Information Sciences 463–464 (2018) 323–343

Contents lists available at ScienceDirect

Information Sciences

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

A two-tiered 2D visual tool for assessing classifier performance

Giuliano Armano*, Alessandro Giuliani

DIEE – Dept. of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi 09123, Cagliari, Italy

ARTICLE INFO

Article history:
Received 5 March 2017
Revised 20 May 2018
Accepted 20 June 2018
Available online 21 June 2018

Keywords:
Classifier performance measures
Confusion matrices
ROC curves
Coverage plots

ABSTRACT

In this article, a new kind of 2D tool is proposed, namely $\langle \varphi, \delta \rangle$ diagrams, able to highlight most of the information deemed relevant for classifier building and assessment. In particular, accuracy, bias and break-even points are immediately evident therein. These diagrams come in two different forms: the first is aimed at representing the phenomenon under investigation in a space where the imbalance between negative and positive samples is not taken into account, the second (which is a generalization of the first) is able to visualize relevant information in a space that accounts also for the imbalance. According to a specific design choice, all properties found in the first space hold also in the second. The combined use of φ and δ can give important information to researchers involved in the activity of building intelligent systems, in particular for classifier performance assessment and feature ranking/selection.

1. Introduction

Many proposals have been made for measuring the performance of classification tasks (see [15,23,26] for extensive information on this matter). There are also several graphical representations and tools for model evaluation. In particular, *ROC* curves [3,13,25] are a 2D visual tool widely acknowledged as the default choice for assessing the intrinsic behavior of a classifier. Various aspects of *ROC* curves have been extensively studied by the machine learning community, including: (i) isometrics of relevant measures [16,27,28], (ii) cost curves [9–11], and (iii) confidence bands [22]. Several proposals have been made with the aim of extending the descriptive power of *ROC* curves, including those on detection error tradeoff [24], on instance-varying costs [14], on adapting *ROC* curves for regression tasks [20], on the relationship between *ROC* analysis and Bayesian models [6], and on explicitly representing the imbalance (e.g., coverage plots [17]). The area under the curve has been actively investigated as well (see, for instance, [4,5,19,21]), for it is considered a good estimate of the discriminant power of a classifier. A further category of measures is aimed at checking the propension to classify inputs as belonging to the positive or negative class (bias) and to what extent the training set affects performance (variance). See also [8] for more information on bias and variance.

The vast majority of performance measures is affected by the imbalance between positive and negative samples. This concept can be given in numerical terms using the ratio between the amount of negative and positive samples (i.e., *class ratio*). If one wants to assess the intrinsic properties of a classifier, adopting a measure that accounts for the class ratio may not be a reliable choice. In fact, although many practical problems are typically unbalanced, most of the existing performance

* Corresponding author.
E-mail address: armano@diee.unica.it (G. Armano).

https://doi.org/10.1016/j.ins.2018.06.052
0020-0255/

Table 1
Summary table on the most relevant aspects concerning $\langle \varphi, \delta \rangle$ measures and diagrams.

Kind of diagram	Notation	Based on	Class ratio	Focus on
Standard	$\langle \varphi, \delta \rangle$	*Unbiased* measures	No	*Intrinsic* properties
Generalized	$\langle \varphi_b, \delta_b \rangle$	*Biased* measures	Yes	*Actual* properties

measures (e.g., accuracy) become progressively meaningless with increasing or decreasing class ratio (e.g., [12,18]). This fact may be worsened by a lack of statistical significance of experimental results, which may hold for minority test samples. While no practical solution able to contrast the latter issue exists, the former is typically dealt with by adopting a pair of measures (e.g., precision and recall, or specificity and sensitivity). Besides, also *ROC* diagrams follow this approach, the default choice being false positive rate (i.e., 1 – specificity) on the x axis and true positive rate (i.e., sensitivity) on the y axis.

In this article, two measures (i.e., φ and δ) are proposed, which allow to assess the performance of classifiers according to a *bias* vs. *accuracy* perspective. These measures are framed in two different kinds of 2D visual tools, i.e., standard and generalized $\langle \varphi, \delta \rangle$ diagrams. The primary goal of the former is to highlight the intrinsic performance of a classifier, regardless of the imbalance of the dataset at hand, whereas the latter have been devised to investigate how the underlying statistics of data affects the behavior of a classifier.[1] To these ends, standard $\langle \varphi, \delta \rangle$ diagrams rely on measures that are not affected by the class ratio, whereas generalized ones account also for the class ratio. In particular, in a standard scenario φ and δ will give information about *unbiased* bias and accuracy, whereas in a generalized scenario φ and δ will give information about *biased* (i.e., actual) bias and accuracy. As the terms "unbiased" and "biased" do not belong to the classical jargon, a full section will be devoted to illustrate the corresponding concepts. So far, let us concentrate on Table 1, which gives a sort of "roadmap" aimed at shedding light on the most relevant aspects concerning $\langle \varphi, \delta \rangle$ measures and diagrams. We are confident on its usefulness for the interested reader.

The remainder of this article is organized as follows: Section 2 points out the existence of unbiased and biased spaces, devoted to highlight intrinsic and actual properties of classifiers. Section 3 introduces φ and δ, first as measures and then as diagrams. This section encompasses also details concerning mutual information, break-even points and relevant isometrics. Section 4 generalizes $\langle \varphi, \delta \rangle$ diagrams, allowing class ratio to be explicitly represented. Section 5 illustrates experimental settings, and Section 6 summarizes some relevant use cases in which $\langle \varphi, \delta \rangle$ diagrams are put into practice. In particular, two main scenarios are described therein: classifier assessment and feature ranking (the fact that $\langle \varphi, \delta \rangle$ measures can also be used to assess features should not be surprising, as a feature can always be considered a simple kind of classifier in itself). Section 7 points out the strengths and weaknesses of this proposal, and Section 8 draws conclusions.

The introductory part of this work partially overlaps the one published in [1]. However, this was purposeful, as this article is intended to become a sort of reference for all researchers that will decide to adopt $\langle \varphi, \delta \rangle$ diagrams for measuring the performance of binary classifiers or for performing feature importance analysis. Any other material, including (i) the semantics of φ and δ axes for both unbiased and biased cases, (ii) a complete study on the relation that holds between $\langle \varphi, \delta \rangle$ measures and mutual information, as well as (iii) isometrics of the most acknowledged performance measures (i.e., accuracy, precision, negative predictive value, sensitivity and specificity), is totally unpublished. Notably, all details about the way relevant equations have been derived can be found in the supplementary material, i.e., in *Appendix A* and in *Appendix B*. The former is devoted to standard $\langle \varphi, \delta \rangle$ measures, whereas the latter to generalized ones.

Not least of all, the reader should also be aware that this is a methodological article, aimed at illustrating and analyzing new measure spaces able to highlight at a glance some classifier or feature properties deemed relevant by the machine learning community. Notwithstanding this perspective, care has been taken to provide a full experimental section, with the goal of giving researchers a flavor of $\langle \varphi, \delta \rangle$ diagrams inherent potential.

2. Unbiased and biased spaces for classifier assessment

Be $\Xi_c(P, N)$ the confusion matrix of a test run in which a classifier $\widehat{c}$ trained on a class c is fed with P positive samples and N negative samples, with a total of M samples. In particular, ξ_{00}, ξ_{01}, ξ_{10}, and ξ_{11} represent true negatives (*TN*), false positives (*FP*), false negatives (*FN*), and true positives (*TP*), respectively. Assuming statistical significance, the confusion matrix, possibly averaged over multiple tests, is expected to give reliable information on the performance of the classifier at hand. Using Bayes decomposition, we can write:

$$\Xi_c(P,N) = M \cdot \underbrace{\begin{bmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{bmatrix}}_{\Omega(c) \approx p(X_c, \widehat{X}_c)} = M \cdot \underbrace{\begin{bmatrix} n & 0 \\ 0 & p \end{bmatrix}}_{\mathcal{O}(c) \approx p(X_c)} \cdot \underbrace{\begin{bmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{bmatrix}}_{\Gamma(c) \approx p(\widehat{X}_c | X_c)} \tag{1}$$

where:

[1] Theoretically, these two tasks are very different. In particular, the former is strictly related to decisions to be taken according to a maximum-likelihood (ML) criterion, whereas the latter on a maximum a posteriori probability (MAP) criterion.

- X_c and $\widehat{X}_c$ are two random variables used to denote oracle and classifier, respectively.
- $\Omega(c)$ is an estimate of the joint probability $p(X_c, \widehat{X}_c)$. In particular, $\omega_{ij} \approx p(e_{ij})$, $i, j = 0, 1$, denotes the joint occurrence of correct classifications ($i = j$) or misclassifications ($i \neq j$). According to the total probability law: $|\Omega| = \sum_{ij} \omega_{ij} = 1$.
- $\mathcal{O}(c)$ represents the behavior of the oracle. As the classifier has been tested with N and P samples, the probability estimates of negative and positive class are $N/M = n$ and $P/M = p$, respectively.
- $\Gamma(c)$ is an estimate of the conditional probability $p(\widehat{X}_c|X_c)$. In particular, $\gamma_{ij} \approx p(\widehat{X}_c = j \mid X_c = i)$, $i, j = 0, 1$, denotes the percent of inputs that have been correctly classified ($i = j$) or misclassified ($i \neq j$) by $\widehat{X}_c$. Note that γ_{00}, γ_{01}, γ_{10}, and γ_{11} are in fact the *rate of* true negatives (tn), false positives (fp), false negatives (fn) and true positives (tp). According to the total probability law, applied to a space of conditional probabilities: $\gamma_{00} + \gamma_{01} = \gamma_{10} + \gamma_{11} = 1$. Hence: $|\Gamma| = \sum_{ij} \gamma_{ij} = 2$.

Eq. (1) highlights the presence of two separate perspectives. The former, related to Ω, is biased by the class ratio, so that the measures defined therein are suited to give information about the *actual* performance of classifiers. The latter, related to Γ, is *not biased* by the class ratio, and everything goes as if negative and positive data were perfectly balanced. Hence, the measures defined therein can give information about the intrinsic performance of classifiers. Several proposals have been made over time to establish a bridge between the two spaces, aimed in particular at turning biased measures into unbiased ones –see, for instance, the work of Flach [16,20]. In our view, a simple yet clear strategy for defining performance measures, framed within a biased or an unbiased perspective, depends on the probability estimate used as reference. In particular, Ω can be used to define relevant measures in a biased space, whereas Γ can be used to define relevant measures in an unbiased space.

The most common definitions of accuracy (a), precision (π), negative predictive value ($\overline{\pi}$), sensitivity (ρ) and specificity ($\overline{\rho}$), given focusing on the joint probability estimate Ω, follow:

$$\begin{cases} a = \dfrac{trace(\Omega)}{|\Omega|} = \dfrac{\omega_{00} + \omega_{11}}{1} = n \cdot \gamma_{00} + p \cdot \gamma_{11} = \dfrac{\sigma \cdot tn + tp}{\sigma + 1} \\ \pi = \dfrac{\omega_{11}}{\omega_{11} + \omega_{01}} = \dfrac{p \cdot \gamma_{11}}{p \cdot \gamma_{11} + n \cdot \gamma_{01}} = \left(1 + \sigma \cdot \dfrac{fp}{tp}\right)^{-1} \\ \overline{\pi} = \dfrac{\omega_{00}}{\omega_{00} + \omega_{10}} = \dfrac{n \cdot \gamma_{00}}{n \cdot \gamma_{00} + p \cdot \gamma_{10}} = \left(1 + \dfrac{1}{\sigma} \cdot \dfrac{fn}{tn}\right)^{-1} \\ \rho = \dfrac{\omega_{11}}{\omega_{11} + \omega_{10}} = \gamma_{11} = tp \\ \overline{\rho} = \dfrac{\omega_{00}}{\omega_{00} + \omega_{01}} = \gamma_{00} = tn \end{cases} \tag{2}$$

Eq. (2) highlights the dependence of accuracy, precision and negative predictive value from the class ratio, only specificity and sensitivity being unbiased.

Turning biased measures into unbiased ones is now trivial. One should just substitute Γ to Ω in Eq. (2). Hence, using the subscript "u" to distinguish unbiased from biased measures, we can write:[2]

$$\begin{cases} a = \dfrac{trace(\Gamma)}{|\Gamma|} = \dfrac{\gamma_{00} + \gamma_{11}}{2} = \dfrac{tn + tp}{2} \\ \pi = \dfrac{\gamma_{11}}{\gamma_{11} + \gamma_{01}} = \left(1 + \dfrac{fp}{tp}\right)^{-1} \\ \overline{\pi} = \dfrac{\gamma_{00}}{\gamma_{00} + \gamma_{10}} = \left(1 + \dfrac{fn}{tn}\right)^{-1} \\ \rho = \dfrac{\gamma_{11}}{\gamma_{11} + \gamma_{10}} = \gamma_{11} \equiv tp \\ \overline{\rho} = \dfrac{\gamma_{00}}{\gamma_{00} + \gamma_{01}} = \gamma_{00} \equiv tn \end{cases} \tag{3}$$

3. Standard $\langle \varphi, \delta \rangle$ diagrams

In a classifier assessment scenario, a researcher is typically interested in understanding to what extent a classifier is able to approximate the oracle and whether it is biased towards the positive or the negative class. In a feature assessment scenario, a researcher is typically interested in assessing to what extent a feature is covariant or contravariant with the positive class, and whether it is characteristic or not for the dataset at hand. As pointed out, the solution proposed in this article addresses the problem from an accuracy-based perspective. In symbols, the following definitions hold (see [1] for further details on this aspect):

$$\begin{cases} \varphi = tp + fp - 1 = \rho - \overline{\rho} \\ \delta = tp - fp = \rho + \overline{\rho} - 1 \end{cases} \tag{4}$$

[2] Note that the unbiased definition of sensitivity and specificity actually ends up with the usual definitions, as they are intrinsically unbiased.

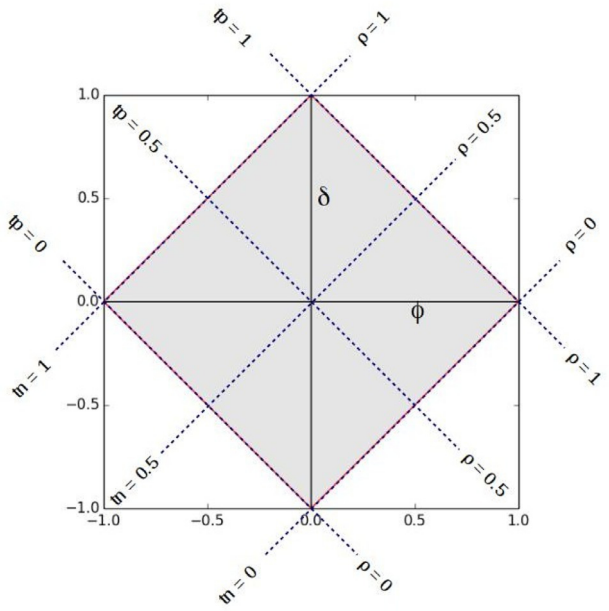


Fig. 1. Shape of the $\varphi - \delta$ space: the diamond centered in (0,0) delimits the area of admissible value pairs for φ and δ (it is easy to verify that, by construction, the diamond area is 2).

Eq. (4) defines standard φ and δ measures, which can be used to assess classifier performance or feature importance. Deployed in a 2D orthogonal view, φ and δ give rise to $\langle \varphi, \delta \rangle$ diagrams. Due to the constraints imposed by $\overline{\rho}$ and ρ, these diagrams have a diamond shape, whose borders are identified by the equation $|\varphi| + |\delta| = 1$. Fig. 1 shows a $\langle \varphi, \delta \rangle$ diagram, together with some isometrics of specificity and sensitivity, i.e., $tn = \overline{\rho} = \{0, 0.5, 1\}$) and $tp = \rho = \{0, 0.5, 1\}$. Note that, by construction, the area of admissible values in the $\langle \varphi, \delta \rangle$ space is 2, as any side of the diamond has length $\sqrt{2}$.

3.1. Semantics of the $\langle \varphi, \delta \rangle$ space

The semantics of the δ axis is strictly related to the *unbiased accuracy*. In fact, δ can be given in terms of the unbiased accuracy as follows:

$$\delta = \rho + \overline{\rho} - 1 = tp + tn - 1 = 2 \cdot \frac{tp + tn}{2} - 1 = 2 \cdot a_u - 1 \tag{5}$$

Hence, the findings highlighted by Eq. (5) allow to state that δ is able to measure the intrinsic discriminant capability of a classifier.

To analyze the semantics of the φ axis, let us start from the fact that the bias can be defined as the expected value of the difference between classifier ($\widehat{X}_c$) and oracle (X_c). In symbols:

$$\text{bias} = \mathbb{E}\left[\widehat{X}_c - X_c\right] = \mathbb{E}\left[\widehat{X}_c\right] - \mathbb{E}[X_c]$$

Having information about the percent of negative and positive samples, $\mathbb{E}\left[\widehat{X}_c\right]$ and $\mathbb{E}[X_c]$ can be rewritten as:

$$\begin{cases} \mathbb{E}\left[\widehat{X}_c\right] & \approx (p - n) + 2\, n \cdot fp - 2\, p \cdot fn \\ \mathbb{E}[X_c] & \approx (p - n) \end{cases}$$

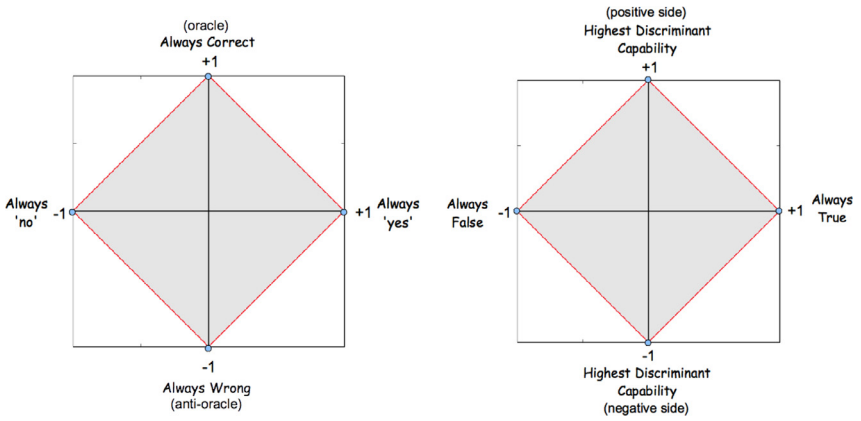
Hence:

$$\text{bias} = \mathbb{E}\left[\widehat{X}_c - X_c\right] = \mathbb{E}\left[\widehat{X}_c\right] - \mathbb{E}[X_c] \approx 2\, n \cdot fp - 2\, p \cdot fn \tag{6}$$

Eq. (6) represents an estimate of the *bias* of a classifier, measured over the confusion matrix that describes the outcomes of the experiments performed with the available data (see also [1] for further details on this aspect). Imposing $n = p = 0.5$, Eq. (6) can be rewritten as:

$$\text{bias} = \mathbb{E}\left[\widehat{X}_c - X_c\right]\Big|_{n=p=0.5} \approx fp - fn = tp - tn = \rho - \overline{\rho} \equiv \varphi \tag{7}$$

Eq. (7) highlights that φ is an estimate of the unbiased bias of a classifier. Indeed, when the performance of a classifier measured over a test set **D** lies on the positive semiplane of φ, one can argue that the classifier has a bias towards the positive class and vice versa.



(a) Classifier-oriented view (b) Feature-oriented view

Fig. 2. Relevant points in the $\langle \varphi, \delta \rangle$ space from a classifier- and feature-oriented view (left- and right-hand side, respectively).

To sum up, a standard $\langle \varphi, \delta \rangle$ diagram is a *bias-accuracy* diagram, where φ represents (an estimate of) the unbiased bias and δ represents the unbiased accuracy remapped in $[-1, +1]$. As a consequence, classifiers with $\varphi \approx 0$ and $|\delta| \approx 1$ are expected to show very good or very bad performance, depending to the sign of δ. Conversely, classifiers with $|\varphi| \approx 1$ are expected to show very bad performance.[3]

Fig. 2a points out that the diamond corners denote in fact well known kinds of classifiers. In particular, the upper corner (characterized by $\rho = \overline{\rho} = 1$) represents a classifier that is always correct (i.e., the *oracle*). Conversely, the lower corner (characterized by $\rho = \overline{\rho} = 0$) represents a classifiers that is always wrong (let us call it *anti-oracle*). As for left and right corners, they both denote dummy classifiers. In particular, the one at the left-hand corner (characterized by $\overline{\rho} = 1$ and $\rho = 0$) always takes pessimistic decisions, considering any sample as belonging to the negative class, whereas the one at the right-hand corner (characterized by $\overline{\rho} = 0$ and $\rho = 1$) always takes optimistic decisions, considering any sample as belonging to the positive class. It is clear that the above behaviors refer to ideal cases. However, they are approximately conserved in proximity of the corners (the closer, the better).

Similar considerations can be made for features. Fig. 2b points out that the highest discriminant capability for a feature occurs at upper and lower corners. In particular, the upper corner denotes a feature that is completely covariant with the positive class, whereas the lower corner denotes a feature that is completely contravariant with the positive class. In either case, the discriminant capability of the feature at hand is the highest. A feature laying at the left- / right-hand corner is in fact always false / true. In either case, any such feature would be completely irrelevant for the classification task.

3.2. $\langle \varphi, \delta \rangle$ Diagrams vs. mutual information and break-even points

It can be shown that the φ axis, defined by the equation $\delta = 0$, denotes the locus of points for which the *mutual information* between classifier and oracle is minimum (i.e., these points have maximum entropy), whereas the δ axis, defined by the equation $\varphi = 0$, denotes the locus of *break-even* points.

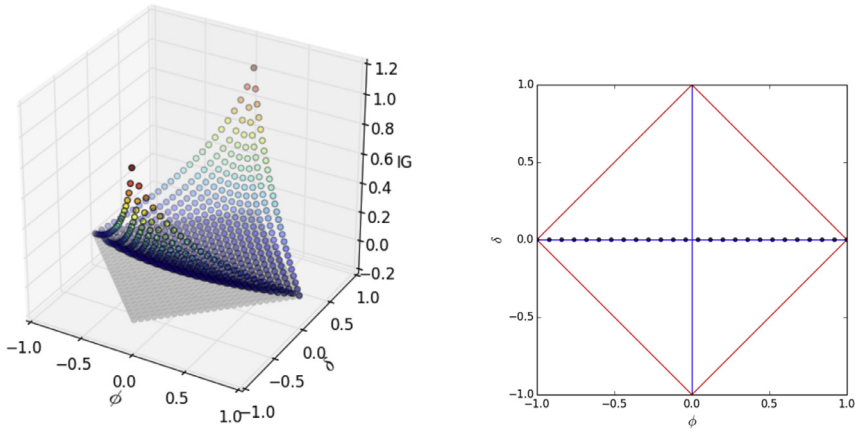
To start analyzing the property concerning mutual information let us recall its definition, given in terms of the random variables associated to the classifier at hand and to the oracle. In symbols:

$$I(X_c; \widehat{X}_c) = H(X_c) - H(X_c|\widehat{X}_c) \tag{8}$$

This information-theoretic measure actually denotes the information gain obtained by applying the classifier to the given problem. One should expect the information gain to be maximum in correspondence of oracle and anti-oracle, whereas a null information gain is expected for dummy classifiers (no matter whether a pessimistic or an optimistic strategy holds). Also the origin of axes is expected to carry a null information gain, as in that point $\rho = \overline{\rho} = 0.5$. In fact, the whole φ axis is characterized by a null information gain. Fig. 3 graphically illustrates this aspect, highlighting that all points for which the information gain is minimum lay on the φ axis. In particular, the left-hand side of the figure shows the whole 3D plot of $I(X_c; \widehat{X}_c)$, whereas the right hand side highlights its minima.

As for break-even points, in general they must fulfill the equation $\pi \equiv \rho$. Enforcing this constraint in an unbiased space, i.e., imposing that $\pi = \rho$, allows to show that break-even points are located on the δ axis, which is characterized by the

[3] A notable exception occurs when the imbalance is in favor of the bias. For instance, a classifier operating on a test set with $\sigma \gg 1$ would have a good performance in the event it associates the negative class to each submitted sample. However, this would only depend on the imbalance of data rather than on the intrinsic behavior of the classifier.



(a) Mutual information over $\langle \varphi, \delta \rangle$ (b) Mutual information minima

Fig. 3. The information gain evaluated over the whole $\langle \varphi, \delta \rangle$ space is reported on the left, whereas the locus of points for which the information gain is minimum is reported on the right (a 25x25 grid has been used in both cases).

Table 2
Summary table concerning isometrics and loci of points in a standard $\langle \varphi, \delta \rangle$ space for unbiased and biased measures (shorthands: $\alpha = 1/\pi - 1$ and $\overline{\alpha} = 1/\overline{\pi} - 1$).

Isometrics of	As unbiased measure	As biased measure
Specificity	$\delta = \varphi - 1 + \overline{\rho}$	–
Sensitivity	$\delta = -\varphi - 1 + \rho$	–
Accuracy	$\delta = 2a - 1$	$\delta = \frac{\sigma - 1}{\sigma + 1} \cdot \varphi + 2a - 1$
Precision	$\delta = (2\pi - 1) \cdot (\varphi + 1)$	$\delta = \frac{\sigma - \alpha}{\sigma + \alpha} \cdot (\varphi + 1)$
Neg Pred Value	$\delta = (1 - 2\overline{\pi}) \cdot (\varphi - 1)$	$\delta = \frac{\overline{\alpha} \cdot \sigma - 1}{\overline{\alpha} \cdot \sigma + 1} \cdot (\varphi - 1)$
Bias	$\varphi \equiv bias_u$	$\delta = \frac{\sigma + 1}{\sigma - 1} \cdot (\varphi - bias) + 1$
Break-Even Points	$\varphi \equiv 0$ (i.e., δ axis)	$\delta = \frac{\sigma + 1}{\sigma - 1} \cdot \varphi + 1$

equation $\varphi \equiv 0$. In symbols:

$$\pi = \frac{\gamma_{11}}{\gamma_{11} + \gamma_{01}} = \frac{tp}{tp + fp} = \frac{\rho}{\rho + (1 - \overline{\rho})} = \rho \tag{9}$$

Hence:

$$\pi = \rho \rightarrow \rho + (1 - \overline{\rho}) = 1 \rightarrow \rho - \overline{\rho} = 0 \rightarrow \varphi = 0 \tag{10}$$

3.3. Isometrics on standard $\langle \varphi, \delta \rangle$ diagrams

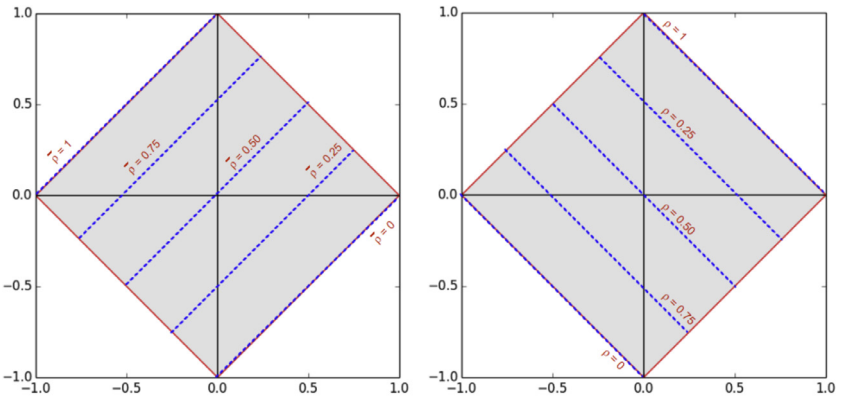
This subsection illustrates how isometrics of specificity, sensitivity, accuracy, precision, negative predictive value and bias are deployed in a standard $\langle \varphi, \delta \rangle$ diagram. For the sake of brevity, the corresponding formulas are summarized in Table 2.

Fig. 4 shows some isometrics of *specificity* and *sensitivity* (left- and right-hand side, respectively). In either case, they are a bundle of straight lines parallel to the contour of a $\langle \varphi, \delta \rangle$ diagram.

Fig. 5 shows some isometrics of *accuracy*. The left-hand side of the figure ($\sigma = 1$) highlights that δ corresponds in fact to the unbiased accuracy remapped in $[-1, +1]$. Note also that the isometrics for $a = 1$ (oracle) and $a = 0$ (anti-oracle) actually reduce to a single point. The right-hand side of the figure ($\sigma = 3$) shows that the slope of accuracy isometrics depend on the class ratio.

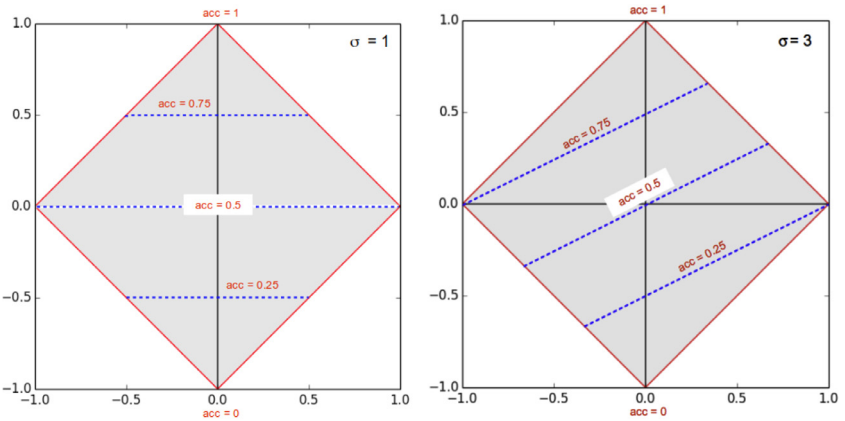
Fig. 6 shows some isometrics of *precision* and *negative predictive value*. In either case, isometrics are a bundle of straight lines, with varying slope, which originate from the left- and right-hand corner of the $\langle \varphi, \delta \rangle$ space. The slope of a line in general depends on σ and π.

Fig. 7 shows some isometrics of *bias*, for $\sigma = 1$ and $\sigma = 3$. The left-hand side of the figure ($\sigma = 1$) highlights that φ corresponds in fact to the unbiased bias.



(a) Isometrics of specificity (b) Isometrics of sensitivity

Fig. 4. Isometrics of specificity and sensitivity for selected values of $\overline{\rho}$ and ρ (i.e., 0., 0.25, 0.5, 0.75, 1.0).



(a) Isometrics of unbiased accuracy (b) Isometrics of accuracy

Fig. 5. Some isometrics of accuracy (i.e., 0., 0.25, 0.5, 0.75, 1.0), with different class ratios: $\sigma = 1$ on the left-hand side and $\sigma = 3$ on the right-hand side (note that when $\sigma = 1$ unbiased and actual accuracy coincide).

4. Generalized $\langle \varphi, \delta \rangle$ diagrams

In this section, a generalization of $\langle \varphi, \delta \rangle$ diagrams is proposed, the corresponding space being named $\langle \varphi_b, \delta_b \rangle$. Notably, all relevant properties identified for $\langle \varphi, \delta \rangle$ diagrams are maintained in the new space, according to a specific design choice. In particular, it is shown (i) that a value measured on the δ_b axis corresponds to the actual accuracy remapped in $[-1, +1]$, (ii) that a value measured on the φ_b axis gives (an estimate of) the actual bias, and (iii) that the δ_b axis is the locus of break-even points.

As we want Eqs. (7) and (5) hold also for the generalized case, the starting point for devising a generalized $\langle \varphi, \delta \rangle$ space is:

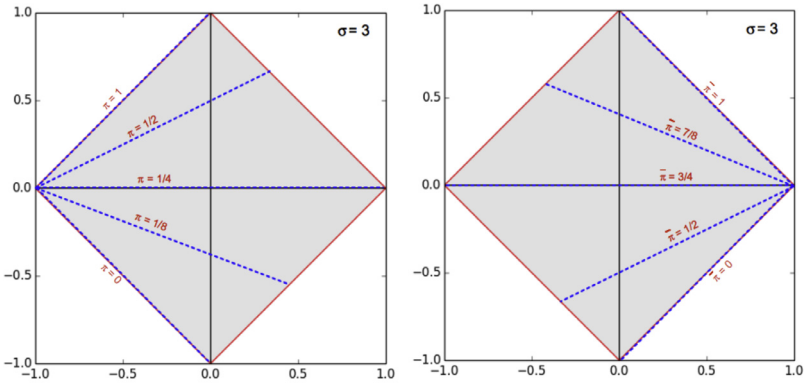
$$\begin{cases} \varphi_b \approx \text{bias} \\ \delta_b \equiv 2\,a - 1 \end{cases} \tag{11}$$

Rewriting Eq. (11) in terms of n, p, $\overline{\rho}$, and ρ yields the following equations, which define the generalized $\langle \varphi, \delta \rangle$ space:

$$\begin{cases} \varphi_b = -2\,n \cdot \overline{\rho} + 2\,p \cdot \rho + 2(n - p) \\ \delta_b = 2\,n \cdot \overline{\rho} + 2\,p \cdot \rho - 1 \end{cases} \tag{12}$$

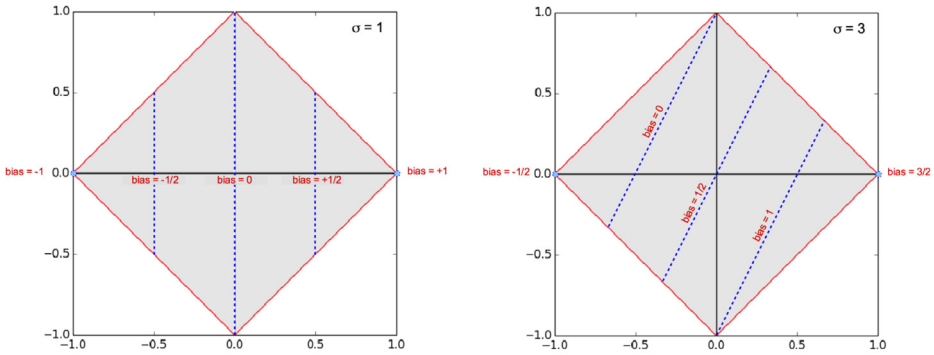
4.1. Semantics of the $\langle \varphi_b, \delta_b \rangle$ space

Fig. 8 illustrates an example of a generalized $\langle \varphi, \delta \rangle$ diagram, with $\sigma = 3$. All relevant points (i.e., oracle, anti-oracle and dummy classifiers) have been highlighted, and the corresponding coordinates reported therein. Furthermore, we know that, *by construction*, bias and accuracy isometrics are straight lines parallel to the δ_b and to the φ_b axis, respectively. As expected,



(a) Isometrics of precision (b) Isometrics of neg predictive value

Fig. 6. Some isometrics of precision and negative predictive value.



(a) Isometrics of unbiased bias (b) Isometrics of bias

Fig. 7. Isometrics of bias for selected values of bias and σ–i.e., $\sigma = 1$ on the left-hand side and $\sigma = 3$ on the right-hand side (recall that when $\sigma = 1$ $\text{bias} \equiv \text{bias}_u$).

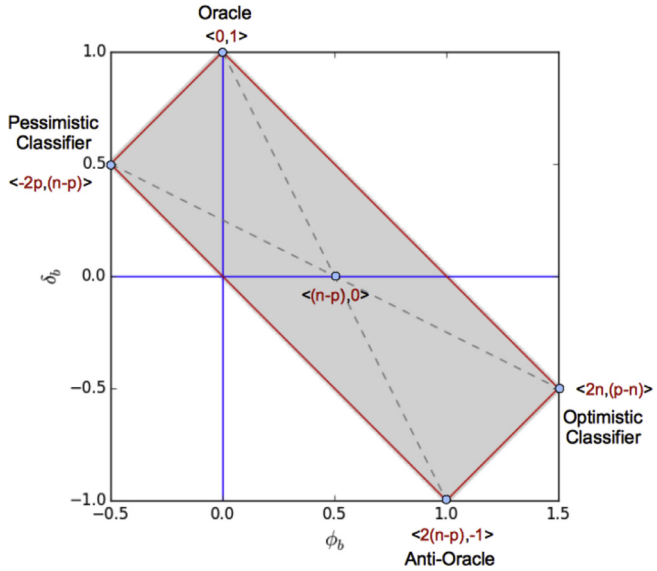
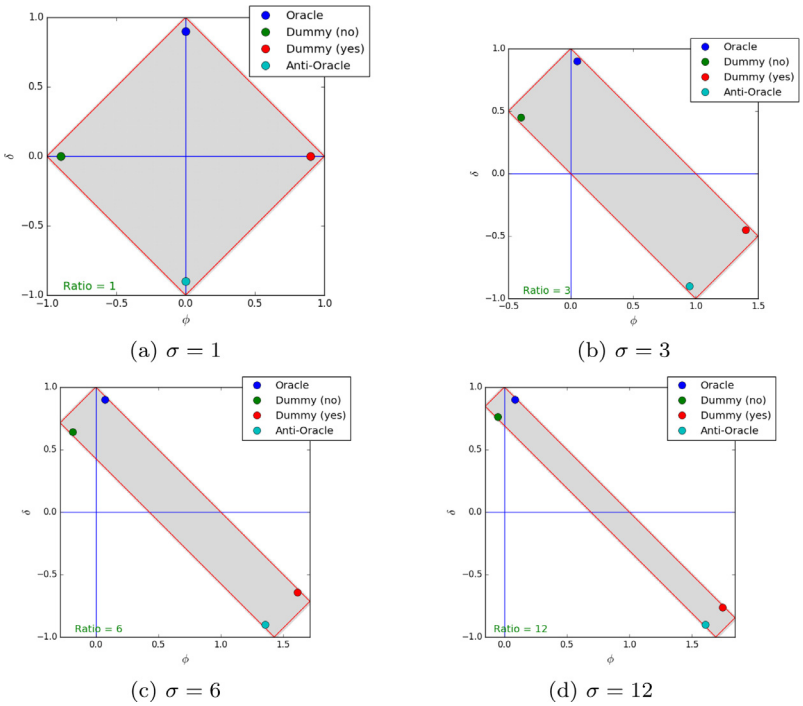


Fig. 8. A $\langle \varphi_b, \delta_b \rangle$ diagram in which $\sigma = 3$ (in this case the diamond area is 3/2).

Table 3
Summary table reporting the coordinates of corners in a generic $\langle \varphi_b, \delta_b \rangle$ space (also the coordinates of φ and δ have been reported for the sake of readability).

Hotspot	Classifier	φ	δ	φ_b	δ_b
TOP	Oracle	0	+1	0	+1
BOTTOM	Anti-Oracle	0	−1	$2\,(n-p)$	−1
LEFT	Dummy (pessimistic)	−1	0	$-2\,p$	$n-p$
RIGHT	Dummy (optimistic)	+1	0	$+2\,n$	$p-n$



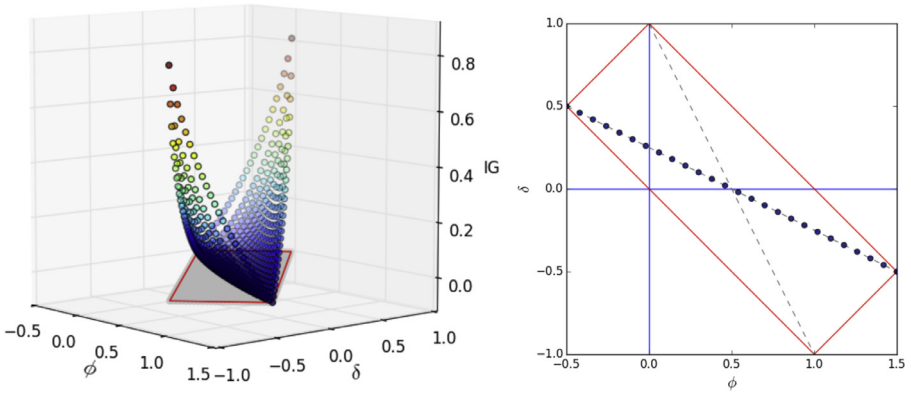
(a) $\sigma = 1$ (b) $\sigma = 3$ (c) $\sigma = 6$ (d) $\sigma = 12$

Fig. 9. Using $\langle \varphi_b, \delta_b \rangle$ diagrams to assess the performance of a classifier in presence of different class ratios (i.e., from left to right and from top to bottom: $\sigma = 1, 3, 6, 12$). In particular, the figures make clear that a dummy classifier may reach "good performance" depending on the class ratio. Note that specular diagrams could be obtained by imposing $\sigma = 1, 1/3, 1/6, 1/12$.

the oracle is still located at $\varphi_b = 0$ and $\delta_b = +1$, whereas the position of the remaining corners changes according to the given class ratio. Table 3 reports the coordinates of corners in a generic $\langle \varphi_b, \delta_b \rangle$ space. The corresponding values for the standard $\langle \varphi, \delta \rangle$ space are also reported for the sake of completeness. Note (i) that the discriminant capability of the anti-oracle is still -1, (ii) that the center of the biased space is not coincident anymore with the origin of axes (it is easy to verify, analytically or using simple geometric considerations, that its coordinates are $\varphi_b = n - p$ and $\delta_b = 0$), and iii) that the area of the rectangle that delimits a $\langle \varphi_b, \delta_b \rangle$ space amounts to $8\,np$ (this happens because the borders of the rectangle now measure $2\,p \cdot \sqrt{2}$ and $2\,n \cdot \sqrt{2}$, e.g., from the oracle to the left-hand corner and from the oracle to the right hand corner, respectively).

Notably, the *apparent* discriminant capability of a dummy classifier might even become close to 1, depending on the characteristics of the dataset at hand. In particular, the pessimistic (optimistic) classifier can reach a very good performance when $n \gg p$ ($p \gg n$). However, most likely, this would be due to the statistics of the dataset at hand rather than to the intrinsic capability of the classifier.

To further investigate this issue, Fig. 9 gives an insight about the potential of $\langle \varphi_b, \delta_b \rangle$ diagrams in the task of assessing the behavior of a classifier with varying class ratio. Four points are highlighted, each representing a relevant category of classifier performance. In particular, two points approximate the behavior of oracle and anti-oracle, whereas the others approximate the behavior of dummy classifiers (optimistic and pessimistic). These diagrams make clear that, with increasing class ratio, the apparent performance of a pessimistic classifier progressively becomes comparable to the one of the oracle. However, here the improvement depends only on the statistics of data that has been imposed rather than on the discriminant capability of the corresponding classifier. Similar considerations hold for decreasing class ratio, the only difference being that –in this case– good performances would be attained by the optimistic classifier.



(a) Mutual information over $\langle \varphi_b, \delta_b \rangle$ (b) Mutual information minima

Fig. 10. The information gain evaluated over the whole $\langle \varphi_b, \delta_b \rangle$ space is reported on the left, whereas the locus of points for which the information gain is minimum is reported on the right (a 25 x 25 grid has been used in both cases).

Table 4
Summary table concerning isometrics and loci of points in the $\langle \varphi_b, \delta_b \rangle$ space.

isometrics of	Unbiased space	Biased space
Specificity	$\delta = \varphi - 1 + 2\,\overline{\rho}$	$\delta_b = \varphi_b + 4\,n \cdot \overline{\rho} - 1 - 2\,(n - p)$
Sensitivity	$\delta = -\varphi - 1 + 2\,\rho$	$\delta_b = -\varphi_b + 4\,p \cdot \rho - 1 + 2\,(n - p)$
Accuracy	$\delta = 2\,a - 1$	$\delta_b = 2\,a - 1$
Precision	$\delta = (2\,\pi - 1) \cdot (\varphi + 1)$	$\delta_b = (2\,\pi - 1) \cdot (\varphi_b + 2\,p) + (n - p)$
Neg Pred Value	$\delta = (1 - 2\,\overline{\pi}) \cdot (\varphi - 1)$	$\delta_b = (1 - 2\,\overline{\pi}) \cdot (\varphi_b - 2\,n) + (p - n)$
Bias	$\varphi \equiv bias_u$	$\varphi_b \equiv bias$
Break-Even	$\varphi \equiv 0$ (i.e., the δ axis)	$\varphi_b \equiv 0$ (i.e., the δ_b axis)

4.2. Generalized $\langle \varphi, \delta \rangle$ diagrams vs. mutual information and break-even points

In generalized $\langle \varphi, \delta \rangle$ diagrams, the minimum information gain holds on the straight line that joins dummy classifiers (i.e., left and right hand corners). This result is in complete agreement with the one seen on standard diagrams, the only difference being that the diagram is now stretched according to the class ratio. Fig. 10 reports mutual information for $\sigma = 3$. In particular, the left-hand side of the figure shows the whole 3D plot of $I(X_c; \widehat{X}_c)$, whereas the right hand side highlights its minima.

As for break-even points, they must fulfill the equation $\pi \equiv \rho$. In symbols:

$$\pi = \frac{p \cdot \gamma_{11}}{p \cdot \gamma_{11} + n \cdot \gamma_{01}} = \frac{p \cdot tp}{p \cdot tp + n \cdot fp} = \frac{p \cdot \rho}{p \cdot \rho + n \cdot (1 - \overline{\rho})} = \rho \tag{13}$$

Hence:

$$\pi = \rho \rightarrow p \cdot \rho + n \cdot (1 - \overline{\rho}) = p \rightarrow p \cdot \rho - n \cdot \overline{\rho} + (n - p) = 0 \rightarrow \varphi_b = 0 \tag{14}$$

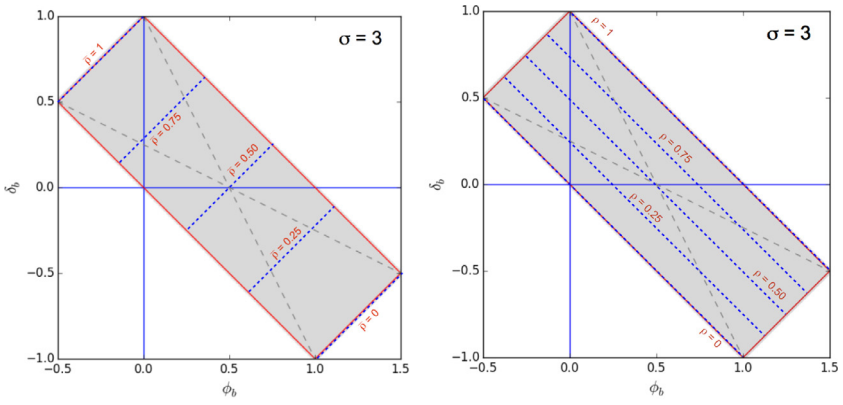
4.3. Isometrics on generalized $\langle \varphi, \delta \rangle$ diagrams

This subsection illustrates how isometrics of specificity, sensitivity, accuracy, precision, negative predictive value and bias are deployed in a generalized $\langle \varphi, \delta \rangle$ diagram. For the sake of brevity, the corresponding formulas are summarized in Table 4. Few differences in terms of equations occur between unbiased and biased isometrics, as equations derived for the generalized space always imply or are equivalent to those derived for the standard one. Nevertheless, to facilitate the comparison, also their graphical representation is given.

Fig. 11 shows some isometrics of *specificity* and *sensitivity* (left- and right-hand side, respectively). In either case, they are a bundle of straight lines parallel to the contour of a $\langle \varphi_b, \delta_b \rangle$ diagram.

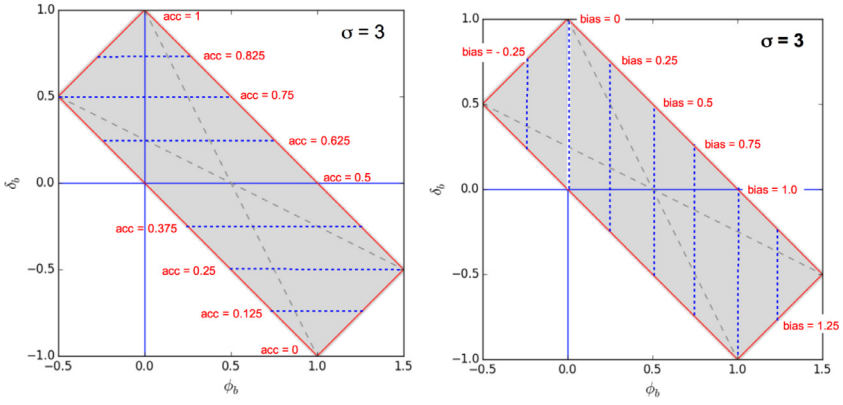
Fig. 12 shows some isometrics of *accuracy* and *bias*. The figure highlights that δ_b corresponds to the accuracy remapped in $[-1, +1]$ (left-hand side), whereas φ_b corresponds to the bias (right-hand side).

Fig. 13 shows some isometrics of *precision* and *negative predictive value*. Note that in both cases the isometrics 0.5 is still parallel to the horizontal axis.



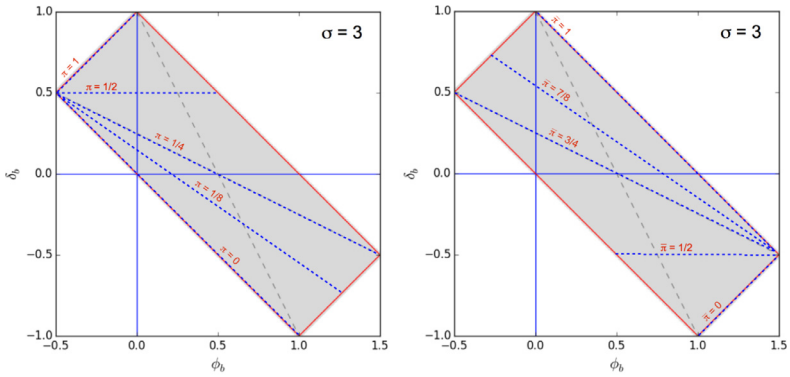
(a) Isometrics of specificity (b) Isometrics of sensitivity

Fig. 11. Isometrics of specificity and sensitivity for selected values of $\overline{\rho}$ and ρ (i.e., 0., 0.25, 0.5, 0.75, 1.0).



(a) Isometrics of accuracy (b) Isometrics of bias

Fig. 12. Some isometrics of accuracy and bias (i.e., 0., 0.25, 0.5, 0.75, 1.0).



(a) Isometrics of precision (b) Isometrics of neg predictive value

Fig. 13. Isometrics of precision and negative predictive value.

5. Experimental settings

We decided to focus mainly on real-world datasets, as they are typically more effective than artificial ones in the task of highlighting the characteristics of a method. However, as the majority of real-world datasets contain non binary features, we had to devise a way for dealing with nominal and floating point features. To provide essential information on this aspect, let us briefly describe the solutions adopted to make these kinds of features compatible with $\langle \varphi, \delta \rangle$ diagrams.

5.1. Nominal features

Dealing with a nominal feature is straightforward, as one may focus on all subsets that originate from its values. In particular, with $\mathcal{P}(.)$ used to denote the powerset operation, the number of significant subsets that can be built over the values of a feature f is $|\mathcal{P}(f)|/2-1$. For example, assuming that f has four values, say *small, medium, large* and *extra_large*, the number of significant subsets would be:

$$|\mathcal{P}(f)|/2-1 = 16/2-1 = 7$$

This result is consistent with the informal appraisal according to which, in this case: (i) singletons make irrelevant the analysis of all subsets with dimension three, e.g., {*small*} makes irrelevant the analysis of {*medium*, *large*, *extra_large*}; and (ii) only half of pairwise subsets are in fact significant, e.g., {*small, medium*} makes irrelevant the analysis of {*large*, *extra_large*}. Most often, binarization is a viable alternative, as the number of values for a nominal feature is typically lower than 10. However, in hard cases, nothing prevents from investigating only subsets with low cardinality, regardless of the actual dimension of the powerset.

5.2. Floating point features

Let us first revisit the strategy adopted for dealing with binary features. In this case, class values are encoded with $+1$ and -1 (positive and negative class, respectively), and the same happens for feature values (with the meaning of *true* and *false*, respectively). In a way, given a binary feature, everything goes as if a full "token" (which amounts to $+1$) were assigned to the set of true / false positives / negatives, depending on the agreement between the feature value (say v) and the class label (say c) for the sample at hand. For instance, with $c=+1$ and $v=-1$, the available token would be entirely used to increment the number of false negatives. Given a feature, the full statistics concerning true / false positives, as well as true / false negatives can be obtained by applying this basic step over all available samples.

The strategy adopted for dealing with floating point features extends the one depicted above by enforcing token sharing. In particular, the function $t(v)$ used for token sharing is:[4]

$$t(v) = \frac{1+c*v}{2} \tag{15}$$

Depending on the sign of c, the cumulative value that accounts for the number of true negatives or true positives is updated according to Eq. (15). In particular, when $c=-1$ the quantity $t(v)$ is assigned to true negatives, whereas the remaining part, i.e., $1-t(v)$, is assigned to false positives. Conversely, when $c=+1$ the quantity $t(v)$ is assigned to true positives, whereas the remaining part is assigned to false negatives. For instance, let us assume that the given sample belongs to the positive class ($c=+1$) and that v, after normalization in $[-1,+1]$, amounts to $+0.8$. This means that v is mainly, though not completely, *covariant* with c. Hence, most of the available token will be assigned to true positives, whereas the remaining part will be assigned to false negatives. In symbols:

$$\begin{aligned} TP &+= 0.9 \quad \text{as} \quad t(v) = (1+1*0.8)/2 = 0.9 \\ FN &+= 0.1 \quad \text{as} \quad 1-t(v) = 1-0.9 = 0.1 \end{aligned} \tag{16}$$

Analogous considerations hold when $c=-1$ (in this case, true negatives and false positives would be affected by the update).

It is easy to show that token sharing reduces to the binary token assignment strategy when $v=\pm 1$. For instance, assuming that $c=+1$ and $v=+1$ (instead of $+0.8$), Eq. (16) would reduce to:

$$\begin{aligned} TP &+= 1. \quad \text{as} \quad t(v) = (1+1*1)/2 = 1. \\ FN &+= 0. \quad \text{as} \quad 1-t(v) = 1-1.0 = 0. \end{aligned} \tag{17}$$

6. Experiments with $\langle\varphi, \delta\rangle$ diagrams

This section reports some relevant use cases aimed at illustrating the expressiveness of $\langle\varphi, \delta\rangle$ diagrams. All datasets used for experiments are publicly available on well-known machine learning web sites.[5] Two separate subsections follow: one focused on classifier assessment and the other on feature assessment.

6.1. Using $\langle\varphi, \delta\rangle$ diagrams for classifier assessment

Let us consider for example the UCI dataset *ionosphere*, which contains radar data collected by a phased array of 16 high-frequency antennas in Goose Bay, Labrador. The targets were free electrons in the ionosphere: "good" radar returns

[4] The proposed token sharing strategy is intentionally *linear*, as the token is split into two proportional parts. Although nothing prevents from devising and implementing alternative non linear strategies, discussing this aspect is far beyond the scope of this article.

[5] The vast majority of datasets has been downloaded from the Machine Learning repository at UCI (http://archive.ics.uci.edu). KEEL (http://sci2s.ugr.es/keel) and OpenML (http://www.openml.org) have also been used as datasets sources.

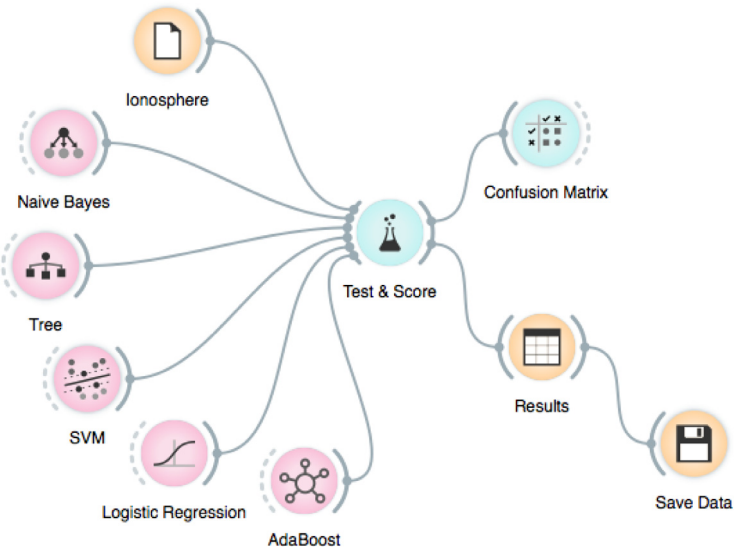
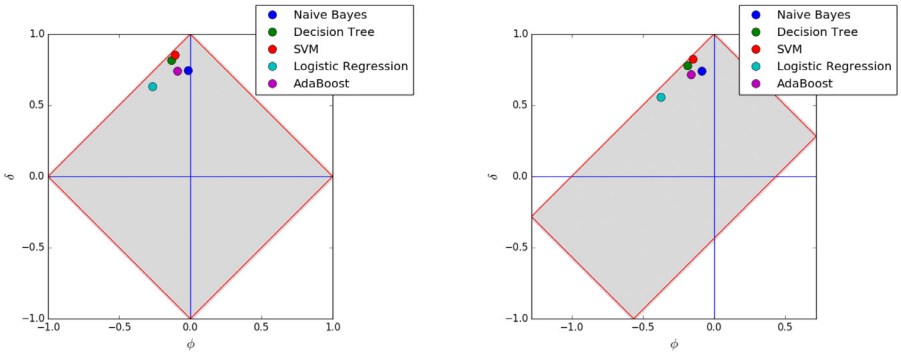


Fig. 14. Workflow of the experiment in which the *ionosphere* dataset has been tested with multiple classifiers (i.e., Naive Bayes, Decision Tree, SVM, Logistic Regression, and AdaBoost) using 10-fold cross-validation.



(a) Ionosphere in $\langle \varphi, \delta \rangle$ (b) Ionosphere in $\langle \varphi_b, \delta_b \rangle$

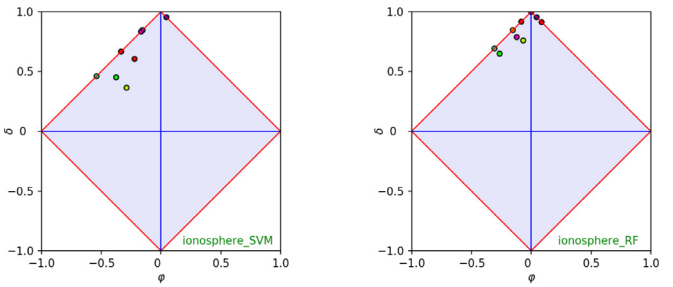
Fig. 15. $\langle \varphi, \delta \rangle$ and $\langle \varphi_b, \delta_b \rangle$ diagrams corresponding to the experiment in which the *ionosphere* dataset has been tested with multiple classifiers (i.e., Naive Bayes, Decision Tree, SVM, Logistic Regression, and AdaBoost) using 10-fold cross-validation. The imbalance for the diagram that lay on the right-hand side of the figure has been arbitrarily set to 1/2.

are those showing evidence of some type of structure in the ionosphere, whereas "bad" returns are those that do not (i.e., signals pass through the ionosphere). The dataset contains 351 samples; 225 good and 126 bad. Assuming that the positive class is "good", this dataset has a ratio $\sigma = 126/225 = 0.56$. Let us assume that one wants to test the behavior of different classifiers (e.g., naive Bayes, decision trees, SVM, logistic regression and adaboost) on the given dataset. Fig. 14 reports the workflow of the experiment, as obtained using the Orange toolbox [7].

The corresponding $\langle \varphi, \delta \rangle$ diagrams, reported in Fig. 15, clearly point out that the best performance (in terms of unbiased and actual accuracy) is attained by SVM. This aspect would not be so clear using *ROC* diagrams, as they are suited to highlight specificity or sensitivity isometrics rather than accuracy or bias isometrics.

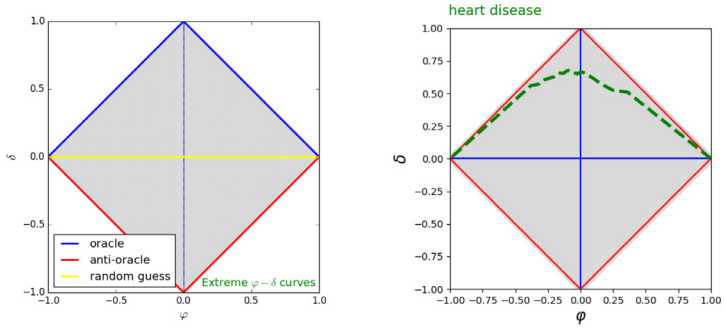
As for bias, Fig. 15a highlights that almost all classifiers, except for Naive Bayes, are expected to show better performances on the *negative* side, meaning that, in absence of imbalance, specificity should be better than sensitivity (a negative value of φ implies that specificity is better than sensitivity). This behavior is further highlighted in Fig. 15b, which results from stretching the unbiased diagram to account for the imbalance of data, according to Eq. (12). As last comment on classifier performance analysis, let us note that the variance of a classifier on the dataset at hand can be estimated by showing the results of all k-fold cross validation runs on a $\langle \varphi, \delta \rangle$ diagram. In so doing, one can easily spot the variance by looking at the scattering over the diagram. Fig. 16 reports the (estimated) variance obtained on the *ionosphere* dataset by running SVM and random forests, both with 10-fold cross validation. The figure makes clear that the variance for SVM is higher than the one obtained by running random forests.

As for classifier tuning, it is usually made according to a MAP criterion, in which the statistics of the input source are known. A typical choice in this kind of experimental setting consists of trying to maximize the overall accuracy of a classifier



(a) SVM: 10-fold runs (b) Random forests: 10-fold runs

Fig. 16. The (estimated) variance obtained by running SVM and random forests on the *ionosphere* dataset (on the left-hand and right-hand side, respectively). Each step of a 10-fold cross validation testing strategy has been reported on a $\langle \varphi, \delta \rangle$ diagram. The scattering of results clearly highlights that the variance observed with SVM is higher than the one observed with random forests.



(a) Extreme (theoretical) $\langle \varphi, \delta \rangle$ curves (b) Real-world $\langle \varphi, \delta \rangle$ curve

Fig. 17. Three *extreme* $\langle \varphi, \delta \rangle$ curves, i.e., oracle (in *blue*), anti-oracle (in *red*), and random guess (in *yellow*) are reported on the left-hand side. The right-hand side reports a $\langle \varphi, \delta \rangle$ curve (in green) that results from testing a naive Bayes classifier on the *heart disease* dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by tuning up the classifier parameters. Among the available parameters, let us concentrate on the classification threshold, which controls the way positive inputs are discriminated from negative ones.[6] This parameter may be used to tune the decision process according to the class ratio. The starting point in a strategy for threshold optimization requires the equivalent of a *ROC* curve, say $\langle \varphi, \delta \rangle$ *curve* hereinafter. Drawing it is straightforward, the obvious difference being that the curve would lay down in a bias vs. accuracy space rather than in a specificity vs. sensitivity space.

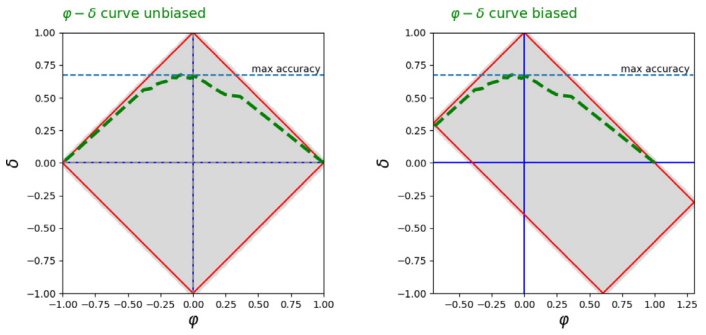
To better illustrate this aspect, let us take a look at Fig. 17. On the left-hand side, three *extreme* $\langle \varphi, \delta \rangle$ curves are drawn, which correspond to: oracle (in blue), anti-oracle (in red), and random guesser (in yellow). On the right-hand side, a real $\langle \varphi, \delta \rangle$ curve is drawn, which represents the behavior of a naive Bayes classifier tested on the well known real-world *heart disease* UCI dataset. Recalling that accuracy isometrics are straight lines parallel to the φ axis, drawing a $\langle \varphi, \delta \rangle$ curve allows one to easily detect the maximum value of unbiased accuracy (see Fig. 18a). More importantly, assuming that the imbalance is known, one can draw the $\langle \varphi, \delta \rangle$ curve in a $\langle \varphi_b, \delta_b \rangle$ diagram and check the maximum value of *actual* accuracy that can be attained (see Fig. 18b). The corresponding threshold should be selected as optimal for the given imbalance.

6.2. Using $\langle \varphi, \delta \rangle$ diagrams for feature assessment

The analysis depicted hereinafter, aimed at investigating the so-called feature importance, can be considered a first step in various activities concerning classifier design. Beyond feature ranking, it could play a primary role in feature selection algorithms (as underlying heuristics) or it could be used to perform a preliminary analysis before deciding whether to apply feature reduction or not to the dataset at hand.

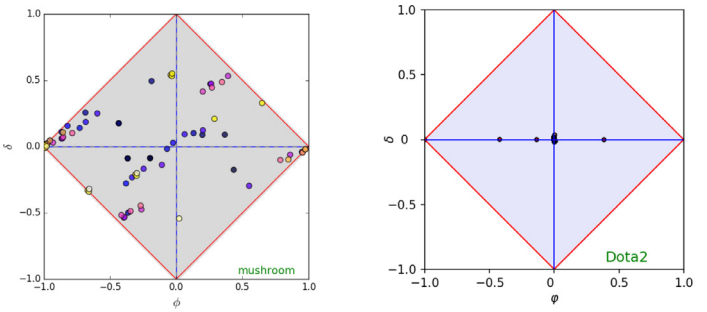
As pointed out, $\langle \varphi, \delta \rangle$ diagrams can also be used to perform feature assessment. In this case, each feature is treated as a single-feature classifier, so that its "performance" can be reported in a $\langle \varphi, \delta \rangle$ space as well. Drawing all resulting points (one point for each feature) gives rise to a kind of "class signature".

[6] Thresholding the output of a classifier is not always feasible, as not all classifier are designed (or implemented) to allow it. When feasible, the default threshold value for a classifier whose output ranges in [0,1] would be 0.5, whereas 0 would be used as default for classifiers whose output ranges in $[-1, +1]$.



(a) Unbiased $\langle \varphi, \delta \rangle$ curve (b) Biased $\langle \varphi, \delta \rangle$ curve

Fig. 18. The $\langle \varphi, \delta \rangle$ curve obtained by running a naive Bayes classifier on the dataset *heart disease* is reported on the left-hand side. The optimal threshold for the classifier at hand on the selected dataset can be easily spotted by looking at the *maximum value of accuracy* found when drawing a $\langle \varphi, \delta \rangle$ curve in a $\langle \varphi_b, \delta_b \rangle$ diagram (right-hand side).



(a) Example of "good" signature (b) Example of "bad" signature

Fig. 19. Examples of "good" and "bad" signatures obtained with two historical UCI datasets: (a) *mushroom*, on mushroom edibility, and (b) *dota2*, which is an online game. For each dataset, the corresponding diagram has been obtained by reporting in the $\langle \varphi, \delta \rangle$ space the performance of its embedded features –as they were in fact single-feature classifiers. Nominal features have been binarized (for the sake of simplicity, only subsets of values with cardinality 1 or 2 have been investigated).

One may wonder about how class signatures can be used to check whether a dataset is expected to be easy or difficult and whether specific signature shapes exist for one or more application fields. Let us separately analyze these aspects.

6.2.1. Identifying easy vs. difficult datasets

A simple strategy for inspecting class signatures consists of checking whether *one or more* features exist, which are covariant or contravariant with the positive class (in either case, the classification task would be easy). On the other hand, class signatures in which *all* features lay close to the φ axis are expected to be difficult to classify. Notably, a class signature can also highlight at a glance which features are rare or characteristic for the dataset at hand. For instance, let us assume that the absence/presence of a word in a document is asserted by a corresponding binary feature. Having the feature mostly false in the document corpus under analysis would mean that the word occurs in very few documents (left-hand corner of a $\langle \varphi, \delta \rangle$ diagram). Conversely, having the feature mostly true would mean that the word occurs in the majority of documents (right-hand corner of a $\langle \varphi, \delta \rangle$ diagram). In either case, the corresponding word would not be helpful for the classification task.

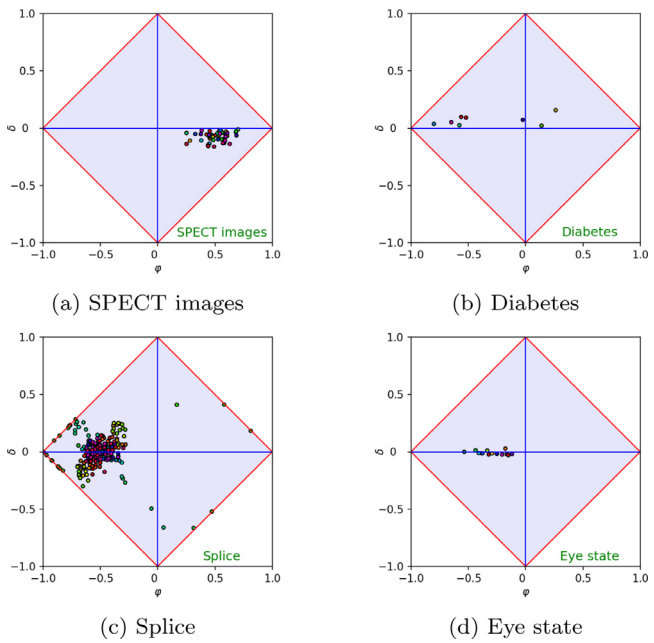
Fig. 19 reports two exemplar cases of class signature, generated from historical datasets downloaded from the UCI machine learning repository. As for the *mushroom* dataset (see Fig. 19a), one may argue that the classification task should be very easy, as many features are spread along the $\langle \varphi, \delta \rangle$ space. In any case, the performance could not be worse than the one associated to the best feature, no matter whether covariant or contravariant with the positive class. In fact, after running a naive Bayes classifier on this datasets the resulting accuracy was 0.98. As for the *dota2* dataset (see Fig. 19b), one may argue that the classification task is expected to be difficult, as all features lay along the φ axis. In fact, after running a naive Bayes classifier on this datasets the resulting accuracy was 0.61. Note, however, that the presence of at least one feature with medium-to-high value of $|\delta|$ *implies* that the dataset at hand is easy, but the converse is not always true.

To further investigate this aspect, several real-world datasets have been selected from various Internet sources. Table 5 reports the selected datasets, divided into four groups, i.e., *life sciences, multimedia, social sciences* and *remote sensing*. Each group encompasses four datasets, listed out with increasing number of samples.

All experiments have been run using the Orange toolbox, which allowed to easily calculate class signature and classification for each dataset. The former task has been performed using a new widget, named $\langle \varphi, \delta \rangle$, which has been embedded

Table 5
Four significant groups of datasets, i.e., *life sciences, multimedia, social sciences* and *remote sensing*, used for experimenting $\langle \varphi, \delta \rangle$ diagrams in the context of feature assessment. Each group encompasses four datasets of different size, with increasing dimensionality.

Life Sciences		**#samples**	**#features**
SPECT images	Heart disease identification by means of SPEC Tomography images	267	44
Diabetes	Data about diabetes retinopathy (from Messidor image dataset)	1151	20
Splice	Splice-junction identification in gene sequences	3190	61
Eye state	EEG measurement with the Emotiv Neuroheadset	14,980	15
Multimedia		**#samples**	**#features**
TechTC	Class taken from the TechTC-100 text collection	163	19,259
Internet ads	Advertisements identification on Internet pages	3279	1558
Phishing	Predicting phishing websites	11,055	68
TV News	News from the TV channel BBC International	17,721	4124
Social sciences		**#samples**	**#features**
Credit Card	Approvals of Australian credit card applications	690	14
German credit	German credit data (loan decision process)	1000	24
Census	US citizens census income (from 1994 data)	48,842	14
FARS	Fatal injuries suffered in motor vehicle traffic crashes	100,968	29
Remote sensing		**#samples**	**#features**
Robot	Force and torque measurements on a robot after failure detection	164	90
Sonar	Bouncing sonar signals (mines vs. rocks identification)	208	60
HAR	Human activity recognition (with cellular phone)	10,299	561
Gas sensors	Gas sensors for home activity monitoring	919,438	11



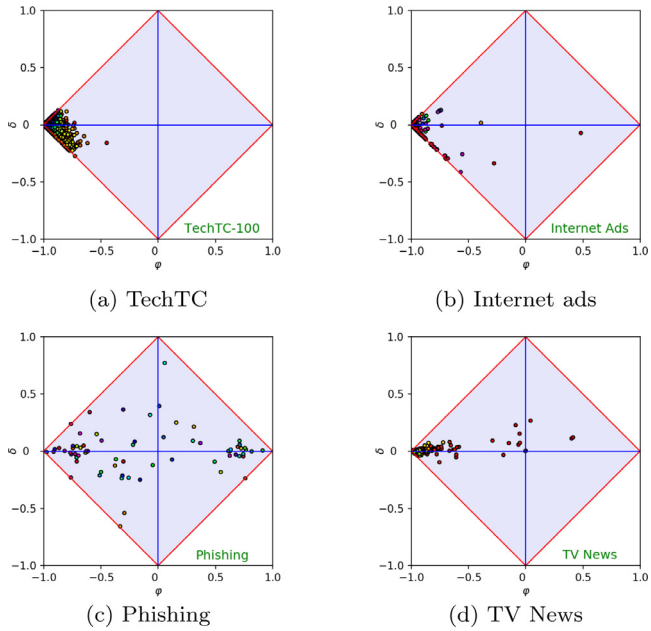
(a) SPECT images (b) Diabetes

(c) Splice (d) Eye state

Fig. 20. Class signatures corresponding to the dataset group *life sciences*.

into Orange. Figs. 20–23 report the class signature of each dataset, according to the group it belongs to. As for classification, it has been performed using a standard widget that embeds random forests, which are widely acknowledged for being very robust and effective (see for example [2] for further details on this aspect).

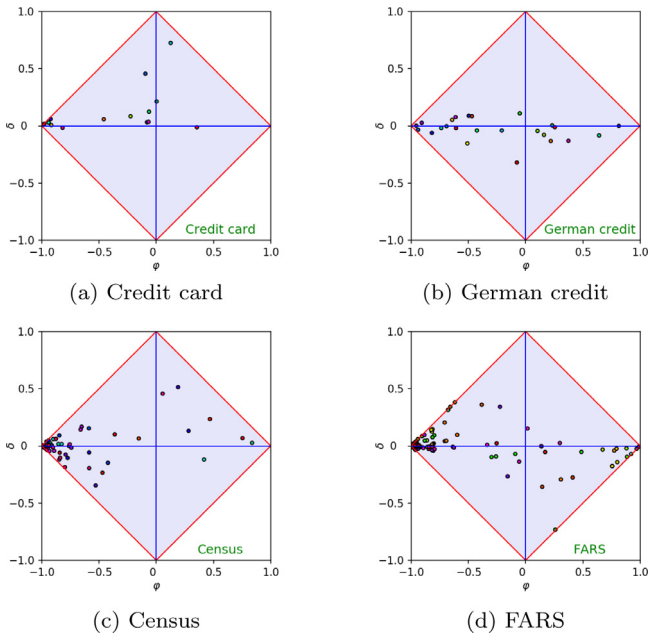
The corresponding experimental results, reported in Table 6, highlight that in most cases a strict correlation holds between classification performance and maximum absolute value of δ. Information about the actual difficulty of the dataset at hand is given in terms of accuracy, specificity and sensitivity. All cases in which the classification performance fulfills the behavior suggested by the class signature have been labeled with a black circle, while weak correlations have been labeled with a white circle. Weak correlations have been observed only on the *Census* and *FARS* datasets. Apparently, in the former case the correlation is high. However, the reported high performance is probably due to the high value of imbalance rather than on the intrinsic capability of the classifier at hand. In fact, most likely, the classifier has been "driven" towards negative samples by the high value of imbalance. As a result, sensitivity is much lower than specificity. However, this drawback did not have a great impact on the actual accuracy, this time thanks to the class ratio. In the latter case, the reported behavior is probably due to the high number of samples compared to the relatively low number of features.



(a) TechTC (b) Internet ads

(c) Phishing (d) TV News

Fig. 21. Class signatures corresponding to the dataset group *multimedia*.



(a) Credit card (b) German credit

(c) Census (d) FARS

Fig. 22. Class signatures corresponding to the dataset group *social sciences*.

6.2.2. Relationship between application field and class signatures

A further issue that has been investigated is whether the shape of the class signature is affected or not by a specific application field. In fact, no evident correlation between class signature shape and application field has been observed on the selected groups of datasets, the only exception being for text categorization tasks. The membership of a dataset to a group being in general not a sufficient condition for shaping up a characteristic signature in the $\langle \varphi, \delta \rangle$ space came as no surprise. As for the "exception" found with text categorization, this is certainly due to the fact that a language lies behind the scenes while analyzing text / document samples. In particular, many studies acknowledge that a generic word occurs in a document corpus according to the *Zipf*'s power law [29]. This law states that the frequency of a word in a document corpus and its position in the corresponding frequency table (i.e., its frequency rank) are approximately inversely proportional. As a consequence, when checking the occurrence of a word in a dataset of documents, few high-ranking words are expected

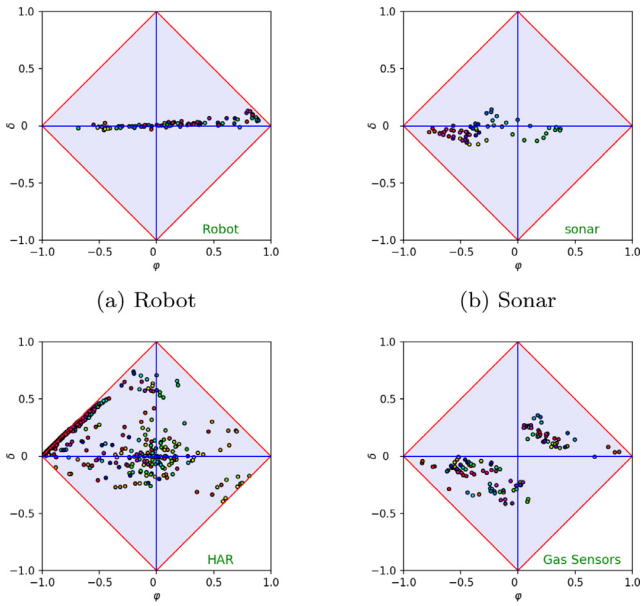


Fig. 23. Class signatures corresponding to the dataset group *remote sensing*.

Table 6
The table reports maximum absolute value of δ and classification scores for the selected datasets. The black circles on the right highlight datasets whose classification performance fulfills the behavior suggested by their class signature, whereas weak correlations have been labeled with a white circle. The table content clearly shows that in almost all cases medium or high values of $|\delta|$ guarantee an easy or very easy classification task.

Life sciences	Neg/Pos	**Max** $\lvert\delta\rvert$	a	$\overline{\rho}$	ρ	
SPECT images	0.26	0.16	0.78	0.25	0.91	•
Diabetes	0.88	0.16	0.77	0.87	0.57	•
Splice	0.93	0.66	0.97	0.98	0.96	•
Eye state	1.23	0.03	0.59	0.56	0.61	•
Multimedia	Neg/Pos	**Max** $\lvert\delta\rvert$	a	$\overline{\rho}$	ρ	
TechTC	1.06	0.27	0.77	0.67	0.86	•
Internet ads	6.16	0.30	0.95	0.97	0.84	•
Phishing	0.80	0.77	0.97	0.96	0.98	•
TV News (BBC dataset)	1.10	0.27	0.84	0.87	0.81	•
Social sciences	Neg/Pos	**Max** $\lvert\delta\rvert$	a	$\overline{\rho}$	ρ	
Credit card	1.25	0.72	0.86	0.88	0.83	•
German credit	2.33	0.32	0.76	0.90	0.44	•
Census	15.12	0.51	0.85	0.93	0.62	○
FARS	0.77	0.73	0.84	0.84	0.85	○
Remote sensing	Neg/Pos	**Max** $\lvert\delta\rvert$	a	$\overline{\rho}$	ρ	
Robot	0.66	0.13	0.80	0.69	0.87	•
Sonar	1.14	0.16	0.72	0.79	0.63	•
HAR	1.24	0.74	1.00	1.00	1.00	•
Gas sensors	0.22	0.42	0.99	0.95	1.00	•

to lay on the right side of a $\langle\varphi, \delta\rangle$ diagram, whereas many low-ranking words are expected to lay on the left side of the diagram. This particular shape, which can be found in all datasets of documents (including Internet web pages), is a direct consequence of the *Zip*'s power law. Note that, in this application field $\langle\varphi, \delta\rangle$ diagrams can be useful also in the task of detecting the so-called *stop words*, as they are constrained to appear on the right side of the diagram. A full confirmation of this hypothesis can be obtained by looking at the *multimedia* group. Indeed, the corresponding class signatures obey to the Zipf's law, the only apparent exception being the *Phishing* dataset. In fact, in this dataset a feature selection has been performed in a pre-processing step, with the goal of retaining only significant features. As a result, most of the rare and

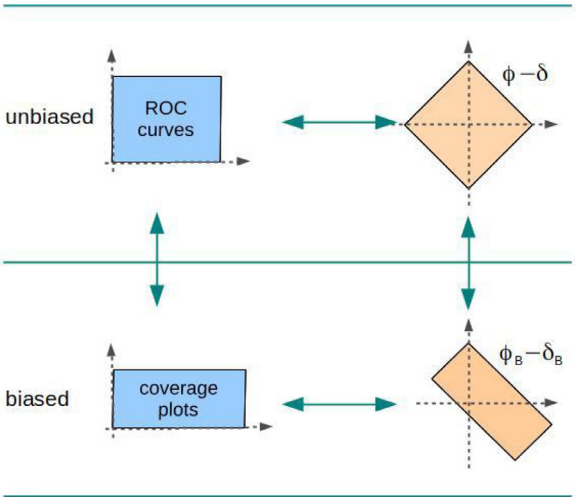


Fig. 24. Overview of the relationships between *ROC* curves and coverage plots, on one hand, and $\langle \varphi, \delta \rangle$ diagrams and $\langle \varphi_b, \delta_b \rangle$, on the other hand.

characteristic words have been removed therein. However, the class signature clearly shows that the pre-processing step could be further improved, as several rare and characteristic words have not been removed.[7]

7. Strengths and weaknesses of this proposal

$\langle \varphi, \delta \rangle$ diagrams come in two different forms. Similarly to what happens for *ROC* curves, standard $\langle \varphi, \delta \rangle$ diagrams are aimed at representing the phenomenon under investigation in a space where the class ratio is not taken into account. Framed along the same perspective of coverage plots, generalized $\langle \varphi, \delta \rangle$ diagrams are able to visualize relevant information in a space that accounts also for the class ratio.

In either form, $\langle \varphi, \delta \rangle$ diagrams allow a researcher to see at a glance bias and accuracy, while *ROC* curves and coverage plots do not.

As for standard $\langle \varphi, \delta \rangle$ diagrams, the unbiased versions of accuracy and bias of a classifier are immediately visible therein, whereas it is not immediate to spot them in *ROC* diagrams. Maximum entropy and break-even points are also clearly visible, as they correspond to the φ and to the δ axis, respectively. Notably, the x-axis is an entropic watershed, being the locus of points for which the mutual information is zero, whereas the y-axis is at the same time the locus of points for which specificity equals sensitivity and the locus of break-even points. Hence, by construction, one can immediately see whether a classifier tends to perform better on negative rather than on positive data or vice versa.

As for the generalized $\langle \varphi, \delta \rangle$ diagrams, they have been devised to let researchers analyze the actual behavior of a classifier, which in general depends also on the statistics of data. Notably, this space is not only suited to assess the performance of a classifier in presence of the statistics on input data. It is also suited to perform *any* sort of "guess if" concerning the behavior of a classifier for specific class ratios. Here, the semantics of the x-axis and of the y-axis are still related to bias and accuracy, this time according to their classical definitions. In particular, isometrics of accuracy are still straight lines parallel to the x-axis, whereas isometrics of bias are still straight lines parallel to the y-axis. The locus of break-even points is still coincident with the φ_b axis, whereas the entropic watershed is stretched according to the given class ratio. The crossing between x-axis and y-axis gives also information about to what extent the performance of a classifier is affected by the class ratio. In fact, whereas the regions identified by the crossing have the same area in presence of perfect balancing, these regions may undergo a great change depending on the imbalance (the greater the imbalance, the more evident the change). This phenomenon may give rise to experimental results deemed (or shown as) good, which are in fact related to the statistics of input data rather than on the way a classifier is able to generalize on the given task. Moreover, *classifier tuning* is straightforward with $\langle \varphi_b, \delta_b \rangle$ diagrams, as the best threshold can be easily found by first searching for the global maximum of accuracy in a $\langle \varphi_b, \delta_b \rangle$ curve and then retrieving the corresponding threshold.

As for the dependence between $\langle \varphi, \delta \rangle$ diagrams (in either form), on the one hand, and *ROC* curves and coverage plots, on the other hand, Fig. 24 clearly highlights their mutual relationships. Which kind of diagram should be selected for assessment depends on the focus of the analysis. *ROC* diagrams and coverage plots are suitable to analyze data according to specificity and / or sensitivity, whereas $\langle \varphi, \delta \rangle$ and $\langle \varphi_b, \delta_b \rangle$ diagrams are the most proper choice for a researcher interested in bias and / or accuracy. $\langle \varphi, \delta \rangle$ diagrams are also one of the best choices for investigating feature importance. In particular, to the time of writing, the concept of class signature is the preserve of $\langle \varphi, \delta \rangle$ diagrams.

[7] Recall that the unbiased accuracy depends on δ according to the formula: $a = (\delta + 1)/2$).

As a final remark, let us note that all software libraries that implement $\langle \varphi, \delta \rangle$ measures and their rendering as diagrams have been implemented using the Python programming language. The available software also includes a visual inspector that allows to perform any sort of "guess if" about the imbalance, as the class ratio on the dataset at hand were in fact changed, together with further code aimed at performing analyses on mutual information, entropy and Gini index. All cited software is available, as alpha release, at GitHub (https://github.com/garmano/phiDelta.git), under the GNU GPL licence.

8. Conclusions and future work

In this article, two measures have been proposed, i.e. φ and δ, which have been framed in both unbiased and biased spaces. For each space, a corresponding 2D visual environment has been devised and implemented, aimed at facilitating the task of assessing the properties of binary classifiers and of performing feature importance analysis. Isometrics and loci of points have been studied first in the standard (unbiased) space and then the generalized (biased) space. By construction, relevant isometrics (in particular, bias and accuracy) share the same behavior in both spaces. With the help of a rich experimental section, it has been shown that $\langle \varphi, \delta \rangle$ diagrams (in either form) can be used: (i) to assess the bias / accuracy of a classifier over a single or multiple runs; (ii) to assess the variance of a classifier; (iii) to find the right threshold to be applied on the classifier at hand; and (iv) to estimate the difficulty of the current classification task (by analyzing its features).

As for future work, the usefulness of standard and generalized $\langle \varphi, \delta \rangle$ diagrams in the implementation of effective classifiers will be further investigated. In particular, we are studying how the class ratio affects the bias of learning algorithms, from a theoretical and an experimental perspective. Furthermore, we are trying to extend the token sharing strategy to evaluate the pairwise correlation between features, this task being often a mandatory step for implementing effective feature selection algorithms. The possibility of embedding the proposed measures in these algorithms is also under study.

Acknowledgments.

This research work has been supported by LR7 2007 grant number: F71J11000590002 (Investment Funds for Basic Research) and by PIA 2010 grant number: 1492-118/2013 (Integrated Subsidized Packages), both funded by the local government of Sardinia. Further support to this work has been given by the DAAD-MIUR Joint Mobility Program, year 2015–16. The authors wish to thank Lorenza Saitta, Dominik Heider and Ursula Neumann for their support in discussing and developing the ideas reported in this article.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ins.2018.06.052

References

[1] G. Armano, A direct measure of discriminant and characteristic capability for classifier building and assessment, Inf. Sci. (Ny) 325 (2015) 466–483.
[2] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[3] A. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159.
[4] T. Calders, S. Jaroszewicz, Efficient AUC optimization for classification, in: Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases, 2007, pp. 42–53.
[5] C. Cortes, M. Mohri, Auc optimization vs. error rate minimization, in: Proceedings of the NIPS, 2003.
[6] L.A. Dalton, Optimal ROC-based classification and performance analysis under Bayesian uncertainty models, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (2016) 719–729.
[7] J. Demšar, et al., Orange: data mining toolbox in python, J. Mach. Learn. Res. 14 (2013) 2349–2353.
[8] P. Domingos, A unified bias-variance decomposition for zero-one and squared loss, in: Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI '00), 2000.
[9] C. Drummond, R.C. Holte, Explicitly representing expected cost: an alternative to ROC representation, in: Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD'00), 2000, pp. 198–207.
[10] C. Drummond, R.C. Holte, What ROC curves cannot do (and cost curves can), in: Proceedings of the ROCAI, 2004.
[11] C. Drummond, R.C. Holte, Cost curves: an improved method for visualizing classifier performance, Mach. Learn. 65(1) (2006) 95–130.
[12] W. Elazmeh, N. Japkowicz, S. Matwin, A framework for comparative evaluation of classifiers in the presence of class imbalance, 3rd Int. Workshop on ROC Analysis in Machine Learning (ROCML-2006), 2006, pp. 25–32.
[13] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. (special issue: ROC analysis in pattern recognition) 27(8) (2006) 861–874.
[14] T. Fawcett, Roc graphs with instance-varying costs, Pattern Recognit. Lett. (special issue: ROC analysis in pattern recognition) 27(8) (2006) 882â891.
[15] P. Flach, Machine Learning – The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012.
[16] P.A. Flach, The geometry of ROC space: understanding machine learning metrics through ROC isometrics, in: Proceedings of the Twentieth International Conference on Machine Learning (ICML '03), AAAI Press, 2003, pp. 194–201.
[17] J. Fürnkranz, P.A. Flach, Roc 'n' rule learning – towards a better understanding of covering algorithms, Mach. Learn. 58 (1) (2005) 39â77.
[18] X. Guo, Y. Yin, C. Dong, G. Yang, G. Zhou, On the class imbalance problem, in: Proceedings of the ICNC'08 – Fourth International Conference on Natural Computation, IEEE, 2008, doi:10.1109/ICNC.2008.871.
[19] 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.
[20] J. Hernández-Orallo, P. Flach, C. Ferri, A unified view of performance metrics: translating threshold choice into expected classification loss, J. Mach. Learn. Res. 13 (2012) 2813–2869.
[21] J. Huang, C.X. Ling, Using auc and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (2005) 299–310.
[22] S.A. Macskassy, F.J. Provost, Confidence bands for ROC curves: Methods and an empirical study, in: Proceedings of the ROCAI, 2004.
[23] M. Majnik, Z. Bosnic, ROC analysis of classifiers in machine learning: a survey, Intell. Data Anal. 17(3) (2013) 531–558, doi:10.3233/IDA-130592.

[24] A.F. Martin, G.D.T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, in: Proceedings of the Eurospeech '97, 1997, pp. 1899–1903.
[25] L.A.C. Millard, P.A. Flach, J.P.T. Higgins, Rate-constrained Ranking and the Rate-weighted AUC, in: Proceedings of the Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 386–403.
[26] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45(4) (2009) 427–437.
[27] S. Vanderlooy, I.G. Sprinkhuizen-Kuyper, E.N. Smirnov, An analysis of reliable classifiers through ROC isometrics, in: 3rd Int. Workshop on ROC Analysis in Machine Learning (ROCML-2006), 2006, pp. 55–62.
[28] S. Vanderlooy, I.G. Sprinkhuizen-Kuyper, E.N. Smirnov, H.J. van den Herik, The ROC isometrics approach to construct reliable classifiers, Intell. Data Anal. 13 (2009) 3–37.
[29] G. Zipf, Human Behavior and the Principle of Least Effort, Addison Wesley, 1949.