Information Sciences 325 (2015) 466–483

Contents lists available at ScienceDirect

Information Sciences

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

A direct measure of discriminant and characteristic capability for classifier building and assessment

Giuliano Armano*

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

ARTICLE INFO

Article history:
Received 6 September 2014
Revised 5 April 2015
Accepted 6 July 2015
Available online 15 July 2015

Keywords:
Classifier performance measures
Feature ranking/selection
Confusion matrices

ABSTRACT

Performance measures are used in various stages of the process aimed at solving a classification problem. Unfortunately, most of these measures are in fact *biased*, meaning that they strictly depend on the class ratio – i.e. on the imbalance between negative and positive samples. After pointing to the source of bias for the best known measures, novel unbiased measures are defined which are able to capture the concepts of discriminant and characteristic capability. The combined use of these measures can give important information to researchers involved in machine learning or pattern recognition tasks, in particular for classifier performance assessment and feature selection.

1. Introduction

How to assess the performance of a classifier and the importance of features are key issues in the process of classifier building and assessment. Although framed in the same process, in principle (and in practice) these research topics follow different perspectives. For this reason, short summaries concerning the corresponding issues and proposals will be separately illustrated.

1.1. Assessing classifier performance

As for classifier performance assessment, there are a number of measures which are well known in the machine learning and pattern recognition communities. Let us recall, in particular: accuracy (strictly related to the error rate), precision, sensitivity (also called recall), specificity, F_1 and Matthews Correlation Coefficient (*MCC*). These measures share a common source, as they are all derived from confusion matrices (also called contingency tables). Other well known measures include Mean Square Error [6], cross-entropy [20], and AUC (e.g., [17]). Relevant work devised to shed light on the characteristics of the above measures include [34], [22], [39], and [24]. There are also many graphical representations and tools for model evaluation, such as ROC curves, a 2D visual environment widely acknowledged as the default choice for assessing the intrinsic behavior of a classifier (see, for instance, [5] and [13]), ROC isometrics [16], and cost curves [11]. More information on graphical methods for classifier performance assessment can be found in [32].

A further category of measures is aimed at assessing to which extent the classifier at hand is "keen" to classify inputs as belonging to the main or alternate category (the bias) and to which extent a classifier varies its performance depending on the datasets used for testing (the variance). The interested reader can consult, for instance, the work of Domingos [10] for more information on the issues related to bias and variance.

* Tel.: +39 706755758.
E-mail address: armano@diee.unica.it, armanodiee@gmail.com

http://dx.doi.org/10.1016/j.ins.2015.07.028
0020-0255/

Relevant issues arise also when tests occur in the presence of imbalance between negative and positive samples. In that case, measuring the overall accuracy (or the overall error) gives poor information about the underlying process enforced by the classifier at hand, any such measure being typically affected by the imbalance between data (the more the imbalance is, the less significant the measure is). This fact may be further worsened by a lack of statistical significance of experimental results, which may hold for minority test samples. While no practical solution exists which is able to contrast the latter issue, the former is usually dealt with by adopting a combination of measures, usually a pair, devised to assess the performance beyond the fact that test data are unbalanced.[1] Precision and recall, on one hand, and specificity and sensitivity, on the other hand, are typical examples of this strategy. Also ROC diagrams follow this approach, the default choice for their axes being false positive rate (i.e., 1 − specificity) on the x axis and true positive rate (i.e., sensitivity) on the y axis. A further strategy for assessing classifiers, often adopted in the presence of unbalanced data samples, consists of defining a single compound measure defined on top of other ones. F_1 and *MCC* are both examples of this strategy.

Unfortunately, regardless of the adopted strategy, most of the existing measures are in fact *biased*, meaning that they strictly depend on the class ratio – i.e. on the imbalance between positive and negative samples. However, the adoption of biased measures can only be recommended when the statistics of input data is available. In the event one wants to assess the *intrinsic* properties of a classifier, or other relevant aspects in the process of classifier building and assessment, the adoption of biased measures may not be a reliable choice. For this reason, in the literature, some proposals have been made to introduce unbiased measures – see in particular the work of Flach [16].

1.2. Ranking/selecting features

Several techniques have been devised to support the process of ranking/selecting features according to their importance. Besides classical approaches, e.g., Fisher linear discriminant analysis [15] and Pearson correlation coefficient [14], a number of proposals have been made over time. With the goal of providing a better understanding of the underlying mechanisms, let us summarize the proposals according to two different perspectives.

Starting from the definition given in [4], Guyon and Elisseeff [21] divide feature selection methods in three broad groups: filter, wrapper and embedded methods. The corresponding definitions concentrate on the dependence between the selection method and the underlying learning algorithm L: (i) filters are used independently from L, (ii) embedded methods are used inside L, and (iii) wrappers use L as a black box. Note that the ordering adopted while recalling these groups is not accidental. In fact, filter methods are usually the fastest, as they consist of proper pre-processing activities, while wrapper methods are expected to be the most expensive in terms of computing resources, as the learning algorithm is repeatedly called with the aim of scoring (subsets of) features according to their predictive power. Embedded methods lay in between, being slower than filters but faster than wrappers.

An alternative grouping strategy can be found in the work of Dash and Liu [8]. Borrowing relevant categories from [3] and [4], the authors divide evaluation functions in five groups: distance, information-theoretic, dependence, consistency, and classifier error rate. Also known as separability, divergence, or discrimination measures, distance measures (and their counterpart, i.e., similarity measures) suggest that a feature should be preferred to another when it induces a greater difference between class conditional probabilities. Notable examples in this category are LDA [19] and ICA [26]. Similarity/dissimilarity measures can also be used to encode the sample space with the aim of imposing some useful constraints therein. Recent work in this category includes LFDA [37], sparse LDA [33] and CDA [30]. Information-theoretic measures, typically aimed at evaluating the entropy or the mutual information of features, have been adopted in various feature ranking or feature subset assessment methods (see, for instance, [2], [27], [12], [35] and [7]). Dependence/correlation measures quantify the ability to predict the value of one variable from the value of another variable. They may be used to measure the dependence between/among input features (typically feature pairs) and to identify whether a correlation exists between a feature/a set of features and the desired output. Beyond the classical correlation measures (e.g., Pearson correlation coefficient or cosine similarity), more recent proposals in this field include CCA (e.g., [23]) and dCOV/dCOR [38]. Consistency measures evaluate the distance of a feature subset from the consistent state. As a subset of features may be more or less distant from the consistent state, a consistency based algorithm typically accepts or rejects a feature subset depending on an inconsistency threshold – usually set by the user. Among other proposals made in this field, let us recall [9], [40], [1] and [36]. Classifier error rate delegates to a classifier the assessment of feature subsets. Relevant proposals that fall in this group are [25], [31] and [29].

It is worth pointing out that a bridge between the cited classifications of feature selection methods can be easily found. Indeed, the first four types of evaluation measures are typically framed according to a *filter* perspective, due to their potential independence from the classifier at hand. However, nothing prevents from devising heuristics *embedded* by a learning algorithm that make use of any of those methods (e.g. the information gain heuristics adopted by decision trees). Classifier error rate measures (the fifth group) coincide "de facto" with the *wrappers* category.

[1] Note that getting information about the intrinsic performance of a classifier could be an issue also in presence of complete knowledge about the statistics of data the classifier is expected to handle. In fact, high imbalance in a dataset may not allow to check whether the adopted classifier is in fact able to discriminate or simply puts into practice a "dummy" strategy – recognizing all or the majority of samples as belonging to the most populated category.

1.3. Aim of this proposal

In this paper a pair of unbiased measures is proposed, able to capture the concepts of *discriminant* and *characteristic* capability. The former is expected to measure to which extent positive samples can be separated from the negative ones, whereas the latter is expected to measure to which extent positive and negative samples can be grouped together. After giving pragmatic definitions of these measures, their semantics is discussed for binary classifiers and binary features. An analysis focusing on the combined use of the corresponding measures in form of 2D diagrams is also made.

The remainder of the paper is organized as follows: after introducing the concept of normalized confusion matrix (obtained by applying Bayes decomposition to any given confusion matrix), in Section 2 a brief analysis of the best known measures is performed, pointing out that most of them are in fact biased. Section 3 firstly revisits the cited work of Flach, concerning the definition of unbiased measures and then a proposal for turning Matthews Correlation Coefficient (*MCC* hereinafter) into an unbiased measure is made. Section 4 introduces novel measures devised to assess the discriminant and characteristic capability of binary classifiers or binary features. An analysis concerning to which extent the behavior of a classifier/feature can be investigated using the proposed measures in combination is also made therein. Section 5 reports experiments aimed at highlighting the potential of 2D diagrams drawn using the proposed measures. Section 6 highlights the strengths and weaknesses of this paper and Section 7 draws conclusions.

2. Analysis of the best known measures

As the concept of confusion matrix is central in this paper, limiting our attention to binary problems, let us preliminarily illustrate the notation that has been adopted – also because it slightly differs from the most acknowledged one.

Let us assume that c and $\bar{c}$ denote the *main* and the *alternate* classes, respectively. Moreover, with Ξ generic confusion matrix, ξ_{ij}, $(i, j = 0, 1)$ denote the number of samples that have been correctly classified ($i = j$) or misclassified ($i \neq j$). 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. As a confusion matrix can also be used to evaluate the degree of agreement between a class and a binary feature, in this case the index j would denote the presence (1) or absence (0) of the binary feature, with i still representing the class of the sample at hand. In other words, ξ_{ij} gives the number of samples that belong to the class with index i whose truth value for the feature under analysis corresponds to index j.

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). With $\widehat{X}_c$ and X_c random variables that account for the output of classifier and oracle, the joint probability $p(X_c, \widehat{X}_c)$ is proportional, through M, to the expected value of $\Xi_c(P, N)$. In symbols:

$$E[\,\Xi_c(P, N))\,] = M \cdot p(X_c, \widehat{X}_c) \tag{1}$$

Note that Eq. (1) is a shorthand that summarizes all specific dependencies that hold between the observed number of occurrences and the corresponding probability, with varying i and j. In fact, we assume that ξ_{ij} reports the outcomes of an experiment according to an underlying process that accounts also for the statistics of data (including the prior probability of main and alternate class). According to this insight, we can write: $\xi_{ij} \approx M \cdot p(e_{ij})$, where e_{ij} represents the event $\langle X_c = i, \widehat{X}_c = j \rangle$. In other words, ξ_{ij} reports the number of times the event e_{ij} has been observed, given M samples.

Assuming statistical significance, the confusion matrix obtained from a single test (or, better, averaged over multiple tests in which the values for P and N are left unchanged) gives us reliable information on the performance of the classifier at hand. In symbols:

$$\Xi_c(P, N) \approx M \cdot p(X_c, \widehat{X}_c) = M \cdot p(X_c) \cdot p(\widehat{X}_c | X_c) \tag{2}$$

In so doing, we assume that the transformation performed by the classifier can be isolated from the inputs it processes, at least from a statistical perspective. Hence, the confusion matrix for a given set of inputs can be written as the product between a term that accounts for the number of positive and negative instances, on one hand, and a term that represents the expected recognition/error rate of the classifier, on the other hand. In symbols:

$$\Xi_c(P, N) = M \cdot \underbrace{\begin{bmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{bmatrix}}_{\Omega(c) \approx p(X_c, \widehat{X}_c)} = M \cdot \underbrace{\begin{bmatrix} n & 0 \\ 0 & p \end{bmatrix}}_{\mathcal{O}(c) \approx p(X_c)} \cdot \underbrace{\begin{bmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{bmatrix}}_{\Gamma(c) \approx p(\widehat{X}_c | X_c)} \tag{3}$$

where:

- $\Omega(c)$ 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: $\sum_{ij} \omega_{ij} = 1$.
- $\mathcal{O}(c)$ represents the behavior of the oracle, under the hypothesis that the classifier has been tested with N and P samples, so that $p(\bar{c}) \approx N/M = n$ and $p(c) \approx P/M = p$.
- $\Gamma(c)$, called *normalized confusion matrix* hereinafter, is an estimate of the conditional probability $p(\widehat{X}_c | X_c)$. In particular, $\gamma_{ij} \approx p(\widehat{X}_c = j \mid X_c = i)$ – with $i, j = 0, 1$ – 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

Table 1
Relevant quantities used to characterize, or derived from, confusion matrices (M denotes the total number of samples).

Literal	Denotes	Relevant formulas
N, P	Number of neg and pos samples	$N + P = M$
$\widehat{N}, \widehat{P}$	Number of samples classified as neg and pos	$\widehat{N} + \widehat{P} = M$
TN, FP	Number of true neg and false pos	$TN + FP = N$
FN, TP	Number of false neg and true pos	$FN + TP = P$
n, p	Percent of neg and pos samples	$n + p = 1$
$\widehat{n}, \widehat{p}$	Percent of samples classified as neg and pos	$\widehat{n} + \widehat{p} = 1$
tn, fp	True neg rate and false pos rate	$tn + fp = 1$
fn, tp	False neg rate and true pos rate	$fn + tp = 1$
σ	Imbalance between neg and pos samples	$\sigma = N/P = n/p$
a	Accuracy	$a = (TN + TP)/M$
$\overline{\rho}, \rho$	Specificity and sensitivity (recall)[1]	$\overline{\rho} = TN/N \equiv tn$ $\rho = TP/P \equiv tp$
$\overline{\pi}, \pi$	Negative predictive value and precision	$\overline{\pi} = TN/\widehat{N}$ $\pi = TP/\widehat{P}$

[1] For the sake of readability, the notations adopted for specificity and negative predictive value are similar to the ones used for their positive-side counterparts (i.e., sensitivity and precision, respectively), the only exception being a bar over the symbol.

(fp), false negatives (fn), and true positives (tp), respectively. Let us point out in advance that $\gamma_{00} \equiv tn \equiv \overline{\rho}$ (specificity) and $\gamma_{11} \equiv tp \equiv \rho$ (sensitivity). According to the total probability law (applied to a space of conditional probabilities): $\gamma_{00} + \gamma_{01} = \gamma_{10} + \gamma_{11} = 1$.

The separation between inputs and the intrinsic behavior of a classifier reported in Eq. (3) suggests an interpretation that recalls the concept of transfer function, where a set of inputs is applied to $\widehat{c}$. In fact, Eq. (3) highlights the separation of the optimal behavior of a classifier from the deterioration introduced by its actual filtering capabilities. In particular, $\mathcal{O} \approx p(X_c)$ represents the *optimal behavior* obtainable when $\widehat{c}$ acts as an *oracle*, whereas $\Gamma \approx p(\widehat{X}_c \,|X_c)$ represents the *expected deterioration* caused by the actual characteristics of the classifier. Hence, under the assumption of statistical significance of experimental results, any confusion matrix can be divided in terms of optimal behavior and expected deterioration using the Bayes theorem. As already pointed out, a different interpretation holds for confusion matrix subscripts when they are used to investigate binary features. In this case i still denotes the actual class, whereas j denotes the truth value of the binary feature (with 0 and 1 made equivalent to *false* and *true*, respectively). However, as a binary feature can always be thought of as a very simple classifier whose output reflects the truth value of the feature in the given samples (also called *single feature classifier*), all definitions and comments concerning classifiers can be applied to binary features as well.

Before examining the best known measures deemed useful for pattern recognition and machine learning according to the above perspective, let us take a look at the list of relevant symbols used throughout the paper, reported in Table 1. All definitions given therein are quite standard; nevertheless, some specific aspects deserve a comment. Table 1 points out that $n + p = 1$ and $\widehat{n} + \widehat{p} = 1$. The former equation is easy to verify, as n and p basically represent prior probabilities, which require the total amount of outcomes be equal to 1 according to the total probability law. As for the latter, let us recall that $\widehat{n}$ and $\widehat{p}$ account for the percent of samples acknowledged as belonging to the alternate and main class, respectively. In formulas: $\widehat{n} = n \cdot tn + p \cdot fn$ and $\widehat{p} = n \cdot fp + p \cdot tp$. Hence:

$$\widehat{n} + \widehat{p} = n \cdot tn + p \cdot fn + n \cdot fp + p \cdot tp \tag{4}$$

$$= n \cdot (tn + fp) + p \cdot (tp + fn) \equiv n + p = 1 \tag{5}$$

Recall that $tn + fp = fn + tp = 1$, as the normalized confusion matrix Γ represents a conditional probability space, in which the total probability law applies to rows. Recall also that: $tn \equiv \gamma_{00}, fp \equiv \gamma_{01}, fn \equiv \gamma_{10}$, and $tp \equiv \gamma_{11}$.

As for the imbalance between prior probabilities (i.e., the class ratio, defined as $\sigma = N/P = n/p$), it can be used to denote n and p. In fact, n and p are not independent (as $n + p = 1$) and can be represented in term of class ratio as follows:

$$n + p = 1 \Rightarrow p \cdot \left(\frac{n}{p} + 1\right) = 1 \quad \Rightarrow \quad \sigma + 1 = \frac{1}{p}$$

Hence:

$$p = \frac{1}{\sigma + 1} \quad \text{and} \quad n = 1 - p = \frac{\sigma}{\sigma + 1}$$

After pointing to the peculiarities of the lexicon, the classical definitions for accuracy (a), precision (π), and recall (ρ) can now be given in terms of false positives rate (fp), true positives rate (tp) and class ratio (σ) as follows:

$$a = \frac{trace(\Omega)}{|\Omega|} = \frac{\omega_{00} + \omega_{11}}{1} = \frac{\sigma \cdot (1 - \gamma_{01}) + \gamma_{11}}{\sigma + 1} = \frac{\sigma \cdot (1 - fp) + tp}{\sigma + 1}$$

$$\pi = \frac{\omega_{11}}{\omega_{11}+\omega_{01}} = \left(1+\sigma\cdot\frac{\gamma_{01}}{\gamma_{11}}\right)^{-1} = \left(1+\sigma\cdot\frac{fp}{tp}\right)^{-1}$$
$$\rho = \frac{\omega_{11}}{\omega_{11}+\omega_{10}} = \gamma_{11} = tp \tag{6}$$

Eq. (6) highlights the dependence of accuracy and precision from the class ratio, only recall being unbiased. Note that the expression concerning accuracy has been obtained taking into account that $p+n=1$ implies $p=1/(\sigma+1)$ and $n=\sigma/(\sigma+1)$.

As for the most popular compound measures, i.e. F_1 and *MCC*, they are both biased. The dependence of F_1 from the class ratio is evident, being defined also in terms of precision. The membership of *MCC* in the group of biased measures is unclear if one starts from the usual definition, reported below for the sake of completeness:

$$MCC = \frac{TN\cdot TP - FP\cdot FN}{\sqrt{P\cdot N\cdot \widehat{P}\cdot \widehat{N}}} \tag{7}$$

However, with $(tn\cdot tp - fp\cdot fn) \stackrel{def}{=} \Delta$, we can write:

$$MCC = \frac{N\cdot P\cdot \Delta}{\sqrt{P\cdot N\cdot \widehat{P}\cdot \widehat{N}}} = \sqrt{\frac{N\cdot P}{\widehat{P}\cdot \widehat{N}}}\cdot\Delta = \sqrt{\frac{\pi}{\rho}\cdot\frac{\overline{\pi}}{\overline{\rho}}}\cdot\Delta \tag{8}$$

where π and $\overline{\pi}$ denote precision and negative predictive value, while ρ and $\overline{\rho}$ sensitivity (i.e. recall) and specificity. Note that Δ is in fact the determinant of the normalized confusion matrix Γ, as $tn\cdot tp - fp\cdot fn \equiv \gamma_{00}\cdot\gamma_{11} - \gamma_{01}\cdot\gamma_{10}$. Eq. (8) clearly highlghts that also *MCC* is biased, as it is defined in terms of precision and negative predictive value.

3. Turning biased measures into unbiased ones

As pointed out, when the goal is to asssess the intrinsic properties of a classifier or a feature, biased measures do not appear to be a proper choice, leaving room for alternative definitions aimed at dealing with the imbalance between negative and positive samples. In this section, some proposals made in this direction are briefly recalled. Moreover, a proposal on how to reformulate MCC as unbiased measure is also made.

In [16], Flach gave definitions of some unbiased measures starting from classical ones. It is worth noting that the formulas of accuracy, precision and recall proposed therein can also be obtained by simply substituting in Eq. (6) the joint probability Ω with the normalized confusion matrix Γ. Hence, we can write (the subscript u is used to distinguish unbiased from biased measures):

$$a_u = \frac{trace(\Gamma)}{|\Gamma|} = \frac{\gamma_{00}+\gamma_{11}}{2} = \frac{1-fp+tp}{2}$$
$$\pi_u = \frac{\gamma_{11}}{\gamma_{01}+\gamma_{11}} = \left(1+\frac{fp}{tp}\right)^{-1}$$
$$\rho_u = \frac{\gamma_{11}}{\gamma_{11}+\gamma_{10}} = tp \equiv \rho \tag{9}$$

Of course, also F_1 and *MCC* can be reformulated accordingly, being compound measures defined in terms of other well-known measures. In particular, unbiased versions of F_1 and *MCC* can be obtained by simply substituting π and $\overline{\pi}$ with their unbiased counterparts. As for F_1, this substitution strategy has already been proposed by Flach, whereas the unbiased version of *MCC* proposed in this paper is:

$$MCC_u = \sqrt{\frac{\pi_u}{\rho_u}\cdot\frac{\overline{\pi}_u}{\overline{\rho}_u}}\cdot\Delta \tag{10}$$

4. Definition of measures able to account for discriminant and characteristic capability

To my knowledge, no satisfactory definitions have been given so far able to account for the need of capturing the potential of a model according to its discriminant and characteristic capability. With the goal of filling this gap, let us spend few words on the expected behavior of any such measures.

4.1. Preliminary analysis and corresponding constraints

Without loss of generality, let us assume the measures be defined in $[-1,+1]$. As for the discriminant capability, we expect its value be close to $+1$ when a classifier or feature partitions a given set of samples in strong accordance with the corresponding class labels. Conversely, the measure is expected to be close to -1 when the partitioning occurs in strong discordance with the class label. As for the characteristic capability, we expect its value be close to $+1$ when a classifier or feature tend to cluster most of the samples as if they were in fact belonging to the main class. Conversely, the measure is expected to be close to -1

Table 2
Expected behavior of measures aimed at measuring discriminant and characteristic capability.

Notation	Intended to measure	Domain	Expected behavior
δ	Discriminant capability	$[-1,+1]$	$\delta \approx \pm 1 \Leftrightarrow \varphi \approx 0$
φ	Characteristic capability	$[-1,+1]$	$\varphi \approx \pm 1 \Leftrightarrow \delta \approx 0$

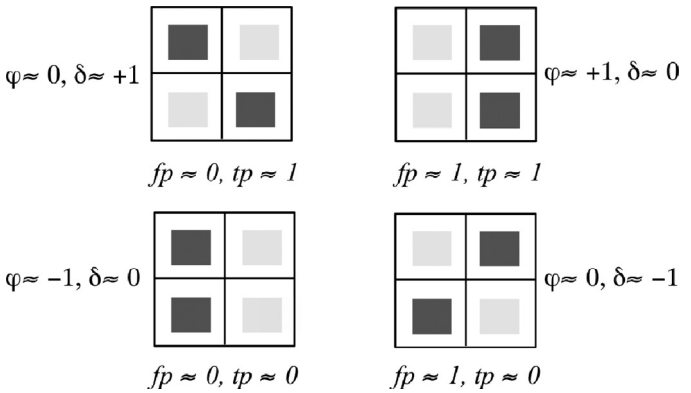


Fig. 1. Relevant cases for δ and φ.

when most of the samples are clustered as belonging to the alternate class.[2] An immediate consequence of the desired behavior is that the above properties are not independent. In other words, regardless of their definition, the measures devised to assess discriminant and characteristic capabilities of a classifier or feature (say δ and φ, hereinafter) are expected to show an orthogonal behavior. In particular, when the absolute value of one measure is about 1 the other should be close to 0 and vice versa (see also Table 2, which summarizes domains and expected behavior of δ and φ).

Let us now characterize δ and φ with more details, focusing on classifiers only (similar considerations can also be made for features):

- $fp \approx 0$ and $tp \approx 1$ – In this case we expect $\delta \approx +1$ and $\varphi \approx 0$, meaning that the classifier is able to partition the given samples almost in complete accordance with the class labels.
- $fp \approx 1$ and $tp \approx 1$ – In this case we expect $\delta \approx 0$ and $\varphi \approx +1$, meaning that almost all samples are in fact recognized as belonging to the main class label.
- $fp \approx 0$ and $tp \approx 0$ – In this case we expect $\delta \approx 0$ and $\varphi \approx -1$, meaning that almost all samples are in fact recognized as belonging to the alternate class label.
- $fp \approx 1$ and $tp \approx 0$ – In this case we expect $\delta \approx -1$ and $\varphi \approx 0$, meaning that the classifier partitions the domain space almost in complete discordance with the class labels (however, this ability can still be used for classification purposes by simply turning the classifier output into its opposite).

To better understand the concepts above, the reader may also refer to Fig. 1, which graphically illustrates the dependence of δ and φ from the corresponding confusion matrix, with varying tp and fp (light gray denotes few hits and dark gray many hits).

The determinant of the normalized confusion matrix is the starting point for giving proper definitions of δ and φ which are able to satisfy the constraints and boundary conditions discussed above. It can be rewritten as follows:

$$\begin{aligned}\Delta &= \gamma_{00}\cdot\gamma_{11} - \gamma_{01}\cdot\gamma_{10} \\ &= tn\cdot tp - fp\cdot fn = tn\cdot tp - (1-tn)\cdot(1-tp) \\ &= tn\cdot tp - 1 + tn + tp - tn\cdot tp = tp + tn - 1 \\ &= tp - fp \equiv \rho + \overline{\rho} - 1\end{aligned} \tag{11}$$

When $\Delta = 0$, the classifier under assessment has no discriminant capability whereas $\Delta = +1$ and $\Delta = -1$ correspond to the highest discriminant capability, from the positive and negative side, respectively. It is clear that the simplest definition of δ is to *make it coincident with* Δ, as the latter has all the desired properties required by the discriminant capability measure. Fig. 2 reports the isometric curves drawn for different values of δ with varying tp ad fp.

[2] It is worth noting that the definition of characteristic capability proposed in this paper is in partial disagreement with the classical concept of "characteristic property" acknowledged by most of the machine learning and pattern recognition researchers. The classical definition focuses only on samples that belong to the main class, whereas the conceptualization adopted in this paper applies to all samples. The motivation of this choice should become clearer later on.

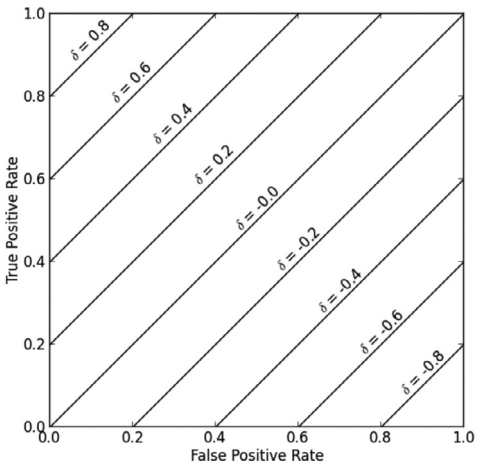


Fig. 2. Isometric plotting of the discriminant capability δ with varying false and true positive rate.

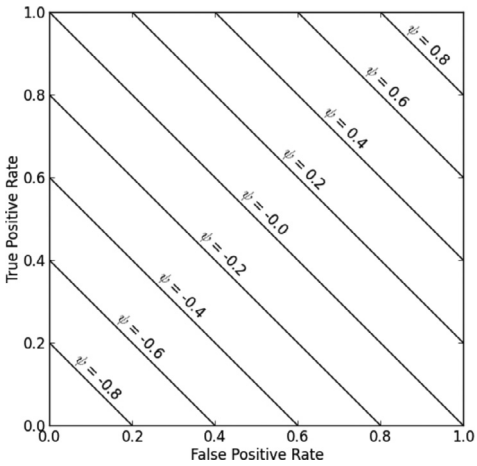


Fig. 3. Isometric plotting of the characteristic capability φ with varying false and true positive rates.

As for φ, considering the definition of δ and the constraints that must apply to a measure intended to assess the characteristic capability, the following definition appears appropriate, being dual with respect to δ also from a syntactic point of view:

$$\varphi = tp + fp - 1 = \rho - \overline{\rho} \tag{12}$$

As expected, the isometric curves of φ, reported in Fig. 3, are orthogonal with respect to the ones drawn for δ. This is a clear indicator that the two measures can be taken in combination for investigating properties of classifiers or features. The run of a classifier over a specific test set, different runs of a classifier over multiple test sets, and the statistics about the presence/absence of a feature on a specific dataset are all examples of potential use cases. However, while reporting information about classifier or feature properties in $\varphi - \delta$ diagrams, one should be aware that the $\varphi - \delta$ space is constrained by a rhomboidal shape, whose borders are identified by the equation $|\varphi| + |\delta| = 1$. This shape depends on the constraints that apply to δ, φ, *tp*, and *fp*. In particular, as $\delta = tp - fp$ and $\varphi = tp + fp - 1$, the following relations hold:

$$\delta = -\varphi + (2 \cdot tp - 1) = +\varphi + (2 \cdot fp + 1) \tag{13}$$

Considering *fp* and *tp* as parameters, we can easily draw the corresponding isometric curves in the $\varphi - \delta$ space. Fig. 4 shows their behavior for $tp = \{0, 0.5, 1\}$ and for $fp = \{0, 0.5, 1\}$. As the definitions of δ and φ are given as linear transformations over *tp* and *fp*, it is not surprising that the isometric curves of *fp* and *tp* drawn in the $\varphi - \delta$ space are again straight lines.

4.2. Semantics of the $\varphi - \delta$ space for classifiers

The discriminant capability of binary classifiers is strictly related to the *unbiased accuracy*, which in turn can be given in terms of *unbiased error* (say e_u). The following equivalences make explicit the relation between a_u, e_u and δ:

$$a_u = \frac{tn + tp}{2} = \frac{1 + \delta}{2} = 1 - \frac{1 - \delta}{2} = 1 - \frac{fp + fn}{2} = 1 - e_u \tag{14}$$

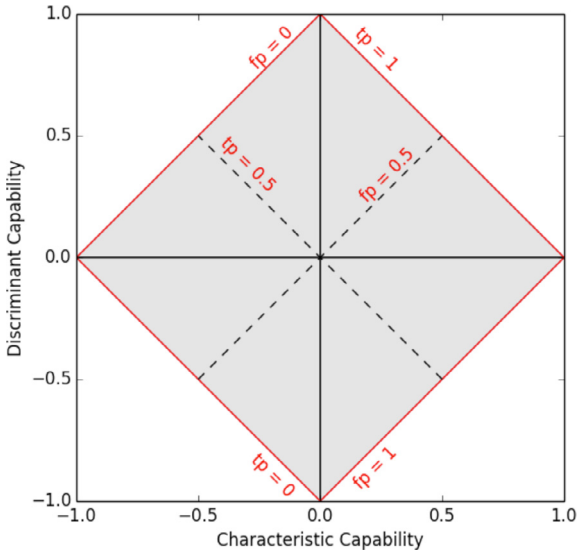


Fig. 4. Shape of the $\varphi - \delta$ space: the diamond centered in (0,0) delimits the area of admissible value pairs.

It is worth pointing out that the actual discriminant capability of a classifier is not a redefinition of accuracy (or error), as a classifier may still have high discriminant capability also in the presence of high unbiased error. Indeed, as already pointed out, a low-performance classifier can be easily transformed into a high-performance one by simply turning its output into its opposite. Thanks to the "turning-into-opposite" trick, the actual discriminant capability of a classifier could in fact be made coincident with the absolute value of δ. However, for reasons related to the informative content of $\varphi - \delta$ diagrams, we still take apart the discriminant capability observed from the positive side from the one observed on the negative side.

As for the characteristic capability, let us preliminarily note that, in presence of statistical significance, the expected values of oracle (X_c) and classifier ($\widehat{X}_c$) can be written as follows:

$$\begin{aligned} E[X_c] &\approx \frac{1}{M} \cdot (P - N) = (p - n) \\ E[\widehat{X}_c] &\approx \frac{1}{M} \cdot \left(\widehat{P} - \widehat{N}\right) = (p - n) + 2 \cdot n \cdot fp - 2 \cdot p \cdot fn \end{aligned} \tag{15}$$

Hence, the difference in terms of expected values between oracle and classifier is:

$$E[X_c - \widehat{X}_c] = E[X_c] - E[\widehat{X}_c] \approx -2 \cdot n \cdot fp + 2 \cdot p \cdot fn \tag{16}$$

It is easy to show that Eq. (16) actually represents an estimate of the *bias* of a classifier, measured over the confusion matrix that describes the outcomes of the experiments performed with the given test set(s). Let us verify it, starting from the classical definition given in Friedman [18]:

$$\text{bias } \hat{f}(\mathbf{X}) = f(\mathbf{X}) - E[\hat{f}(\mathbf{X})] \tag{17}$$

where $f(\mathbf{X})$ corresponds to the output of an oracle over a set $\mathbf{X}$ of samples and $E[\hat{f}(\mathbf{X})]$ to the expected value over $\mathbf{X}$ of the corresponding classifier. Assuming, without loss of generality, that the output label of a binary classifier is in $\{-1, +1\}$, it is easy to show that – in presence of statistical significance – the following relations hold:

$$\begin{aligned} f(\mathbf{X}) &= \frac{1}{|\mathbf{X}|} \cdot \sum_{s \in \mathbf{X}} h(s) = p - n \\ E[\hat{f}(\mathbf{X})] &= \frac{1}{|\mathbf{X}|} \cdot \sum_{s \in \mathbf{X}} \hat{h}(s) = (p - n) + 2 \cdot n \cdot fp - 2 \cdot p \cdot fn \end{aligned} \tag{18}$$

where:

- $h(s)$ and $\hat{h}(s)$ denote the output of the oracle and the output of the classifier at hand on a sample s;
- n and p denote, as usual, the percent of negative and positive samples – respectively.[3]

[3] For the sake of simplicity, here we assume a deterministic behavior for the classifier at hand. In so doing, the expected value over $\hat{f}(\mathbf{X})$ can be disregarded. However it is easy to verify that, in presence of statistical significance, the formulation of bias in terms of false and true positive rates still holds, regardless of the hypothesis made about the behavior of the classifier (i.e., deterministic or nondeterministic).

To show how the results reported in Eq. (18) can be obtained, let us assume (i) that $\mathbf{X}^+$ and $\mathbf{X}^-$ denote the subset of samples that belong to main and alternate class, respectively, and that (ii) $\hat{\mathbf{X}}^+$ and $\hat{\mathbf{X}}^-$ denote the subset of samples that have been *classified as* belonging to main and alternate class, respectively. Hence, we can write:

$$\begin{aligned}
f(\mathbf{X}) &= \frac{1}{|\mathbf{X}|}\cdot\sum_{s\in\mathbf{X}} h(s)\\
&= \frac{1}{|\mathbf{X}|}\cdot\left(\sum_{s\in\mathbf{X}^-}(-1)+\sum_{s\in\mathbf{X}^+}(+1)\right)\\
&= -\frac{N}{M}+\frac{P}{M}\\
&= -n+p\\
E[\hat{f}(\mathbf{X})] &= \frac{1}{|\mathbf{X}|}\cdot\sum_{s\in\mathbf{X}} \hat{h}(s)\\
&= \frac{1}{|\mathbf{X}|}\cdot\left(\sum_{s\in\hat{\mathbf{X}}^-}(-1)+\sum_{s\in\hat{\mathbf{X}}^+}(+1)\right)\\
&= -\frac{TN+FN}{M}+\frac{TP+FP}{M}\\
&= -\frac{N\cdot tn+P\cdot fn}{M}+\frac{P\cdot tp+N\cdot fp}{M}\\
&= -(n\cdot tn+p\cdot fn)+(p\cdot tp+n\cdot fp)\\
&= -n\cdot tn-p\cdot fn+p\cdot tp+n\cdot fp\\
&= -n\cdot(1-fp)-p\cdot fn+p\cdot(1-fn)+n\cdot fp\\
&= (p-n)+2\cdot n\cdot fp-2\cdot p\cdot fn
\end{aligned} \tag{19}$$

In conditions of perfect balancing (i.e., when $n=p=1/2$), Eq. (17) can be rewritten as:

$$\text{bias}\ \hat{f}(\mathbf{X}) = 1-tp-fp \equiv -\varphi \tag{20}$$

It is worth noting that the trailing minus sign that occurs in Eq. (20) depends only on the classical definitions of bias, which typically measure to which extent the oracle differs from the classifier at hand on the given data or data distribution. This is definitely a minor point when the bias is used to calculate mean squared errors, as it appears squared in the corresponding formulas. However, in the event one is interested at identifying to which extent the classifier at hand *differs from* the oracle, a better formulation for the bias would be something like $E[\hat{f}(\mathbf{X})]-f(\mathbf{X})$, making the φ axis in fact coincident with the (unbiased) bias. In particular, when the performance of a classifier measured over a test set $\mathbf{X}$ lies on the positive semiplane of φ, one can argue that the classifier has a positive bias towards the main class (+1) and vice versa (as usual, the strength of the assumption depends on to which extent the test is statistically significant).

Summarizing, when framed in a classifier-oriented view, a $\varphi-\delta$ diagram is in fact a *bias-error* (or *accuracy-error*) diagram, as φ represents the unbiased bias, whereas δ is linearly related to the unbiased accuracy (or, equivalently, to the unbiased error). Fig. 6 points out this aspect, under the assumption that the generic point P_0 reports the performance of a classifier, measured on a given test set, in terms of φ and δ. As a consequence, classifiers with $\varphi\approx 0$ and $\delta\approx 1$ are expected to show very good performance, with loose dependencies on the actual imbalance of data. Conversely, classifiers with $|\varphi|\approx 1$ are expected to show very bad performance.[4] To better illustrate these concepts, let us refer to Fig. 5, which points out that the diamond corners are in fact well known kinds of classifiers. In particular, the upper corner (characterized by $\rho=\overline{\rho}=1$) denotes a classifier that is always correct (i.e., an *oracle*). Conversely, the lower corner (characterized by $\rho=\overline{\rho}=0$) denotes a classifier that is always wrong (let us call it *anti-oracle*). As for the left and the right corners, they both denote *dummy classifiers*. In particular, the one on the left side always takes pessimistic decisions (considering any sample as belonging to the alternate class), whereas the one on the right side always takes optimistic decisions (considering any sample as belonging to the main 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).

As final remarks, let us point out that the φ axis (defined by the equation $\delta=0$) can also be characterized as the locus of points for which the mutual information between classifier and oracle is minimum (i.e., they are points of maximum entropy), whereas the δ axis (defined by the equation $\varphi=0$) denotes the locus of points called breakeven points. As for the former property, the starting point for analyzing it is the definition of mutual information $I(\widehat{X}_c;X_c)$ between the random variables associated to the classifier (i.e., $\widehat{X}_c$) and to the oracle (i.e., X_c). In symbols:

$$I(\widehat{X}_c;X_c) = H(X_c)-H(X_c|\widehat{X}_c) \tag{21}$$

[4] 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 alternate class to each submitted sample. However, this would only depend on the imbalance of data, rather than on the intrinsic behavior of the classifier.

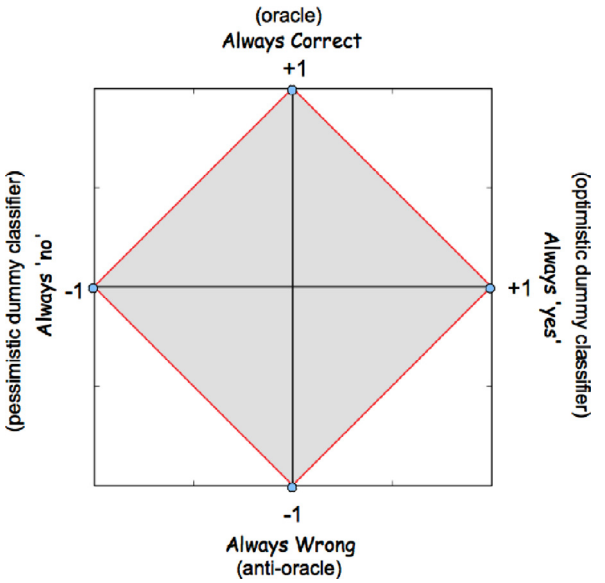


Fig. 5. Relevant points in the $\varphi - \delta$ space, when framed in a classifier-oriented view.

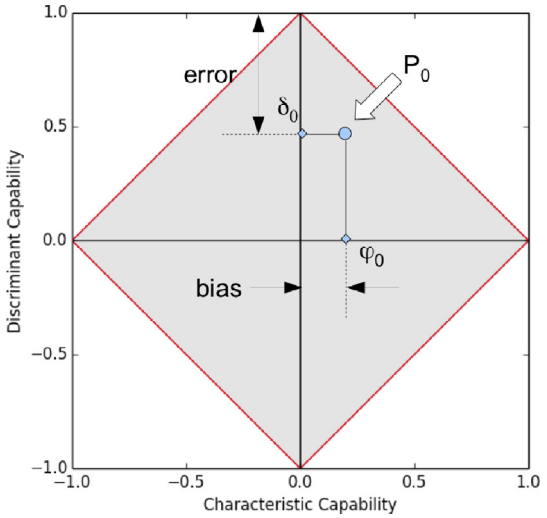


Fig. 6. A point in the $\varphi - \delta$ space, framed within a classifier-oriented view.

Applied to this specific setting, this information-theoretic measure actually gives the information gain obtained by applying the classifier at hand to the given problem. Of course, 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 is enforced). Also the origin of axes (on which $\rho = \overline{\rho} = 1/2$) is expected to carry a null information gain, as in that point the classifier basically takes random guesses. In fact, the whole φ axis is characterized by a null information gain. Fig. 7 graphically illustrates this aspect, highlighting that all points for which the information gain is minimum (i.e., the entropy is maximum) lay in fact on the φ axis. In particular, the left hand side of the figure shows the whole 3D plot of $I(\widehat{X}_c; X_c)$, whereas the right hand side highlights its minima.

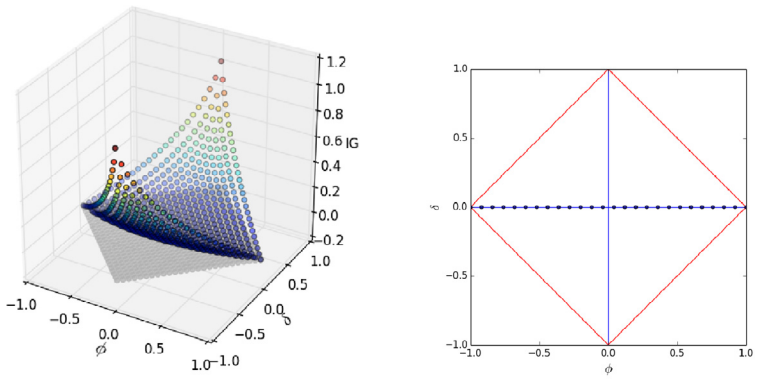
As for breakeven points, let us recall that they are characterized by the fact that $\pi \equiv \rho$. When $\sigma = 1$, precision is defined as:

$$\pi = \frac{TP}{\widehat{p}} = \frac{TP}{TP + FP} \equiv \frac{tp}{tp + \sigma \cdot fp} = \frac{tp}{tp + fp} \tag{22}$$

Enforcing the constraint that identifies a breakeven point, we can write:

$$\rho = \pi \Rightarrow tp + fp = 1 \Rightarrow \rho + (1 - \overline{\rho}) = 1 \Rightarrow \rho - \overline{\rho} = 0 \Rightarrow \varphi = 0 \tag{23}$$

where the equation $\varphi = 0$ of course identifies the δ axis.



(a) Mutual information over $\varphi - \delta$ (b) Mutual information minima

Fig. 7. The mutual information (i.e., the information gain) evaluated over the whole $\varphi - \delta$ space is reported on the left, whereas the locus of points for which the information gain is minimum is reported on the right (a 25×25 grid has been used in both cases).

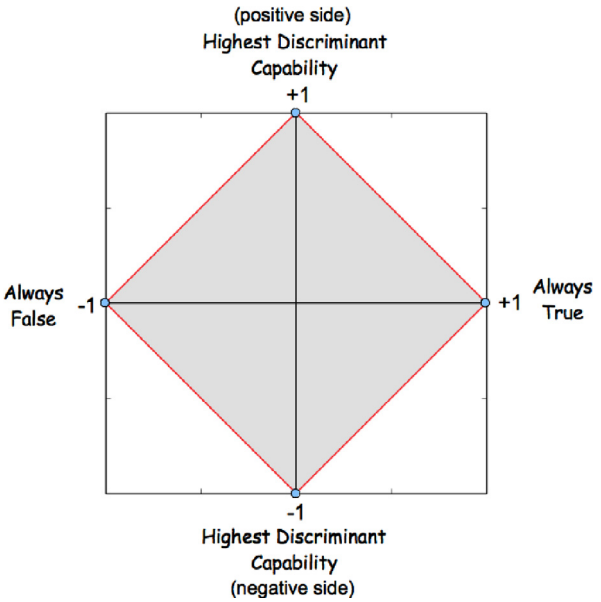


Fig. 8. Relevant points in the $\varphi - \delta$ space, when framed in a feature-oriented view.

4.3. *Semantics of the $\varphi - \delta$ space for features*

As for binary features, δ measures to which extent a feature is able to partition the given samples in accordance ($\delta \approx +1$) or in discordance ($\delta \approx -1$) with the main class. In the former case the feature is *covariant* with the main class, whereas in the latter it is covariant with the alternate class (hence, it is *contravariant* with the main class). In either case, any such feature would have *high discriminant capability*, the importance of the feature being mainly related to a high absolute value of δ. However, as already pointed out for classifiers, instead of considering the absolute value of δ as a measure of discriminant capability, we take apart the value observed on the positive side from the one observed on the negative side for reasons related to the informative content of $\varphi - \delta$ diagrams. As for φ, it measures to which extent the feature at hand is pervasive for the given dataset. A high positive value of φ indicates that the feature is mainly true along positive and negative samples, whereas a high negative value indicates that the feature is mainly false in the dataset – in both cases, regardless of the class label of samples. Fig. 8 graphically illustrates the semantics of a $\varphi - \delta$ diagram, when framed in a feature-oriented view. In particular, the upper corner (characterized by $\rho = \overline{\rho} = 1$) denotes a perfect agreement between the feature and the main class. Conversely, the lower corner (characterized by $\rho = \overline{\rho} = 0$) denotes a complete agreement between the feature and the alternate class. As for left and right corners, they point to the case in which the feature is always false (left corner) or true (right corner).

The borders that delimits the $\varphi - \delta$ space deserve a final remark, as they may carry important information. For example, let us concentrate on the point, in a $\varphi - \delta$ diagram, for which the feature at hand (say f) is in complete agreement with the main

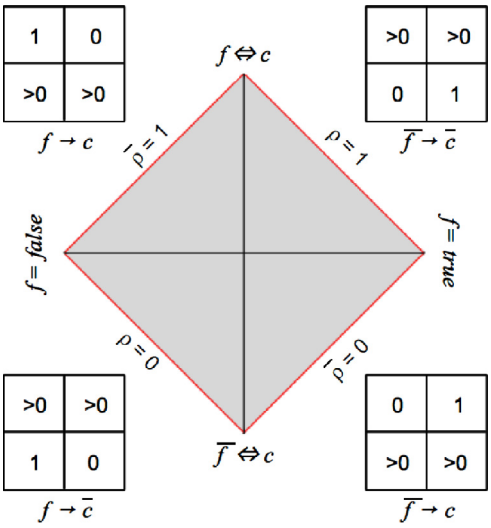


Fig. 9. Semantics of borders in the $\varphi - \delta$ space (feature-oriented view).

category. By definition, this point, coincident with the upper corner of a $\varphi - \delta$ diagram, has the highest discriminant capability and a null characteristic capability. Hence, as at the upper corner of the diamond both specificity and sensitivity reach their maximum value, we can state that f implies c and vice versa (in symbols: $f \Leftrightarrow c$). If we move our point down along the left border, one of the implications becomes progressively weaker, while the other remains unaltered. In particular, as any point located in the upper-left border has $\overline{\rho} = 1$ with varying ρ (see also Fig. 4), the implication $c \to f$ becomes progressively weaker, whereas the implication $f \to c$ maintains its strength.[5] Fig. 9 reports all implications that lay aside borders and includes, for the sake of readability, also the content of the corresponding normalized confusion matrices (for the sake of simplicity, f and $\overline{f}$ are used to denote the truth values of the feature at hand). In particular, looking at the upper left border, one may notice that the corresponding normalized confusion matrix is characterized by $\overline{\rho} = 1$ (i.e., $tn = 1$ and $fp = 0$). This means that no false positives exist for f according to the statistics evaluated on training data. Hence, having to classify a sample for which f is true, one can assert that the sample belongs to the main class, as no false positives have been observed on training data. Similar considerations can be made on the remaining borders.

As a consequence, the borders of a $\varphi - \delta$ diagram may give information about the (optional) presence of strict implications between feature values and main or alternate class. In particular, the diamond borders allow to assert the following implications: $f \to c$ (upper-left), $\overline{f} \to \overline{c}$ (upper-right), $f \to \overline{c}$ (lower-left), $\overline{f} \to c$ (lower-right). In all these cases, we can state that f is *supporting* the classification.

5. Experiments

Some experiments have been performed with the aim of assessing the potential of $\varphi - \delta$ diagrams. The underlying scenario is text categorization, with source documents taken from Reuters Corpus Vol. 1 [28] and from DMOZ (The Open Directory Project).[6] In these scenarios, we expect terms important for categorization to appear at the upper or lower corner of the $\varphi - \delta$ diamond, in correspondence with high values of $|\delta|$. Moreover, terms that occur barely on documents are expected to appear at the left hand corner (high negative values of φ), whereas the so-called *stop words* are expected to appear at the right hand corner (high values of φ). It is worth recalling that, during experiments, particular emphasis has been given to the identification of discriminant terms and stop words.

As for classifiers' performance, we expect that categories with a $\varphi - \delta$ "signature" in which several terms (at least one) have medium to high values for $|\delta|$ are easy to classify. The (unbiased) bias of the learned classifiers and their stability (i.e., their variance) are expected to depend also on the signature.

5.1. Experiments with datasets extracted from the Reuters Corpus Vol. 1

Categories C21 (production/services), C31 (markets/marketing), E11 (economic performance), and M11 (equity markets) have been selected from the Reuters Corpus. The corresponding alternate categories have been derived considering siblings (e.g., the alternate category for C21 is the union of C22, C23 and C24). Note that, in accordance with the Zipf's law [41], most of the terms

[5] Let us note, however, that this assumption becomes weaker while approaching the left corner side. In this case, the feature is so pervasively false on the given dataset that the implication $f \to c$ may not get enough support from training data.

[6] http://www.dmoz.org.

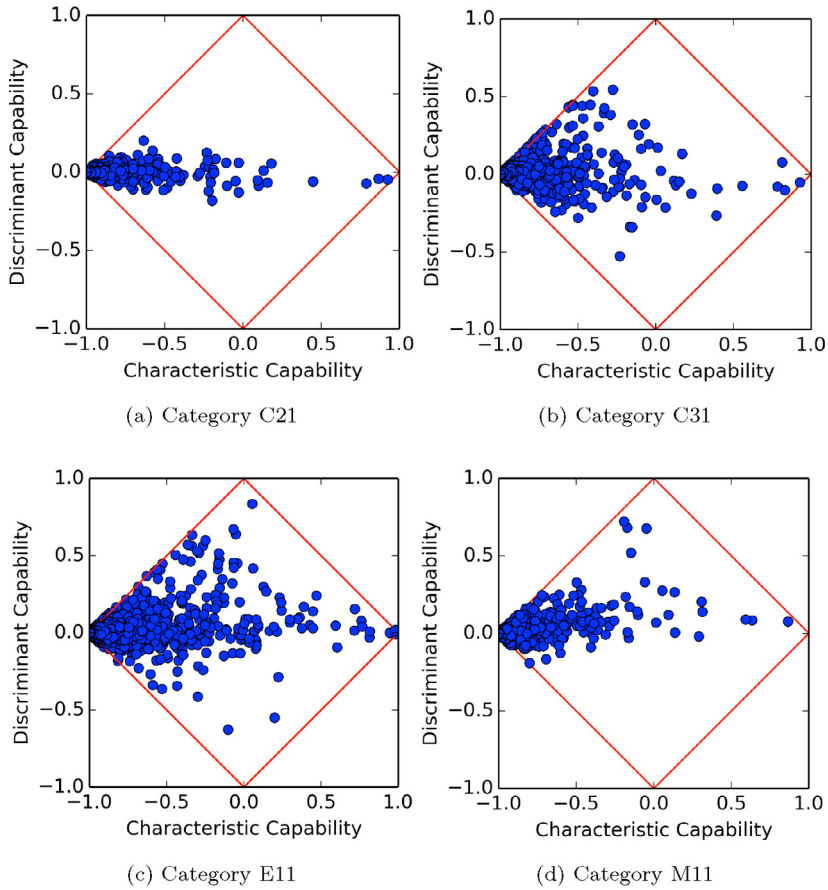


Fig. 10. Diagrams reporting the position of terms within $\varphi - \delta$ diagrams (i.e., the signatures) of Reuters categories C21, C31, E11 and M11.

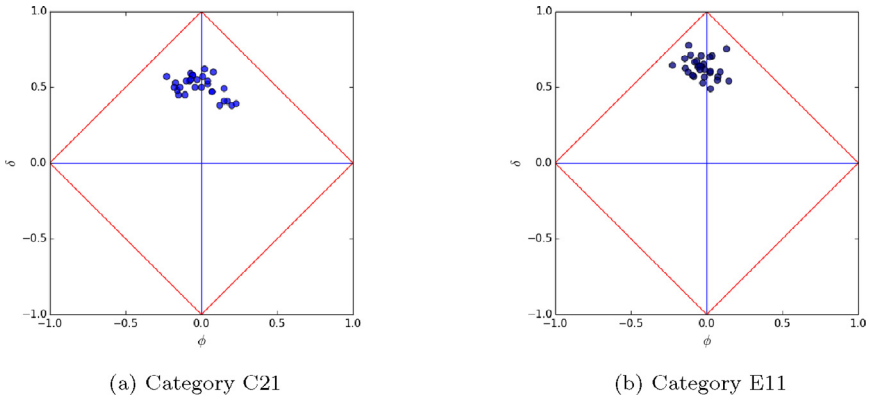


Fig. 11. Performance, in terms of $\varphi - \delta$ diagrams, of 30 naive Bayes classifiers trained on Reuters categories C21 and E11.

are located at the left hand corner of the constraining diamond. Fig. 10 plots the signatures obtained for the cited categories. Looking at the drawings, it appears that C21 is expected to be the most difficult category to predict, as no terms with a significant value of $|\delta|$ exist for it. On the contrary, samples of E11 appear to be relatively easy to classify, as several terms exist with $|\delta| \geq 0.5$ and also considering that many supporting terms (in particular, located at the left-upper corner) can help the classification process.

This conjecture is confirmed after training 30 naive Bayes classifiers using only terms t whose characteristic capability satisfies the constraint $|\varphi(t)| < 0.7$. For each class, train and test samples have been randomly extracted (with hold out strategy) at each run. Training data and test data have been kept balanced for the sake of simplicity. Fig. 11 reports the signatures of classifiers trained (a) on category C21 and its siblings and (b) on category E11 and its siblings. The figure clearly points out that, as expected,

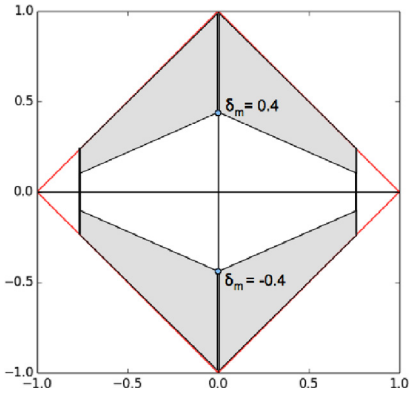


Fig. 12. Active areas (in gray) from which terms can be selected according to a constraint that enforces a trade-off between the need for selecting high values of $|\delta|$ and the need to identify supporting terms.

Table 3
Summary of experimental results concerning Reuters (the standard deviation gives information about the variance of results).

#Terms	φ_M	δ_m	Cat	Acc (%)	Spec	Sens	Bias	S.dev.
20,000	1.0	0.0	C21	73.6	0.716	0.756	0.040	0.151
			E11	80.5	0.818	0.793	−0.025	0.106
2000	0.9	0.1	C21	75.5	0.782	0.729	−0.053	0.012
			E11	81.5	0.834	0.796	−0.082	0.156
500	0.8	0.2	C21	72.6	0.762	0.691	−0.071	0.104
			E11	75.3	0.776	0.731	−0.044	0.100
100	0.7	0.4	C21	69.7	0.718	0.677	−0.041	0.107
			E11	71.2	0.741	0.684	−0.057	0.090

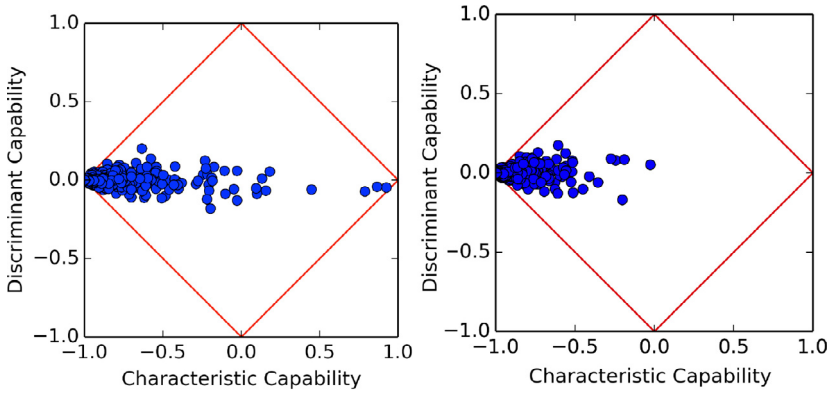
the average (unbiased) accuracy obtained on E11 is higher than the one obtained on C21. Besides, $\varphi - \delta$ diagrams point out that also variance and bias of classifiers trained for category C21 are apparently worse than those measured on classifiers trained for category E11.

To investigate to which extent $\varphi - \delta$ measures can support feature selection, the Reuters vocabulary to be used for training and testing categories C21 and E11 has been progressively reduced. The full dictionary consists of about 20,000 terms. In the first experiment it has been reduced of one order of magnitude, while in the second and the third experiments the set of terms has been reduced to about 500 and 100. At each trial, a preliminary selection has been performed according to the following criteria: $|\varphi| < \varphi_M$ and $|\varphi| + |\delta|/\delta_m \geq 1$, where φ_M and δ_m denote the *maximum* admissible value of φ and the *minimum* admissible value for δ, respectively.[7] It is worth pointing out that, among many other selection strategies that could be adopted, the above choice tries to find a trade-off between the need for selecting high values of $|\delta|$ and the need for identifying supporting terms which lay at (or in proximity of) the borders. Fig. 12 highlights (in gray) the constrained areas used for selecting about 100 terms. A further selection has been performed with the goal of limiting the number of terms to the amount decided for each experiment (namely, about 2000, 500, and 100). In this case, words have been ordered and selected according to their discriminant capability (i.e., $|\delta|$), in descending order –until the wanted number was reached.

Table 3 clearly shows that the average (unbiased) accuracy obtained while training classifiers with about 2000 terms is actually better than the one obtained with the full set of terms, whereas it slopes gently with the number of selected terms on the subsequent experiments.

A further investigation has been performed on stop words. Let us recall that the right hand side of each drawing is expected to highlight stop words. This specific aspect is shown in Fig. 13, in which the documents belonging to category C21 and to its siblings have been (a) left unchanged or (b) preprocessed with the goal of removing stop words. It is worth noting, though, that terms at the right hand corner do not necessarily represent only typical stop words (i.e. common articles, nouns, conjunctions, verbs and adverbs). Rather, also category-dependent stop words may be located in that area. In particular, limiting our attention to the first 10 terms ranked according to their φ value, only classical stop words occur for categories C21 and C31, whereas category-dependent stop words occur for categories E11 (i.e., "percent" and "year") and M11 (i.e., "market", "trade", and "percent").

[7] The selection criteria have been intentionally left "loose" to permit the selection of a not negligible number of terms also for category C21, whose "signature", as already pointed out, is particularly poor of discriminant terms.



(a) Category C21 with stop words (b) Category C21 without stop words

Fig. 13. Two $\varphi - \delta$ diagrams for Reuters category C21, whose documents (a) have been left unchanged or (b) preprocessed with the goal of removing stop words.

5.2. Experiments with datasets extracted from DMOZ

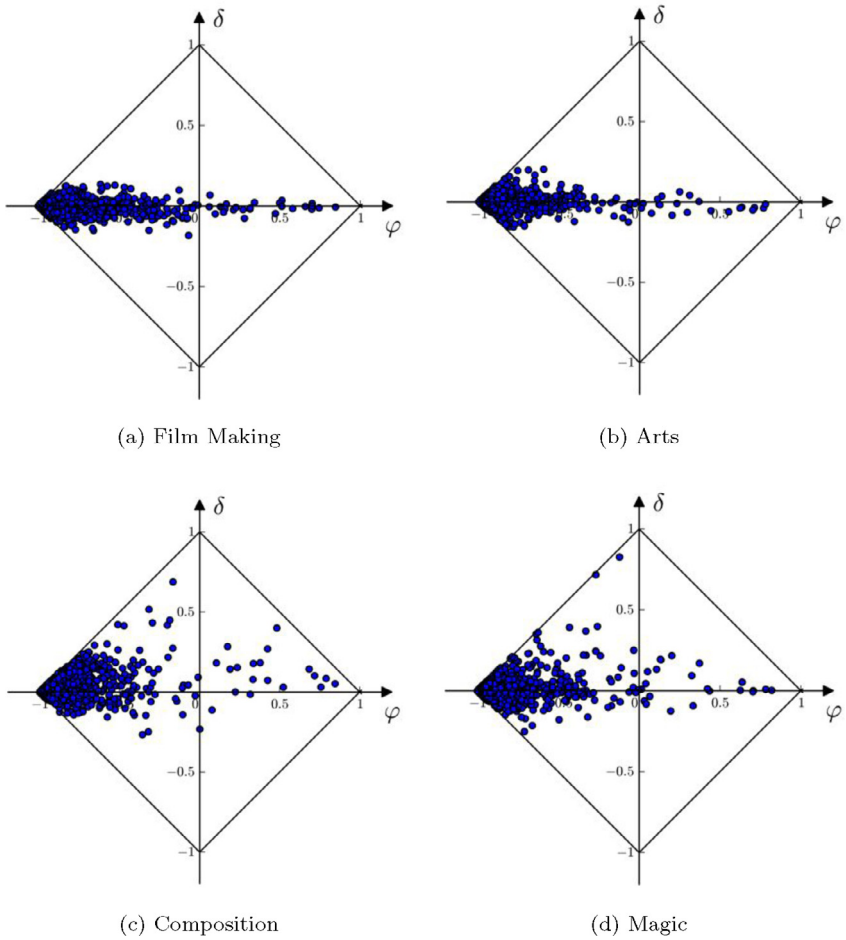
Further experiments have been performed on datasets extracted from DMOZ. In particular, the following categories have been taken into account:

- Filmmaking (from *Top:Arts:Movies:Filmmaking*),
- Arts (from *Top:Arts*),
- Composition (from *Top:Arts:Music:Composition*),
- Magic (from *Top:Arts:Performing Arts:Magic*).

As done for the Reuters' datasets, also in this case alternate categories have been obtained selecting negative samples from the siblings of the categories under analysis. Fig. 14 reports the signatures corresponding to *Film Making*, *Arts*, *Composition* and *Magic*, whereas Fig. 15 reports the performance obtained with 30 decision trees. In this case, training data have been intentionally kept unbalanced, with the goal of checking whether this choice would affect the bias of the classifiers. In particular, the ratio between negative and positive samples has been set to 10. It is worth noting that the Zip'f law is experimentally verified also in this case, as most of the terms occur at the left-hand side of the diagrams. In all cases, the signatures extracted from the categories under analysis are clearly correlated with the performances of the corresponding classifiers (the more words are scattered in the $\varphi - \delta$ space, the better). The figures point out that also the variance of classifiers is strongly correlated with the signatures of the selected categories. In fact, Film Making and Arts appear dramatically worse than Composition and Magic – both in terms of performance and of variance. As for the bias, it appears relatively immune from the quality of the signatures and from the ratio between negative and positive samples.

6. Strengths and weaknesses of this proposal

Apart from the analysis of existing measures and the definition of an unbiased *MCC*, the paper has been mainly concerned with the definition of two novel measures deemed useful in the task of developing and accessing machine learning and pattern recognition algorithms and systems. All in all, there is no magic in the given definitions. In fact, the $\varphi - \delta$ space is basically obtained by rotating the $fp - tp$ space of $\pi/4$. Although this is not a dramatic change of perspective, it is clear that the $\varphi - \delta$ space allows to analyze *at a glance* the most relevant properties of classifiers or features. In particular, the unbiased versions of accuracy (δ value) and bias (φ value) of a classifier are immediately visible on the $\varphi - \delta$ space. Maximum entropy and breakeven points are also clearly visible, as they correspond to the φ and to the δ axis, respectively. An estimate of the variance of a classifier can be easily assessed by just reporting the results of several experiments in the $\varphi - \delta$ space (see, for instance, Figs. 11 and 15, which clearly point out to which extent the performance of individual classifiers change along experiments). All the above measures are completely independent from the imbalance of data by construction, as the $\varphi - \delta$ space is defined on top of unbiased measures (i.e., ρ and $\overline{\rho}$). This aspect is very important for classifier assessment, making it easier to compare the performance obtained on different test data, regardless of the imbalance between negative and positive samples. Summarizing, the $\varphi - \delta$ space for classifiers can be actually thought of as a *bias* vs. *accuracy* (or *error*) space, whose primary uses can be: (i) assessing the accuracy of a classifier over a single or multiple runs, looking at its δ axis; (ii) assessing the bias of a classifier over a single or multiple runs, looking at the φ axis; (iii) assessing the variance of a classifier, looking at the scattering of multiple runs on the $\varphi - \delta$ space. As for binary features, an insight about the potential of $\varphi - \delta$ diagrams in the task of assessing their importance has been given in Section 5. In particular, let us recall that the most important features related to a given domain are expected to have high values of $|\delta|$, whereas not important ones are expected to have high values of $|\varphi|$. Also supporting features can be easily spotted, as they typically lay down along the borders.



(a) Film Making (b) Arts

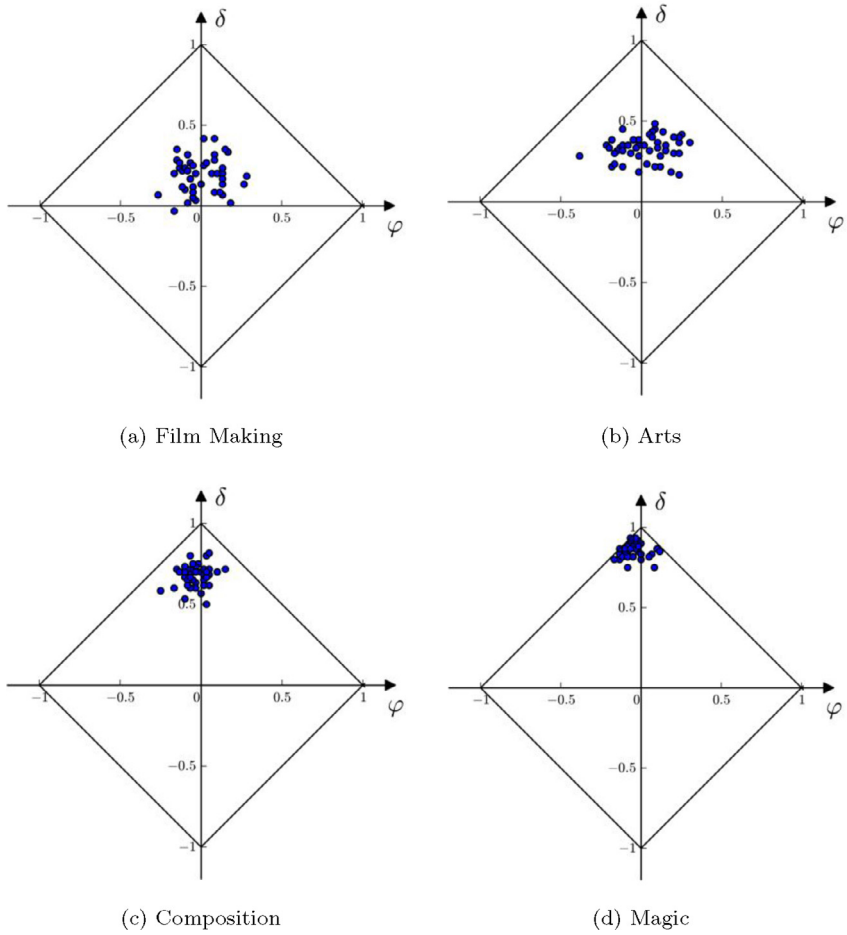
(c) Composition (d) Magic

Fig. 14. Diagrams reporting the position of terms within $\varphi - \delta$ diagrams (i.e., the signatures) of DMOZ categories Film Making, Arts, Composition and Magic (ordered according to the expected difficulty).

It is worth mentioning that alternative definitions could also be given in the $\varphi - \delta$ space for further relevant properties studied in other settings – e.g., performance or isometric curves in ROC diagrams, AUC, and Gini's coefficient. Although these aspects are beyond the scope of this paper, let us spend few words on ROC curves. It is easy to verify that when a classifier has a null discriminant capability the corresponding ROC curve reported to a $\varphi - \delta$ diagram would be constrained to the φ axis (which is also the locus of points with maximum entropy). On the other hand, the ROC curve of a classifier acting as an oracle would coincide with the borders of the surrounding diamond for which $\delta \geq 0$.

7. Conclusions and future work

After discussing and analyzing some issues related to the best known measures used in pattern recognition and machine learning, first the definition of an unbiased MCC has been given. Then, two novel measures have been proposed, i.e. δ and φ, intended to assess discriminant and characteristic capabilities of binary classifiers and binary features. The proposed measures are unbiased and are obtained as linear transformations of false and true positive rates. Moreover, the corresponding isometric curves show that they are orthogonal. The applications of $\varphi - \delta$ diagrams to pattern recognition and machine learning problems are manifold, ranging from feature ranking and feature selection to classifier performance assessment. Some experiments performed in text categorization settings confirm the usefulness of the proposal. As for future work, the properties of terms in scenarios such as hierarchical text categorization and taxonomy building will be investigated using δ and φ diagrams. An extension of δ and φ to multilabel categorization problems with multivalued features is also under study, together with the possibility of embedding the proposed measures in various algorithms, including those devised to deal with feature subset selection. A generalization of the $\varphi - \delta$ diagrams able to visualize the actual behavior of a classifier/set of features in the presence of an unbalanced data is under way.



(a) Film Making (b) Arts

(c) Composition (d) Magic

Fig. 15. Diagrams reporting the performance of classifiers with $\varphi - \delta$ diagrams of DMOZ categories Film Making, Arts, Composition and Magic (as expected, the performance of classifiers follow the ordering of signatures).

Acknowledgments

This work has been supported by LR7 2007 (Investment Funds for Basic Research and by PIA 2010 (Integrated Subsidized Packages – both funded by the local government of Sardinia. The author wishes to thank Lorenza Saitta for her support in discussing and developing the ideas reported in this paper. The author also wishes to thank Dario Deledda for having performed part of the experiments.

References

[1] A. Arauzo-Azofra, J.M. Benitez, J.L. Castro, Consistency measures for feature selection, J. Intell. Inf. Syst. 30 (3) (2011) 273–292.
[2] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (1994) 537–550.
[3] M. Ben-Bassat, An introduction to variable and feature selection, in: P. Krishnaiah, L. Kanal (Eds.), Handbook of Statistics, North-Holland, Amsterdam, 1982, pp. 773–791.
[4] A. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.
[5] A. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159.
[6] G.W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1950) 1–3.
[7] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.
[8] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (2003) 155–176.
[9] M. Dash, H. Liu, H. Motoda, Consistency based feature selection, in: Proceedings of the 4th Pacific Asia Conference on Knowledge Discovery and Data Mining, PAKDD, 2000, pp. 98–109.
[10] P. Domingos, A unified bias-variance decomposition for zero-one and squared loss, in: Proceedings of the 7th National Conference on Artificial Intelligence, AAAI, 2000.
[11] C. Drummond, R.C. Holte, Explicitly representing expected cost: an alternative to ROC representation, in: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, KDD, 2000, pp. 198–207.
[12] W. Duch, K. Grabczewski, K.G. abczewski, T. Winiarski, J. Biesiada, A. Kachel, Feature selection based on information theory, consistency and separability indices, in: Proceedings of the 9th Conference on Neural Information Processing, 2002.
[13] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.

[14] R.A. Fisher, Notes on regression and inheritance in the case of two parents, Proc. R. Soc. Lond. 58 (1895) 240–242.
[15] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[16] P.A. Flach, The geometry of ROC space: understanding machine learning metrics through ROC isometrics, in: Proceedingsof the 20th International Conference on Machine Learning, ICML, AAAI Press, 2003, pp. 194–201.
[17] P.A. Flach, J. Hernandez-Orallo, C. Ferri, A coherent interpretation of AUC as a measure of aggregated classification performance, in: Proceedings of the 28th International Conference on Machine Learning, ICML, Morgan Kaufmann, 2011, p. 657664.
[18] J. Friedman, On bias, variance, 0/1–loss and the curse of dimensionality, Data Min. Knowl. Discov. 1 (1997) 55–77.
[19] K. Fukunaga, Introduction to Statistical Pattern Classification, Academic Press, 1990.
[20] I.J. Good, Rational decisions, J. R. Stat. Soc. (Ser. B) 14 (1952) 107114.
[21] I. Guyon, A. Elisseeff, Pattern recognition and reduction of dimensionality, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[22] D.J. Hand, Construction and Assessment of Classification Rules, John Wiley and Sons Inc, 1997.
[23] D.R. Hardoon, S. Szedmak, J.S. Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.
[24] 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.
[25] G. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Proceedings of the 11th International Conference on Machine Learning, ICML, AAAI Press, 1996, pp. 121–129.
[26] C. Jutten, J. Herault, Blind separation of sources, Signal Process. 24 (1991) 1–10.
[27] D. Koller, M. Sahami, Toward optimal feature selection, in: Proceedings of the 13th International Conference on Machine Learning, ICML, AAAI Press, 1996, pp. 284–292.
[28] D. Lewis, Y. Yang, T. Rose, F. Li, RCV1: a new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397.
[29] H. Liu, R. Setiono, Feature selection and classificationa probabilistic wrapper approach, in: Proceedings of the 9th International Conference on Industrial and Engineering Applications of AI and ES, AAAI Press, 1996, pp. 419–424.
[30] Y. Ma, S. Lao, E. Takikawa, M. Kawade, Discriminant analysis in correlation similarity measure space, in: Z. Ghahramani (Ed.), Proceedings of the 24th Conference in Machine Learning, ICML, ACM International Conference Proceeding Series, 227, ACM, 2007, pp. 577–584.
[31] A.W. Moore, M. Lee, Efficient algorithms for minimizing cross validation error, in: Proceedings of the 11th International Conference on Machine Learning, ICML, Morgan Kaufmann, 1994, pp. 190–198.
[32] R.C. Prati, G.E.A.P.A. Batista, M.C. Monard, A survey on graphical methods for classification predictive performance evaluation, IEEE Trans. Knowl. Data Eng. 23 (2011) 1601–1618.
[33] Z. Qiao, L. Zhou, J.Z. Huang, Effective linear discriminant analysis for high dimensional, low sample size data, in: Proc. of the World Congress on Engineering, WCE, 2, 2008, pp. 1070–1075.
[34] V.V. Raghavan, G.S. Jung, P. Bollmann, A critical investigation of recall and precision as measures of retrieval system performance, ACM Trans. Inf. Syst. 7 (1989) 205–229.
[35] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524.
[36] K. Shin, X. Xu, Consistency-based feature selection, in: Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2009, pp. 342–350.
[37] M. Sugiyama, Local fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of the 23rd International Conference on Machine Learning, ICML, ACM, New York, NY, USA, 2006, pp. 905–912, doi:10.1145/1143844.1143958.
[38] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Stat. 35 (6) (2007) 2769–2794.
[39] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Stat. 32 (1) (2003) 56–85.
[40] Z. Zhao, H. Liu, Searching for interacting features, in: Proceedings of International Joint Conference on Artificial Intelligence, 2007, pp. 1156–1161.
[41] G. Zipf, Human Behavior and the Principle of Least Effort, Addison Wesley, 1949.