Measuring Discriminant and Characteristic Capability for Building and Assessing Classifiers

Giuliano Armano, Francesca Fanni and Alessandro Giuliani

Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi I09123, Cagliari, Italy
{armano,francesca.fanni,alessandro.giuliani}@diee.unica.it

Abstract. Performance metrics are used in various stages of the process aimed at solving a classification problem. Unfortunately, most of these metrics 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 most acknowledged metrics, novel unbiased metrics are defined, able to capture the concepts of discriminant and characteristic capability. The combined use of these metrics can give important information to researchers involved in machine learning or pattern recognition tasks, such as classifier performance assessment and feature selection.

1 Introduction

Several metrics are used in pattern recognition and machine learning in various tasks concerning classifier building and assessment. An important category of these metrics is related to confusion matrices. Accuracy, precision, sensitivity (also called recall) and specificity are all relevant examples [5] of metrics that belong to this category. As none of the above metrics is able to give information about the process under assessment in isolation, two different strategies have been adopted so far for assessing classifier performance or feature importance: i) devising single metrics on top of other ones and ii) identifying proper pairs of metrics able to capture the wanted information. The former strategy is exemplified by F_1 [6] and *MCC* (Matthews Correlation Coefficient) [4], which are commonly used in the process of model building and assessment. Typical members of the latter strategy are sensitivity vs. specificity diagrams, which allow to draw relevant information (e.g., ROC curves [1]) in a Cartesian space. Unfortunately, regardless from the strategies discussed above, most of the existing metrics 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 metrics 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 evaluation, the adoption of biased metrics does not appear a reliable choice. For this reason, in the literature, some proposals have been made to introduce unbiased metrics –see in particular the work of Flach [2]. In this paper a pair of unbiased metrics 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 metrics, their semantics is discussed for binary classifiers and binary features. An analysis focusing on the combined use of the corresponding metrics in form of Cartesian 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 most acknowledged metrics is performed, pointing out that most of them are in fact biased. Section 3 introduces novel metrics devised to measure the discriminant and characteristic capability of binary classifiers or binary features. Section 4 reports experiments aimed at pointing out the potential of Cartesian diagrams drawn using the proposed metrics. Section 5 highlights the strengths and weaknesses of this paper and Section 6 draws conclusions.

2 Background

As the concept of confusion matrix is central in this paper, let us preliminarily illustrate the notation adopted for its components (also because the adopted notation slightly differs from the most acknowledged one). When used for classifier assessment, the generic element ξ_{ij} of a confusion matrix Ξ accounts for the number of samples that satisfy the property specified by the subscripts. Limiting our attention to binary problems, in which samples are described by binary features, let us assume that 1 and 0 identify the presence and the absence of a property.

In particular, let us denote with $\Xi_c(P,N)$ the confusion matrix of a run in which a classifier $\widehat{c}$, trained on a category 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 $\Xi_c(P,N)$.

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. In symbols:

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

In so doing, we assume that the transformation performed by $\widehat{c}$ 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 $\widehat{c}$, on the other hand. In symbols:

$$\Xi_c(P,N) = M \cdot \underbrace{\begin{bmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{bmatrix}}_{\Omega(c)\approx p(X_c,\widehat{X}_c)} = M \cdot \underbrace{\begin{bmatrix} n & 0 \\ 0 & p \end{bmatrix}}_{\mathcal{O}(c)\approx p(X_c)} \cdot \underbrace{\begin{bmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{bmatrix}}_{\Gamma(c)\approx p(\widehat{X}_c|X_c)} \tag{2}$$

where:

- $\omega_{ij} \approx p(X_c = i\ ,\ \widehat{X}_c = j),\ i,j = 0,1$, denotes the joint occurrence of correct classifications ($i = j$) or misclassifications ($i \neq j$). According to the total probability law: $\sum_{ij} \omega_{ij} = 1$.
- p is the percent of positive samples and n is the percent of negative samples.
- $\gamma_{ij} \approx p(\widehat{X}_c = j \mid X_c = i),\ i,j = 0,1$, denotes the percent of inputs that have been correctly classified ($i = j$) or misclassified ($i \neq j$) by $\widehat{X}_c$. $\gamma_{00}, \gamma_{01}, \gamma_{10}$, and γ_{11} respectively denote the *rate* of true negatives, false positives, false negatives, and true positives. According to the total probability law: $\gamma_{00} + \gamma_{01} = \gamma_{10} + \gamma_{11} = 1$. An estimate of the conditional probability $p(\widehat{X}_c|X_c)$ for a classifier $\widehat{c}$ that accounts for a category c will be called *normalized confusion matrix* hereinafter.

The separation between inputs and the intrinsic behavior of a classifier reported in Equation (2) suggests an interpretation that recalls the concept of transfer function, where a set of inputs is applied to $\widehat{c}$. In fact, Equation (2) highlights the separation of the optimal behavior of a classifier from the deterioration introduced by its actual filtering capabilities. In particular, $\mathscr{O} \approx p(X_c)$ 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.

A different interpretation holds for confusion matrix subscripts when they are used to investigate binary features. In this case i still denotes the actual category, 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 though of as a very simple classifier whose classification output reflects the truth value of the feature in the given samples, all definitions and comments concerning classifiers can be applied to binary features as well.

Let us now examine the most acknowledged metrics deemed useful for pattern recognition and machine learning according to the above perspective. The classical definitions for accuracy (a), precision (π), and recall (ρ) can be given in terms of false positives rate (fp), true positives rate (tp) and class ratio (the imbalance between negative and positive samples, σ) as follows:

$$\begin{aligned} a &= \frac{trace(\Omega)}{|\Omega|} = \frac{\omega_{00} + \omega_{11}}{1} = \frac{\sigma \cdot (1 - \gamma_{01}) + \gamma_{11}}{\sigma + 1} = \frac{\sigma \cdot (1 - fp) + tp}{\sigma + 1} \\ \pi &= \frac{\omega_{01}}{\omega_{01} + \omega_{11}} = \left(1 + \sigma \cdot \frac{\gamma_{01}}{\gamma_{11}}\right)^{-1} = \left(1 + \sigma \cdot \frac{fp}{tp}\right)^{-1} \\ \rho &= \frac{\omega_{11}}{\omega_{11} + \omega_{10}} = \gamma_{11} = tp \end{aligned} \tag{3}$$

Equation (3) 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 pointed out, when the goal is to assess the intrinsic properties of a classifier or a feature, biased metrics do not appear a proper choice, leaving room for alternative definitions aimed at dealing with the imbalance between negative and positive samples.

In [2], Flach gave definitions of some unbiased metrics starting from classical ones. In practice, unbiased metrics can be obtained from classical ones by setting the imbalance σ to 1. In the following, if needed, unbiased metrics will be denoted using the subscript u.

3 Definition of Novel Metrics

To our 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 metrics intended to measure them. Without loss of generality, let us assume the metrics 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 metric 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 category. Conversely, the metric is expected to be close to -1 when most of the samples are clustered as belonging to the alternate category.[1] An immediate consequence of the desired behavior is that the above properties are not independent. In other words, regardless from their definition, the metrics devised to measure discriminant and characteristic capability 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 metric is about 1 the other should be close to 0.

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 $tp \approx 1$ – We expect $\delta \approx +1$ and $\varphi \approx 0$, meaning that the classifier is able to partition the samples almost in complete accordance with the class labels.
- $fp \approx 1$ and $tp \approx 1$ – We expect $\delta \approx 0$ and $\varphi \approx +1$, meaning that almost all samples are recognized as belonging to the main class label.
- $fp \approx 0$ and $tp \approx 0$ – We expect $\delta \approx 0$ and $\varphi \approx -1$, meaning that almost all samples are recognized as belonging to the alternate class label.
- $fp \approx 1$ and $tp \approx 0$ – We expect $\delta \approx -1$ and $\varphi \approx 0$, meaning that the classifier is able to partition 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).

The determinant of the normalized confusion matrix is the starting point for giving proper definitions of δ and φ able to satisfy the constraints and boundary conditions

[1] 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 only focuses 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.

discussed above. It can be rewritten as follows:

$$\begin{aligned}\Delta &= \gamma_{00}\cdot\gamma_{11} - \gamma_{01}\cdot\gamma_{10} = \gamma_{00}\cdot\gamma_{11} - (1-\gamma_{00})\cdot(1-\gamma_{11}) \\ &= \gamma_{00}\cdot\gamma_{11} - 1 + \gamma_{11} + \gamma_{00} - \gamma_{00}\cdot\gamma_{11} = \gamma_{11} + \gamma_{00} - 1 \\ &= \rho + \overline{\rho} - 1 \equiv tp - fp\end{aligned} \tag{4}$$

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 to* Δ, as the latter has all the desired properties required by the discriminant capability metric.

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

$$\varphi = \rho - \overline{\rho} = tp + fp - 1 \tag{5}$$

Figure 1 reports the isometric curves drawn for different values of δ and φ, respectively, with varying tp and fp.

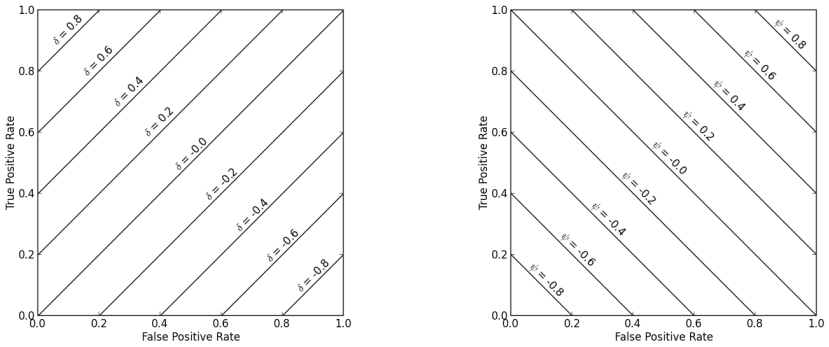

Fig. 1: Isometric plotting of δ and φ with varying false and true positive rate.

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. 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:

$$\delta = -\varphi + (2\cdot tp - 1) = +\varphi + (2\cdot fp + 1) \tag{6}$$

Considering fp and tp as parameters, we can easily draw the corresponding isometric curves in the $\varphi - \delta$ space. Figure 2 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.

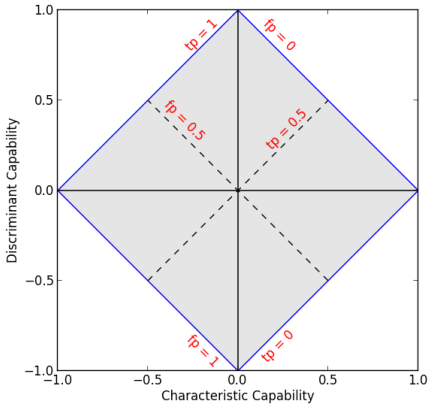

Fig. 2: Shape of the $\varphi - \delta$ space: the rhombus centered in (0,0) delimits the area of admissible value pairs.

Semantics of the $\varphi - \delta$ space for classifiers. As for binary classifiers, their discriminant capability 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:

$$a_u = \frac{tn + tp}{2} = \frac{1 + \delta}{2} = 1 - \frac{1 - \delta}{2} = 1 - \frac{fp + fn}{2} = 1 - e_u \tag{7}$$

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 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, we can write:

$$\begin{aligned} E[X_c] &\approx \frac{1}{M} \cdot (P-N) = (p-n) \\ E[\widehat{X}_c] &\approx \frac{1}{M} \cdot \left(\widehat{P}-\widehat{N}\right) = (p-n) + 2 \cdot n \cdot fp - 2 \cdot p \cdot fn \end{aligned} \tag{8}$$

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

$$E[X_c - \widehat{X}_c] = E[X_c] - E[\widehat{X}_c] \approx -2 \cdot n \cdot fp + 2 \cdot p \cdot fn \tag{9}$$

According to Friedman [3], it is easy to show that Equation (9) actually represents an estimate of the *bias* of a classifier, measured over the confusion matrix that describes the outcomes of the experiments performed on the test set(s). Summarizing, in a $\varphi - \delta$ diagram used for assessing classifiers, the δ-axis and the φ-axis represent the unbiased accuracy and the unbiased bias, respectively. It is worth pointing out that a high positive value of δ means that the classifier at hand approximates the behavior of an *oracle*, whereas a high negative value approximates the behavior of a classifier that is almost always wrong (say *anti-oracle* when $\delta = -1$). Conversely, a high positive value of φ denotes a *dummy classifier* that almost always consider input items as belonging to the main category, whereas a high negative value denotes a *dummy classifier* that almost always consider input items as belonging to the alternate category.

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 ($\delta \simeq +1$) or in discordance ($\delta \simeq -1$) with the main class label. In either case, the feature has high discriminant capability. 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. On the other hand, φ measures to which extent the feature at hand is spread over 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 –regardless of the class label of samples.

4 Experiments

Some experiments have been performed with the aim of assessing the potential of $\varphi - \delta$ diagrams. In our experiments we use a collection in which each document is a webpage. The dataset is extracted from the DMOZ taxonomy[2]. Let us recall that DMOZ is the collection of HTML documents referenced in a Web directory developed in the Open Directory Project (ODP). We choose a set of 174 categories containing about 20000 documents, organized in 36 domains.

In this scenario, we expect terms important for categorization appear at the upper or lower corner of the $\varphi - \delta$ rhombus, in correspondence with high values of $|\delta|$. As

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

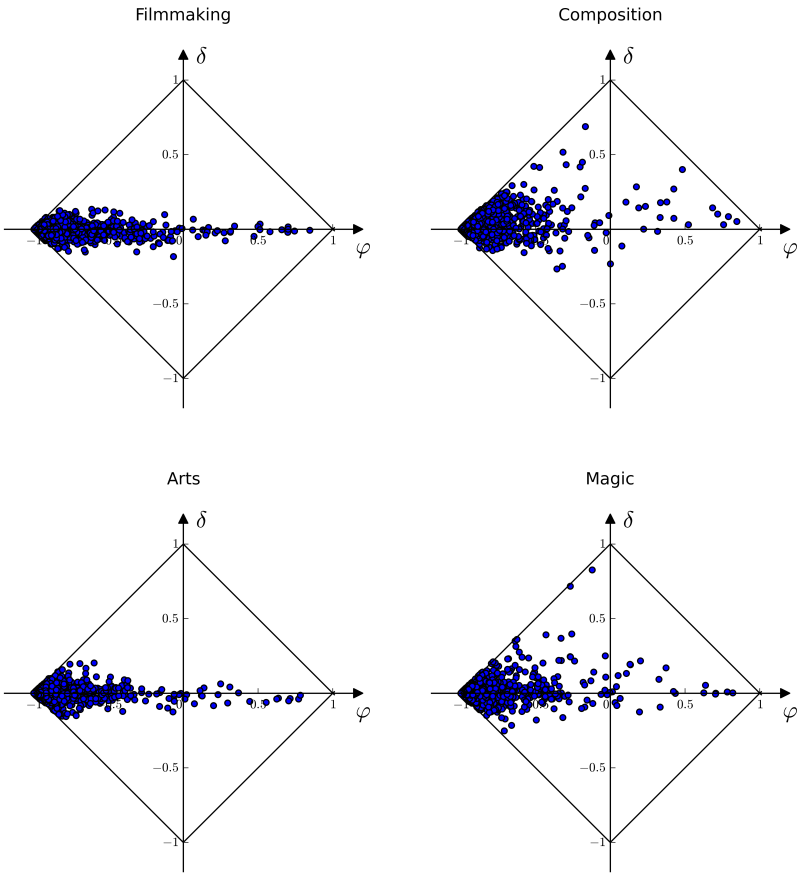

Fig. 3: Position of terms within $\varphi - \delta$ diagrams for the selected DMOZ's categories.

for the characteristic capability, terms that occur barely on documents are expected to appear at the left hand corner (high negative values of φ), while the so-called *stopwords* are expected to appear at the right hand corner (high values of φ).

Experiments have been focusing on the identification of discriminant terms and stopwords. Figure 3 plots the "signatures" obtained for DMOZ's categories *Filmmaking, Composition, Arts,* and *Magic*. Alternate categories have been derived considering the corresponding siblings. Note that, in accordance with the Zipf's law [7], most of the words are located at the left hand corner of the constraining rhombus. Looking at the drawings, it appears that *Filmmaking* and *Arts* are expected to be the most difficult categories to predict, as no terms with a significant value of $|\delta|$ exist for it. On the contrary, documents of *Composition* and *Magic* appear to be relatively easy to classify, as several terms exist with significant discriminant value. This conjecture is confirmed after training 50 decision trees using only terms t whose characteristic capability satisfies the

constraint $|\varphi(t)| < 0.4$. For each category, test samples have been randomly extracted at each run, whereas the remainder of the samples trained the classifiers.

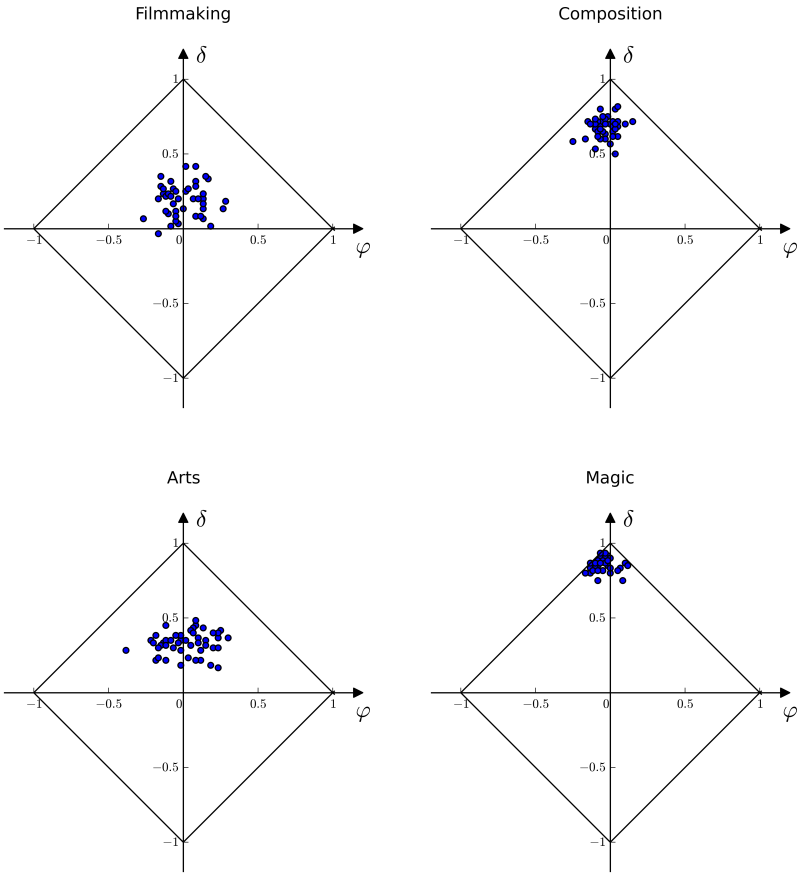

Fig. 4: Four diagrams reporting the classification results.

Figure 4 reports the signatures of classifiers. The figure clearly points out that, as expected, the average (unbiased) accuracies obtained on categories *Composition* and *Magic* are higher than the ones obtained on categories *Filmmaking* and *Arts*. Besides, $\varphi - \delta$ diagrams point out that also variance and bias of classifiers trained for categories *Filmmaking* and *Arts* are apparently worse than those measured on classifiers trained for categories *Composition* and *Magic*.

5 Strengths and Weaknesses of This Proposal

Apart from the analysis of existing metrics, the paper has been mainly concerned with the definition of two novel metrics deemed useful in the task of developing and assessing 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) accuracy and the (unbiased) bias of a classifier are immediately visible on the vertical and horizontal axis of a $\varphi - \delta$ space, respectively. Moreover, an estimate of the variance of a classifier can be easily investigated by just reporting the results of several experiments in the $\varphi - \delta$ space (see, for instance, Figure 4, which clearly points 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 metrics (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 from 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 4. 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 $|\varphi|$. Moreover, in the special case of text categorization, stopwords are expected to occur at the right hand corner of the rhombus that constrains the $\varphi - \delta$ space.

It is worth mentioning that alternative definitions could also be given in the $\varphi - \delta$ space for other relevant properties, e.g., ROC curves and AUC (or 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 random guessing for a classifier would constrain the ROC curve to the φ axis, whereas the ROC curve of a classifier acting as an oracle would coincide with the positive border of the surrounding rhombus.

6 Conclusions and Future Work

After discussing and analyzing some issues related to the most acknowledged metrics used in pattern recognition and machine learning, two novel metrics have been proposed, i.e. δ and φ, intended to measure discriminant and characteristic capability for binary classifiers and binary features. They 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 selection

to classifier performance assessment. Some experiments performed in a text categorization setting confirm the usefulness of the proposal. As for future work, the properties of terms in a scenario of hierarchical text categorization will be investigated using δ and φ diagrams. A generalization of δ and φ to multilabel categorization problems with multivalued features is also under study.

Acknowledgments. This work has been supported by LR7 2009 - Investment funds for basic research (funded by the local government of Sardinia).

References

1. Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, 30(7):1145–1159, July 1997.
2. Peter A. Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *in Proceedings of the Twentieth International Conference on Machine Learning*, pages 194–201. AAAI Press, 2003.
3. Jerome H. Friedman and Usama Fayyad. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.
4. B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.
5. Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229, July 1989.
6. C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.
7. George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.