:

1. IAM < 0 : We know that the number of instances that the classifier predicts incorrectly on average is higher than the number it predicts correctly. This means the performance is poor.
2. IAM > 0 : We know that the number of instances that the classifier predicts correctly on average is higher than the number it does not. This means that the performance could be promising depending on the magnitude.

Similarly, how to interpret the magnitude is given below.

1. The magnitude in either positive or negative direction shows how well/bad the classification fits the data.
2. Zero or close to zero values show that the number of correct and incorrect predictions are the same or very close.

2.4.2. Illustrative examples for the IAM

We calculate the IAM for all the examples given in Section 2.2. The results are given in Tables 9 and 10.

Insights:

1. With multiple metric computation scheme the CBA is the lowest metric for both classifiers and ML_1 with higher CBA is selected. The IAM metric confirms that (0.14>0.10 in favor of ML_1).

Table 9
IAM value for examples 1, 2 and 4.

Example 1a	Example 1b	Example 2a	Example 2b
0.14	0.10	−0.175	−0.185

Table 10
IAM value for examples 3.

ML_5	ML_6	ML_7
−0.38	−0.37	−0.50

Table 11
Balanced dataset.

Class	1	2	3
Frequency	100	100	100

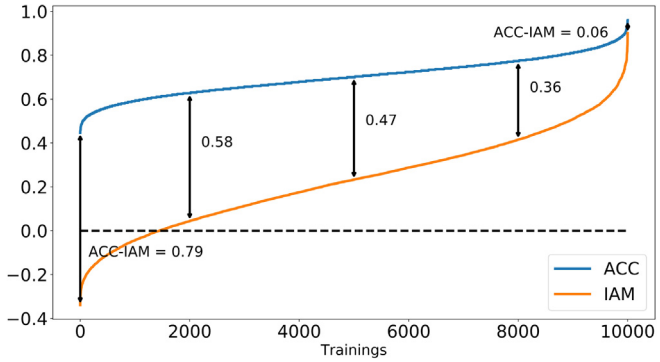

Fig. 1. The ACC, CBA and IAM values in the dataset given in Table 9.

2. Although the CBA metric is a conservative metric itself (always lower than the precision and recall), we realize that it is not always lower than the ACC. The IAM for ML_3 is higher and should be selected as the better model (no need for multiple metrics).
3. In example 3, the IAM could help us make the model selection decision easier. The IAM for ML_6 has the highest value and should be selected as the best classifier.

3. Further analysis of IAM

We study the IAM under different circumstances to evaluate what additional benefits it could offer. For that, we study the IAM in balanced, highly skewed, with many classes and binary datasets.

3.1. IAM and balanced datasets

To compare the relative changes of ACC and IAM on a balanced set, we assume a balanced dataset given in Table 11. We train a classifier under different settings and present the overall accuracy and IAM in Fig. 1. It is shown that both metrics are consistent, with the IAM to be slightly more sensitive — see the diverging gap. The lowest point in the ACC is close to 0.5 (half-way down) where it is below zero for the IAM (we naturally expected 0). It might be safe to say that the IAM is more conservative than ACC.

3.2. IAM and extremely skewed sets

The given dataset in Table 12 is extremely skewed. Training a classifier under different settings has given different accuracies in Fig. 2. The two lines provide opposite narratives from the performance. Needless to explain that the ACC line here is misleading.

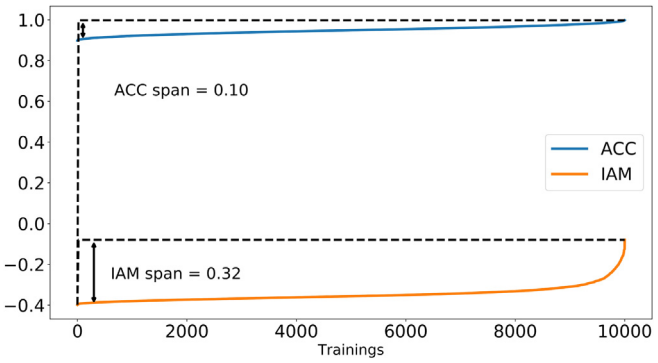

Fig. 2. The ACC, CBA and IAM values in the dataset given in Table 10.

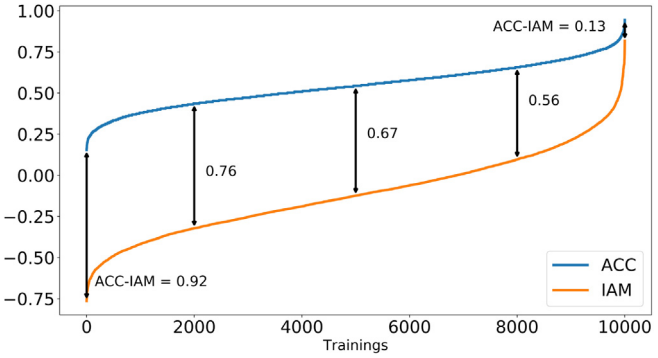

Fig. 3. The ACC, CBA and IAM values in the dataset given in Table 10.

Table 12
Heavily skewed dataset.

Class	1	2	3
Frequency	1,000,000	100	100

Table 13
A dataset with Large number of classes.

Class	1	2	3	4	5	6	7	8	9	10
Frequency	1000	60	50	50	20	20	10	5	2	2

Table 14
Binary class dataset.

Class	1	2
Frequency	5200	4800

Table 15
Distribution of glass types.

Class	1	2	3	4	5	6	7
Frequency	70	76	17	0	13	9	29

It is also shown that the ACC is almost insensitive (close to a flat line). The IAM has shown a larger span (0.32) and could offer more help in model tuning and model selection.

3.3. IAM and imbalance sets with large number of classes

The given dataset in Table 13 has 10 classes and is heavily imbalance. Training a classifier under different settings has given different accuracy values in Fig. 3. We can see the span of changes from the best to worst performance in the IAM is larger (−0.75,0.75) but both metrics are consistent.

Table 16
Accuracy values for glass dataset (Max values in columns are highlighted in bold and min values in rows are underlined).

Classifier	Multiple metric approach					Selected	IAM	Selected
	ACC	CBA	MAR	MAP	F-score			
KNN	0.651	0.450	0.671	0.511	0.542		−0.109	
KNN-RUS	0.480	0.435	0.517	0.512	0.490		−0.129	
KNN-ROS	0.613	0.527	0.696	0.627	0.627		0.054	
LR	0.569	0.366	0.401	0.598	0.411		−0.248	
LR-RUS	0.451	0.453	0.649	0.533	0.536		−0.097	
LR-ROS	0.493	0.497	0.635	0.532	0.563		−0.005	
SVM	0.571	0.340	0.472	0.671	0.438		−0.288	
SVM-RUS	0.334	0.295	0.445	0.608	0.359		−0.409	
SVM-ROS	0.334	0.258	0.465	0.466	0.334		−0.480	
RF	0.783	**0.681**	0.723	0.781	**0.746**	✓	**0.342**	✓
RF-RUS	0.640	0.556	0.715	0.593	0.631		0.111	
RF-ROS	0.693	0.661	**0.779**	0.741	0.741		0.321	
ANN	0.541	0.362	0.481	**0.828**	0.371		−0.388	
ANN-RUS	0.427	0.299	0.472	0.412	0.384		−0.403	
ANN-ROS	0.513	0.348	0.348	0.810	0.348		−0.401	
GBM	**0.786**	0.626	0.697	0.807	0.675		0.263	
GBM-RUS	0.346	0.347	0.484	0.451	0.416		−0.301	
GBM-ROS	0.720	0.642	0.697	0.685	0.683		0.284	

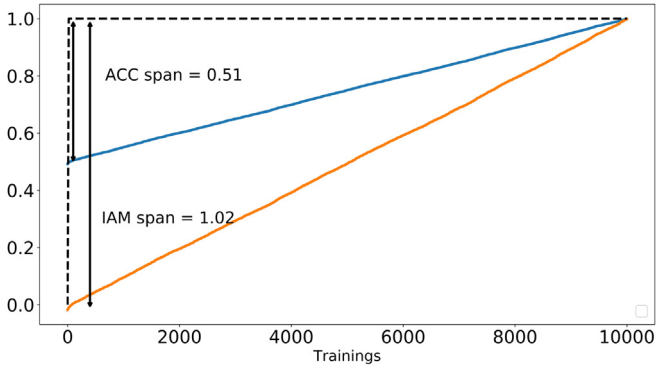

Fig. 4. The AUC and IAM values in the dataset given in Table 11.

3.4. IAM and binary classification

As a final experiment, we compute the IAM for a classifier trained for a binary classification problem and compare it against the AUC (Area Under ROC Curve). The ROC curve is a two-dimensional representation of classifier performance. To compare classifiers, ROC performance could be reduced to a single scalar value representing expected performance commonly calculated as the area under the ROC curve [18]. The AUC is a very popular metric for performance evaluation of classifiers on classifying binary datasets. But it is not as popular for multi-class problems due to several complexities such as the issue of combining multiple pairwise discriminability values explained in [18]. The given dataset in Table 14 has 2 classes. Training a classifier under different settings has given different AUC and IAM values in Fig. 4. Examining Fig. 4 shows that IAM is consistent with the AUC and the main difference is on the scale of the two metrics.

3.5. IAM implementation on multi-class datasets

We use two real-world multi-class datasets, glass identification dataset [20], and satimage dataset [21] to show how IAM can be used for model selection compared to multiple metric approach.

The glass datasets is imbalance and has 11 attributes. The characteristic of the attributes can be accessed in [20]. The target variable (glass type) has 7 classes whose distribution is given in Table 15.

Table 17
Distribution of satimage target variable.

Class	1	2	3	4	5	6	7
Frequency	1531	703	1356	625	707	0	1508

We train and test six classifiers on the glass dataset using the libraries provided in [22] . The classifiers are k-nearest neighbors (KNN), logistic regression (LR), support vector classifier (SVC), random forests (RF), neural network (ANN) and gradient boosting machine (GBM). In addition, we train each algorithm with random under-sampling (-RUS) and random over-sampling (-ROS) schemes which are two popular methods to improve the training quality in imbalance datasets [6,7]. The results are summarized in Table 16. The accuracy results are obtained via 5-fold cross-validation. Based on the results (average accuracies), we note that the CBA metric is the lowest metric for majority of the classifiers but not for all. Nevertheless, we choose the CBA to identify the best performing classifier. The CBA column in Table 16 shows that the RF has the highest CBA value, and therefore should be selected as the best classifier. If we decided to use the IAM column for model selection, we would come to the same conclusion as the RF offers the highest IAM value. This example shows that when we use the IAM, we are able to use it as a solo measure and do not have to compute multiple other metrics unless we choose to — the case that does not hold true for five other metrics.

In another example, the satimage database consists of the multi-spectral values of pixels in 3×3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood is used. The dataset has 36 attributes and is imbalance. The target variable has 7 classes whose distribution is given in Table 17. Similarly, we train and test the same classifiers on staimage dataset. The results are summarized in Table 18. Both multiple metric approach and IAM computations point to GBM classifier as the best performing model. The CBA measure that offers the lowest accuracy among multiple metrics has the highest value (max[min(row)] value) for the GBM model. The IAM is also the highest for the GBM. Note that the GBM dominates all other classifiers on all metrics.

4. Conclusions and future work

In this paper, we asked two research questions about usefulness of introducing a new metric for evaluating classifiers in multi-class imbalance datasets and about additional benefits that this new metric could offer. We responded these questions by

Table 18
Accuracy values for satimage dataset (Max values in columns are highlighted in bold and min values in rows are underlined).

Classifier	Multiple metric approach					Selected	IAM	Selected
	ACC	CBA	MAR	MAP	F-score			
KNN	0.898	0.858	0.888	0.899	0.884		0.817	
KNN-RUS	0.890	0.847	0.892	0.873	0.871		0.794	
KNN-ROS	0.892	0.851	0.897	0.897	0.897		0.815	
LR	0.890	0.850	0.880	0.901	0.880		0.793	
LR-RUS	0.884	0.850	0.900	0.861	0.889		0.779	
LR-ROS	0.891	0.851	0.888	0.889	0.888		0.798	
SVM	0.861	0.830	0.864	0.870	0.863		0.661	
SVM-RUS	0.839	0.816	0.843	0.844	0.841		0.632	
SVM-ROS	0.856	0.834	0.861	0.861	0.859		0.669	
RF	0.910	0.901	0.910	0.909	0.909		0.868	
RF-RUS	0.906	0.900	0.916	0.903	0.901		0.866	
RF-ROS	0.909	0.900	0.902	0.908	0.901		0.857	
ANN	0.903	0.889	0.904	0.914	0.902		0.778	
ANN-RUS	0.900	0.858	0.906	0.902	0.901		0.717	
ANN-ROS	0.898	0.852	0.903	0.901	0.897		0.704	
GBM	**0.921**	**0.916**	**0.921**	**0.921**	**0.921**	✓	**0.892**	✓
GBM-RUS	0.916	0.902	0.907	0.902	0.906		0.885	
GBM-ROS	0.918	0.903	0.913	0.918	0.911		0.887	

proposing a new accuracy metric, IAM, that offers certain benefits compared to other widely used accuracy metrics in imbalanced datasets which are listed below:

1. The IAM is introduced primarily to be used on multi-class imbalance problems, but as provided evidence it can very well be used in binary and balanced datasets.
2. In proposing this metric, we intended to ensure that the IAM is simple to use, and offers easy interpretation as we believed the issues of simplicity and interpretability are among the most important factors encouraging or discouraging the analysts for/from using a metric.
3. The IAM is similar to overall accuracy and class balance accuracy metrics but distinctly measures how well a classifier is expected not to classify a random instance in incorrect classes. This quality makes the IAM a conservative metric.
4. We expect the analysts to use the IAM as a bottom-line accuracy. That is, the analysts do not have to compute multiple metrics to make model selection decisions unless they choose to. Also, when the IAM is promising, there is no dispute that the classifier is performing well. The opposite also holds true.
5. Since the IAM is built up on top of the existing metrics, it is consistent with them.

One enhancements to our work would be to compare IAM against less popular metrics introduced in information theory such as Matthew's correlation coefficient, relative classifier information, and confusion entropy.

CRediT authorship contribution statement

Ebrahim Mortaz: Conceptualization, Analysis, Developments, Write up and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N.V. Chawla, Data mining for imbalanced datasets: An overview, in: Data Mining and Knowledge Discovery Handbook, Springer, 2009, pp. 875–886.
[2] S. Wang, X. Yao, Multiclass imbalance problems: Analysis and potential solutions, IEEE Trans. Syst. Man Cybern. B 42 (4) (2014) 1119–1130.
[3] B. Liu, G. Tsoumakas, Dealing with class imbalance in classifier chains via random undersampling, Knowl.-Based Syst. (2019) 105292.
[4] L. Guillaum, F. Nogueira, A.K. Christos, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (1) (2017) 559–563.
[5] C. Zhang, J. Bi, S. Xu, E. Ramentol, G. Fan, B. Qiao, H. Fujita, Multi-imbalance: An open-source software for multi-class imbalance learning, Knowl.-Based Syst. 174 (2019) 137–143.
[6] Y. Yan, M. Tan, Y. Xu, J. Cao, M. Ng, H. Min, Q. Wu, Oversampling for imbalanced data via optimal transport, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 5605–5612.
[7] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[8] C.X. Ling, V.S. Sheng, Q. Yang, Test strategies for cost-sensitive decision trees, IEEE Trans. Knowl. Data Eng. 18 (8) (2006) 1055–1067.
[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man Cybern. C (Applications and Reviews) 42 (4) (2011) 463–484.
[10] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (4) (2009) 427–437.
[11] J. Shreve, H. Schneider, O. Soysal, A methodology for comparing classification methods through the assessment of model stability and validity in variable selection, Decis. Support Syst. 52 (1) (2011) 247–257.
[12] L. Mosley, A balanced approach to the multi-class imbalance problem, Ph.D. Dissertation, Iowas State University, 2013.
[13] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[14] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach. Learn. 45 (2) (2001) 171–186.
[15] Y. Sun, M.S. Kamel, Y. Wang, Boosting for learning multiple classes with imbalanced class distribution, in: Sixth International Conference on Data Mining (ICDM'06), IEEE, 2006, pp. 592–602.
[16] B.W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochim. Biophys. Acta (BBA)-Protein Structure 405 (2) (1975) 442–451.
[17] J.-M. Wei, X.-J. Yuan, Q.-H. Hu, S.-Q. Wang, A novel measure for evaluating classifiers, Expert Syst. Appl. 37 (5) (2010) 3799–3809.
[18] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[19] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recognit. 91 (2019) 216–231.
[20] D. Dua, C. Graff, UCI machine learning repository, 2017, University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml.
[21] J. Vanschoren, J.N. van Rijn, B. Bischl, L. Torgo, Openml: Networked science in machine learning, SIGKDD Explorations 15 (2) (2013) 49–60, http://dx.doi.org/10.1145/2641190.2641198.
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.