F-Measure Curves for Visualizing Classifier Performance with Imbalanced Data

Roghayeh Soleymani[1], Eric Granger[1], and Giorgio Fumera[2](✉)

[1] Laboratoire d'imagerie, de vision et d'intelligence artificielle,
École de technologie supérieure, Université du Québec, Montreal, Canada
`rSoleymani@livia.etsmtl.ca`, `Eric.Granger@etsmtl.ca`
[2] Pattern Recognition and Applications Lab,
Department of Electrical and Electronic Engineering, University of Cagliari,
Cagliari, Italy
`fumera@diee.unica.it`

Abstract. Training classifiers using imbalanced data is a challenging problem in many real-world recognition applications due in part to the bias in performance that occur for: (1) classifiers that are often optimized and compared using unsuitable performance measurements for imbalance problems; (2) classifiers that are trained and tested on a fixed imbalance level of data, which may differ from operational scenarios; (3) cases where the preference of correct classification of classes is application dependent. Specialized performance evaluation metrics and tools are needed for problems that involve class imbalance, including scalar metrics that assume a given operating condition (skew level and relative preference of classes), and global evaluation curves or metrics that consider a range of operating conditions. We propose a global evaluation space for the scalar F-measure metric that is analogous to the cost curves for expected cost. In this space, a classifier is represented as a curve that shows its performance over all of its decision thresholds and a range of imbalance levels for the desired preference of true positive rate to precision. Experiments with synthetic data show the benefits of evaluating and comparing classifiers under different operating conditions in the proposed F-measure space over ROC, precision-recall, and cost spaces.

Keywords: Class imbalance · Performance visualization tools
F-measure

1 Introduction

Evaluating performance is a critical step in classifier design and comparison. Classification accuracy is the most widely used performance metric, also used as the objective function of many state-of-the-art learning algorithms (e.g., support

L. Pancioni et al. (Eds.): ANNPR 2018, LNAI 11081, pp. 165–177, 2018.
https://doi.org/10.1007/978-3-319-99978-4_13

vector machines). However, when data from different classes are imbalanced, it favours the correct classification of the majority classes at the expense of high misclassification rates for the minority ones. This is an issue in many detection problems where samples of the class of interest ("positive" or "target" class) are heavily outnumbered by those of other ("negative" or "non-target") classes. The widely used ROC curve (which plots the true positive rate vs the false positive rate for two-class classification problems), is not suitable for imbalanced data either, since it is independent of the level of imbalance. The alternative Precision-Recall (PR) curve is more suitable than ROC space, since precision is sensitive to imbalance; however, the performance of a given classifier under different imbalance levels corresponds to different PR curves, which makes it difficult to evaluate and compare classifiers.

Alternatively, scalar performance metrics like the expected cost (EC) and the F-measure (widely used in information retrieval) are typically employed when data is imbalanced. Since they seek different trade-offs between positive and negative samples, the choice between them is application-dependent. EC allows to indirectly address class imbalance by assigning different misclassification costs to positive and negative samples. Two graphical techniques have recently been proposed to easily visualize and compare classifier performance in terms of EC under all possible operating conditions: cost curves (CC) [3] and Brier curves (BC) [5]. The F-measure, recently analyzed by many researchers [2,12–14] is defined as the weighted harmonic mean of precision and recall, and thus evaluates classifier performance using a weight that controls the relative importance of recall (i.e., the true positive rate) and precision, which is sensitive to class imbalance. However, no performance visualization tool analogous to CC or BC exists for the F-measure. One may use the PR space to this aim, but the isometrics of the F-measure in PR space are hyperbolic [7,9], which does not allow to easily evaluate classifiers under diverse operating conditions.

This paper introduces F-measure curves, a global visualization tool for the F-measure analogous to CC. It consists in plotting the F-measure of a given classifier versus two parameters – the level of imbalance and the preference between recall and precision – and allows to visualize and compare classifier performance in class imbalance problems for different decision thresholds, under different operating conditions. In this space, a crisp classifier corresponds to a curve that shows its F-measure over all possible imbalance levels, for a desired level of preference between recall and precision. A soft classifier corresponds to the upper envelope of such curves for all possible decision thresholds. This space allows to compare classifiers more easily than in the PR space for a given operating condition, analogously to CC or BC vs the ROC space. For a given preference level between precision and recall, one classifier may outperform another over all skew levels, or only for a specific range, which can be determined both analytically and empirically in the proposed space, as with the CC space. To clarify the benefits of the proposed space, experiments are performed on synthetic data.

2 Performance Metrics and Visualization Tools

In many real-world applications, the distribution of data is imbalanced [10]; correctly recognizing positive samples is the main requirement, while avoiding excessive misclassification of negative samples can also be important. If application requirements are given by misclassification costs, misclassification of positive samples usually exhibits a higher cost, which "indirectly" addresses class imbalance. Otherwise, assigning different "fictitious" costs to misclassifications of positive and negative samples can be an indirect means to achieve the same goal. Several performance metrics have been proposed so far for applications involving imbalanced classes [1,6,8,11,15]. This section provides a review of these metrics in terms of their sensitivity to imbalance, focusing on global spaces that consider different operating conditions and preference weights.

Scalar Performance Metrics. We focus on two-class problems, although some metrics can also be applied in multi-class cases. Let $P(+)$ and $P(-)$ be the prior probability of the positive and negative class, and $\lambda = {}^{P(-)}/_{P(+)}$ the class skew. From a given data set with n_+ positive and n_- negative samples, $P(+)$ can be estimated as ${}^{n_+}/_{(n_+ + n_-)}$, and similarly for $P(-)$, whereas λ can be estimated as ${}^{n_-}/_{n_+}$. As in [3], we focus on evaluating classifier performance as a function of the prior of the positive class when the classifier is deployed, which can be different than in the training and testing sets; accordingly, from now on we will use $P(+)$ (and $P(-)$) to denote the class prior during classifier deployment (use). Since this value is unknown during classifier design, we will evaluate classifier performance across all possible $P(+)$ values.

Classifier performance on a given data set can be summarized by its confusion matrix, in terms of the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) counts. Let N_+ and N_- the number of samples classified as positive and negative, respectively. The corresponding rates are defined as $TPR = TP/n_+$, $FNR = FN/n_+$, $TNR = TN/n_-$ and $FPR = FP/n_-$.

Several scalar metrics can be defined from the above rates. The widely used error rate, defined as $(FP + FN)/(n_+ + n_-)$, is biased towards the correct classification of the negative (majority) class, which is not suitable to imbalanced data. When costs can be associated to classification outcomes (either correct or incorrect), the expected cost (EC) is used; denoting as C_{FN} and C_{FP} the misclassification costs of positive and negative samples (usually the cost of correct classifications is zero), EC is defined as:

$$EC = \mathrm{FNR} \cdot P(+) \cdot C_{\mathrm{FN}} + \mathrm{FPR} \cdot P(-) \cdot C_{\mathrm{FP}} \tag{1}$$

When data is imbalanced, usually $C_{\mathrm{FN}} > C_{\mathrm{FP}}$, which can also avoid the bias of the error probability toward the negative class. Accordingly, by setting suitable fictitious costs, EC can also be used to deal with class imbalance even if misclassification costs are not precisely known or difficult to define. However, as $C_{\mathrm{FN}}/C_{\mathrm{FP}}$ increases, minimizing EC increases TPR at the expense of increasing FPR, which may be undesirable.

In information retrieval applications the complementary metrics Precision (Pr) and Recall (Re) are often used, instead: Re corresponds to TPR, whereas Pr is defined as $TP/(TP + FP)$ or TP/N_+. Pr depends on both TP and FP, and drops severely when correct classification of positive class is attained at the expense of a high fraction of misclassified negative samples, as can be seen by rewriting Pr as:

$$Pr = \frac{\frac{TP}{n_+}}{\frac{TP}{n_+} + \frac{FP}{n_+} \times \frac{n_-}{n_-}} = \frac{TPR}{TPR + \lambda FPR}. \tag{2}$$

This is useful to reveal the effect of class imbalance, compared to EC.

Pr and Re can be combined into the F-measure scalar metric [16], defined as their weighted harmonic mean:

$$F_\alpha = \frac{1}{\alpha \frac{1}{\mathrm{Pr}} + (1 - \alpha)\frac{1}{\mathrm{Re}}}, \tag{3}$$

where $0 < \alpha < 1$. By rewriting α as $(1 + \beta^2)^{-1}$, $\beta \in [0, +\infty)$, F_α can be rewritten as:

$$F_\beta = \frac{(1 + \beta^2)\mathrm{Pr} \cdot \mathrm{Re}}{\beta^2 \mathrm{Pr} + \mathrm{Re}} = \frac{(1 + \beta^2)\mathrm{TP}}{(1 + \beta^2)\mathrm{TP} + \mathrm{FP} + \beta^2 \mathrm{FN}}. \tag{4}$$

When $\alpha \to 0$, $F_\alpha \to \mathrm{Re}$, and when $\alpha \to 1$, $F_\alpha \to \mathrm{Pr}$. Note that the sensitivity of the F-measure to the positive and negative classes can be adjusted by tuning α. This measure can be preferable to EC for imbalanced data, since it weighs the relative importance of TPR (i.e., Re) and Pr, rather than TPR and FPR.

Other metrics have been used, or specifically proposed, for class imbalance problems, although they are currently less used than EC and the F-measure [6,8].

Global Evaluation Curves. In many applications it is desirable for the classifier to perform well over a wide range of operating conditions, i.e., the misclassification costs or the relative importance between Pr and Re, and the class priors. Global curves depict the trade-offs between different evaluation metrics under different operating conditions, without reducing them to an incomplete scalar measure.

The ROC curve is widely used for two-class classifiers: it plots TPR vs FPR as a function of the decision threshold. A classifier with a specific threshold corresponds to a point in ROC space; a potentially optimal classifier lies on the ROC convex hull (ROCCH) of the available points, regardless of operating conditions. The best thresholds correspond to the upper-left point, corresponding to the higher TPR and the lower FPR (see Fig. 4(a)). A drawback of the ROC space is that it does not reflect the impact of imbalance, since TPR and FPR do not depend on class priors [4]. The performance of a classifier for a given skew level can be indirectly estimated in terms of EC, since in ROC space, each operating condition corresponds to a set of isoperformance lines with identical slope. An optimal classifier for a given operating condition is found by intersecting the ROCCH with the upper-left isoperformance line.

When Pr and Re are used, their trade-off across different decision thresholds can be evaluated by the precision-recall (PR) curve, which plots Pr vs Re. The PR curve is sensitive to class imbalance, given its dependence on Pr. However, different operating conditions (skew levels) lead to different PR curves, which makes classifier comparison difficult. Moreover, differently from ROC space, the convex hull of a set of points in PR space has no clear meaning [7]. If the F-measure is used, its isometrics can be analytically obtained in PR space, analogously to EC isometrics in ROC space; however they are hyperbolic [7,9], which makes it difficult to visualize classifier performance over a range of decision thresholds, skew levels, and preference of Pr to Re. In the case of EC this problem has been addressed by the CC visualization tool, described below, and by its BC extension. Inspired by CC, we propose in Sect. 3 an analogous visualization tool for the F-measure.

Expected Costs Visualization Tools. CCs [3] are used to visualize EC over a range of misclassification costs and skew levels. More precisely, CCs visualize the normalised EC (NEC), which is defined as EC divided by the maximum possible value of EC; the latter value turns out to be $P(+)C_{\mathrm{FN}} + P(-)C_{\mathrm{FP}}$, and NEC can be written as:

$$NEC = (\mathrm{FNR} - \mathrm{FPR})PC(+) + \mathrm{FPR} \in [0, 1], \tag{5}$$

where PC(+) is the "probability times cost" normalization term, which is defined as:

$$PC(+) = \frac{P(+) \cdot C_{\mathrm{FN}}}{P(+) \cdot C_{\mathrm{FN}} + P(-)C_{\mathrm{FP}}} \in [0, 1]. \tag{6}$$

CCs are obtained by depicting NEC versus $PC(+)$ on a $[0, 1] \times [0, 1]$ plot, which is named "cost space". Note that $NEC = \mathrm{FPR}$, if $PC(+) = 0$, and $NEC = \mathrm{FNR} = 1 - \mathrm{TPR}$, if $PC(+) = 1$. The always positive and always negative classifiers correspond to two lines connecting points (1,0) to (0,1), and (0,0) to (1,1), respectively, in the cost space. The operating range of a classifier is the set of operating points for which it dominates both these lines [3]. By defining:

$$m = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}} + C_{\mathrm{FN}}}, \text{ where } 0 < m \leq 1 \tag{7}$$

m can be seens as weighing the importance of both classes, and Eq. (6) can be rewritten as:

$$PC(+) = \frac{(1/m - 1) \cdot P(+)}{(1/m - 2) \cdot P(+) + 1} \tag{8}$$

The CCs of two classifiers C_i and C_j may cross: in this case each classifier outperforms the other for a certain range of operating points.

Interestingly, there is a point-line duality between CC and ROC space: a point in ROC space is a line in cost space, and vice versa. The lower envelope of cost lines corresponds to the ROCCH in ROC space. In cost space quantitatively evaluating classifier performance for given operating conditions does not require geometric constructions as in ROC space, but only a quick visual inspection [3]. This helps users to easily compare classifiers to the trivial classifiers, to select between them, or to measure their difference in performance [3].

BCs [5] are a variant of CCs – they visualize classifier performance assuming that the classifier scores are estimates of the posterior class probabilities, without requiring optimal decision threshold for a given operating condition.

No performance visualization tools analogous to CCs or BCs exist for the F-measure: defining and investigating such a space is the subject of the next section.

3 The F-Measure Space

We propose a visualization tool analogous to CC for evaluating and comparing the F-measure of one or more classifiers under different operating conditions, i.e., class priors and α. To this aim we rewrite the F-measure from Eq. (3) to make the dependence on $P(+)$ and α explicit:

$$F_\alpha = \frac{\mathrm{TPR}}{\alpha(\mathrm{TPR} + \lambda \cdot \mathrm{FPR}) + (1-\alpha)} \tag{9}$$

$$= \frac{{}^1\!/_\alpha \mathrm{TPR}}{{}^1\!/_\alpha + {}^1\!/_{P(+)}\mathrm{FPR} + \mathrm{TPR} - \mathrm{FPR} - 1} \tag{10}$$

In contrast to the EC of Eqs. (1) and (10) indicates that F_α cannot be written as a function of a single parameter. However, since our main focus is performance evaluation under class imbalance, we consider the F-measure as a function of $P(+)$ only, for a fixed α value. Accordingly, we define the F-measure curve of a classifier as the plot of F_α as a function of $P(+)$, for a given α.

F-Measure Curve of a Classifier. For a crisp classifier defined by given values of TPR and FPR, the F-measure curve is obtained by simply plotting F_α as a function of $P(+)$, for a given α, using Eq. (10). Equation (10) implies that, when $P(+) = 0$, $F_\alpha = 0$, and when $P(+) = 1$, $F_\alpha = \mathrm{TPR}/(\alpha(\mathrm{TPR} - 1) + 1)$. It is easy to see that, when TPR > FPR (which is always the case for a non-trivial classifier), F_α is an increasing and concave function of $P(+)$. For different values of α one gets a family of curves. For $\alpha = 0$ we have $F_\alpha = \mathrm{TPR}$, and for $\alpha = 1$ we have $F_\alpha = \mathrm{Pr}$. Thus, for any fixed $\alpha \in (0, 1)$, each curve starts at $F_\alpha = 0$ for $P(+) = 0$, and ends in $F_\alpha = \mathrm{Pr}$ for $P(+) = 1$. By computing $\mathrm{d}F_\alpha/\mathrm{d}\alpha$ from Eq. (10), one also obtains that all curves (including the one for $\alpha = 0$) cross when $P(+) = FPR/(FPR - TPR + 1)$. Figure 1 shows an example for a classifier with $\mathrm{TPR} = 0.8$ and $\mathrm{FPR} = 0.15$, and for five α values. CCs are also shown for comparison.

Consider now changing the decision threshold for a given soft classifier and a given α value. Whereas a point in ROC space corresponds to a line in cost space, it corresponds to a (non-linear) curve in F-measure space. As the decision

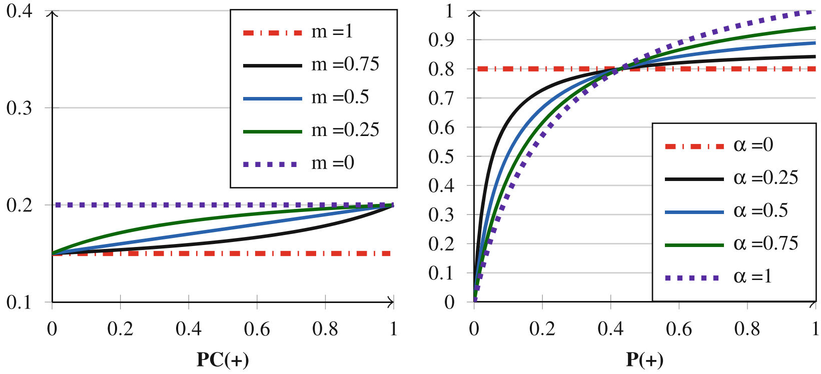


Fig. 1. Cost curves (left) and F-measure curves (right) for a given classifier with TPR = 0.8 and FPR = 0.15, for different values of m and α. Note that for all values of $P(+)$: (1) for $m = 0$, $EC = 1 - TPR$, (2) for $m = 1$, $EC = FPR$, (3) for $\alpha = 0$, $F_\alpha = TPR$.

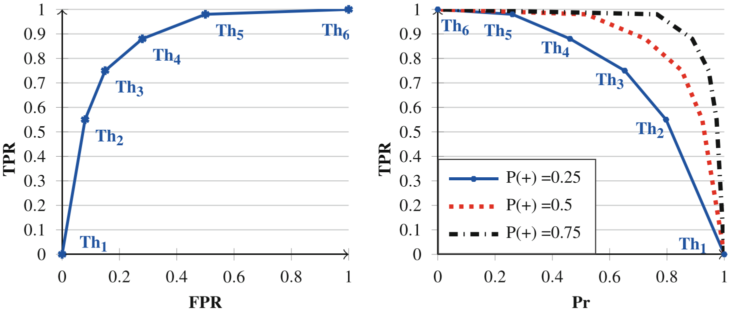
threshold changes, one obtains a curve in ROC space, a family of lines in cost space, and a family of curves in F-measure space. More precisely, as the decision threshold increases (assuming that higher classifier scores correspond to a higher probability of the positive class), the ROC curve starts at $TPR = 0$ and $FPR = 0$, and proceeds towards $TPR = 1$ and $FPR = 1$. For a given value of α, the corresponding F-measure curves move away from the Y axis and get closer to the diagonal line connecting the lower-left point $P(+) = 0, F_\alpha = 0$ to the upper-right point $P(+) = 1, F_\alpha = 1$. An example is shown in Fig. 2. For any given operating condition (i.e., value of $P(+)$), only one decision threshold provides the highest F_α. Accordingly, the upper envelope of the curves that correspond to the available pairs of (TPR, FPR) values shows the best performance of the classifier with the most suitable decision threshold for each operating condition.

Comparing Classifiers in the F-Measure Space. Consider two classifiers with given values of (TPR_i, FPR_i) and (TPR_j, FPR_j), and a fixed value of α. From Eq. (10) one obtains that, if $\mathrm{FPR}_j < \mathrm{FPR}_i$ and $\mathrm{TPR}_j < \mathrm{TPR}_i$, or when $\mathrm{FPR}_j > \mathrm{FPR}_i$ and $\mathrm{TPR}_j > \mathrm{TPR}_i$, then the F-measure curves cross in a *single* point characterized by:

$$P^*_{\mathrm{i,j}}(+) = \frac{\mathrm{FPR}_i \cdot \mathrm{TPR}_j - \mathrm{FPR}_j \cdot \mathrm{TPR}_i}{(1 - {}^1/_\alpha)(\mathrm{TPR}_j - \mathrm{TPR}_i) + \mathrm{FPR}_i \cdot \mathrm{TPR}_j - \mathrm{FPR}_j \cdot \mathrm{TPR}_i}. \tag{11}$$

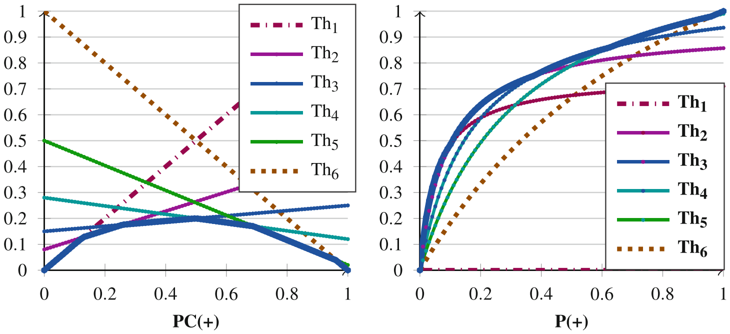
It is also easy to analytically determine which of the classifiers outperform the other for lower or higher $P(+)$ values than $P^*_{\mathrm{i,j}}(+)$. If the above conditions do not hold, one of the classifiers dominates the other for all values of $P(+) > 0$; the detailed conditions under which $F^j_\alpha > F^i_\alpha$ or $F^j_\alpha < F^i_\alpha$ are not reported here for the sake of simplicity, but can be easily obtained as well. Examples of the two cases above are shown in Fig. 3.

In general, given any set of crisp classifiers, the best one for any given $P(+)$ value can be analytically determined in terms of the corresponding TPR and



(a) ROC space

(b) PR space



(c) Cost space

(d) F-measure space

Fig. 2. A soft classifier in ROC space (ROCCH), inverted PR space (for three values of $P(+)$), cost space ($m = 0.5$) and F-measure space ($\alpha = 0.5$), for six threshold values $\mathrm{Th}_1 > \mathrm{Th}_2 > \ldots > \mathrm{Th}_6$ corresponding to $\mathrm{TPR}_1 = 0, 0.55, 0.75, 0.88, 0.98, 1$, and $\mathrm{FPR}_1 = 0, 0.08, 0.15, 0.28, 0.5, 1$. The upper envelope of the cost and F-measure curves is shown as a thick, blue line. (Color figure online)

FPR values, and can be easily identified by the corresponding F-measure curve. Similarly, the overall performance of two or more soft classifiers can be easily compared by visually comparing the upper envelopes of their F-curves.

An example of the comparison of two soft classifiers, with six different threshold values, is shown in Fig. 4, where C_1 is the same as in Fig. 2. In ROC space, the ROCCH of C_1 and C_2 cross on a single point around $FPR = 0.3$. The lower envelopes of the corresponding CCs cross around $PC(+) = 0.7$, and thus C_1 and C_2 perform the same for approximately $0.6 < PC(+) < 0.7$, whereas C_1 outperforms C_2 for $PC(+) < 0.6$. When the F-measure is used, comparing C_1 and C_2 for different skew levels in PR space is more difficult, instead, as shown by the corresponding (inverted) PR curves. This task is much easier in the F-measure space; in this example it can be seen that the upper envelopes of the F-measure curves of C_1 and C_2 cross: C_2 outperforms C_1 for $P(+) < 0.4$, they perform the same for $0.4 < P(+) < 0.6$, and C_1 outperforms C_2 for $P(+) > 0.6$.

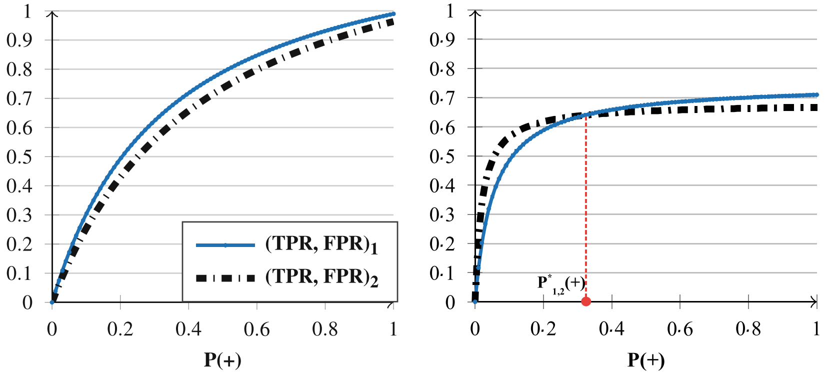


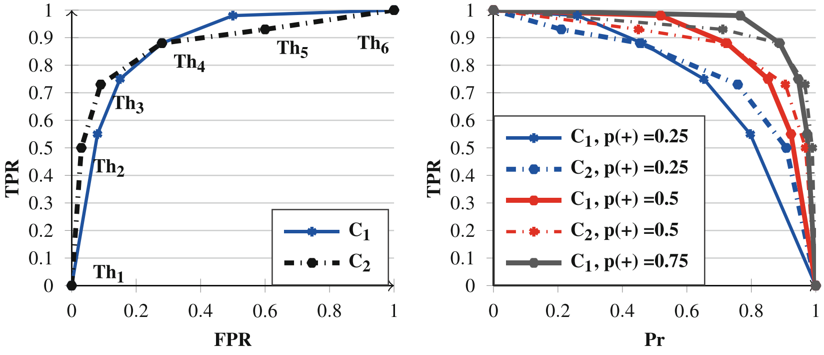
Fig. 3. F-measure curves of two classifiers, for $\alpha = 0.5$. Left: $(TPR_1, FPR_1) = (0.98, 0.5)$, $(TPR_2, FPR_2) = (0.93, 0.6)$: C_1 dominates C_2. Right: $(TPR_1, FPR_1) = (0.55, 0.08)$, $(TPR_2, FPR_2) = (0.5, 0.03)$: the two curves cross at the $P^*_{1,2}(+)$ value of Eq. (11) shown in red. (Color figure online)

These example shows that comparing the F-measure of two (or more) classifiers over all skew levels in F-measure space is as easy as comparing their EC in cost space.

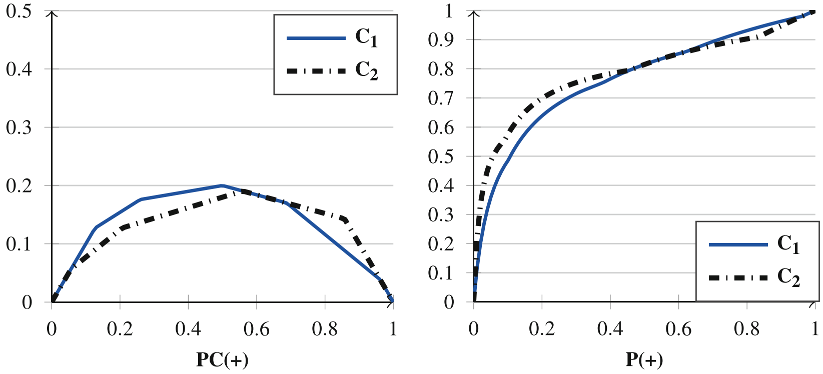
Selecting the Best Decision Threshold or the Best Classifier. ROC curves can be used to set parameters like the optimal decision threshold, or to select the best classifier, for a given operating condition. To this aim, when the EC is used as the performance measure, the ROCCH of the classifier(s) is found and the optimal classifier (or parameter value) is selected by intersecting the upper-left EC iso-performance line corresponding to the given operating condition with the ROCCH. This process is easier in cost space, where the operating condition is shown on the X axis. Analogously, when the F-measure is used, this process is easier in the F-measure space than in PR space. For this purpose, the classifier(s) can be evaluated during design on a validation set (or on different validation sets with different imbalance levels, if the imbalance level during operation is unknown); then, during operation, the imbalance level of the data is estimated and the classification system is adapted based on its performance in cost or F-measure space.

4 Synthetic Examples

We give an example of classifier performance evaluation and comparison in F-measure space and, for reference, in ROC, PR, and cost spaces. In particular, we show how the effect of class imbalance can be observed using these global visualization tools. To this aim we generate a non-linear, 2D synthetic data set: the negative class is uniformly distributed, and surrounds the normally distributed positive class with mean $\mu_+ = (0.5, 0.5)$ and standard deviation $\sigma_+ = 0.33$. The



(a) ROC space (b) PR space

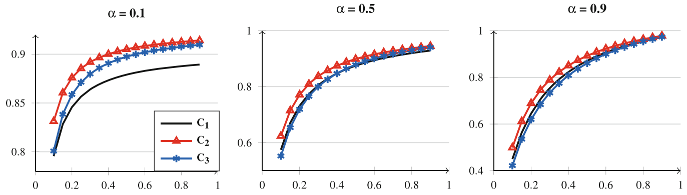


(c) Cost space (d) F-measure space

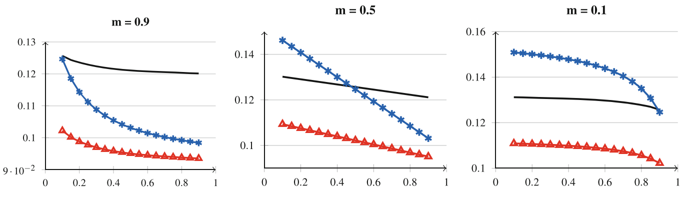
Fig. 4. Comparison between two soft classifiers (C_1 is the same as in Fig. 2) with six threshold values in ROC space, inverted PR space, cost space ($m = 0.5$) and F-measure space ($\alpha = 0.5$). Note that in cost and F-measure spaces the lower and upper envelopes of the curves corresponding to the six threshold values are shown, respectively.

class overlap is controlled by the minimum distance $\delta = 0.15$ of negative samples to μ_+. We consider three classifiers: Naive Bayes (C_1), 5-NN (C_2), and RBF-SVM (C_3). We draw 2000 samples from each class ($M^- = M^+ = 2000$), and use half of them for balanced training. To visualize classifier performance under different operating conditions, we consider different imbalance levels for testing (which simulates the classifier deployment phase). To this aim, we draw from the remaining 2000 samples different testing data subsets of fixed size equal to 1000. The number of testing samples from both classes is chosen as follows: for $P(+) < 0.5$, $M_+ = 500$, $M_- = \lambda M_+$, where $\lambda \in \{0.1, \ldots, 0.9\}$ with a step of 0.05; for $P(+) > 0.5$, $M_- = 500$, $M_+ = \lambda M_-$, with λ chosen in the same way; for $P(+) = 0.5$, $M_+ = M_- = 500$.

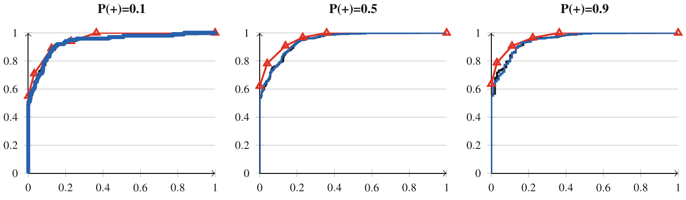
The performance of the three crisp classifiers, using a decision threshold of 0.5, is first compared in F-measure and cost spaces in Figs. 5a and b, for $0.1 < P(+) < 0.9$, $\alpha = 0.1, 0.5, 0.9$, and $m = 0.1, 0.5, 0.9$. It can be seen that



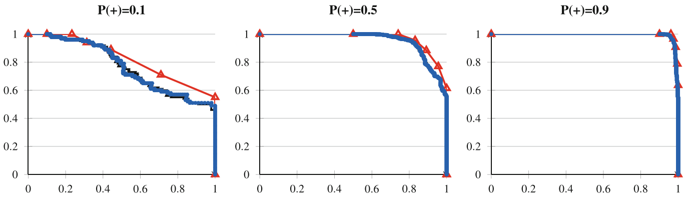
(a) F-measure space: threshold=0.5, three different values of α.



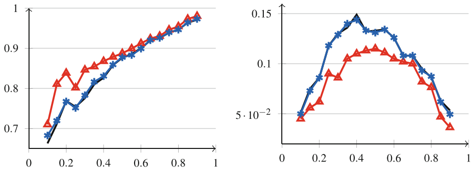
(b) Cost space: threshold=0.5, three different values of *m*.



(c) ROC space: three different values of $P(+)$.



(d) PR space: three different values of $P(+)$.



(e) F-measure space, $\alpha = 0.5$. (f) Cost space, $m = 0.5$.

Fig. 5. Performance comparison among Naive Bayes (C_1), 5-NN (C_2) and RBF-SVM (C_3) in different spaces.

some of the corresponding curves cross, depending on α and m: in this case each classifier outperforms the other for a different range of values of $PC(+)$ or $P(+)$; these ranges can be easily determined analytically. The performance of the same, soft classifiers across different decision thresholds is then compared in ROC and PR spaces for three values of $P(+) = 0.1, 0.5, 0.9$ (Figs. 5c and d), and, for all possible values of $P(+)$, in cost and F-measure spaces (Figs. 5e and f).

As expected, ROC space is not affected by the degree of class imbalance, i.e., by changes in $P(+)$. In PR space each value of $P(+)$ leads to a different curve for a given classifier, instead, but visual comparison of the corresponding F-measure is very difficult: indeed this would require to draw also the hyperbolic iso-performance lines, and anyway only a small, finite number of both $P(+)$ and F_α values can be considered in this space, which does not allow a complete comparison. In cost and F-measure spaces the performance of each classifier for all possible values of $P(+)$ is visualized by a single curve, instead, for a given value of m (in cost space) or α (in F-measure space). In these spaces visual comparison of the corresponding performance measure is very easy, and can be carried out for all possible operating conditions (i.e., values $P(+)$). In this example, from Figs. 5e and f one can conclude that, in terms of both EC and F-measure, C_1 and C_3 perform nearly equally across all operating conditions. Moreover, classifier C_2 dominates both C_1 and C_3 for all values of $P(+)$; however the amount by which C_2 outperforms them is very small in terms of the F-measure, when $P(+)$ is higher than about 0.6, and in terms of EC, when $P(+)$ is around 0.7.

5 Conclusions

In this paper, we reviewed the main existing scalar and global measures and visualization tools for classifier performance evaluation, focusing on class imbalance. Then we proposed a new, specific visualization tool for the scalar F-measure, which is widely used for class imbalance problems, filling a gap in the literature.

Similarly to cost curves, the proposed F-measure curves allow to easily evaluate and compare classifier performance, in terms of the F-measure, across all possible operating conditions (levels of class imbalance) and values of the decision threshold, for a given preference weight between precision and recall. This space can be used to select the best decision threshold for a soft classifier, and the best soft classifier among a group, for a given operating condition. In ongoing research, we are investigating how to use the F-measure space for the design of classifier ensembles that are robust to imbalance, and to adapt learning algorithms to class imbalance.

References

1. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: ICML, pp. 233–240 (2006)
2. Dembczynski, K.J., Waegeman, W., Cheng, W., Hüllermeier, E.: An exact algorithm for F-measure maximization. In: NIPS, pp. 1404–1412 (2011)
3. Drummond, C., Holte, R.C.: Cost curves: an improved method for visualizing classifier performance. Mach. Learn. **65**(1), 95–130 (2006)
4. Fawcett, T.: An introduction to ROC analysis. Pattern. Recognit. Lett. **27**(8), 861–874 (2006)
5. Ferri, C., Hernández-orallo, J., Flach, P.A.: Brier curves: a new cost-based visualisation of classifier performance. In: ICML, pp. 585–592 (2011)

6. Ferri, C., Hernández-Orallo, J., Modroiu, R.: An experimental comparison of performance measures for classification. Pattern. Recognit. Lett. **30**(1), 27–38 (2009)
7. Flach, P., Kull, M.: Precision-recall-gain curves: PR analysis done right. In: NIPS, pp. 838–846 (2015)
8. Garcıa, V., Mollineda, R., Sánchez, J.: Theoretical analysis of a performance measure for imbalanced data. In: ICPR, pp. 617–620 (2010)
9. Hanczar, B., Nadif, M.: Precision-recall space to correct external indices for biclustering. In: ICML, pp. 136–144 (2013)
10. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Prog. in AI **5**(4), 221–232 (2016)
11. Landgrebe, T.C., Paclik, P., Duin, R.P.: Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In: ICPR, vol. 4, pp. 123–127 (2006)
12. Lipton, Z.C., Elkan, C., Naryanaswamy, B.: Optimal Thresholding of classifiers to maximize F1 measure. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS (LNAI), vol. 8725, pp. 225–239. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44851-9_15
13. Parambath, S.P., Usunier, N., Grandvalet, Y.: Optimizing F-measures by cost-sensitive classification. In: NIPS, pp. 2123–2131 (2014)
14. Pillai, I., Fumera, G., Roli, F.: Designing multi-label classifiers that maximize F measures: state of the art. Pattern. Recognit. **61**, 394–404 (2017)
15. Prati, R.C., Batista, G.E., Monard, M.C.: A survey on graphical methods for classification predictive performance evaluation. IEEE Trans. KDE **23**(11), 1601–1618 (2011)
16. Van Rijsbergen, C.: Information retrieval: theory and practice. In: Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems, pp. 1–14 (1979)