Pattern Recognition 91 (2019) 216–231

Contents lists available at ScienceDirect

Pattern Recognition

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

The impact of class imbalance in classification performance metrics based on the binary confusion matrix

Amalia Luque[a,*], Alejandro Carrasco[b], Alejandro Martín[a], Ana de las Heras[a]

[a] Dpto. Ingeniería del Diseño. Escuela Politécnica Superior. Universidad de Sevilla. Virgen de África, 7. 41011 Sevilla, Spain
[b] Dpto. Tecnología Electrónica. Escuela Politécnica Superior. Universidad de Sevilla. Virgen de África, 7. 41011, Sevilla, Spain

ARTICLE INFO

Article history:
Received 4 September 2018
Revised 22 December 2018
Accepted 22 February 2019
Available online 28 February 2019

Keywords:
Classification
Performance measures
Imbalanced datasets
Class Balance Metrics

ABSTRACT

A major issue in the classification of class imbalanced datasets involves the determination of the most suitable performance metrics to be used. In previous work using several examples, it has been shown that imbalance can exert a major impact on the value and meaning of accuracy and on certain other well-known performance metrics. In this paper, our approach goes beyond simply studying case studies and develops a systematic analysis of this impact by simulating the results obtained using binary classifiers. A set of functions and numerical indicators are attained which enables the comparison of the behaviour of several performance metrics based on the binary confusion matrix when they are faced with imbalanced datasets. Throughout the paper, a new way to measure the imbalance is defined which surpasses the Imbalance Ratio used in previous studies. From the simulation results, several clusters of performance metrics have been identified that involve the use of Geometric Mean or Bookmaker Informedness as the best null-biased metrics if their focus on classification successes (dismissing the errors) presents no limitation for the specific application where they are used. However, if classification errors must also be considered, then the Matthews Correlation Coefficient arises as the best choice. Finally, a set of null-biased multi-perspective Class Balance Metrics is proposed which extends the concept of Class Balance Accuracy to other performance metrics.

1. Introduction

In recent years, the scientific community working on classification algorithms has shown an increasing interest in the challenges that arise when imbalanced datasets are considered. Several overviews on these issues have been addressed in [2,17,21,24]. In these analyses, the Synthetic Minority Over-sampling Technique (SMOTE) [9] and the AdaBoost [39] are highlighted as general-purpose solutions, although algorithms of a more specific nature can also be found, either as general-purpose [3], as problem-oriented [15] or as classifier-oriented [33]. Algorithms that address the imbalance problem in multi-label classification [8] or that use advanced classifiers such as extreme learning machines [41] can also be found. An up-to-date comparison of these techniques addressing a specific problem can be found in [40].

A key aspect of these methods involves the determination of the classification performance, not only in order to assess the final result, but also to obtain a figure which has to be optimized by tuning the classifier parameters. However, there is no single way to select the *best* algorithm as any algorithm can obtain good results in one class but poor scores in other classes. For this reason, several metrics are usually considered, which permits the polyhedral characteristics of the classification performance to be viewed from different points of views.

The impact of class imbalance on classification performance metrics has therefore become a major issue. Several authors have addressed this topic by showing a few examples of this impact on accuracy [10,22] and on several other metrics [5,13,20]. To the best of our knowledge, only one systematic (albeit limited) study has been published that is not simply based on examples, [25].

The quantitative research on classification performance metrics has traditionally been tackled either by using a collection of known and widely available datasets [1,4], or by randomly simulating the classifier results [27,34]. A mixture of random simulated variations on known datasets is employed in Jeni et al. [25].

* Corresponding author.
E-mail addresses: amalialuque@us.es (A. Luque), acarrasco@us.es (A. Carrasco), ammartin@us.es (A. Martín), adelasheras@us.es (A. de las Heras).

https://doi.org/10.1016/j.patcog.2019.02.023

In order to overcome the effect of imbalance on performance metrics, several solutions have been suggested, among which the most cited is the use of the Class Balanced Accuracy (*CBA*) [1,6,32]. Approaches such as relevance-based evaluation [5], the Normalized Precision Rate [13], the Index of Balanced Accuracy [18,28,29], and the multiclass performance score (MPS) [23] have also been proposed.

In this paper, an extensive and systematic study is undertaken of the impact of class imbalance on classification performance metrics. Several dozen performance metrics can be found in the scientific literature, some based on a threshold, others based on probabilities, while yet others are based on ranks [14]. However, the most widely employed metrics are those based on the confusion matrix [38], where the multi-class case is usually reduced to a set of binary cases using the One-versus-All or the One-versus-One approach [31]. For these reasons, our research is focused on classification performance metrics based on the binary confusion matrix.

The rest of the paper is organized as follows. Section 2 presents the methodology employed to measure the impact of imbalance in performance metrics, thereby formally defining the confusion matrix (Section 2.1) and the metrics based thereon (Section 2.2). Section 2.3 proposes a new figure for the quantification of the class imbalance, and several functions and indicators of the aforementioned impact of imbalance are defined in Section 2.4. The application to a set of classification performance metrics based on the binary confusion matrix is presented in Section 3. The discussion and conclusion of these results are addressed in Section 4.

2. Methodology

2.1. Definition of the confusion matrix

Consider a dataset $\mathcal{D} = \{d_1, d_2, \cdots, d_m\}$ made up of m elements where d_k represents the k-th element. Let Θ be a set of C classes $\Theta = \{\theta_1, \theta_2, \cdots, \theta_C\}$ where θ_i defines the i-th class. The classifier $\mathcal{C}$ operating on d_k (the k-th element of the dataset $\mathcal{D}$) assigns a label θ_j and estimates that this element belongs to the j-th class, that is, $d_k \xrightarrow{\mathcal{C}} \theta_j$ or $\mathcal{C}(d_k) = \theta_j$, while it really belongs to the i-th class θ_i, thereby causing a misclassification (a confusion) when $i \neq j$.

Let $\mathcal{A} = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$ be the set of actual classes corresponding to the dataset $\mathcal{D}$, where α_k is the actual class of the element d_k. Furthermore, let $\mathcal{E} = \{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m\}$ be the set of classes estimated by the classifier $\mathcal{C}$ for each element in $\mathcal{D}$, where ε_k is the estimated class of the element d_k. The performance of $\mathcal{C}$ can be assessed using a measuring function $\mathcal{M}$, which assigns a metric $\mu \in \mathbb{R}$ to the pair $(\mathcal{A}, \mathcal{E})$, that is, $(\mathcal{A}, \mathcal{E}) \xrightarrow{\mathcal{M}} \mu$.

In this paper, we will focus on metrics based on the confusion matrix, which represents one of the most common methods to present the results obtained by a classifier, and is defined as

$$CM \equiv \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1C} \\ m_{21} & m_{22} & \dots & m_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ m_{C1} & m_{C2} & \dots & m_{CC} \end{bmatrix}. \tag{1}$$

In this expression, m_{ij} represents the number of elements actually belonging to the i-th class (θ_i) but that are classified as members of the j-th class (θ_j). In the context of our research, it is better to describe the term m_{ij} in relation to the total number of elements m_i belonging to the i-th class (θ_i). By denoting λ_{ij} as the ratio m_{ij}/m_i, the confusion matrix can be rewritten as

$$CM = \begin{bmatrix} \lambda_{11}m_1 & \lambda_{12}m_1 & \dots & \lambda_{1C}m_1 \\ \lambda_{21}m_2 & \lambda_{22}m_2 & \dots & \lambda_{2C}m_2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{C1}m_C & \lambda_{C2}m_C & \dots & \lambda_{CC}m_C \end{bmatrix}, \tag{2}$$

which can be expressed as the Hadamard (element-wise) product of two matrices in the form

$$CM = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1C} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{C1} & \lambda_{C2} & \dots & \lambda_{CC} \end{bmatrix} \circ \begin{bmatrix} m_1 & m_1 & \dots & m_1 \\ m_2 & m_2 & \dots & m_2 \\ \vdots & \vdots & \ddots & \vdots \\ m_C & m_C & \dots & m_C \end{bmatrix}. \tag{3}$$

In the binary case, that is, when the number of classes is $C = 2$, then the confusion matrix can be written as

$$CM \equiv \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}. \tag{4}$$

In general, one of the classes is called the "Positive" class and the other is named the "Negative" class. Therefore, the confusion matrix can be rewritten according to this new terminology as

$$CM \equiv \begin{bmatrix} m_{PP} & m_{PN} \\ m_{NP} & m_{NN} \end{bmatrix}. \tag{5}$$

The elements of this matrix are named with the following convention: m_{PP}, "True Positive" (*TP*); m_{PN}, "False Negative" (*FN*); m_{NP}, "False Positive" (*FP*); and m_{NN}, "True Negative" (*TN*).

The total number of positive m_P and negative m_N elements in $\mathcal{D}$ meet that they sum m, the total number of elements, that is, $m_P + m_N = m$. Furthermore, it is also true that the number of elements correctly classified in class P (m_{PP}), and the number of elements misclassified in that class P (m_{PN}), adds up to the number of elements in the positive class (m_P), that is, $m_{PP} + m_{PN} = m_P$. Similarly, it can be stated that $m_{NP} + m_{NN} = m_N$. The confusion matrix can therefore be written as

$$CM = \begin{bmatrix} m_{PP} & m_P - m_{PP} \\ m_N - m_{NN} & m_{NN} \end{bmatrix}, \tag{6}$$

which can also be formulated in terms of the λ_{ij} ratios as

$$CM = \begin{bmatrix} \lambda_{PP}m_P & \lambda_{PN}m_P \\ \lambda_{NP}m_N & \lambda_{NN}m_N \end{bmatrix} = \begin{bmatrix} \lambda_{PP}m_P & (1-\lambda_{PP})m_P \\ (1-\lambda_{NN})m_N & \lambda_{NN}m_N \end{bmatrix}. \tag{7}$$

Additionally, the total number of elements in $\mathcal{D}$ estimated by $\mathcal{C}$ as positive (despite their actual class), e_P, and those estimated as negative, e_N, can be written as

$$\begin{aligned} e_P &= m_{PP} + m_{NP} = \lambda_{PP}m_P + (1-\lambda_{NN})m_N. \\ e_N &= m_{NN} + m_{PN} = \lambda_{NN}m_N + (1-\lambda_{PP})m_P. \end{aligned} \tag{8}$$

They also add up to the total number of elements, $e_P + e_N = m$. The definitions regarding the confusion matrix are summarized in Fig. 1.

2.2. Metrics based on the binary confusion matrix

Based on the binary confusion matrix, numerous performance metrics have been proposed [19,27,30,34,37]. For our study, the focus is placed on 10 of these metrics, which are summarized in Table 1. All these metrics, take values within the [0, 1] range, except the last three (*MCC*, *BM*, and *MK*), whose ranges lie within the $[-1, 1]$ interval. For comparison purposes, these metrics are used herein in their normalized version (*MCCn*, *BMn*, and *MKn*). By naming a metric defined within the $[-1, 1]$ interval as μ, it can be normalized within the [0, 1] range by the expression

$$\mu_n \equiv \frac{\mu + 1}{2}. \tag{9}$$

		Predicted Class		
		P	***N***	**Instances**
Actual Class	***P***	**TP** $\lambda_{PP} m_P$	**FN** $(1-\lambda_{PP}) m_P$	m_P
	N	**FP** $(1-\lambda_{NN}) m_N$	**TN** $\lambda_{NN} m_N$	m_N
Estimations		e_P	e_N	m

Fig. 1. Confusion matrix for binary classification.

Table 1
Definition of classification performance metrics.

Symbol	Metric	Defined as
SNS	Sensitivity	$\frac{TP}{TP+FN}$
SPC	Specificity	$\frac{TN}{TN+FP}$
PRC	Precision	$\frac{TP}{TP+FP}$
NPV	Negative Predictive Value	$\frac{TN}{TN+FN}$
ACC	Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
F_1	F_1 score	$2\frac{PRC \cdot SNS}{PRC+SNS}$
GM	Geometric Mean	$\sqrt{SNS \cdot SPC}$
MCC	Matthews Correlation Coefficient	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
BM	Bookmaker Informedness	$SNS + SPC - 1$
MK	Markedness	$PPV + NPV - 1$

It can easily be shown (see supplementary material in the electronic version of this paper) that all these metrics can be expressed as a function $\mu = \mu(\lambda_{PP}, \lambda_{NN}, \pi_P, \pi_N)$, where π_P is the ratio of positive elements in the dataset (m_P/m) and, analogously, where $\pi_N \equiv m_N/m$. Furthermore, for balanced classes, when $\pi_P = \pi_N = 0.5$, them the metrics can be formulated as $\mu = \mu(\lambda_{PP}, \lambda_{NN})$. These functions are depicted in Fig. 2 as a heat map for each metric.

The best classifier achieves a value of $\lambda_{PP} = 1$ (all the positive elements are correctly classified as positive) and, also a value of $\lambda_{NN} = 1$ (all the negative elements are correctly classified as negative), corresponding to the upper right-hand-side corner in the graphic. Instead, the worst classifier ($\lambda_{PP} = 0$, $\lambda_{NN} = 0$) corresponds to the lower left-hand-side corner of the graphic.

Although only performance metrics based on the confusion matrix are considered, a marginal approach to Receiver Operating Characteristics (ROC) analysis [16] can also be carried out. In this analysis, the Area Under Curve (*AUC*) is commonly used as a performance metric. However, for classifiers offering only a label (and not a set of scores for each label), or when a single threshold is used on scores, the value of *AUC* and *BMn* are the same [36]. Therefore, in the forthcoming sections, whenever *BMn* is mentioned it could also be understood as *AUC*.

2.3. Defining class imbalance

The concept of class imbalance is relatively clear: it arises when the dataset has a different number of elements in positive and negative classes. However, its formalization is far from being univocally accepted. For instance, in [18,29], the imbalance is characterized by the dominance (*Dom*) or prevalence relationship between the positive class and the negative class, and is defined as $TPR - TNR$. This value is later employed to compensate performance metrics affected by the imbalance problem. However, the dominance is not exactly a measure of the imbalance in the dataset because it considers imbalance in the outcomes of the classifier.

Other authors formalize this concept by using the entropy [12] or, more commonly, the proportion between positive and negative instances (formalized as 1: *X*) [13] which is similar to the imbalance ratio (*IR*) defined as m_P/m_N [1], also called skew [25]. This value lies within the [0, ∞] range, having a value $IR = 1$ in the balanced case.

Other authors formalize this concept by using the proportion between positive and negative instances (formalized as 1: *X*) [13] or, which is similar, the imbalance ratio (*IR*) defined as m_P/m_N [1], which is also called skew [25]. This value lies within the [0, ∞] range, having a value $IR = 1$ in the balanced case.

In this paper, it is preferred to feature the imbalance with a value within the $[-1, 1]$ range, while reserving the 0 value for when the classes are perfectly balanced (lack of imbalance). For this purpose, we propose the imbalance coefficient δ as

$$\delta = \delta_P \equiv 2\pi_P - 1 = 2\frac{m_P}{m} - 1. \tag{10}$$

For the negative class, the coefficient $\delta_N = 2\pi_N - 1$ is also defined. The sum of these coefficients is

$$\delta_P + \delta_N = 2\pi_P - 1 + 2\pi_N - 1 = 2(\pi_P + \pi_N) - 2 = 0. \tag{11}$$

Hence $\delta_N = -\delta_P = -\delta$. From (10), the value of π_P can be obtained as

$$\pi_P = \frac{1+\delta}{2}. \tag{12}$$

Moreover, the value of π_N can be derived from (11)

$$\pi_N = \frac{1-\delta}{2}. \tag{13}$$

Therefore, the metrics $\mu = \mu(\lambda_{PP}, \lambda_{NN}, \pi_P, \pi_N)$ can be redefined as $\mu = \mu(\lambda_{PP}, \lambda_{NN}, \delta)$. It is clear that the value of the metric μ depends not only on the classifier's performance, but also on the imbalance δ.

It can easily be derived that the relationship between the imbalance ratio (*IR*) and the imbalance coefficient (δ) is

$$IR = \frac{1+\delta}{1-\delta}. \tag{14}$$

2.4. Assessing the impact of imbalance

In order to assess the impact of the imbalance in a certain metric, its value (μ_b) for the balanced case (when $\delta = 0$) is first considered.

$$\mu_b \equiv \mu|_{\delta=0} = \mu(\lambda_{PP}, \lambda_{NN}, 0) = \mu_b(\lambda_{PP}, \lambda_{NN}). \tag{15}$$

This definition is later employed to propose a family of metrics where the effect of the imbalance is dismissed. In the scientific literature, a few examples of these metrics can be found, as in the case of Class Balance Accuracy (*CBA*) [35]. On generalizing this approach, the metrics μ_b are called Class Balance Metrics (*CBM*). Table 2 summarizes the equations for each metric, both in the class imbalance and balance cases. These results are derived in the supplementary material of the paper, available online.

With these definitions, it is now possible to quantify the impact of imbalance, by using the bias of the metric which is defined as

$$B_\mu \equiv \mu - \mu_b = \mu(\lambda_{PP}, \lambda_{NN}, \delta) - \mu_b(\lambda_{PP}, \lambda_{NN}). \tag{16}$$

Table 3 summarizes the definition of bias for each metric. These results are derived in the supplementary material of the paper, available online.

As can be observed, bias depends on three variables: $B_\mu = B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)$. In order to study this function, a 4-dimensional space is required. Its representations are first tackled using heat volumes (3D), where each point in the 3-dimensional (λ_{PP}, λ_{NN}, δ) space has a bias-dependent colour.

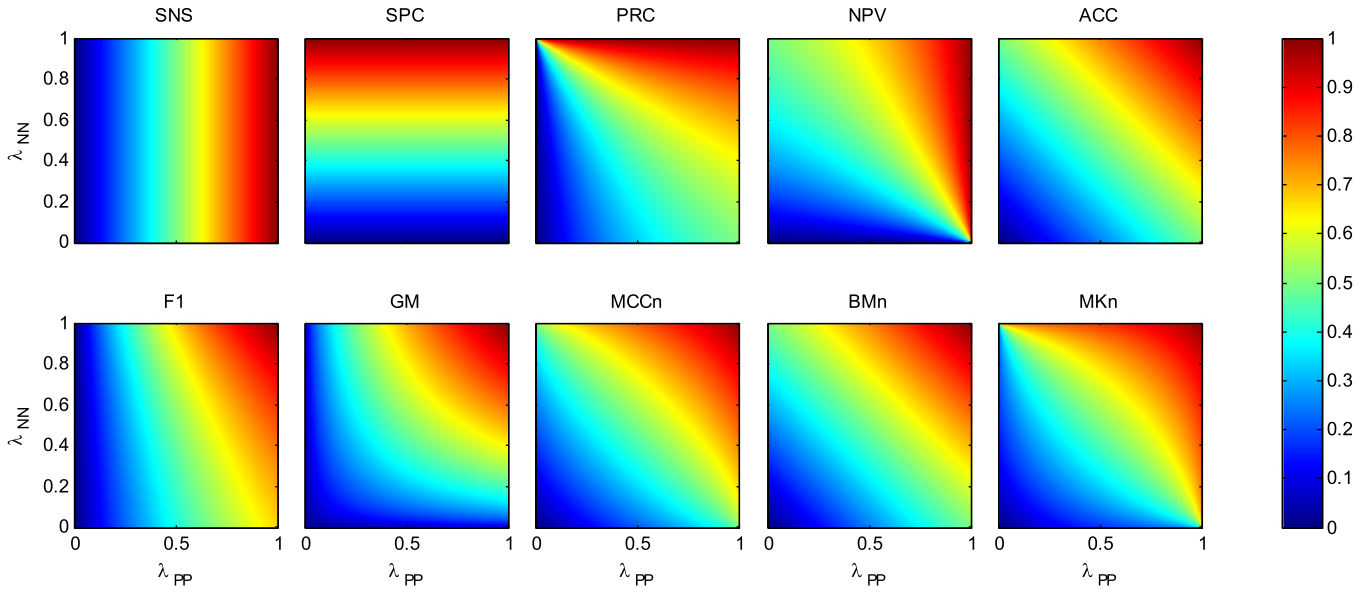

Fig. 2. Heat maps for metrics with balanced classes.

Table 2
Classification performance metrics as a function of imbalance.

Metrics	Class imbalance metrics $\mu(\lambda_{PP}, \lambda_{NN}, \delta)$	Class Balance metrics $\mu_b(\lambda_{PP}, \lambda_{NN})$
SNS	λ_{PP}	λ_{PP}
SPC	λ_{NN}	λ_{NN}
PRC	$\frac{\lambda_{PP}(1+\delta)}{\lambda_{PP}(1+\delta)+(1-\lambda_{NN})(1-\delta)}$	$\frac{\lambda_{PP}}{\lambda_{PP}+(1-\lambda_{NN})}$
NPV	$\frac{\lambda_{NN}(1-\delta)}{\lambda_{NN}(1-\delta)+(1-\lambda_{PP})(1+\delta)}$	$\frac{\lambda_{NN}}{\lambda_{NN}+(1-\lambda_{PP})}$
ACC	$\lambda_{PP}\frac{1+\delta}{2}+\lambda_{NN}\frac{1-\delta}{2}$	$\frac{\lambda_{PP}+\lambda_{NN}}{2}$
F_1	$\frac{2\lambda_{PP}(1+\delta)}{(1+\lambda_{PP})(1+\delta)+(1-\lambda_{NN})(1-\delta)}$	$\frac{2\lambda_{PP}}{2+\lambda_{PP}-\lambda_{NN}}$
GM	$\sqrt{\lambda_{PP}\cdot\lambda_{NN}}$	$\sqrt{\lambda_{PP}\cdot\lambda_{NN}}$
MCCn	$\frac{1}{2}\left(\frac{\lambda_{PP}+\lambda_{NN}-1}{\sqrt{[\lambda_{PP}+(1-\lambda_{NN})\frac{1-\delta}{1+\delta}][\lambda_{NN}+(1-\lambda_{PP})\frac{1+\delta}{1-\delta}]}}+1\right)$	$\frac{1}{2}\left(\frac{\lambda_{PP}+\lambda_{NN}-1}{\sqrt{[\lambda_{PP}+(1-\lambda_{NN})][\lambda_{NN}+(1-\lambda_{PP})]}}+1\right)$
BMn	$\frac{\lambda_{PP}+\lambda_{NN}}{2}$	$\frac{\lambda_{PP}+\lambda_{NN}}{2}$
MKn	$\frac{1}{2}\left(\frac{1+\delta}{(1+\delta)+\frac{1-\lambda_{NN}}{\lambda_{PP}}(1-\delta)}+\frac{1-\delta}{(1-\delta)+\frac{1-\lambda_{PP}}{\lambda_{NN}}(1+\delta)}\right)$	$\frac{1}{2}\left(\frac{1}{1+\frac{1-\lambda_{NN}}{\lambda_{PP}}}+\frac{1}{1+\frac{1-\lambda_{PP}}{\lambda_{NN}}}\right)$

Table 3
Bias of performance metrics due to class imbalance.

Metrics	Bias $B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)$
SNS	0
SPC	0
PRC	$\frac{1+\delta}{(1+\delta)+\frac{1-\lambda_{NN}}{\lambda_{PP}}(1-\delta)}-\frac{1}{1+\frac{1-\lambda_{NN}}{\lambda_{PP}}}$
NPV	$\frac{1-\delta}{(1-\delta)+\frac{1-\lambda_{PP}}{\lambda_{NN}}(1+\delta)}-\frac{1}{1+\frac{1-\lambda_{PP}}{\lambda_{NN}}}$
ACC	$\frac{\delta}{2}(\lambda_{PP}-\lambda_{NN})$
F_1	$\frac{2\lambda_{PP}(1+\delta)}{(1+\lambda_{PP})(1+\delta)+(1-\lambda_{NN})(1-\delta)}-\frac{2\lambda_{PP}}{2+\lambda_{PP}-\lambda_{NN}}$
GM	0
MCCn	$\frac{\lambda_{PP}+\lambda_{NN}-1}{2\sqrt{[\lambda_{PP}+(1-\lambda_{NN})\frac{1-\delta}{1+\delta}][\lambda_{NN}+(1-\lambda_{PP})\frac{1+\delta}{1-\delta}]}}-\frac{\lambda_{PP}+\lambda_{NN}-1}{2\sqrt{[\lambda_{PP}+(1-\lambda_{NN})][\lambda_{NN}+(1-\lambda_{PP})]}}$
BMn	0
MKn	$\frac{1}{2}\left(\frac{1+\delta}{(1+\delta)+\frac{1-\lambda_{NN}}{\lambda_{PP}}(1-\delta)}-\frac{1}{1+\frac{1-\lambda_{NN}}{\lambda_{PP}}}+\frac{1-\delta}{(1-\delta)+\frac{1-\lambda_{PP}}{\lambda_{NN}}(1+\delta)}-\frac{1}{1+\frac{1-\lambda_{PP}}{\lambda_{NN}}}\right)$

Alternatively, B_μ is also represented as a set of heat maps (or contour graphs). Each heat map (2D) represents the metric bias for a certain fixed value of the imbalance (let us say δ_0), thereby making bias dependent on two variables (λ_{PP}, λ_{NN}) and on one constant (δ_0). Hence, $B_\mu = B_\mu(\lambda_{PP}, \lambda_{NN}, \delta_0)$.

Nevertheless, drawing conclusions regarding a 4-dimensional space is, in most cases, a challenging task. Several partial prospects are therefore proposed that reduce the bias function dimensionality. In this respect, the first approach involves considering the metric bias for classifiers whose performance is singularly located on the (λ_{PP}, λ_{NN}) plane, thereby obtaining a set of bias indicators which are generically denoted as $\sigma B_\mu(\delta)$. The singular classifiers and their formulations are proposed in Table 4.

Table 4
Bias indicators for singular classifiers.

Singular classifier	Symbol	$\sigma B_\mu(\delta)$
Worst classifier	$wcB_\mu(\delta)$	$\lim_{\varepsilon\to 0} B_\mu(\varepsilon, \varepsilon, \delta)$
Best classifier	$bcB_\mu(\delta)$	$\lim_{\varepsilon\to 0} B_\mu(1-\varepsilon, 1-\varepsilon, \delta)$
Worst-positive classifier	$wpcB_\mu(\delta)$	$\lim_{\varepsilon\to 0} B_\mu(\varepsilon, 1-\varepsilon, \delta)$
Worst-negative classifier	$wncB_\mu(\delta)$	$\lim_{\varepsilon\to 0} B_\mu(1-\varepsilon, \varepsilon, \delta)$
Medium classifier	$mcB_\mu(\delta)$	$B_\mu(0.5, 0.5, \delta)$

An alternative way to reduce the dimensionality of B_μ is through the consideration that λ_{PP} and λ_{NN} are randomly and uniformly distributed within the [0, 1] range. Bias can therefore be seen for each value of the imbalance coefficient δ as a random variable $B_\mu(\delta)$, which is characterized by its probability density function ($pdf[B_\mu(\delta)]$). Additionally, several local statistical indicators can be defined, which are generically denoted as $\psi B_\mu(\delta)$. The term *local* attributed to these indicators (summarized in Table 5) means that they are defined for each value of δ.

Definitions in Tables 4 and 5 have introduced several prospects of bias, all of which depend on the imbalance δ. They are gener-

Table 5
Local statistical indicators on bias.

Local statistical indicator	Symbol	$\psi B_\mu(\delta)$
Mean	$mB_\mu(\delta)$	$\int\int B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)\, d\lambda_{PP}\, d\lambda_{NN}$
Standard deviation	$sdB_\mu(\delta)$	$\sqrt{\iint [B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) - mB_\mu(\delta)]^2\, d\lambda_{PP}\, d\lambda_{NN}}$
Root-Mean-Square Bias	$rmsB_\mu(\delta)$	$\sqrt{\iint B_\mu^2(\lambda_{PP}, \lambda_{NN}, \delta)\, d\lambda_{PP}\, d\lambda_{NN}}$
Maximum absolute value	$maxaB_\mu(\delta)$	$\max_{0 \le \lambda_{PP} \le 1,\ 0 \le \lambda_{NN} \le 1} \lvert B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)\rvert$
Skewness	$skB_\mu(\delta)$	$\frac{1}{sdB_\mu^3(\delta)} \sqrt{\iint [B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) - mB_\mu(\delta)]^3\, d\lambda_{PP}\, d\lambda_{NN}}$
Kurtosis	$kB_\mu(\delta)$	$\frac{1}{sdB_\mu^4(\delta)} \sqrt{\iint [B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) - mB_\mu(\delta)]^4\, d\lambda_{PP}\, d\lambda_{NN}}$

Table 6
Global statistical indicators on bias.

Global statistical indicator	Symbol	$\Psi B_\mu(\delta)$
Mean	MB_μ	$\frac{1}{2} \iiint B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)\, d\lambda_{PP}\, d\lambda_{NN}\, d\delta$
Standard deviation	SDB_μ	$\sqrt{\frac{1}{2} \iiint [B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) - MB_\mu]^2\, d\lambda_{PP}\, d\lambda_{NN}\, d\delta}$
Root-Mean-Square Bias	$RMSB_\mu$	$\sqrt{\frac{1}{2} \iiint B_\mu^2(\lambda_{PP}, \lambda_{NN}, \delta)\, d\lambda_{PP}\, d\lambda_{NN}\, d\delta}$
Maximum absolute value	$MAXAB_\mu$	$\max_{0 \le \lambda_{PP} \le 1,\ 0 \le \lambda_{NN} \le 1,\ 0 \le \delta \le 1} \lvert B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)\rvert$
Skewness	SKB_μ	$\frac{1}{SDB_\mu^3} \sqrt{\frac{1}{2} \iiint [B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) - MB_\mu]^3\, d\lambda_{PP}\, d\lambda_{NN}\, d\delta}$
Kurtosis	KB_μ	$\frac{1}{SDB_\mu^4} \sqrt{\frac{1}{2} \iiint [B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) - MB_\mu(\delta)]^4\, d\lambda_{PP}\, d\lambda_{NN}\, d\delta}$

ically denoted as $xB_\mu(\delta) = \{\sigma B_\mu(\delta), \psi B_\mu(\delta)\}$. It should be observed that these are not numbers but functions. In order to obtain a single value derived from these functions, we can make the hypothesis that δ is randomly and uniformly distributed within the $[-1, 1]$ range. Mean values of each function can therefore be computed as

$$\overline{xB}_\mu \equiv \frac{1}{2} \int_{-1}^{1} xB_\mu(\delta)\ d\delta. \tag{17}$$

Another way to reduce the generic function $xB_\mu(\delta)$ to a single value is by focusing on its value for extremely positive-imbalanced datasets where $\delta \to 1$. A generic value can therefore be obtained through the expression

$$xB_\mu^{\varepsilon P} \equiv \lim_{\varepsilon \to 0} xB_\mu(1 - \varepsilon). \tag{18}$$

The corresponding negative counterpart is defined as

$$xB_\mu^{\varepsilon N} \equiv \lim_{\varepsilon \to 0} xB_\mu(-1 + \varepsilon). \tag{19}$$

A global regard of bias can also be undertaken by considering that λ_{PP} and λ_{NN} are randomly and uniformly distributed within the [0, 1] range, and also that δ lies within the $[-1, 1]$ range. Bias can now be seen as a random variable B_μ that is independent of the imbalance coefficient δ. Bias can therefore be characterized by its probability density function ($pdf[B_\mu]$). Based on this overall *pdf*, several global statistical indicators can be defined, which are generically denoted as ΨB_μ and are summarized in Table 6.

Definitions in Eqs. (17)–(19) and in Table 6 have introduced several single-valued indicators regarding bias which will generically be denoted as $XB_\mu(\delta) = \{\overline{xB}_\mu, xB_\mu^{\varepsilon P}, xB_\mu^{\varepsilon N}, \Psi B_\mu(\delta)\}$.

Throughout this subsection, several function and single-valued indicators have been introduced to assess the impact of dataset imbalance on classification performance metrics. A summary of these indicators is depicted in Fig. 3.

3. Results

3.1. Performance metric bias function

The methods described in Section 2 will now be applied to the ten metrics defined in Table 1. As explained above, bias depends on three variables, that is, $B_\mu = B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)$ and its formulation for the selected metrics is also shown in Table 3. The first approach for their representation is based on the heat volumes as depicted in Fig. 4, where each point in the 3-dimensional (λ_{PP}, λ_{NN}, δ) space has a bias-dependent colour.

In certain performance metrics (for instance in *MCCn*), bias has low values for many points in the (λ_{PP}, λ_{NN}, δ) space. In these cases, the expressive power of the whole range of colours is not completely exploited. It is therefore better to select the colour of each point not directly based on bias, but on the relative value of bias within the range of values for their corresponding metric. The colour-map is then rescaled to show the relative bias with the value -1 corresponding to the minimum bias, and the value $+1$ to the maximum. The result is depicted in Fig. 5, where the range of each metric is shown in its corresponding subplot title.

As explained above, B_μ can also be represented as a set of heat maps. Each heat map represents the metric bias for a certain fixed value of the imbalance $B_\mu = B_\mu(\lambda_{PP}, \lambda_{NN}, \delta_0)$. The result for precision (*PRC*) is portrayed in Fig. 6. Similar graphics can be obtained for the remaining metrics.

Now let us suppose that the value of δ is known, for instance $\delta = \delta_0 = 0.95$. Therefore $B_\mu = B_\mu(\lambda_{PP}, \lambda_{NN}, \delta_0)$ depends on only two variables (λ_{PP} and λ_{NN}) and can be represented as a heat map. The results for each metric are shown in Fig. 7 where the colours represent the absolute value of bias.

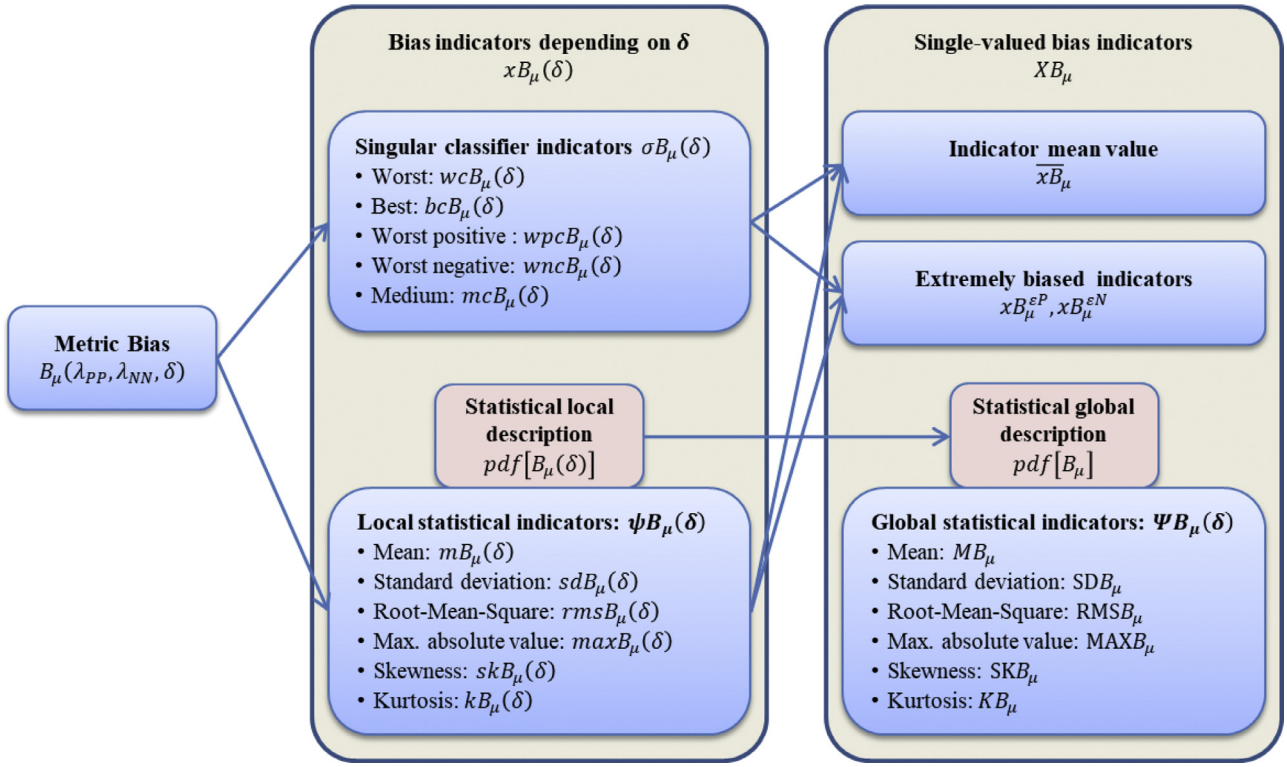

Fig. 3. Functions and single-valued indicators assessing bias in performance metrics due to class imbalance.

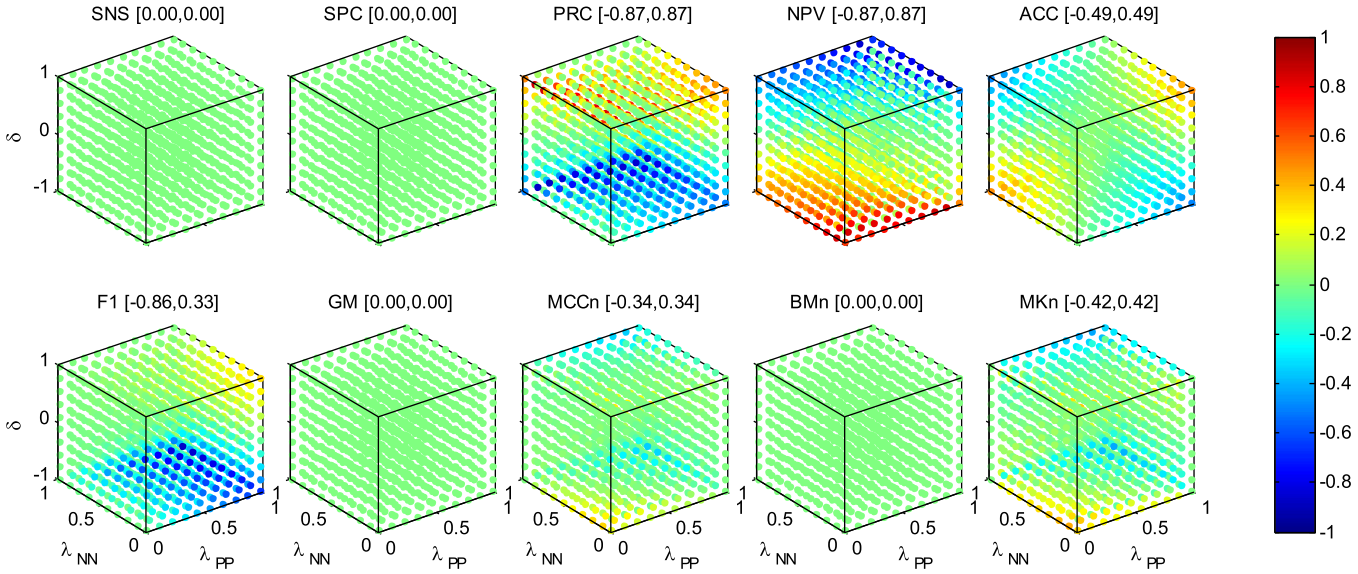

Fig. 4. Heat volumes of bias for each performance metric.

This information can also be presented in the form of contour graphs as in Fig. 8 where the colours represent the absolute value of bias.

3.2. Bias indicators depending on δ

In the previous section, several bias indicators depending only on the imbalance coefficient δ were defined. As mentioned therein, the first approach is to consider classifiers whose performance is singularly located on the (λ_{PP}, λ_{NN}) plane. The results obtained for each bias indicator and performance metric $\sigma B_\mu(\delta)$ are summarized in Table 7, where unbiased performance metrics (*SNS, SPC, GM, BMn*) have been omitted.

Table 7
Bias indicators for singular classifiers: $\sigma B_\mu(\delta)$.

	PRC	*NPV*	*ACC*	F_1	*MCCn*	*MKn*
$wcB\mu$	0	0	0	0	0	0
$bcB\mu$	0	0	0	0	0	0
$wpcB\mu$	$\delta/2$	$-\delta/2$	$-\delta/2$	0	0	0
$wncB\mu$	$\delta/2$	$-\delta/2$	$\delta/2$	$\frac{2(1+\delta)}{3+\delta} - \frac{2}{3}$	0	0
$mcB\mu$	$\delta/2$	$-\delta/2$	0	$\frac{1+\delta}{2+\delta} - \frac{1}{2}$	0	0

It can be observed that only four types of non-null indicators appear. Their dependence on δ is plotted in Fig. 9.

Alternatively, it can be assumed that λ_{PP} and λ_{NN} are randomly and uniformly distributed across the [0, 1] range and the bias

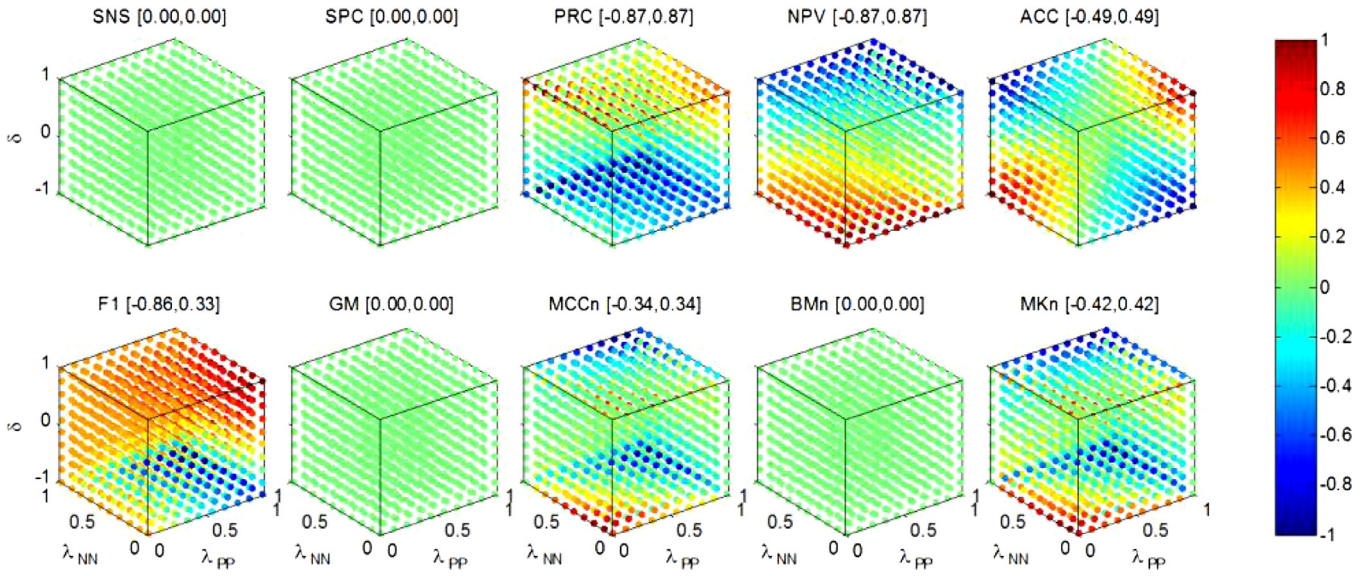

Fig. 5. Heat volumes of relative bias for each performance metric.

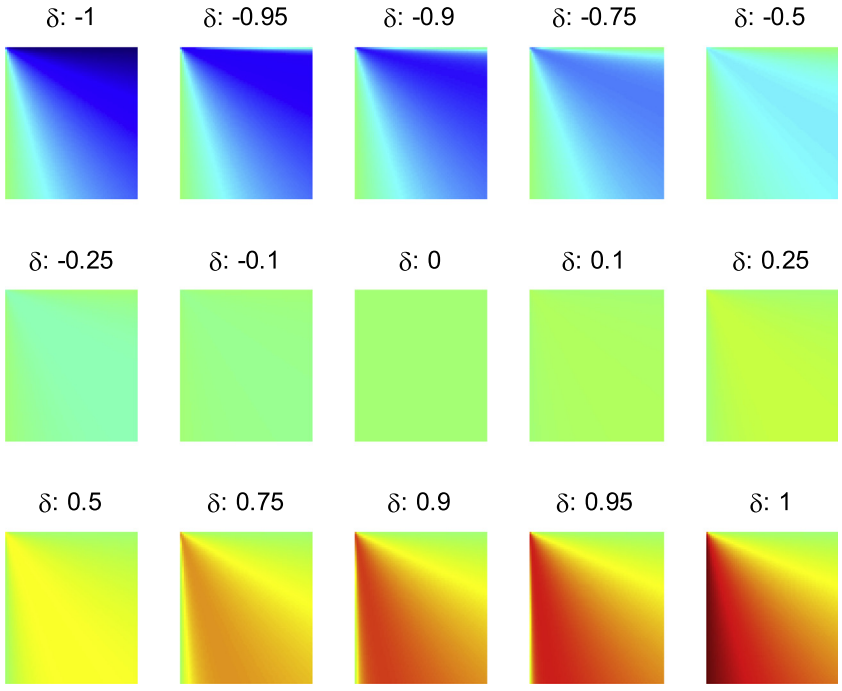

Fig. 6. Set of heat maps of bias for precision.

$B_\mu(\delta)$ can then be statistically characterized. First, the probability density function ($pdf[B_\mu(\delta)]$) is derived for each performance metric. The results are shown in Fig. 10, where, for each δ and each bias B_μ, the values of *pdf* are shown using different colours (dark blue corresponds to $pdf = 0$).

Additionally, several local statistical indicators have been defined (see Table 5) and these are generically denoted as $\psi B_\mu(\delta)$. The results obtained for each performance metric are depicted in Fig. 11. For easier reading, unbiased performance metrics (*SNS, SPC, GM, BMn*) have been omitted. Moreover, *NPV* presents symmetric behaviour to *PRC* and has also been disregarded from the graphs for the sake of simplicity.

3.3. Single-valued bias indicators

As has already been pointed out, there are various ways to obtain single-valued bias indicators. First, we consider bias functions $\sigma B_\mu(\delta)$ as summarized in Table 7. By assuming that δ is randomly and uniformly distributed within the $[-1, 1]$ range, the mean values of each measure can be computed. The results for each performance metric are shown in Table 8.

Now let us consider bias functions $\psi B_\mu(\delta)$ depicted in Fig. 11. By applying the same method, mean values of these measures can also be obtained. Results for each metric are shown in Table 9.

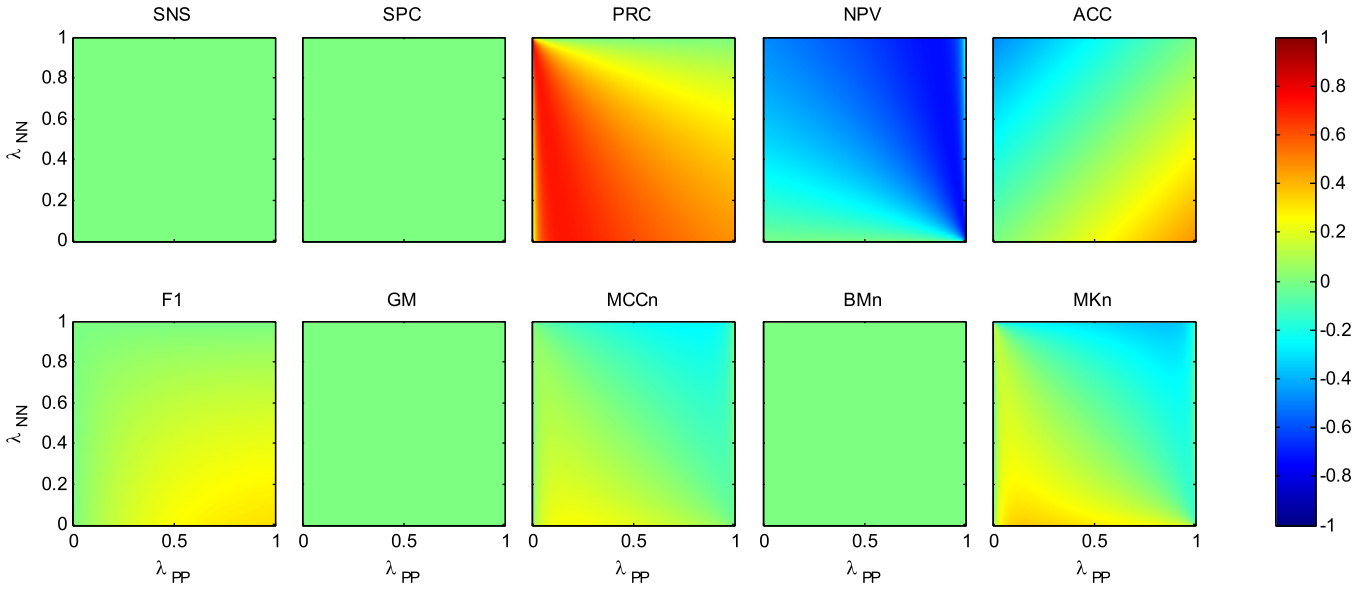

Fig. 7. Heat maps of bias for each performance metric ($\delta = 0.95$).

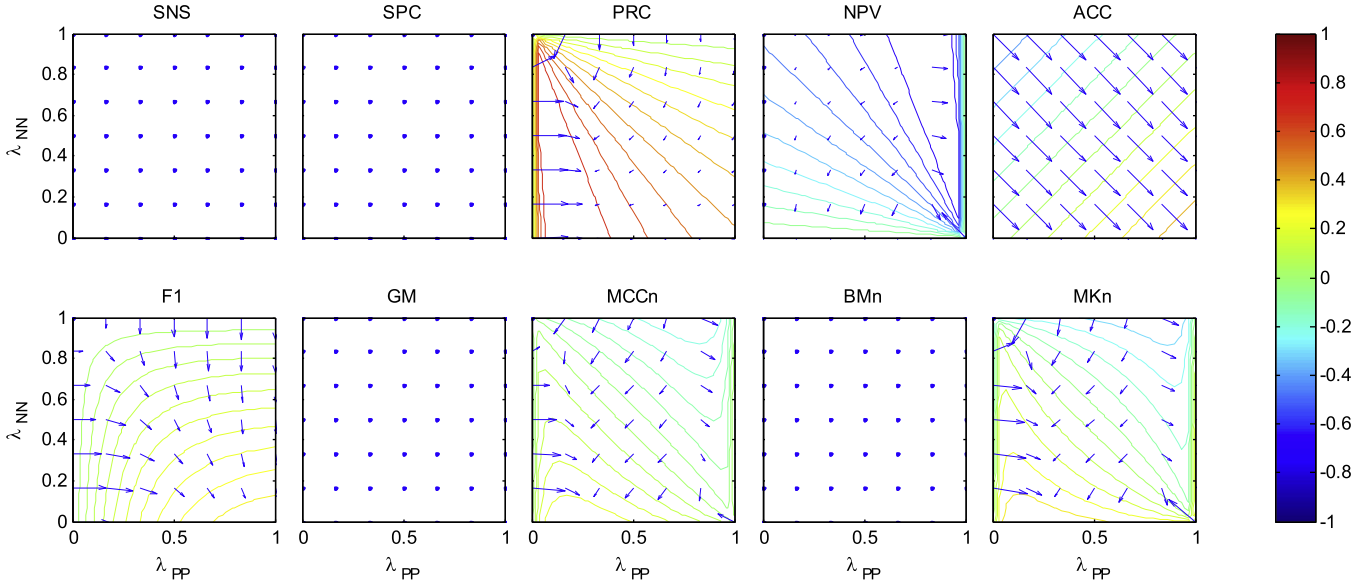

Fig. 8. Contour graphs of bias for each performance metric ($\delta = 0.95$).

Table 8
Mean values of bias functions for singular classifiers $\sigma B_\mu(\delta)$.

	Symbol	*PRC*	*NPV*	*ACC*	F_1	*MCCn*	*MKn*
1	$\overline{wcB}_\mu$	0	0	0	0	0	0
2	$\overline{bcB}_\mu$	0	0	0	0	0	0
3	$\overline{wpcB}_\mu$	0	0	0	0	0	0
4	$\overline{wncB}_\mu$	0	0	0	−0.053	0	0
5	$\overline{mcB}_\mu$	0	0	0	−0.049	0	0

Table 9
Mean values of bias of local statistical indicators $\psi B_\mu(\delta)$.

	Symbol	*PRC*	*NPV*	*ACC*	F_1	*MCCn*	*MKn*
6	$\overline{mB}_\mu$	0	0	0	−0.041	0	0
7	$\overline{sdB}_\mu$	0.082	0.082	0.102	0.066	0.038	0.066
8	$\overline{rmsB}_\mu$	0.228	0.228	0.102	0.135	0.038	0.066
9	$\overline{maxaB}_\mu$	0.308	0.308	0.25	0.244	0.090	0.154
10	$\overline{skB}_\mu$	0	0	0	0.129	0	0
11	$\overline{kB}_\mu$	0.080	0.080	−0.6	−1.093	0.006	0.028

Instead of computing the mean of a function, single-valued bias indicators can also be obtained by focusing on their value for extremely imbalanced datasets. By first considering the bias metrics of singular classifiers ($\sigma B_\mu(\delta)$), their results are shown in Table 10.

Regarding the extremely imbalanced case for local statistical indicators ($\psi B_\mu(\delta)$), their results are shown in Table 11.

For a global view of bias, let us consider that λ_{PP} and λ_{NN} are randomly and uniformly distributed across the [0, 1] range while δ lies within the $[-1, 1]$ range, and then bias B_μ is statistically characterized. First, the probability density function ($pdf(B_\mu)$) is derived for each performance metric. The results are shown in Fig. 12, where *NPV* has been omitted from the graph because it shows symmetric behaviour to *PRC* and has the same *pdf*.

Finally, several global statistical indicators based on $pdf(B_\mu)$ have been defined (see Table 6), which are generically denoted as ΨB_μ. The results obtained for each performance metric are shown in Table 12.

Table 10
Bias indicators on singular classifiers ($\sigma B_\mu(\delta)$) for extremely imbalanced datasets.

		Symbol	*PRC*	*NPV*	*ACC*	F_1	*MCCn*	*MKn*
Extremely	12	$wcB_\mu^{\varepsilon P}$	0.667	0	0	0	0.211	0.333
positive-	13	$bcB_\mu^{\varepsilon P}$	0	−0.667	0	0	−0.211	−0.333
imbalanced	14	$wpcB_\mu^{\varepsilon P}$	0.5	−0.5	−0.5	0	0	0
	15	$wncB_\mu^{\varepsilon P}$	0.5	−0.5	0.5	0.333	0	0
	16	$mcB_\mu^{\varepsilon P}$	0.5	−0.5	0	0.167	0	0
Extremely	17	$wcB_\mu^{\varepsilon N}$	0	0.667	0	0	−0.211	0.333
negative-	18	$bcB_\mu^{\varepsilon N}$	−0.667	0	0	−0.5	0.211	−0.333
imbalanced	19	$wpcB_\mu^{\varepsilon N}$	−0.5	0.5	0.5	0	0	0
	20	$wncB_\mu^{\varepsilon N}$	−0.5	0.5	−0.5	−0.667	0	0
	21	$mcB_\mu^{\varepsilon N}$	−0.5	0.5	0	−0.5	0	0

Table 11
Bias measures on local statistical indicators ($\psi B_\mu(\delta)$) for extremely imbalanced datasets.

		Symbol	*PRC*	*NPV*	*ACC*	F_1	*MCCn*	*MKn*
Extremely	22	$mB_\mu^{\varepsilon P}$	0.5	−0.5	0	0.137	0	0
positive-	23	$sdB_\mu^{\varepsilon P}$	0.238	0.238	0.204	0.088	0.213	0.226
imbalanced	24	$rmsB_\mu^{\varepsilon P}$	0.554	0.554	0.204	0.163	0.213	0.226
	25	$maxaB_\mu^{\varepsilon P}$	1	1	0.5	0.333	0.5	0.5
	26	$skB_\mu^{\varepsilon P}$	0	0	0	0.244	0	0
	27	$kB_\mu^{\varepsilon P}$	−0.651	−0.651	−0.6	−1.043	−0.790	−1.014
Extremely	28	$mB_\mu^{\varepsilon N}$	−0.5	0.5	0	−0.477	0	0
negative-	29	$sdB_\mu^{\varepsilon N}$	0.238	0.238	0.204	0.241	0.213	0.226
imbalanced	30	$rmsB_\mu^{\varepsilon N}$	0.554	0.554	0.204	0.534	0.213	0.226
	31	$maxaB_\mu^{\varepsilon N}$	1	1	0.5	1	0.5	0.5
	32	$skB_\mu^{\varepsilon N}$	0	0	0	0.168	0	0
	33	$kB_\mu^{\varepsilon N}$	−0.651	−0.651	−0.6	−0.933	−0.790	−1.014

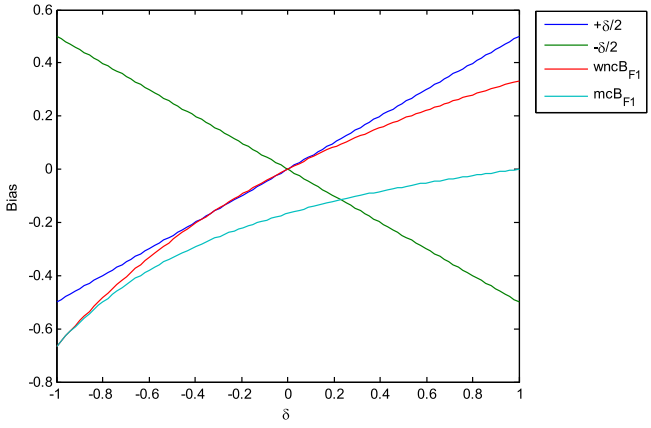

Fig. 9. Four types of bias indicators for singular classifiers.

Table 12
Global statistical indicators of bias ΨB_μ for each performance metric.

	Symbol	*PRC*	*NPV*	*ACC*	F_1	*MCCn*	*MKn*
34	MB_μ	0	0	0	−0.041	0	0
35	SDB_μ	0.271	0.271	0.118	0.169	0.055	0.086
36	$RMSB_\mu$	0.271	0.271	0.118	0.174	0.055	0.086
37	$MAXAB_\mu$	1	1	0.5	1	0.5	0.5
38	SKB_μ	0	0	0	−1.269	0	0
39	KB_μ	−0.046	−0.046	1.32	2.043	6.9	3.061

In this section, 39 single-valued bias indicators have been considered, which can also be presented in a graphical form, as shown in Fig. 13. Here, the bias indicators $\overline{\sigma B_\mu(\delta)}$, detailed in Table 8, have been omitted, since all these indicators are null (except for F_1 score) and therefore their comparison would be meaningless.

3.4. Symmetry of bias functions

In order to categorize the bias functions for each performance metric, it is convenient to study their symmetry. For this purpose, let us begin by considering the Matthews Correlation Coefficient (*MCC*), since this coefficient is particularly clear. Its heat map for an imbalance coefficient $\delta = 0.95$ is shown in the upper left-hand-side plot of Fig. 14. A first anti-clockwise 90° rotation on the $(\lambda_{PP}, \lambda_{NN})$ plane is performed, and the result is shown in the upper right-hand-side plot. The result of a second 90° rotation is shown in the lower right-hand-side graph. Finally, the sign of the bias values is changed, as shown in the lower left-hand-side plot. It can be observed that the result coincides with the original heat map.

The 180° rotation enables the bias to be defined on a new set of axes $(\Lambda_{PP}, \Lambda_{NN})$, which are related to the original $(\lambda_{PP}, \lambda_{NN})$ through the expressions $\Lambda_{PP} = 1 - \lambda_{PP}$; $\Lambda_{NN} = 1 - \lambda_{NN}$. This symmetry can therefore be formalized as

$$\begin{aligned} B_{MCC}(\lambda_{PP}, \lambda_{NN}, \delta) &= -B_{MCC}(\Lambda_{PP}, \Lambda_{NN}, \delta) \\ &= -B_{MCC}(1 - \lambda_{PP}, 1 - \lambda_{NN}, \delta). \end{aligned} \tag{20}$$

Hence, the *MCC* bias function shows an order-2 (180°) rotational odd symmetry (or anti-symmetry) on the $(\lambda_{PP}, \lambda_{NN})$ plane. Furthermore, this bias function shows symmetry with respect to the principal diagonal on the $(\lambda_{PP}, \lambda_{NN})$ plane if the sign of δ is inverted, that is,

$$B_{MCC}(\lambda_{NN}, \lambda_{PP}, -\delta) = B_{MCC}(\lambda_{PP}, \lambda_{NN}, \delta). \tag{21}$$

This dual behaviour is called Type I symmetry. Bias functions for *ACC* and *MK* also exhibit this type of symmetry.

When the bias of precision (*PRC*) is considered, no symmetry on the $(\lambda_{PP}, \lambda_{NN})$ plane can be found. However, it exhibits a symmetry in the $(\lambda_{PP}, \lambda_{NN}, \delta)$ space as can be observed in Fig. 15 where its heat volume for an imbalance coefficient $\delta = 0.95$ is shown in the upper left-hand-side plot. First, a mirror symmetry, with respect to the $\delta = 0$ plane, is performed and the result is

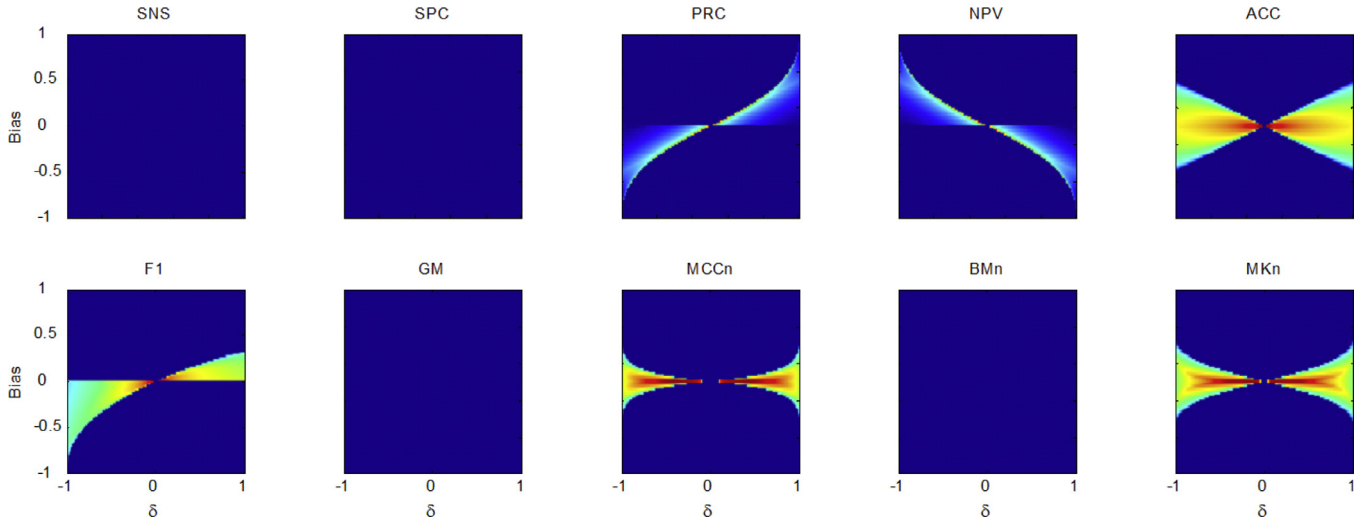

Fig. 10. Probability density function $pdf[B_\mu(\delta)]$ for each performance metric.

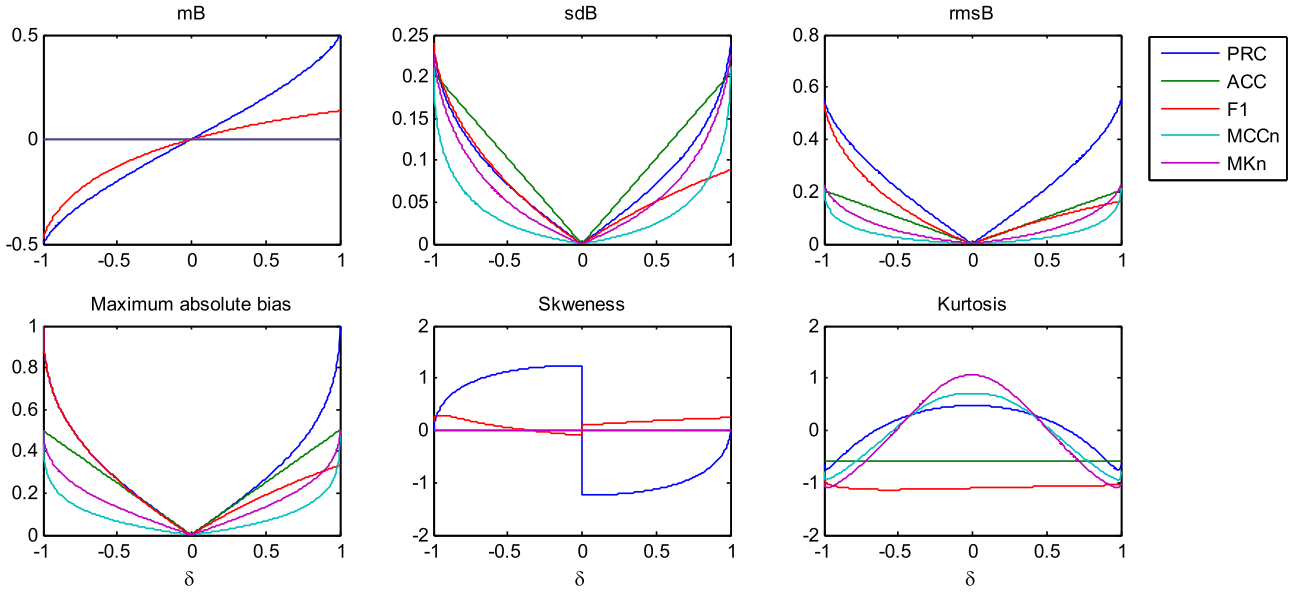

Fig. 11. Local statistical indicators of bias $\psi B_\mu(\delta)$ for each performance metric.

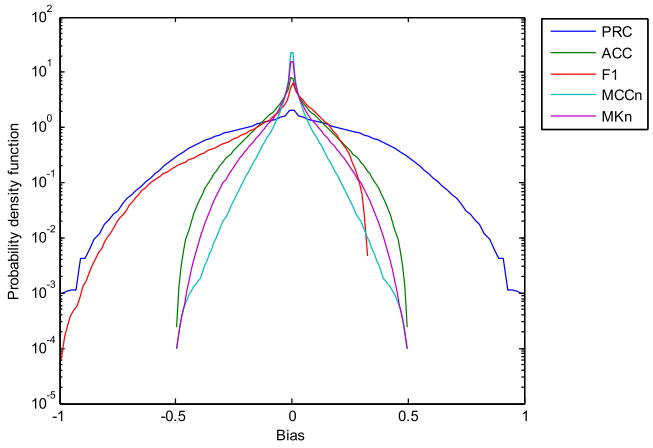

Fig. 12. Probability density function $pdf(B_\mu)$ for each performance metric.

shown in the upper right-hand-side plot. The sign of the bias values is then changed and the results are shown in the lower right-hand-side plot. Finally, a second mirror symmetry is performed, this time with respect to the anti-diagonal plane drawn in the third plot. The result is shown in the lower left-hand-side plot. It can be observed that the result coincides with the original heat volume.

The double mirror symmetry enables the bias to be defined on a new set of axes (Λ_{PP}, Λ_{NN}, Δ), which are related to the original set (λ_{PP}, λ_{NN}, δ) through the expressions $\Lambda_{PP} = 1 - \lambda_{NN}$; $\Lambda_{NN} = 1 - \lambda_{PP}$; $\Delta = -\delta$. This symmetry can hence be formalized as

$$\begin{aligned} B_{PRC}(\lambda_{PP}, \lambda_{NN}, \delta) &= -B_{PRC}(\Lambda_{PP}, \Lambda_{NN}, \Delta) \\ &= -B_{PRC}(1 - \lambda_{NN}, 1 - \lambda_{PP}, -\delta). \end{aligned} \tag{22}$$

Therefore, the *PRC* bias function shows a double mirror odd symmetry (or anti-symmetry) in the (λ_{PP}, λ_{NN}, δ) space. This behaviour is called Type II symmetry. Bias functions for each metric (except F_1 score) exhibit this type of symmetry.

Additionally, the symmetry of each *pdf* can be measured through the skewness statistics.

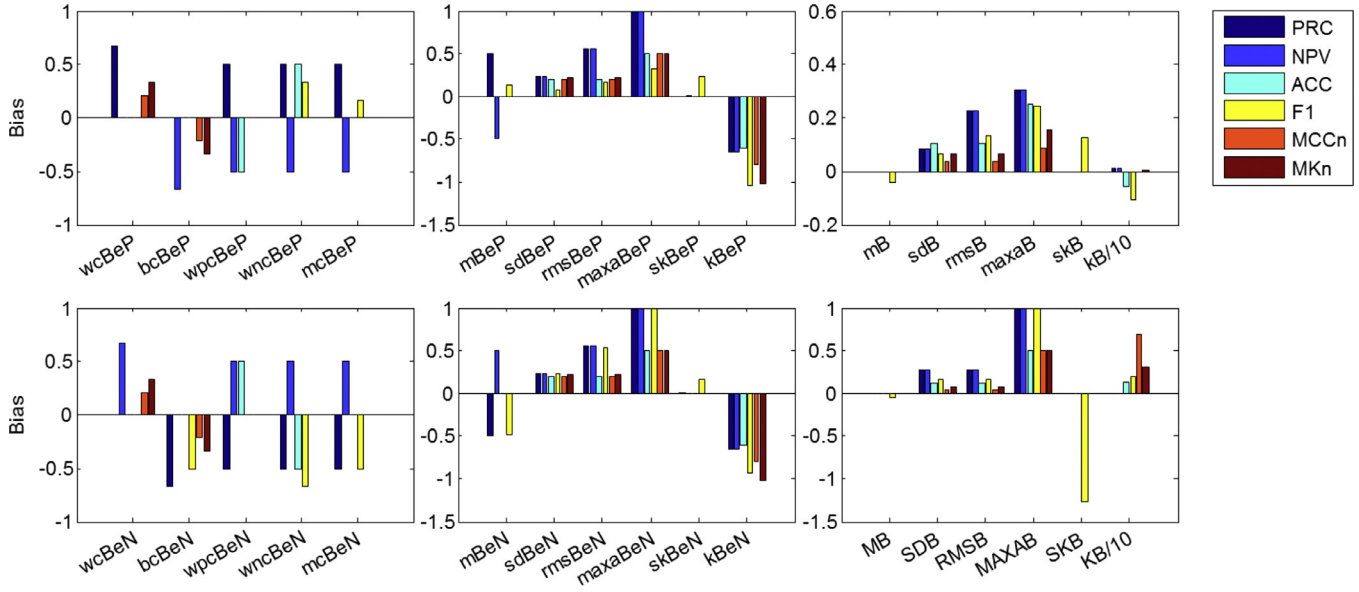

Fig. 13. Single-valued bias indicators for each performance metric.

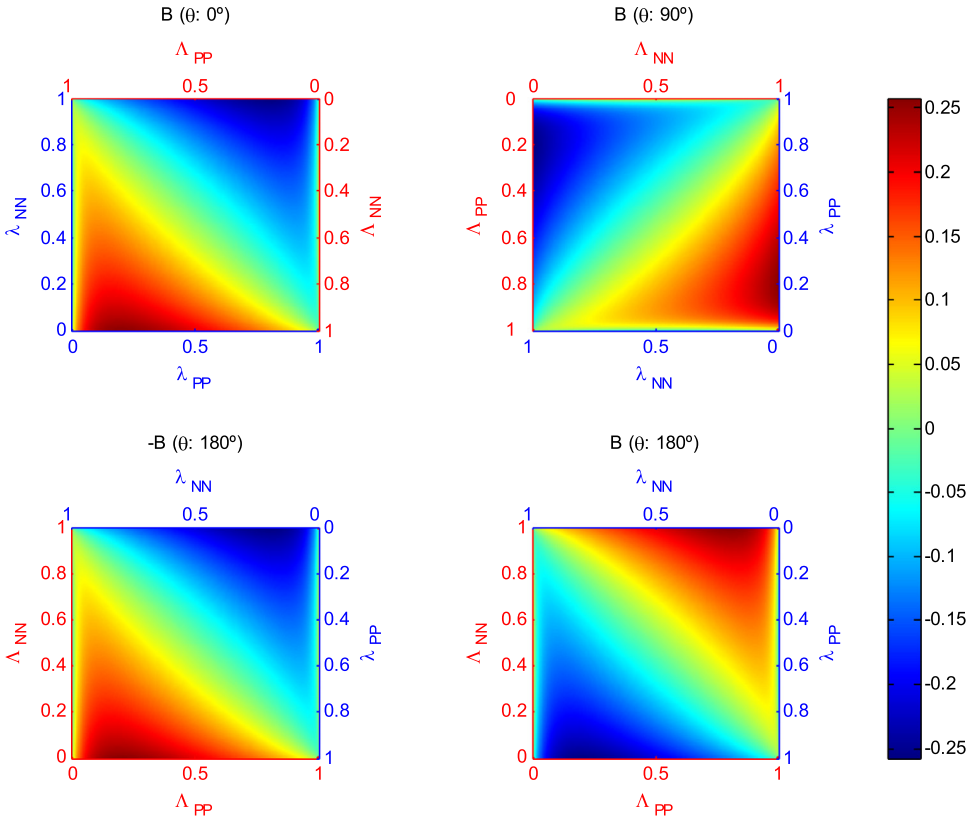

Fig. 14. Study of symmetry for $B_{MCC}(\lambda_{NN}, \lambda_{PP})$ with $\delta = 0.95$.

3.5. Clustering performance metrics based on their bias

In previous sections, the impact of imbalance on ten classification performance metrics is studied. Based on their bias, the performance metrics can now be grouped into several clusters. In order to perform this clustering, the 39 single-valued bias indicators are considered. That is, each performance metric is to be featured by a point in an $\mathbb{R}^{39}$ space.

To tackle the issue of how to visualize such a high-dimensional vector, a reduction of dimensionality to a plane (2D) must be undertaken. The first approach (Fig. 16-A) involves the selection of 2 highly significant bias indicators and the projection of the points on that plane. Our selection is of $RMSB_{\mu}$, which is an indicator of the mean global bias, and of $rmsB_{\mu}^{\varepsilon P}$, which is a mean gauge of bias for extremely positive-imbalanced datasets. In this graphic, it can be observed that: bias for *SNS*, *SPC*, *GM* and *BMn* performance metrics are at the same point; bias for *ACC*, *MCCn* and *MKn* are very close to each other; bias for F_1 is not far from these metrics; and, finally, bias for *PRC* and *NPV* are at the same point but distant from the other metrics.

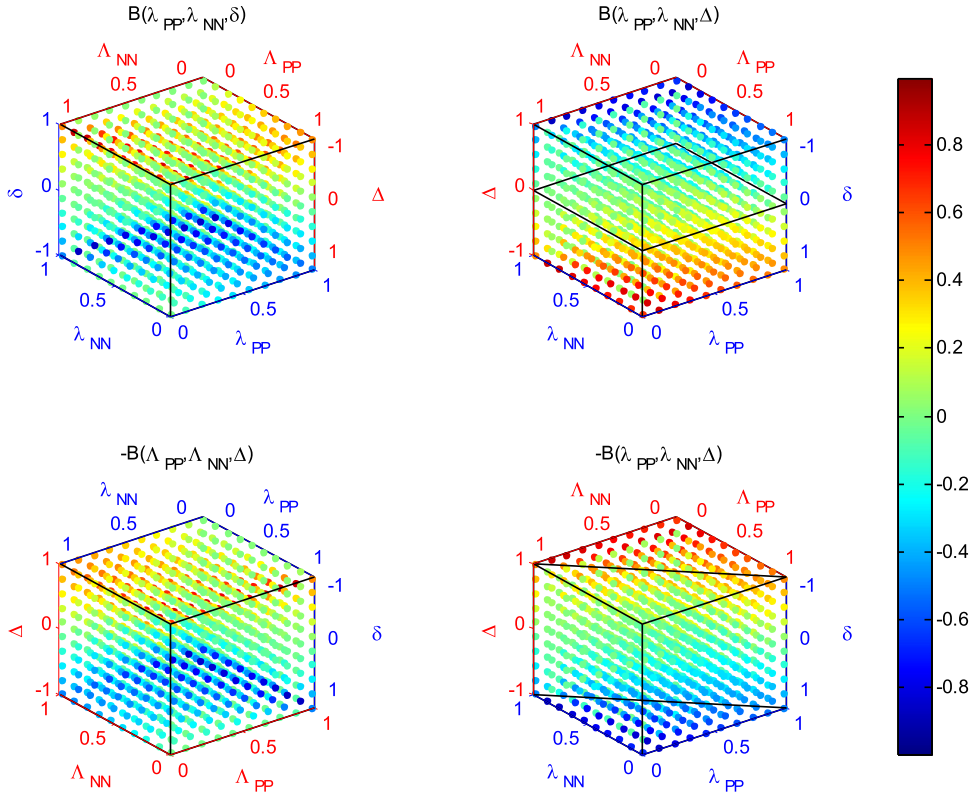

Fig. 15. Study of symmetry for $B_{PRC}(\lambda_{NN}, \lambda_{PP}, \delta)$ with $\delta = 0.95$.

Following these criteria, either 3 or 4 clusters could be established.

However, a more in-depth insight shows that *PRC* and *NPV* have symmetric behaviour for many bias indicators. They have appeared together in the previous graph because the selected indicators ($RMSB_\mu$ and $rmsB_\mu^{\varepsilon P}$) compute a squared mean, which hides their symmetric characteristics. In order to overcome this issue, the dimensionality reduction can be made by selecting a different pair of bias indicators. In Fig. 16-B, *MAXAB* is a global indicator of the absolute maximum value of bias (and still hides the symmetry), and $mB_\mu^{\varepsilon P}$is another mean gauge of bias for extremely positive-imbalanced datasets that reveals the symmetry. Five clusters now clearly appear.

An alternative to the previous somewhat arbitrary and reductionist selection of the pair of bias indicators involves the consideration of the full set of indicators and the performance of some kind of bidimensional reduction. Principal Component Analysis (PCA), shown in Fig. 16-C [26], and Multi-dimensional Scaling (MDS), shown in Fig. 16-D [7], are employed as the techniques for this reduction.

Each panel on Fig. 16 represents a different bidimensional perspective of highly dimensional ($\mathbb{R}^{39}$) set of points. Therefore, slightly different clustering may arise in any of them. But considering all the panels, it can be seen that the following 5 clusters appear:

I. Cluster comprised of *SNS, SPC, GM* and *BMn* with metrics having null bias.

II. Cluster comprised of *ACC, MCCn* and *MKn* with the following features:
- Bias has Type I symmetry, that is, $B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) = -B_\mu(1-\lambda_{PP}, 1-\lambda_{NN}, \delta)$ and $B_\mu(\lambda_{NN}, \lambda_{PP}, -\delta) = B_\mu(\lambda_{PP}, \lambda_{NN}, \delta)$.
- Bias *pdf* has null skewness.
- Bias values are moderate: 0.5 for maximum ($MAXAB_\mu$, $maxaB_\mu^{\varepsilon P}$ and $maxaB_\mu^{\varepsilon N}$); and about 0.2 for average bias in extremely imbalanced datasets ($rmsB_\mu^{\varepsilon P}$ and $rmsB_\mu^{\varepsilon N}$).
- Sign of bias does not depend on sign of imbalance.

III. Cluster (with 2 subclusters) comprised of *PRC* and *NPV* with the following features:
- Bias has Type II symmetry, that is, $B_\mu(\lambda_{PP}, \lambda_{NN}, \delta) = -B_\mu(1-\lambda_{NN}, 1-\lambda_{PP}, -\delta)$.
- Bias *pdf* has null skewness.
- Bias values are high: 1 for maximum ($MAXAB_\mu$, $maxaB_\mu^{\varepsilon P}$ and $maxaB_\mu^{\varepsilon N}$); and about 0.5 for average bias in extremely imbalanced datasets ($rmsB_\mu^{\varepsilon P}$ and $rmsB_\mu^{\varepsilon N}$).
- The relationship between the sign of bias and the sign of imbalance establishes 2 subclusters.
  A. *PRC*, with bias and imbalance having the same sign.
  B. *NPV*, with bias and imbalance having the opposite sign.

IV. Cluster comprised by F_1 with the following features:
- Bias has no symmetry.
- Bias *pdf* has non-null skewness.
- Bias values are low for positive imbalance: 0.33 for maximum ($maxaB_\mu^{\varepsilon P}$); and approximately 0.15 for average bias in extremely imbalanced datasets ($rmsB_\mu^{\varepsilon P}$).
- Bias values are high for negative imbalance: 1 for maximum ($maxaB_\mu^{\varepsilon N}$); and approximately 0.5 for average bias in extremely imbalanced datasets ($rmsB_\mu^{\varepsilon N}$).
- Sign of bias and sign of imbalance are the same.

Clustering information is summarized in Table 13.

Another way to represent how performance metrics are grouped according to the bias behaviour is by drawing a dendrogram. To this end, the full set of bias indicators is employed to feature each performance metric. The distances between the metrics are then computed in the space of the $\mathbb{R}^{39}$ bias indicators. These

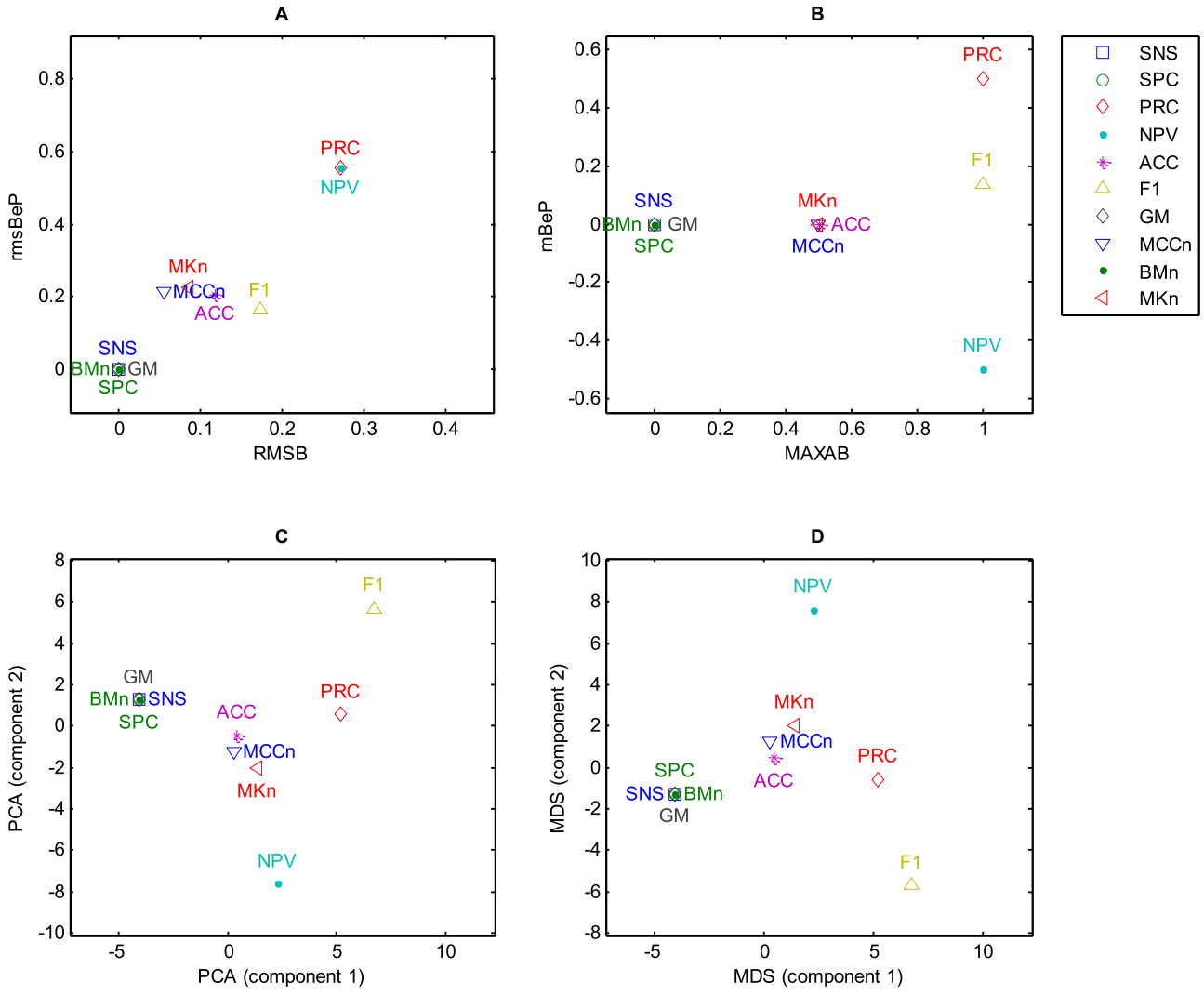

Fig. 16. Bidimensional representation of performance metrics according to their bias indicators.

Table 13
Clusters of performance metrics attending to their bias.

		I		II		III.A		III.B		IV	
		SNS SPCGM BM		ACC MCC MK		PRC		NPV		F1	
$\delta > 0$	$maxaB_{\mu}^{\varepsilon P}$	Null	0	Medium	0.5	High	1	High	1	Low	0.333
	$rmsB_{\mu}^{\varepsilon P}$		0		~ 0.2		0.554		0.554		0.163
$\delta < 0$	$maxaB_{\mu}^{\varepsilon N}$		0		0.5		1		1	High	1
	$\mathrm{rmsB}_{\mu}^{\varepsilon N}$		0		~ 0.2		0.554		0.554		0.534
Symmetry	Type I	Yes		Yes		Yes		Yes		No	
	Type II	Yes		Yes		No		No		No	
	Skewness	0		0		0		0		$\neq 0$	
$sgn(B)$ vs. $sgn(\delta)$		=		Independent		=		$\neq$		=	

distances are employed to gauge how separated the metrics are, as shown in Fig. 17. Once again, 5 clusters clearly appear.

4. Discussion

The first issue to be discussed is the comparison between the imbalance ratio (*IR*), defined by several authors to quantify imbalance, and the imbalance coefficient (δ) as proposed in this paper. Although they are both valid indicators of the degree to which the datasets are imbalanced, we prefer δ because it is defined within the $[-1, +1]$ bounded range with the balanced case ($\delta = 0$) in the middle of the segment (and hence it is symmetric); whilst *IR* is defined within the $[0, +\infty]$ unbounded range with the balanced case $IR = 1$, which is clearly asymmetric. In order to obtain symmetric behaviour based on *IR*, the logarithm of *IR* could be used (*LIR*), whose range is $[-\infty, +\infty]$ with the balanced case ($LIR = 0$) in the middle; however, its range still remains unbounded. Fig. 18 shows an example of a local statistical indicator of bias (*rmsB*) as a function of the imbalance using δ (left-hand-side) and *IR* in logarithmic scale (right-hand-side).

A practical application of the above results is that the bias's mean value of every performance metric (and other related statistics) can be computed using the equations in Table 6, and their results for the ten studied metrics are shown in Table 12.

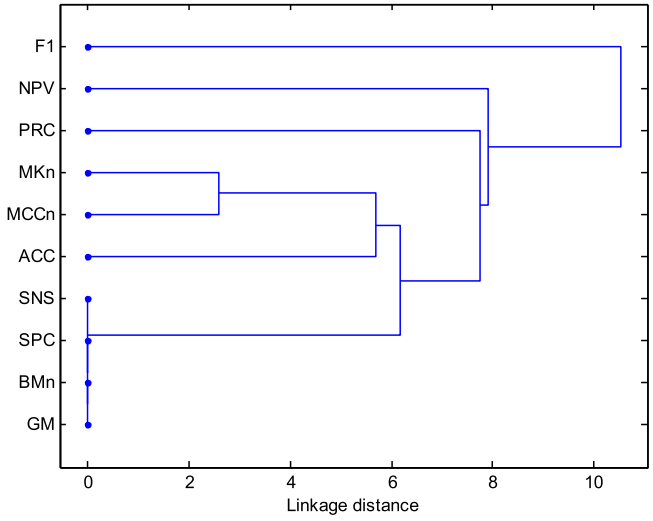

Fig. 17. Dendrogram of performance metrics according to their bias measures.

In the cases where the dataset is known, the expected bias of every performance metric can be computed. Indeed, the dataset determine the value of the imbalance coefficient δ and, for that value, the bias (its root-mean-square value) can be obtained as it is shown in Fig. 18. The full set of bias's statistics can be computed using the equations in Table 5 and their results are depicted in Fig. 11.

Additionally if the classifier results on that dataset are also known (that is, the values of λ_{PP} and λ_{NN}) the bias's exact value of every performance metric can be computed using the equations in Table 3.

Let us now focus on the bias functions. From the above results it is clear that the best performance metrics, that is, those with no bias due to imbalance, are those of sensitivity (*SNS*) and specificity (*SPC*). These metrics can be considered one-dimensional (or partial) performance metrics however, since they take into account only the results on either the positive (*SNS*) or the negative (*SPC*) class, but not both.

Null bias is also shown by two metrics directly depending on *SNS* and *SPC*: geometric mean (*GM*) and bookmaker informedness (*BM* or single-threshold *AUC*). These solve the one-dimensionality problem of the *SNS* and *SPC* metrics by considering either their arithmetic (*BM*) or geometric (*GM*) mean. Although these two metrics constitute good alternatives to be used with imbalanced datasets, they have the drawback of focusing on only the classification successes (λ_{PP} and λ_{NN}), and fail to directly consider the classification errors (λ_{PN} and λ_{NP}).

Table 14
Behaviour of performance metrics with imbalanced datasets.

Cluster	Metric	Bias	$RMSB_\mu$	Focus on classes (Positive, Negative)	Focus on results (Successes, Errors)
I	*SNS*	Null	0	P	S
	SPC	Null	0	N	S
	GM	Null	0	P & N	S
	BM	Null	0	P & N	S
II	*ACC*	Medium	0.118	P & N	S
	MCC	Medium	0.055	P & N	S & E
	MK	Medium	0.086	P & N	S & E
III	*PRC*	High	0.271	P & N	S & E
	NPV	High	0.271	P & N	S & E
IV	F_1	High	0.174	P & N	S & E

The second best (lowest biased) cluster of metrics is that which is comprised of accuracy (*ACC*), Matthews correlation coefficient (*MCC*), and markedness (*MK*). These all have a global (not partial) perspective, since classification results on both positive and negative classes are considered. From among these 3 metrics, *ACC* focuses only on the classification successes, which is a drawback and, additionally, has the highest bias (except when extreme balanced datasets are used). In this cluster, the lowest bias is shown by *MCC* with moderate values (lower than 0.2 in the normalized version) for almost every value of the imbalance coefficient.

Finally, the metrics in the third and fourth clusters, precision (*PRC*), negative predictive value (*NPV*), and F_1 score (F_1), are highly biased and should be avoided for use in imbalanced datasets. Table 14 summarizes the behaviour of performance metrics with imbalanced datasets.

As a practical conclusion, when dealing with imbalanced datasets, *GM* and *BM* are the best performance metrics if their focus on successes (dismissing the errors) presents no limitation for the specific application where they are used. However, if classification errors must also be considered, then *MCC* arises as the best choice.

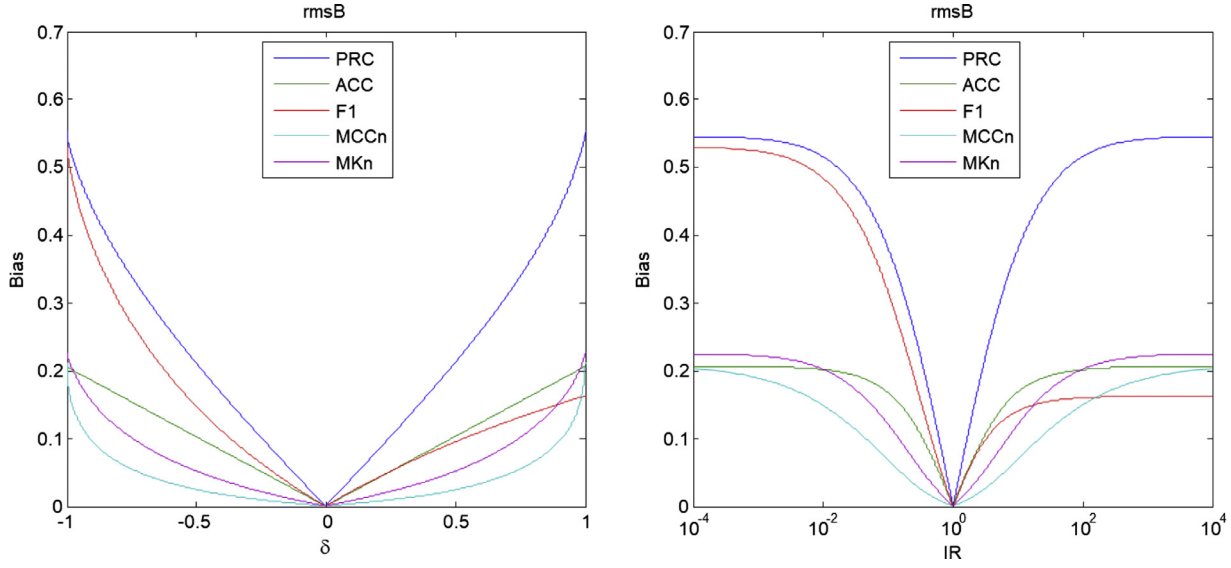

Fig. 18. Comparison of δ and *IR* to measure the imbalance.

Concordant results have been obtained in other studies, although most of these previous results were only shown and not exhaustively explored and quantified. The weakness of *ACC* in imbalanced problems has been signalled by many authors [10,22]. Furthermore, the use of *PRC* has been extensively discouraged [5,13,20]. F_1 score, which depends on *PRC*, is also indirectly dismissed by those authors and directly rejected by Jeni, Cohn & De La Torre [25]. Most of the literature on this issue does not select any performance metrics, thereby limiting their study to a mere indication that they are biased. A few authors also suggest that the best choice is the *MCC* metric [4,11].

Probably the most cited solution to overcome the effect of imbalance on performance metrics is the use of the Class Balanced Accuracy (*CBA*). In the terminology used throughout this paper, this is the accuracy for the classifier operating on a balanced dataset (ACC_b). Nevertheless, according to our study, this idea can be extended to the remaining performance metrics, by obtaining their balanced counterpart (μ_b), generally called Class Balance Metrics (*CBM*), as formulated in the last column of Table 2. These extensions permit different null-bias perspectives to be used in the assessment of the results obtained by a classifier in the imbalanced case.

The Class Balance version of the ten studied metrics (μ_b) will show a null bias and, therefore, a value of $RMSB_\mu = 0$ (columns 3 and 4 in Table 14). As every metric is now unbiased, choosing any of them should be based on other criteria, for instance, on their symmetry (column 1), their focus on classes (column 5) or their focus on results (last column). These values remain unaltered with respect to their biased counterpart.

The behaviour of each Class Balance Metric is shown in Fig. 2 which can also be used as a guide for the selection of the metric.

5. Conclusions

In this paper, an extensive and systematic study of the impact of class imbalance on classification performance metrics is undertaken. To the best of our knowledge, no quantitative and complete study of this issue has been previously published.

To characterize the disparity between classes the Imbalance Coefficient has been defined: a new measure which surpasses the Imbalance Ratio or the Entropy used in previous studies.

Throughout our analysis several practical procedures to determine the bias's quantitative value of a metric have been derived, either for a general case, for a certain dataset or for a given experiment (pair of classifier and dataset).

From the simulation results, a quantitatively justified guide to select performance metrics in the presence of imbalance classes has been developed. In our analysis, several clusters of performance metrics have been identified that involve the use of Geometric Mean or Bookmaker Informedness as the best null-biased metrics if their focus on classification successes (dismissing the errors) present no limitation for the specific application where they are used. However, if classification errors must also be considered, then the Matthews Correlation Coefficient arises as the best choice.

Finally, a set of null-biased multi-perspective Class Balance Metrics is proposed which extends the concept of Class Balance Accuracy to other performance metrics.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2019.02.023.

References

[1] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, … A. Hussain, Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study, IEEE Access 4 (2016) 7940–7957.
[2] G.E. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explor. *Newslett.* 6 (1) (2004) 20–29.
[3] C. Beyan, R. Fisher, Classifying imbalanced data sets using similarity based hierarchical decomposition, Pattern Recognit. 48 (5) (2015) 1653–1672.
[4] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PloS One 12 (6) (2017) e0177678.
[5] P. Branco, L. Torgo, R.P. Ribeiro, Relevance-based evaluation metrics for multi–class imbalanced domains, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Cham, Springer, 2017, May, pp. 698–710.
[6] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE, 2010, August, pp. 3121–3124.
[7] R. Caruana, A. Niculescu-Mizil, Data mining in metric space: an empirical analysis of supervised learning performance criteria, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, August, pp. 69–78.
[8] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, Addressing imbalance in multilabel classification: measures and random resampling algorithms, Neurocomputing 163 (2015) 3–16.
[9] 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.
[10] N.V. Chawla, Data mining for imbalanced datasets: An Overview, in: Data Mining and Knowledge Discovery Handbook, Springer US, 2005, pp. 853–867.
[11] D. Chicco, Ten quick tips for machine learning in computational biology, BioData Min. 10 (1) (2017) 35.
[12] R.A. Dara, M.S. Kamel, N. Wanas, Data dependency in multiple classifier systems, Pattern Recognit. 42 (7) (2009) 1260–1273.
[13] S. Daskalaki, I. Kopanas, N. Avouris, Evaluation of classifiers for an uneven class distribution problem, Appl. Artif. Intell. 20 (5) (2006) 381–417.
[14] C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognit. Lett. 30 (1) (2009) 27–38.
[15] A. Fernández, S. del Río, N.V. Chawla, F. Herrera, An insight into imbalanced Big Data classification: outcomes and challenges, Complex Intell. Syst. 3 (2) (2017) 105–120.
[16] P.A. Flach, The geometry of ROC space: understanding machine learning metrics through ROC isometrics, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 194–201.
[17] V. Ganganwar, An overview of classification algorithms for imbalanced datasets, Int. J. Emerg. Technol. Adv. Eng. 2 (4) (2012) 42–47.
[18] V. García, R.A. Mollineda, J.S. Sánchez, Index of balanced accuracy: a performance measure for skewed class distributions, in: Iberian Conference on Pattern Recognition and Image Analysis, Berlin, Heidelberg, Springer, 2009, June, pp. 441–448.
[19] J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient, Comput. Biol. Chem. 28 (5–6) (2004) 367–374.
[20] Q. Gu, L. Zhu, Z. Cai, Evaluation measures of the classification performance of imbalanced data sets, in: International Symposium on Intelligence Computation and Applications, Berlin, Heidelberg, Springer, 2009, October, pp. 461–471.
[21] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[22] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, Int. J. Data Min. Knowl. Manage. Process 5 (2) (2015) 1.
[23] T. Kautz, B.M. Eskofier, C.F. Pasluosta, Generic performance measure for multiclass-classifiers, Pattern Recognit. 68 (2017) 111–125.
[24] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (4) (2016) 221–232.
[25] L.A. Jeni, J.F. Cohn, F. De La Torre, Facing imbalanced data–recommendations for the use of performance metrics, in: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference, IEEE, 2013, September, pp. 245–251.
[26] I. Jolliffe, Principal component analysis, in: International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011, pp. 1094–1096.
[27] G. Jurman, S. Riccadonna, C. Furlanello, A comparison of MCC and CEN error measures in multi-class prediction, PLoS One 7 (8) (2012) e41882.
[28] M.S. Kraiem, M.N. Moreno, Effectiveness of basic and advanced sampling strategies on the classification of imbalanced data. a comparative study using classical and novel metrics, in: International Conference on Hybrid Artificial Intelligence Systems, Cham, Springer, 2017, June, pp. 233–245.
[29] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.
[30] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA) 405 (2) (1975) 442–451.
[31] N. Mehra, S. Gupta, Survey on multiclass classification methods, Int. J. Comput. Sci. Inf. Technol. 4 (4) (2013) 572–576.
[32] L. Mosley, A Balanced Approach to the Multi-Class Imbalance Problem, Iowa State University, 2013.

[33] H. Núñez, L. Gonzalez-Abril, C. Angulo, Improving SVM classification on imbalanced datasets by introducing a new bias, J. Classification 34 (3) (2017) 427–443.
[34] D.M. Powers, Evaluation: from precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation, School of Informatics and Engineering, Flinders University, Adelaide, Australia, 2011 Technical Report SIE-07-001.
[35] B. Raman, T.R. Ioerger, Enhancing Learning Using Feature and Example Selection, Texas A&M University, College Station, TX, USA, 2003.
[36] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation, in: Australasian Joint Conference on Artificial Intelligence, Berlin, Heidelberg, Springer, 2006, December, pp. 1015–1021.
[37] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (4) (2009) 427–437.
[38] S.V. Stehman, Selecting and interpreting measures of thematic classification accuracy, Remote Sens. Environ. 62 (1) (1997) 77–89.
[39] Y. Sun, M.S. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognit. 40 (12) (2007) 3358–3378.
[40] X. Yuan, L. Xie, M. Abouelenien, A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data, Pattern Recognit. 77 (2018) 160–172.
[41] W. Zong, G.B. Huang, Y. Chen, Weighted extreme learning machine for imbalance learning, Neurocomputing 101 (2013) 229–242.

Amalia Luque received her Industrial Engineering degree in 2007, Master in Automation in 2010 and her Doctoral Degree in 2014. She has been involved in teaching related to Project Engineering at the University of Seville since 2015. Her main areas of research are business intelligence, data mining and feature extraction.

Alejandro Carrasco received his Computer Engineering degree in 1998 and his Ph.D. in 2003. He is a Lecturer and Researcher at the University of Seville since 1998, has founded a NTBF and directed several R&D projects. His main areas of research are data mining, industrial computing and network security.

Alejandro M. Martín-Gómez, B. Industrial Electronics and Automation Engineer, MS Product and facilities design and development, MS Security and health and PhD Student (manufacturing engineering). He worked across a wide variety of projects of industrial sector. He is assistant professor at University of Seville (Field of Knowledge: Engineering Project).

Ana de las Heras García de Vinuesa is a BSc. Eng Industrial Design, MSc. Ecodesign, and PhD. in Manufacturing and Environmental Engineering. She works as an assistant professor in the Department of Design Engineering, Engineering Project area, at University of Seville, Spain.