F-Measure Curves: A Tool to Visualize Classifier Performance Under Imbalance

Roghayeh Soleymani[a], Eric Granger[a], Giorgio Fumera[b]

[a] *Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), Dept. of Systems Engineering,*
Université du Québec, École de Technologie Supérieure, Montreal, Canada
[b] *Pattern Recognition and Applications Group, Dept. of Electrical and Electronic Engineering*
University of Cagliari, Cagliari, Italy.

Abstract

Learning from imbalanced data is a challenging problem in many real-world machine learning applications due in part to the bias of performance in most classification systems. This bias may exist due to three reasons: (1) classification systems are often optimized and compared using performance measurements that are unsuitable for imbalance problems; (2) most learning algorithms are designed and tested on a fixed imbalance level of data, which may differ from operational scenarios; (3) the preference of correct classification of classes is different from one application to another. This paper investigates specialized performance evaluation metrics and tools for imbalance problem, 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 focus on the case in which the scalar metric F-measure is preferred over other scalar metrics, and propose a new global evaluation space for the F-measure 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 possible imbalance levels for the desired preference of true positive rate to precision. Curves obtained in the F-measure space are compared to those of existing spaces (ROC, precision-recall and cost) and analogously to cost curves. The proposed F-measure space allows to visualize and compare classifiers' performance under different operating conditions more easily than in ROC and precision-recall spaces. This space allows us to set the optimal decision threshold of a soft classifier and to select the best classifier among a group. This space also allows to empirically improve the performance obtained with ensemble learning methods specialized for class imbalance, by selecting and combining the base classifiers for ensembles using a modified version of the iterative Boolean combination algorithm that is optimized using the F-measure instead of AUC. Experiments on a real-world dataset for video face recognition show the advantages of evaluating and comparing different classifiers in the F-measure space versus ROC, precision-recall, and cost spaces. In addition, it is shown that the performance evaluated using the the F-measure of Bagging ensemble method can improve considerably by using the modified iterative Boolean combination algorithm.

Keywords: Pattern Classification, Class Imbalance, Performance Metrics, F-Measure, Visualization Tools, Video Face Recognition

1. Introduction

Evaluating classification performance is an important step for both guiding the learning process, and for comparing different systems. Classification systems are usually trained over a number of iterations, and the direction of the parameter optimization process in each iteration depends on the performance of the classifier(s) during the previous iteration(s). As an example, with Boosting ensemble learning methods, the classifier error of a given iteration affects the sample selection process during the next iteration, as well as the final prediction function. Additionally, after training any classification system, its performance should be objectively compared to alternative systems for the problem at hand.

Performance evaluation is challenging in pattern recognition problems with class imbalance, where the level of imbalance observed in test mode may differ the design data. In this case the most widely used performance metric – classification accuracy – tends to favour the correct classification of the most populated class (or classes). This is an issue in many machine learning applications, where the number of available samples from the minority class of interest ("positive", or "target" class) is heavily outnumbered by other classes, especially in two-class classification problems. Since the objective functions of many standard, state-of-the-art learning algorithms (e.g., support vector machines) seek to maximize unsuitable performance metrics for imbalance, the trained classifiers become biased towards correctly recognizing the majority ("negative" or "non-target" class) at the expense of high misclassification rates for the positive class. On the other hand, the widely used Receiver Operating Characteristic (ROC) curve, area under ROC curve (AUC) and G-mean favour the correct classification of positive samples, at the expense of excessive misclassification of negative samples. The reason is that when data is highly imbalanced, a change in the number of correctly classified positive samples (TP) and the number of misclassified negative samples (FP), reflect in a more significant change in the true positive rate (TPR) compared to the change in false positive rate (FPR).

Precision-Recall (PR) space is in turn more suitable than the ROC space for imbalance problem and plot the performance of the classifier in terms of precision vs. recall (or TPR) [1, 2]. The reason is that, precision is preferred to FPR when classifying imbalanced data because precision measures the proportion of TPR to the FPR multiplied by the skew level. However, PR curves are difficult to analyze when comparing classifiers under different skew levels of data because each skew level of data may result in a different curve and the curves in this space correspond to equal preference of classes. In other words, Pr is plotted against Re without having the flexibility to give more weight to one compared to the other.

To address this issue, other performance metrics like the expected cost (EC) and the F-measure are being used in imbalanced data classification. Such metrics follow different objectives in terms of favouring the correct classification of positive samples, and of avoiding the opposite drawback of allowing excessive misclassification of negative samples. The choice between expected cost (EC) and the F-measure is therefore application-dependent. when expected cost is used to compare classifiers, one can give more importance to TPR and TNR (which is $1-$ FPR) by assigning different cost factors to them. Therefore, when the data is imbalanced, these cost factors can be tuned to give more importance to the minority class and neutralize the effect of imbalance. Two graphical techniques – Cost curves (CC) [3] and Brier curves (BC) [4] – have recently been proposed to easily visualize and compare classifier performance under all possible operational points, i.e. class prior probabilities and misclassification costs (or relative preference of classes). These plots exhibit several advantages over the traditional, well-known ROC plot since ROC curves are independent of class imbalance [2].

The other metric, the F-measure, is widely used metric in information retrieval and class imbalance problems and have been analyzed by many researchers [5, 6, 7, 8]. The main benefit of F-measure is that it compares the performance of the classifier in terms of recall (or TPR) to precision using a factor that controls their relative importance. However, no analogous performance visualization tool exists for the F-measure. We point out that in the Precision-Recall (PR) space, the F-measure is presented as hyperbolic isometrics [9, 10], and does not allow to easily visualize the F-measure of a given classifier under different operational conditions (i.e., different class priors and different preference of precision and recall).

This paper is an extended and improved version of a conference paper presented by the authors [12]. This paper presents the new F-measure curves as a global visualization tool analogous to cost curves for expected cost, which consist of plotting the F-measure value of a given classifier versus two parameters; level of imbalance and the level of preference between recall and precision. It allows to visualize and compare classifier performance in class imbalance problems for different decision thresholds under different operational conditions. To this aim, we rewrite F-measure formula to highlight its dependence on both the class priors and the weights of precision and recall. In this space, a crisp classifier is presented as a curve that shows the performance over a range of possible imbalance levels for the desired level of preference between recall and precision. A soft classifier, therefore, is shown by upper envelope of several curves that correspond to different decision thresholds. Analogously to cost curves, this space allows to compare the classifiers more easily than the ROC space for the given operating condition. For a given preference weight, one classifier may outperform the others in terms of F-measure over all skew levels, or only on a specific range. This range can be determined both analytically and empirically in the proposed F-measure space (and in cost space) based on the values of TPR and FPR of the classifiers. Finally, this space can be used to easily select (1) the decision threshold of the given soft classifier, (2) the best classifier among many of them, and (3) the best combination of a group of classifiers, for the given operating condition.

In summary the main contribution of this paper is proposing a new global performance evaluation space that allows evaluating performance directly, in terms of the scalar F-measure metric. To our knowledge, no performance visualization tool analogous to CC and BC exist for the F-measure. The proposed F-measure space has the following properties.

- Possibility of visualizing the performance of any classifier (soft or crisp) under different imbalance levels of deployment data.
- Possibility of selecting the best threshold of a classifier under the given imbalance level and preference between precision and recall.
- Possibility of comparing more than two classifiers over different decision thresholds and under different imbalance levels of test data with the ability of selecting a preference level between classes.
- Possibility of selecting the best combination of a set of classifiers based on their performance in the F-measure space. As the second contribution, the proposed F-measure space is used to modify the Iterative Boolean Combination (IBC) method to adapt the selection and combination of classifiers in the ensemble for an optimal performance under different operating conditions (imbalance levels).
- The F-measure space is preferred to the ROC and Precision-Recall spaces to compare classifiers under different imbalance levels and preference between classes.

- The F-measure space can be preferable to cost space in some applications when precision-recall is preferred to the misclassification cost like in information retrieval. In addition, the F-measure space is more sensitive to class imbalance and tuning the preference between classes results in a visible difference in performance in the F-measure space compared to the cost space.

To clarify the benefits of the proposed space, two experiments are conducted on a real-world dataset for video-based face recognition dataset. In the first experiment, the behaviour of different classifiers are compared and analyzed using ROC, PR, cost and F-measure spaces. In this experiment, the optimal decision threshold of the soft classifiers and the best classifier among them is selected for the given operating condition. In the second experiment, the Bagging ensemble learning method is optimized and adapted to the given operating condition using the F-measure space by selecting and combining classifiers using a modified version of the Iterative Boolean Combination technique [11] in the F-measure space.

The rest of the paper is organized as follows. Sect. 2 reviews the existing scalar metrics used in, or proposed for class imbalance problems, as well as the global evaluation spaces ROC and PR plots, and the CC and BC visualization tools. We then focus on the F-measure in Sect. 3, where the F-measure space is proposed to visualize performance of different classifiers over a range of possible imbalance levels and with different preference levels of recall to precision. The behaviour of both Cost and F-measure spaces is analyzed in this section. In Sect. 4, experiments on a video-based face recognition dataset are presented and discussed.

2. Review of Performance Metrics and Visualization Tools

Imbalanced data distributions occur in many real-life applications [13], often in two-class problems. In these applications, correctly recognizing samples of the positive class is the main requirement. Avoiding excessive misclassification of negative samples can also be more or less important, depending on the application at hand.

In some applications, the above requirements can be expressed in terms of misclassification costs, where a higher cost is assigned to the misclassification of a positive sample. This also allows one to ”indirectly” take into account class imbalance. Similarly, in other applications assigning different ”fictitious” costs to misclassifications of positive and negative instances can be a (indirect) way to take class imbalance into account. For example, in a medical application like automatic cancer diagnosis the number of positive samples (patients who have cancer) is often much less than negative samples (patients who do not have cancer). In this application, misclassifying a positive sample as a negative i.e., wrongly discharge of a patient who actually has cancer results in a delayed treatment, cost is usually much higher than the one of misclassifying a negative sample, i.e., wrongly suspecting a patient with cancer; the reason is that in the latter case the diagnosis can be corrected in the follow-up tests [14]. Another example is online video surveillance applications, where the objective is to find images of a suspected individual (positive samples) in a public place over a network of cameras. In this application the misclassification of negative samples is tolerable to some extent, since it would be corrected later by a human operator. However, beyond a certain limit, this could waste too much time of the operator, up to missing the person of interest. In this case, a small number of misclassified positive samples could result in saving a relatively higher misclassifications of negative samples. In such cases, the application requirements are better to be determined using a factor that optimizes the trade-off between the correct classification and false alarm rate considering the imbalance level and the desired preference between classes.

Several performance metrics have been used so far for applications with imbalanced classes, and specific metrics have also been proposed. Some reviews of these metrics can be found in the work by Ferri et al. [15], where the behaviour of some scalar performance metrics are analyzed experimentally for several problems including imbalance to find the correlation between these metrics. Garcia et al. [16] and Fernández et al. [17] also compare the performance metrics for imbalance problem. In [1, 2] ROC and PR spaces are compared and the relationship between them is analyzed. ROC, PR and cost spaces are compared in a survey by Prati et al. [18]. In this paper we focus on reviewing these metrics in terms of their sensitivity to imbalance, specifically global spaces that consider different operating conditions and different preference weights. They are reviewed in the following subsections.

2.1. Scalar performance metrics

We focus here on two-class problems, although some of the existing metrics can also be applied to multi-class problems. We denote the prior probability of the positive and negative class with $P(+)$ and $P(-)$, respectively, and the measure of class skew with $\lambda = {}^{P(-)}/_{P(+)}$ (assuming that the positive class is the minority one, we have $\lambda > 1$).

The performance of a two-class classifier on a given set of labelled samples (e.g., a testing set) can be summarized by its confusion matrix. For a given data set (i.e., a testing set), let us denote with n_+ and n_- the number of positive and negative samples, and with N_+ and N_- the number of samples labelled as positive and negative by the classifier at hand. The confusion matrix reports the number of correctly and wrongly classified samples from both classes: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The true positive and false negative rates, denoted respectively as TPR and FNR, are defined as TP/n_+ and FN/n_+; analogously, the true negative and false positive rates (TNR and FPR) are defined as TN/n_- and FP/n_-. These rates are sample-based estimates of the corresponding probabilities (e.g., TPR estimates the probability of correctly classifying positive samples). In particular, in image or document retrieval applications Precision (Pr) and Recall (Re) metrics are used: Re corresponds to TPR, whereas Pr is defined as $TP/(TP+FP)$ or TP/N_+, i.e., the fraction of correctly classified samples among the ones labelled as positive, which estimates the probability that a sample labelled as positive is actually positive.

From the above classification rates, several scalar metrics can be defined. The widely used, standard accuracy is defined as the probability of correctly classifying a random sample, which can be estimated as $A = (TP + TN)/(n_+ + n_-)$. However, in case of imbalanced data, accuracy is biased towards the correct classification of the negative class. In particular, the accuracy of the trivial classifier that labels all samples as negatives tends to one as level of imbalance increases. This makes it not suitable when the correct classification of the positive class is more important, which is often the case in such kind of problem.

A generalization of accuracy (more precisely, of the error probability, defined as $1-A$) is the expected cost (EC), which is used in applications where a cost (in a suitable, application-dependent unit of measurement) can be associated to the classification outcome (either correct or incorrect) of a sample. In the simplest case, the cost of correct classifications is zero, whereas two misclassification costs are associated to the positive and negative classes, denoted respectively as C_{FN} and C_{FP}. EC is then defined as the expected classification cost of a random sample, which amounts to:

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

When data is imbalanced, misclassifying a positive sample is usually more costly than misclassifying a negative one, i.e., $C_{\mathrm{FN}} > C_{\mathrm{FP}}$ (see, e.g., the above example of medical diagnosis). This avoids the bias of classification accuracy toward the negative class. Accordingly, EC can be used also in applications with imbalanced classes where misclassification costs are not precisely known, or even difficult to define: one can design a classifier focused on correctly recognizing the positive class by suitably defining C_{FN} and C_{FP}, and using EC as the objective function to be minimized. On the other hand, as the ratio $C_{\mathrm{FN}}/C_{\mathrm{FP}}$ increases, minimizing EC allows to increase TPR at the expense of increasing FPR, which may be undesirable in some applications.

In information retrieval applications, misclassifications are usually not associated to a cost, and thus EC is not used. To evaluate the effectiveness of a retrieval system, Pr and Re are the most widely used metrics, instead. Note that they measure complementary aspects of retrieval effectiveness: for a given retrieval system, Pr can usually be increased only at the expense of a lower Re, and vice versa. For instance, when a two-class classifier is used to label input samples (e.g., images or documents) as relevant or non-relevant to a given query, changing its decision threshold results in increasing one of the two metrics and in decreasing the other.

In particular, in the case of class imbalance, Pr takes into account both TP and FP, and drops severely when correct classification of positive class is attained at the expense of a high number of misclassified negative samples. This can be seen more clearly by rewriting Pr as follows:

$$Pr = \frac{TPR}{TPR + \lambda FPR} \ . \tag{2}$$

Pr is in general more sensitive to imbalance. The reason is that when data is highly imbalanced, a change in the number of correctly classified positive samples (TP) and the number of misclassified negative samples (FP), reflect in a more significant change in the true positive rate (TPR) compared to the change in false positive rate (FPR). For example, let's consider a case when the number of positive samples is 25 and the number of negative samples is 75. A classifier is tuned in two different ways such that with one setting TP = 22 and FP = 3, while with another setting, TP = 24 and FP = 5. It is observed that for the same change in TP and FP, the change in TPR is 0.12 while the change in FPR is only 0.02.

It is now evident that, as the λ increases, any given increase in FPR results in a higher reduction in Pr. This is an interesting feature compared to EC mentioned above for class imbalance problems. As an example, let's consider two cases where $(TPR, FPR)_1 = (0.88, 0.04)$ and $(TPR, FPR)_2 = (0.88, 0.06)$ and $C_{\mathrm{FN}} = C_{\mathrm{FP}} = 1$. If $\lambda = 3$ (or $P(+) = 0.25$), then $EC_1 = 0.06$, $Pr_1 = 0.88$, and $EC_2 = 0.075$, $Pr_2 = 0.83$. If $\lambda = 4$ (or $P(+) = 0.2$), then $EC_1 = 0.056$, $Pr_1 = 0.846$, and $EC_2 = 0.072$, $Pr_2 = 0.78$. It is observed that the (relative) increase of EC is less significant than the decrease of Pr when data is more imbalanced.

To obtain a single scalar metric from Pr and Re, the F-measure has been proposed in [19]. It is defined as the weighted harmonic mean of Pr and Re:

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

where $0 < \alpha < 1$ is the weight parameter. Note that another form of the F-measure is more commonly used: it is obtained by rewriting the weight as $\alpha = (1+\beta^2)^{-1}$, with $\beta \in [0, +\infty)$, which leads to

$$\begin{aligned} F_\beta &= \frac{(1+\beta^2)\mathrm{Pr}\cdot\mathrm{Re}}{\beta^2\mathrm{Pr}+\mathrm{Re}} && (4)\\ &= \frac{(1+\beta^2)\mathrm{TP}}{(1+\beta^2)\mathrm{TP}+\mathrm{FP}+\beta^2\mathrm{FN}} \,. && (5) \end{aligned}$$

It is easy to see that for $\alpha \to 0$, $F_\alpha \to \mathrm{Re}$, whereas for $\alpha \to 1$, $F_\alpha \to \mathrm{Pr}$; setting $\alpha = 1/2$ one gets the unweighted harmonic mean of Pr and Re, i.e., both metrics are equally weighted. One interesting feature of the F-measure is that its sensitivity to the correct classification of the positive and the negative classes can be adjusted by tuning α.

This measure is useful to compare classifiers when data is imbalanced, especially in information retrieval, since it depends on TPR and Pr rather than TPR and FPR directly. For the same example presented above where $(TPR, FPR)_1 = (0.88, 0.04)$ and $(TPR, FPR)_2 = (0.88, 0.06)$, Table 1 shows how the value of F-measure is different for different values of λ and α.

Table 1: The value of F-measure when $(TPR, FPR)_1 = (0.88, 0.04)$ and $(TPR, FPR)_2 = (0.88, 0.06)$, for different values of λ and α.

		$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$
$\lambda = 3$	F_1	1	1	1
	F_2	0.72	0.71	0.70
$\lambda = 4$	F_1	0.73	0.73	0.72
	F_2	0.67	0.65	0.63

In general, note that each of the metrics TPR, FPR, TNR and FNR, that are directly extracted from the confusion matrix, focuses on the performance of each class individually, and does not allow to evaluate the effect of class imbalance. However, any metric that uses values from both rows of this matrix, like Pr, will be inherently sensitive to imbalance [20].

We finally mention other existing scalar metrics that have been exploited, and new ones that have been specifically proposed, for class imbalance problems, although they are currently less used than EC and the F-measure (some of them include parameters that can be tuned to weigh the correct classification of both classes). A more extensive review of these metrics can be found in [15, 16]. Other scalar metrics that have been used in the class imbalance context include: Matthews correlation coefficient [21], defined as the correlation between the true and the predicted label of a random sample; the geometrical mean [22] either between TNR and Re, or between Pr and Re (the former values the correct classification of both classes equally, whereas the latter is more sensitive to imbalance due to the use of Pr); the arithmetic average of TPR and TNR (macro-averaged accuracy) [15], which values the correct classification of both classes equally; The metrics that have been proposed for class imbalance problems include: a variant of macro-averaged accuracy, mean-class-weighted accuracy [23], that includes a weight to increase the relative importance of the two classes; optimized precision [24], that combines Re and TNR and accounts for the relative number of positive and negative samples; the adjusted geometric mean [25] between

Re and TNR, that aims at increasing Re while keeping the reduction of TNR to a minimum; and the index of balanced accuracy [16], aimed at reducing the effect of the difference in TPR and TNR in metrics that combine them. Variants of the above performance metrics have also been proposed to account for $P(+)$ [16, 10].

2.2. Global Evaluation Curves

In many real-world applications there is no precise knowledge of the operating condition where the classification system will be deployed, i.e., the misclassification costs (when they can be applied) or the relative importance between Pr and Re (the weight α of the F-measure), and the class priors. In this case, it is desirable that the classifier performs well over a wide range of operating conditions, and it is thus useful to evaluate its performance over such a range.

Global curves depict the trade-offs between different evaluation metrics in a multidimensional space under different operating conditions, rather than reducing these aspects to a single scalar measure which gives an incomplete picture of prediction performance. However, global methods are not as easy to interpret and analyze as single scalar values and pose some difficulties in conducting experiments where many data sets are used.

A well known and widely used tool for two-class problems is the Receiver-Operating Characteristic (ROC) curve, which plots TPR vs FPR for a classifier (typically as a function of its decision threshold), allowing to evaluate the trade-off between these measures for different operating conditions, as well as to compare different classifiers. Each classifier with a specific threshold corresponds to a point in the ROC space, and a potentially optimal classifier lies on the convex hull of the available set of points, regardless of operating conditions (misclassification costs and class skew). In Figure 4 (a), the ROC curve of two classifiers are shown. These curves have been approximated using six different values of their decision thresholds. The classifiers located in the most north west area of this space are preferred since they result in higher TPR for the lower FPR. The ROC curve can also be summarized by a scalar metric defined as the area under the ROC curve (AUC), which equals the probability that a random positive sample is given a higher score than a random negative sample.

In problems with class imbalance, a drawback of the ROC space is that it does not reflect the impact of imbalance, since for a given classifier, the TPR and FPR values do not depend on class priors [26]. However, it is possible to estimate the expected performance of the classifier in ROC space for a given skew level of data in terms of EC. In ROC space, each operating condition corresponds to a set of "isoperformance" lines that have the same slope. An optimal classifier for a specific operating condition is found by intersecting the ROC convex hull (ROCCH) with the upper left isoperformance line. To make this process easier, Cost Curves (CC) [3] have been proposed to better visualize the performance of the classifiers over a range of misclassification costs and/or skew levels in terms of EC. This space is further investigated in section 2.3.

In contrast to TPR and FPR, precision (Pr), is sensitive to class imbalance. When Pr and Re, the well-known metrics in information retrieval, are used as the performance metrics, their trade-off across different choices of the classifier's decision threshold can be evaluated by the precision-recall (PR) curve, which is obtained by plotting Pr as a function of Re. Contrary to the ROC curve, the PR curve is sensitive to class imbalance, given its dependence on Pr. However, for different operating conditions (skew levels), different curves are obtained in this space, which makes it difficult to compare the classifiers when a range of operating conditions are considered. A disadvantage with respect to the ROC space is that the convex hull of a set of points in PR space, and the area under the PR curve, have no clear meaning, despite they are used by several authors [10].

As explained in section 2.1, in PR space, a given Pr and Re pair, under a specific operating condition (skew level and desired preference of Re to Pr), can be summarized into a single value metric; F-measure (similarly to EC in ROC space). F-measure isometrics (sets of points that correspond to the same value of F-measure) in PR space are hyperbolic [9, 10]. In fact it is not easy to visualize the performance of a classifier in terms of F-measure over a range of decision thresholds, skew level, and preference of Pr to Re at the same time in PR space. This is analogous to difficulty of visualizing the performance in terms of expected cost in ROC space. In the case of the expected cost this problem has been addressed by proposing the cost (and Brier) curves visualization tools, alternative to the ROC space, described in section 2.3. Inspired by them, we will propose in section 3 an analogous visualization tool for the F-measure. We first review the cost performance visualization tools in section 2.3 and propose a new visualization tool for F-measure in section 3.

2.3. Expected Costs Visualization Tools

In the lifetime of a classifier learned from training data there are at least three or four distinct sets of examples that might each have a different proportion of positive examples. $P_{train}(+)$ is the percentage of positive examples in the dataset used to learn the classifier. $P_{validation}(+)$ is the percentage of positive examples in the dataset that is used to evaluate and tune the parameters of the classification system. Depending on the type and structure of the classification the validation may or may not exist. $P_{test}(+)$ is the percentage of positive examples in the dataset used to put the system to test and build the classifier's confusion matrix. $P_{deploy}(+)$ is the percentage of positive examples when the classifier is deployed (put to use). The ROC curve plots TPR versus FPR computed from the class-conditional probabilities, which are assumed to remain constant for different $P_{deploy}(+)$. However, because we do not necessarily know $P_{deploy}(+)$ at the time we are learning or evaluating the classifier we would like to visualize the classifier's performance across all possible values of $P_{deploy}(+)$. Cost curves [3] do precisely that and estimate the classifier performance in terms of EC across all possible values of $P(+)$. It is $P_{deploy}(+)$ that should be used for $P(+)$, because it is the performance during deployment that we wish to estimate of a classifier.

With $P(-) = 1 - P(+)$, and $\mathrm{FNR} = 1 - \mathrm{TPR}$, EC is normalized to take the maximum value of 1 as:

$$PC(+) = \frac{P(+) \cdot C_{\mathrm{FN}}}{P(+) \cdot C_{\mathrm{FN}} + (1 - P(+))C_{\mathrm{FP}}} \tag{6}$$

$$NEC = (1 - \mathrm{TPR} - \mathrm{FPR})PC(+) + \mathrm{FPR} \tag{7}$$

Cost curve depict the normalized expected cost NEC (Eq. (7)) versus $PC(+)$

$$NEC = \begin{cases} \mathrm{FPR} & \text{if } PC(+) = 0 \\ 1 - \mathrm{TPR} & \text{if } PC(+) = 1 \end{cases} \tag{8}$$

The always positive and always negative classifiers are shown with two lines in the cost space connecting (1,0) to (0,1), and (0,0) to (1,1), respectively. The operating range of a classifier is the set of operating points, for which the classifier dominates both these lines [3]. Note that in the rest of the paper, we use EC instead of NEC.

There is a point-line duality between cost curves and ROC space. A point in ROC space is represented by a line in the cost space and a point in cost space is represented by a line in the ROC space. The lower envelope of cost lines in cost space shape the CC corresponding to the convex hull of all pairs of (TPR and FPR) points in the ROC space (ROCCH). There are some advantages in cost space to visualize the performance of the classifiers compared to ROC space. For reading quantitative performance information from an ROC plot for specific operating conditions, one deals with some geometric constructions using the iso-performance lines and it is difficult to analyse them when inspecting ROC curves with naked eye [3]. In contrast, the classifiers can be analysed easily by a quick visual inspection for given operating conditions. This property of cost space helps the user to easily compare the classifiers to the trivial classifiers, to select between them, or to measure the difference in performance between classifiers for the given operating conditions [3].

Brier curves (BC) [4] 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. Similarly to the cost space BC plots classification error versus C_{FN}, C_{FP} and π assuming a fixed decision threshold equal to cost proportion $c = C_{\mathrm{FN}}/(C_{\mathrm{FN}} + C_{\mathrm{FP}})$ or skew that is defined as $z = c\pi/(c\pi + (1-c)(1-\pi))$ where $\pi = n_+/(n_+ + n_-)$.

Having in mind the advantages of CC and BC to visualize the expected cost of the classifiers, no performance visualization tools analogous to CC and BC exist for the F-measure, and investigating such space is the subject of the next section which, to our knowledge, has not been addressed in literature. There is a first step by Flach [27] ”towards linking threshold choice methods and performance metrics for loss functions based on F-measure.” derived as F-cost curves. F-cost curves plot a non-linear transformation of the F-measure, defined as $2(1-F_\alpha)/F_\alpha$, as a function of a parameter that corresponds to $(1-\alpha)$. In this space the curves which correspond to a single TPR, FPR point are straight lines as in CC. So F-cost curves are visually similar to CC. However this does not allow one to see the behaviour of the F-measure as function of class skew, which is the (different) goal of the proposed F-measure space in this paper.

3. The F-Measure Space

In this section, an alternative visualization tool is proposed that is analogous to CC for evaluating the F-measure of one classifier. This tool is used to compare classifiers under different operating conditions, that correspond to class priors and to the α parameter in the case of the F-measure. To this aim, we start by rewriting the F-measure from Eq. (3) as follows, to make its dependence on $P(+)$ and on α explicit:

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

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

For $\alpha \to 0$, $F_\alpha \to \mathrm{TPR}$, whereas for $\alpha \to 1$, $F_\alpha \to \mathrm{Pr} = \frac{\mathrm{TPR}}{\mathrm{TPR}+({}^1\!/_{p(+)}-1)\mathrm{FPR}}$; setting $\alpha = 1/2$ one gets the unweighted harmonic mean of Pr and Re ($F_\alpha \to \frac{2\mathrm{TPR}}{\mathrm{TPR}+({}^1\!/_{p(+)}-1)\mathrm{FPR}+1}$).

Contrary to the EC of Eq. (1), Eq. (10) shows that the F-measure cannot be written as a function of a single parameter ($PC(+)$ in case of EC) that takes into account both $P(+)$ and α. This means that the F_α values of a classifier should in principle be plotted as a 3D surface as a function of two distinct variables, $P(+)$ and α. However, this would not allow an easy visualization. Therefore, since the main focus of this paper is class imbalance,

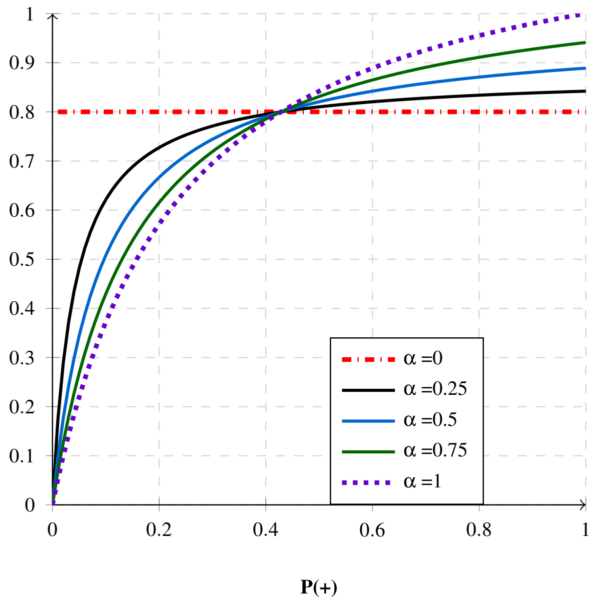


Figure 1: Global F_α curves for a given classifier with TPR=0.8 and FPR=0.15, for different values of α. Note that for all values of $P(+)$: if $\alpha = 0$, then $F_\alpha = TPR$.

in the following we will consider a simpler plot of the F-measure as function of $P(+)$ only, for a fixed value of α.

3.1. F-measure curve of a classifier

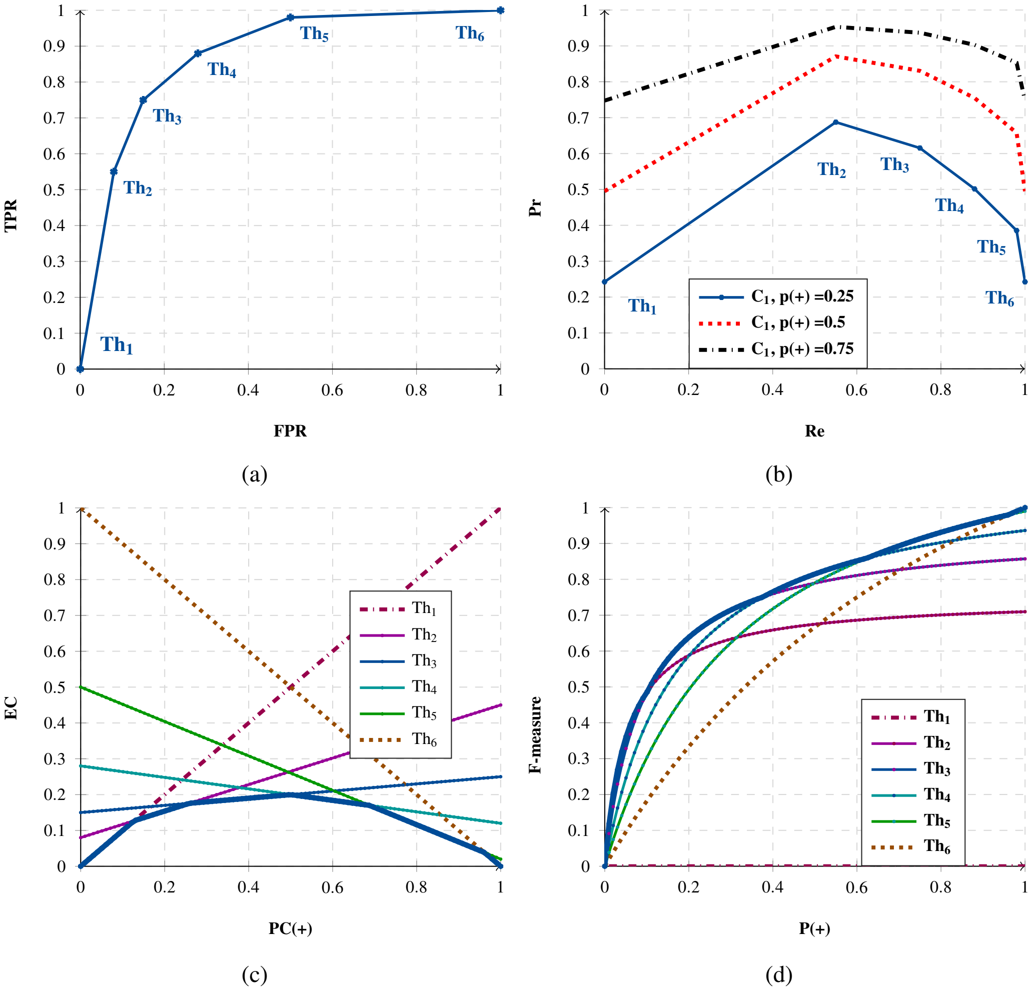
Let us first consider the behaviour of F_α for a given crisp classifier (i.e., for given TPR and FPR values), as a function of $P(+)$. From Eq. (10) one obtains that, when $P(+) = 0$, $F_\alpha = 0$, whereas when $P(+) = 1$, $F_\alpha = \text{TPR}/(\alpha(\text{TPR} - 1) + 1)$. It is then easy to see that the first derivative of F_α with respect to $P(+)$ is strictly positive; the second derivative is strictly negative when TPR $>$ FPR, which is always the case for a non-trivial classifier. Accordingly, the F-measure curve that corresponds to a given classifier is an increasing and concave function of $P(+)$.

For different values of α we get a family of curves. For $\alpha = 0$ we have $F_\alpha = \text{TPR}$ for any value of $P(+)$ and for $\alpha = 1$ we have $F_\alpha = \text{Pr}$. Therefore, for $0 < \alpha < 1$, each curve starts at $F_\alpha = 0$ for $P(+) = 0$ and ends in $F_\alpha = \text{Pr}$ for $P(+) = 1$.

By computing the first derivative of F_α (Eq. 10) with respect to α, for any fixed $P(+)$, one gets that its value is zero for $P(+) = FPR/(FPR - TPR + 1)$, it is negative for smaller $P(+)$ values, and it is positive for higher $P(+)$ values. This means that all curves (including the one for $\alpha = 0$) cross when $P(+) = FPR/(FPR - TPR + 1)$.

An example is shown in Fig. 1, for a classifier with TPR $= 0.8$ and FPR $= 0.15$, and for five values of α.

Consider now the behaviour of the F-measure curve for a given soft classifier and a given α value, when the decision threshold changes. Let us first recall that, as mentioned in Sect. 2.2, a point in the ROC space corresponds to a line in the cost space. Similarly, it corresponds to a (non-linear) curve in the F-measure space. As the decision threshold of a classifier changes, one obtains a curve in ROC space, and a family of lines in cost space. Similarly, one also obtains a family of (non-linear) curves in F-measure space. More precisely, as the decision threshold increases from its maximum to its minimum (assuming that higher classifier scores correspond to a higher probability that the input sample is positive), TPR and FPR start at $TPR = 0$ and $FPR = 0$, and increase towards $TPR = 1$ and $FPR = 1$. For a given value of α, the corresponding curves in F-measure space move away



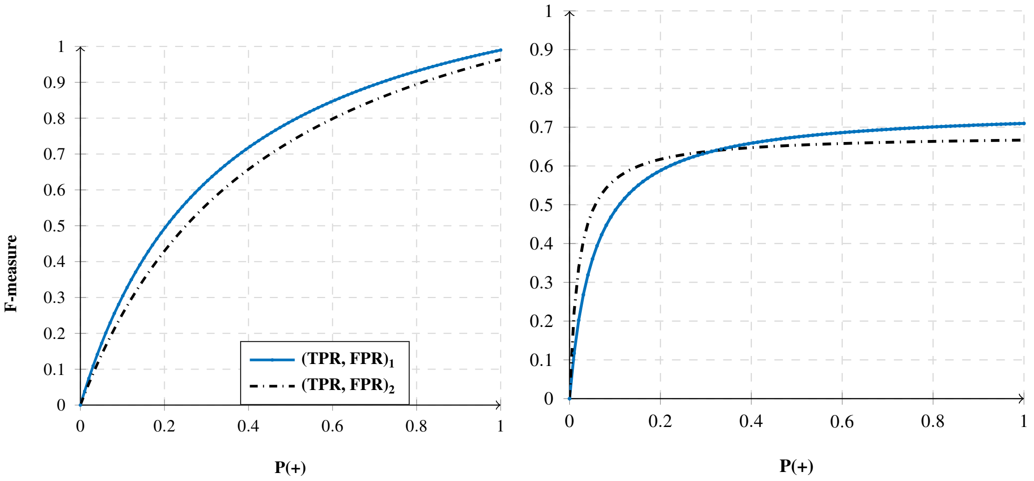
(a)

(b)

(c)

(d)

Figure 2: Performance of one soft classifier in ROC, Precision-Recall, Cost and F-measure spaces with $m = 0.5$ and $\alpha = 0.5$. $\text{TPR}_1 = [10^{-6}, 0.55, 0.75, 0.88, 0.98, 1]$, and $\text{FPR}_1 = [10^{-6}, 0.08, 0.15, 0.28, 0.5, 1]$.



(a) $\text{FPR}_2 > \text{FPR}_1$, $\text{TPR}_2 < \text{TPR}_1$.
$(TPR, FPR)_1 = (0.98, 0.5)$,
$(TPR, FPR)_2 = (0.93, 0.6)$.

(b) $\text{FPR}_2 < \text{FPR}_1$, $\text{TPR}_2 < \text{TPR}_1$.
$(TPR, FPR)_1 = (0.55, 0.08)$,
$(TPR, FPR)_2 = (0.5, 0.03)$.

Figure 3: F-measure curves of two classifiers ($\alpha = 0.5$).

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 $P(+) = 1, F_\alpha = 1$. This behaviour is intuitive: as we move from (0,0) on the ROC curve towards the (1,1), both FPR and TPR are increasing. However, for a non-trivial classifier the increase in TPR becomes less and less greater for a small increase of FPR. In other words, the slope of the tangent line to the ROC curve becomes steeper. The sharper the tangent line to the curve is, the better performance is achieved for the correct classification of positive samples when class skew is high (smaller P(+)).

An example is shown in Fig. 2, we assume a classifier with six different threshold values ($\mathrm{Th}_1 > \mathrm{Th}_2 > \ldots > \mathrm{Th}_6$). From top-left to bottom-right, Fig. 2 shows the convex hull of the corresponding ROC curve, the corresponding precision-recall curves for three values of $P(+)$, cost lines, and the F-measure curves (one for each point in ROC space) with $\alpha = 0.5$. From Eq. 2, corresponding precision to the points in ROC space, could vary based on the skew level of data, as seen in Fig. 2.

For any given operating condition, i.e., for each value of $P(+)$, it is clear that only one of the decision thresholds provides the highest F_α. Accordingly, among the curves that correspond to all the available pairs of (TPR, FPR) values of a soft classifier, their upper envelope shows the best performance of the classifier with the most suitable tuning of decision threshold for each operating condition.

3.2. Comparing classifiers in the F-measure space

We now discuss how two or more classifiers, characterized by given values of (TPR_i, FPR_i) and (TPR_j, FPR_j), etc. , can be compared in the F-measure space, for a fixed value of α. As explained above, the F-measure curve of any classifier starts at $F_\alpha = 0$ for $P(+) = 0$, whereas the one of a classifier characterized by (TPR_i, FPR_i) ends at $F_\alpha^i = TPR_i/[\alpha(TPR_i - 1) + 1]$ when $P(+) = 1$. Note that the latter value depends only on the TPR value, not on FPR.

It is easy to see that $F_\alpha^j > F_\alpha^i$ for all values of $P(+) > 0$, under three different conditions:

(i) when $\mathrm{FPR}_i = \mathrm{FPR}_j$ and $\mathrm{TPR}_j > \mathrm{TPR}_i$;

(ii) when $\mathrm{TPR}_i = \mathrm{TPR}_j$ and $\mathrm{FPR}_j < \mathrm{FPR}_i$;

(iii) when $\mathrm{FPR}_j < \mathrm{FPR}_i$ and $\mathrm{TPR}_j > \mathrm{TPR}_i$.

The analogous conditions under which $F_\alpha^2 < F_\alpha^1$ for all values of $P(+) > 0$ can be easily obtained. As an example, consider two classifiers with $\mathrm{FPR}_1 < \mathrm{FPR}_2$ and $\mathrm{TPR}_1 > \mathrm{TPR}_2$ in Fig. 3(a) ($\alpha = 0.5$). It can be seen that C_2 dominates C_1 for all values of $P(+)$.

Instead, 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 corresponding F-measure curves cross in a *single* point, with the following value of $P^*_{\mathrm{i,j}}(+)$:

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

As an example, consider two classifiers with $\mathrm{FPR}_1 > \mathrm{FPR}_2$ and $\mathrm{TPR}_1 > \mathrm{TPR}_2$ in Fig. 3(b).

In particular, from Eq. (11), we can determine the exact range of $P(+)$ for the given α value when one classifier can outperform the other in terms of F-measure. We see that, if $\mathrm{FPR}_i \cdot \mathrm{TPR}_j = \mathrm{FPR}_j \cdot \mathrm{TPR}_i$, the curves cross only when $P(+) = 0$ ($P^*_{\mathrm{i,j}}(+) = 0$), which means that the classifier with highest TPR and lowest FPR exhibits a higher value of F_α for all values of $P(+) > 0$. If $\mathrm{FPR}_i \cdot \mathrm{TPR}_j \neq \mathrm{FPR}_j \cdot \mathrm{TPR}_i$, the classifier with highest TPR and FPR exhibits a higher value of F_α for $P(+) > P^*_{\mathrm{i,j}}(+)$ than Eq. (11), and the opposite happens for $P^*_{\mathrm{i,j}}(+) < P(+)$.

Therefore, given a set of classifiers characterized by given values of (TPR_i, FPR_i), those that lie on the upper envelope of F-measure curves can be determined accurately because a classifier $C_j(TPR_j, FPR_j)$ dominates $C_i(TPR_i, FPR_i)$ in contributing to the upper envelope F-curve for all $P(+) > P^*_{\mathrm{i,j}}(+)$ if and only if one of the following conditions hold:

(i) $\mathrm{FPR}_j < \mathrm{FPR}_i$ and $\mathrm{TPR}_j > \mathrm{TPR}_i$, $P^*_{\mathrm{i,j}}(+) = 0$;

(ii) $\mathrm{FPR}_j > \mathrm{FPR}_i$ and $\mathrm{TPR}_j > \mathrm{TPR}_i$, $P^*_{\mathrm{i,j}}(+) \neq 0$.

If these conditions hold, the classifiers lie on the ROC convex hull and the upper envelope of F-curves of the given classifiers is obtained from F-curves of all or a subset of classifiers that lie on the ROC convex hull. Note that two classifiers may both correspond to the upper envelope in F-measure space at $P^*_{\mathrm{i,j}}(+)$ if condition (ii) holds.

The overall performance of two or more soft classifiers can be easily compared by comparing the upper envelopes of their F-curves. An example of the comparison of two classifiers is shown in Fig. 4. Classifier C_1 is the same as in Fig. 2. Also for classifier C_2 we consider six different threshold values ($\mathrm{Th}_1 > \mathrm{Th}_2 > \ldots > \mathrm{Th}_6$). In Fig. 4(a), we show the convex hulls of the ROC curves of the two classifiers, which cross on a single point around $FPR = 0.3$. Fig. 4(b) shows the PR curve of these classifiers for three values of $P(+)$. It is difficult to make a statement to compare C_1 and C_2 about their performance for different skew levels of data from their PR curves.

In Fig. 4(c), the lower envelopes of the cost curves are compared. The cost curve of these classifiers cross when $PC(+)$ is close to 0.7, and thus C_1 and C_2 perform the same for approximately $0.6 < PC(+) < 0.7$, and C_1 outperforms C_2 for $PC(+) < 0.6$. From Eq. 6, the classifiers can be compared in this space for any given $P(+)$, C_{FN} and C_{FP}.

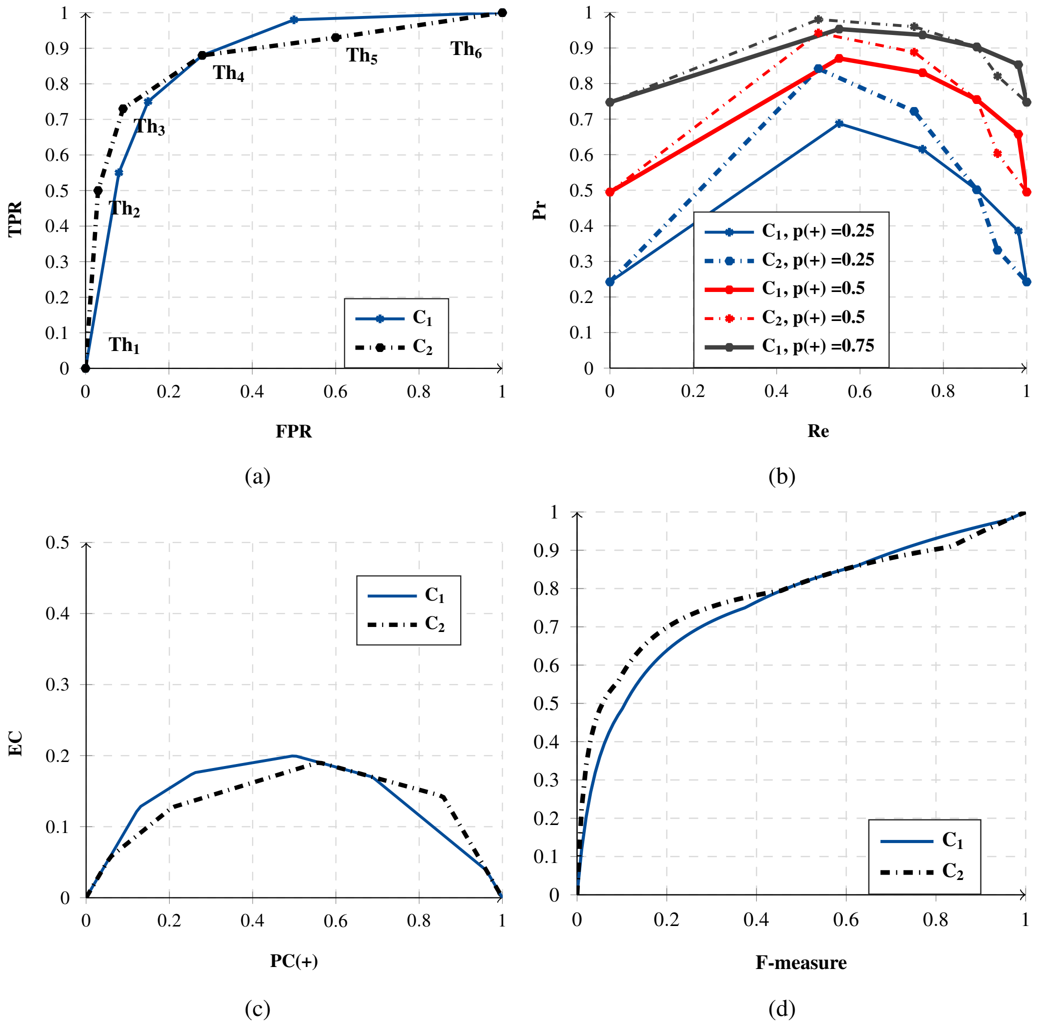
In Fig. 4(d), the upper envelopes of the F-measure curves of C_1 and C_2 are compared for $\alpha = 0.5$. From F-space, C_2 outperforms C_1 for $P(+) < 0.4$, whereas C_1 and C_2 perform the same for $0.4 < P(+) < 0.6$, and C_1 outperforms C_2 for $P(+) > 0.6$. These examples show that comparing the F-measure of two or more classifiers over all skew levels using the F-measure space is as easy as comparing their expected cost using the cost curves.

3.3. F-measure Space vs. Cost Space

As explained in section 2.3, cost space is usually used to compare classifiers' performance in terms of expected cost, given two cost factors C_{FP} and C_{FN}. In this paper we proposed a similar performance visualization space to compare classifiers' performance in terms of the F-measure. In this section, we analyze the similarities and dissimilarities of the proposed F-measure space versus cost space to compare classifiers more closely. To investigate the analogy and the difference between the F-measure space and the cost space we first define a factor m for the expected cost calculation similar to α for the F-measure as follows:

$$m = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}} + C_{\mathrm{FN}}}, \text{ where } 0 < m \leq 1 \tag{12}$$

m can take different values to weigh importance of positive and negative classes. For $m < 0.5$, $C_{\mathrm{FN}} > C_{\mathrm{FP}}$ (similar to $\alpha > 0.5$) and correct classification of positive class is considered more important. For $m = 0.5$, $C_{\mathrm{FP}} = C_{\mathrm{FN}}$ and correct classification of both classes becomes equally important. For $0.5 < m \leq 1$, $C_{\mathrm{FN}} < C_{\mathrm{FP}}$ (similar to $\alpha < 0.5$) and correct classification of positive class becomes less important. It should be noted that α and m are



(a)

(b)

(c)

(d)

Figure 4: Performance of two soft classifiers in ROC, precision-recall, Cost and the F-measure spaces with $m = 0.5$ and $\alpha = 0.5$ ($\text{TPR}_2 = [10^{-6}, 0.5, 0.73, 0.88, 0.93, 1]$, and $\text{FPR}_2 = [10^{-6}, 0.03, 0.09, 0.28, 0.6, 1]$).

actually unrelated, so the pairs of values used in examples in the rest of the paper (where $m = \alpha$) have no particular meaning. From 12 equation 6 is rewritten as:

$$PC(+) = \frac{(1/m - 1) \cdot P(+)}{(1/m - 2) \cdot P(+) + 1} \tag{13}$$

For $m = 0$, $PC(+) = 1$ and $EC = 1 - TPR$, for $m = 0.5$, $PC(+) = P(+)$ and $EC = (1 - TPR - FPR)P(+) + FPR$, and for $m = 1$, $PC(+) = 0$ and $EC = FPR$. Figure 5 shows the expected cost of a single classifiers against $PC(+)$ and $P(+)$ for different values of m. Comparing Figure 5(a) with Figure 1 shows that the sensitivity of the F-measure to the correct classification of the positive and the negative classes can be adjusted by tuning α and therefore a classifier can result in different curves in the F-measure space depending on the value of α. However, tuning m doesn't provide the same effect in the conventional cost space that depicts EC against $PC(+)$.

The cost curves of two classifiers (C_i and C_j) may cross and one classifier may outperform the other for a certain range of operating points, i.e. either before or after the intersection point found as:

$$PC^{*}_{\text{i,j}}(+) = \frac{\text{FPR}_i - \text{FPR}_j}{(\text{TPR}_i - \text{TPR}_j) + \text{FPR}_i - \text{FPR}_j} \tag{14}$$

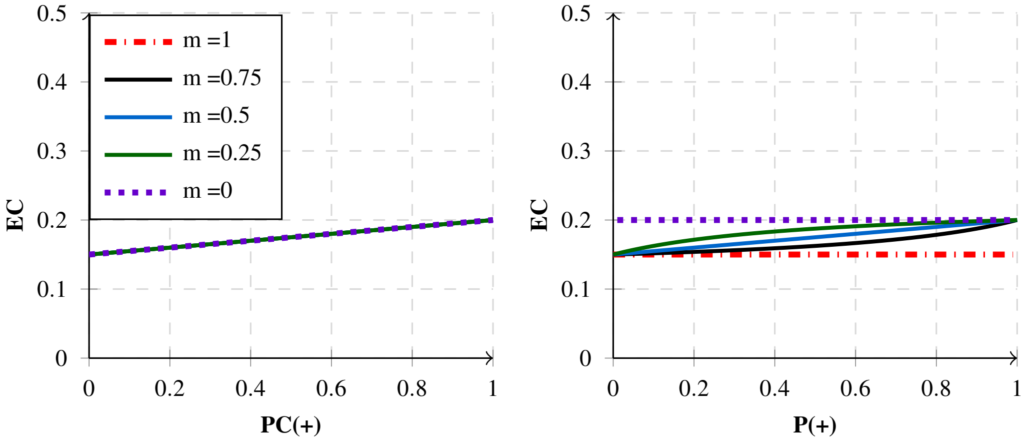


Figure 5: Global expected cost of a given classifier with TPR $= 0.8$ and FPR $= 0.15$, for different values of m against PC$(+)$ and $P(+)$. Note that for all values of $P(+)$: (1) for $m = 0$, $EC = 1 - TPR$, (2) for $m = 1$, $EC = FPR$.

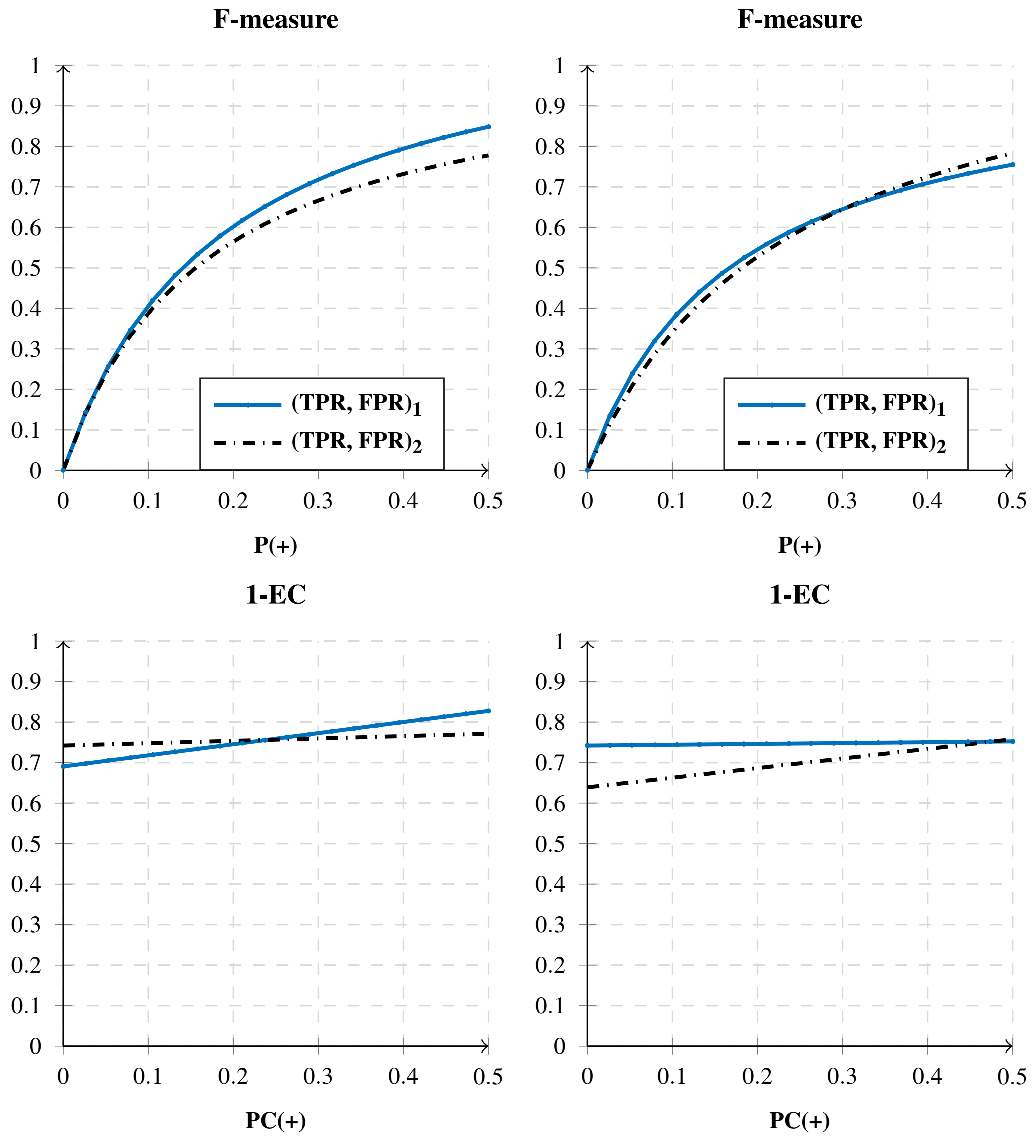
or,

$$P'_{\text{i,j}}(+) = \frac{\text{FPR}_i - \text{FPR}_j}{(\text{TPR}_i - \text{TPR}_j) + (1/(m-1))(\text{FPR}_i - \text{FPR}_j)} \tag{15}$$

The relative performance of a pair of classifiers may differ in the F-measure and cost space and the best classifier for the same $P(+)$ can be different if one uses the EC or the F-measure because the crossing point of the curves in the F-measure and cost spaces can differ. Two examples are shown in Figure 6, where the F-measure and EC is plotted vs. $P(+)$ and $PC(+)$ for (TPR, FPR) pairs to present cases when a crisp classifier dominates another one in CC space whereas their F-measure curves cross, or vice versa. In this example α and m are set to 0.5 (note that when $m = 0.5$, $PC(+) = P(+)$).

The difference between behaviours of the F-measure and cost curves is due to the difference between their sensitivity to the difference between FPR values of classifiers when data is imbalanced ($P(+) < 0.5$). The partial derivatives of EC (Eq. (7)) with respect to FPR is $1 - PC(+)$ for any values of TPR and FPR and therefore it only depends on $P(+)$ and m. However the partial derivative of the F-measure (Eq. (10)) with respect to FPR is $\frac{1-1/p}{TPR}F_\alpha^2$ and is a function of TPR, FPR, α and P(+). Similarly, the partial derivatives of EC with respect to TPR is $-PC(+)$ whereas the partial derivative of the F-measure with respect to TPR is $\frac{\alpha}{TPR}F_\alpha^2 + \frac{1}{TPR}F_\alpha$.

An example is shown in Figure 7 that demonstrate the behaviour of curves in the F-measure and cost spaces when comparing different classifiers with different FPR values and the same TPR value. In Figure 7(a), the F-measure and 1-EC are plotted against $0 < \text{FPR} < 0.5$ for a fixed TPR $= 0.75$, $P(+) = 0.25$, and $\alpha = m = 0.50$. It is observed that the F-measure exhibits a larger decrease than $1 - EC$ as FPR increases for the same value of TPR. Figure 7(b) shows the F-measure and cost curves of classifiers for a fixed TPR $= 0.75$, $\alpha = m = 0.50$ and a range of $0.0001 < P(+) < 0.50$. Comparing the plots show that the difference between F-measure curves of classifiers with the same TPR and different FPR values is more significant than their cost curves, especially for higher imbalance level of data (smaller $P(+)$). For example, the cost curves of classifiers with $FPR = 0.01$ and 0.00 (shown as dashed and dotted curves) almost overlap whereas their F-measure curves exhibit visibly more different behaviour as the $P(+)$ decreases. Very small changes in FPR when data is highly imbalanced means large changes in the number of false positives (or false alarms). Therefore, the F-measure space can be better than



(a) $(TPR, FPR)_1 = (0.96, 0.31)$, $(TPR, FPR)_2 = (0.80, 0.26)$

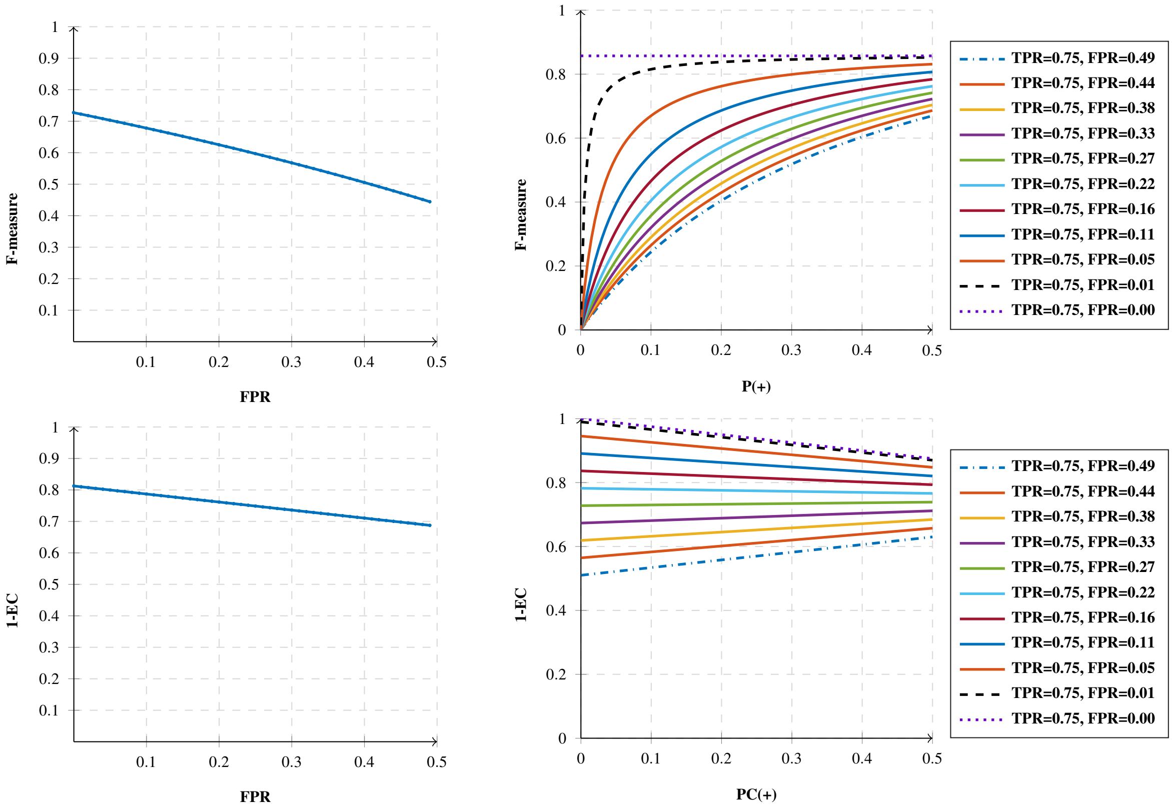
(b) $(TPR, FPR)_1 = (0.76, 0.26)$, $(TPR, FPR)_2 = (0.88, 0.36)$

Figure 6: Comparing the F-measure and cost curves of pairs of classifiers with $\alpha = m = 0.5$ (in this case $PC(+) = P(+)$).

the cost space when the correct classification of positive class is attained at the expense of an excessive number of misclassified negative samples.

3.4. Using F-measure Space to Tune Classification Systems

ROC curves can be used to set the parameters of the system such as the optimal decision threshold or to select the best classifier to gain the best performance for a particular operating condition. For that, ROCCH of the classifier(s) is found and the optimal classifier (or the threshold of the classifier) is selected by intersecting the iso-performance lines corresponding to the given operating condition with the ROCCH at the most upper left side of the curve. If two classifiers (vertices) belong to the desired iso-performance line, there are two optimal classifiers (or threshold values). This process is easier in cost and F-measure spaces because operating condition is shown on the x axis and one can easily find the optimal decision threshold of a soft classifier, or select the best classifier among a group or find the best combination of a set.



(a) F-measure and $1-EC$, $P(+) = 0.25$.

(b) F-measure and $1-EC$ curves for all $P(+)$.

Figure 7: Comparing the sensitivity of the F-measure and cost curves to the difference between FPR values of the classifiers for the given TPR = 0.75 ($\alpha = m = 0.5$).

Two classifiers (vertices) in ROC space that belong to the desired iso-performance correspond to two lines in cost space that intersect in the given operating condition. Similarly, in F-space, two classifiers (vertices) in ROC space correspond to two curves. The cost and F-measure spaces are therefore useful to adapt the classification systems for the given conditions during operations. Three problems can be addressed using this space: (1) tuning the decision threshold of a single classifier; (2) choosing a given classifier between different available ones (each one with a predefined decision threshold); (3) choosing the best combination of a (subset of) available classifiers (each one with a predefined decision threshold).

3.4.1. *Setting an Optimal Decision Threshold or Selecting the Best Classifier*

Each decision threshold of a soft classifier corresponds to a crisp classifier and therefore setting an optimal decision threshold and selecting the best classifier among a group are carried out the same way. ROC curves are ideally suited for setting the optimal decision threshold of a classifier based on Neyman-Pearson criterion. In this case a decision threshold is optimal that corresponds to the maximum TPR for the given acceptable FPR. In cost space a point can be placed on the y-axis representing the criterion (a vertical line crossing the acceptable FPR) and the classifier corresponding the cost line that intersects the y-axis under this point and also participate in forming the lower envelope is the best [3]. The lower bound of expected cost with FPR_{max} and $\text{TPR} = 1$ has $\text{EC}^{min} = (1 - PC(+))\text{FPR}$. In the proposed F-measure space, the upper bound of F-measure value with FPR_{max} and $\text{TPR} = 1$ has $F_{\alpha}^{max} = \frac{1}{1+\alpha(1/P(+)-1)\text{FPR}_{max}}$. Therefore, the best classifier is found as the one that has F_{α} closer to F_{α}^{max} and dominates the others because if $\text{FPR}_1 = \text{FPR}_2$ and $F_{\alpha}^{1} < F_{\alpha}^{2}$, then for any α and $P(+)$, $\text{TPR}_1 < \text{TPR}_2$.

Another criterion to select the best classifier is the operating condition which is the skew level and the preference of classes. In ROC space the best classifier is found as the intersection of ROC convex hull with iso performance lines that correspond to the given operating condition in most up left part of ROC space. In both cost and F-measure spaces the best classifier can be found more easily than ROC space because the skew level ($P(+)$) is shown on the x-axis. The preference between classes (α) is considered in plotting the curves corresponding to classifiers in F-measure space. In cost space the preference between classes (m) is considered along with $P(+)$ in $PC(+)$ on the x-axis.

3.4.2. Choosing the Best Combination of Available Classifiers Using Iterative Boolean Combination

For imbalanced data classification the best single classifier or a combination of a subset of classifiers from a pool may be selected for each operating condition during deployment [11, 28, 29]. For this purpose the classifiers in the pool are tested using validation sets with different imbalance levels after training to find the best (single or an ensemble of) classifier for the specific imbalance level of the validation data. During test, the imbalance level of the test data is estimated (using Hausdorff distance [30]) and the corresponding best (single or an ensemble of) classifier is used for classification of test data. However, note that the goal of CC (and of the F-measure space) is to evaluate classifier performance in terms of skew level at *deployment* time. So, using these tools the performance (at deployment time) for different imbalance levels can be estimated from the testing set performance estimated for a single imbalance level, the same as in training data.

There exists several ways to combine classifiers in either score or decision levels including score averaging, majority voting, learning a meta classifier on either decisions or score, etc. An interesting combination method is Iterative Boolean Combination (IBC) of classifiers [11], which involves selection and combination at the same time. This algorithm starts by combining two classifiers and keeps the combination if it improves ROCCH over the ROC curves of the combined classifiers. In the next iteration, the resulting classifier from the previous iteration is combined with the third classifier and this process continues until all the classifiers are combined. The combination method used in this algorithm is the direct Boolean combination of decisions from pairs of classifiers [1].

In this method, first the decision thresholds of each classifier are selected by sorting the unique scores of the classifier in ascending order. Let's consider Th_t^e ($i = 1, ..., N_c, t = 1, ..., T_e$) as the t^{th} thresholds of the e^{th} classifier c_e (There are total of E classifiers that each have T_e thresholds). Then the decision vector corresponding to each Th_t^e classifier is found as d_t^e. Each element d_t^e is set to 1 when the decision of c_e using the threshold Th_t^e is correct (the class that the sample is assigned to is the same as the true label), and it is set to 0 when the detected class is incorrect. These vectors can then be combined directly with a Boolean functions like $\wedge$ (AND).

In this paper, we optimize the IBC using the F-measure space to select the best combination of classifiers across a range of operating conditions (P(+)) (Alg. 1). A pool of classifiers are trained on a dataset with $P_{train}(+)$ and then tested on the validation set with $P_{validation}(+)$. Then all the decision vectors of each soft classifier are found by varying its decision threshold (lines 1-3 of Algo. 1). After that every single decision vector is combined with the other using each of the Boolean functions $B_1 : a \wedge b, B_2 : \neg a \wedge b, B_3 : a \wedge \neg b, B_4 : \neg(a \wedge b), B_5 : a \vee b, B_6 : \neg a \vee b, B_7 : a \vee \neg b, B_8 : \neg(a \vee b), B_9 : a \oplus b, B_{10} : a \equiv b$. This process results in $\mathrm{N}_D = T_1 + T_2 ... + T_e + 10 \times E \times$

[1] The combination methods and their comparison are presented in Appendix A.

$(E-1) \times T_1 \times T_2 ... \times T_e$ decision thresholds that also includes the original decision vectors of the classifiers in the pool.

In section 3.1, we explained that a soft classifier in the F-mesure space can be shown as the upper envelope of all the curves that correspond to all the available pairs of (TPR, FPR) values of that soft classifier, and the upper envelope shows the best performance of the classifier with the most suitable tuning of decision threshold for each operating condition. Therefore, one can justify that finding the upper envelope of all the F-measure curves of the available pairs of (TPR, FPR) values obtained from a pool of classifiers and different combinations of them results in finding the best collection of classifiers across a range of $P(+)$. We use this idea in lines 13-16 of Alg. 1 to collect the best classifiers or a combination of them for each operating condition.

The worst-case time complexity of combining a pair of classifiers with this algorithm is $O(T_e T_k)$ for each of 10 Boolean operations, given classifiers c_e and c_k have respectively T_e and T_k thresholds [11]. Therefore, for combining E classifiers in design stage of the proposed Boolean combination of classifiers in the F-measure space, the worst-case time complexity of Boolean combination algorithm is $O(T_{\max}^2 E(E-1))$ Boolean operations. Where $T_{\max}$ corresponds to the maximum number of thresholds among E classifiers. During deployment, given a specific operating condition, the worst-case time complexity is $O(T_D)$ (see Alg. 1).

An example is shown in Figure A.12c for combining two soft classifiers in both ROC space and the proposed F-measure space. The AND function alone improves over the performance of C_1 and C_2 for higher imbalance levels (lower $P(+)$) and using Alg. 1 improves the performance over the whole range of operating conditions.

Algorithm 1: Choosing the Best Combination of Classifiers

```
Input: Soft classifiers: $c_e$ , $e = 1, ..., E$
       Validation Set: $\mathbf{V} = \{(\mathbf{x}_i, y_i); i = 1, ..., M\}$, $y_i \in \{0, 1\}$
       Preference between classes: $\alpha$
       Operating points: $P(+) = \{p_j, j = 1, ..., N_{op}\}$
       Boolean functions: $B_1 : a \wedge b, B_2 : \neg a \wedge b, B_3 : a \wedge \neg b, B_4 : \neg(a \wedge b), B_5 : a \vee b, B_6 : \neg a \vee b, B_7 : a \vee \neg b, B_8 :$
       $\neg(a \vee b), B_9 : a \oplus b, B_{10} : a \equiv b$
Output: Combined set of classifiers: $BC$
1  for $e = 1, .., E$ do
2      Test $c_e$ on $\mathbf{V}$ and get back scores set $\mathbf{S}_e$.
3      Define $T_e$ thresholds as the unique values in $\mathbf{S}_e$ and get back decisions $d_i^e (i = 1, ..., T_e)$
4  Define $D_{bc}$ and store all $d_i^e$s.
5  for $l = 1, ..., 10$ do
6      for $e = 1, ..., E$ do
7          for $k = 1, ..., E (k \neq e)$ do
8              for $t = 1, ..., T_e$ do
9                  for $n = 1, ..., T_k$ do
10                     Combine $d_t^e$ and $d_n^k$ using $B_l$ and add the resulting decision vector to $D_{bc}$.
11 Find the F-measure curve of $D_{bc}$ with the size N$_D = T_1 + T_2 ... + T_e + 10 \times E \times (E-1) \times T_1 \times T_2 ... \times T_e$ (as the upper
     envelope of F-measure curves corresponding to all decision vectors in $D_{bc}$) as:
12 for $i = 1, .., N_{op}$ do
13     for $j = 1, .., N_D$ do
14         Find TPR and FPR values from the $j^{th}$ decision vector in $D_{bc}$ and the true labels of the validation data using
             confusion matrix.
15         calculate $f_j = \frac{\alpha^{-1} TPR}{\alpha^{-1} + p_j^{-1} FPR + TPR - FPR - 1}$
16     $F_i = \max\limits_{j=1,...,N_D} f_j$
17     Store the corresponding (l, e, k, t, n) for the $i^{th}$ operating condition and call it $BC_i$ to be used during testing and
         deployment.
```

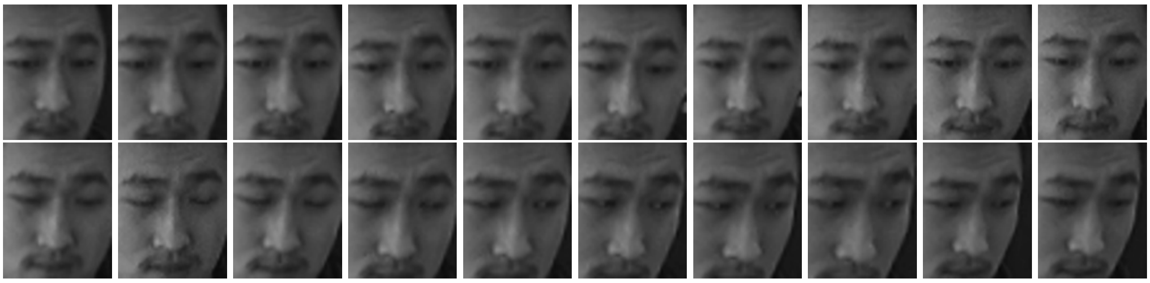


Figure 8: Examples of 21 ROIs captured in a trajectory for individual ID 20110319_0010 and camera 3 (video 2) of the COX dataset.

4. Experiments and Results

The experiments in this paper have been performed for an application of face recognition in video surveillance. In particular, face re-identification is an application that seeks to recognise the facial image of individuals captured over a distributed network of video cameras at different time instants and/or locations. Non-target faces captured in videos under various challenging conditions are compared to those of the target individual using a video-to-video face recognition system. One important challenge in this application is that the number of faces captured from the target individual (positive class) is typically limited and greatly outnumbered by non-target ones (negative class) [31, 32, 33, 34]. We use the COX Face dataset [35] that is a dataset for face recognition in video surveillance [35] and contains videos from 1000 participants captured with 4 cameras under different capture conditions. The faces are detected and each extracted facial region of interest (ROI) is converted to grey scales and normalised to a common size of 64×64 in this dataset. Some examples of ROIs captured in a trajectory[2] from this dataset is shown in Figure 8. Then, multi-resolution gray-scale and rotation invariant Local Binary Patterns (LBP) [37] histograms have been extracted as features. The local image texture for LBP has been characterized with 8 neighbours on a 1 radius circle centred on each pixel. Finally, a feature vector with the length of 59 has been obtained for each ROI.

The experiments in this paper consists of two parts. In the first part (Section 4.1), three classifiers are trained and tested on the COX video dataset and their performance is visualized in ROC, PR , EC, and F-measure spaces. Then, given the operating condition (imbalance level and preference between classes), the F-measure space is used for (1) selecting a single best classifier among them, (2) setting an optimal decision threshold of each classifier. In the second part (Section 4.2), the Bagging ensemble learning [38] method (with RBF-SVM base classifiers) for imbalance is adapted to the given operating conditions using F-measure space by selecting and combinubg of a subset of classifiers using a modified version of the IBC technique proposed in this paper (see Algo. 1).

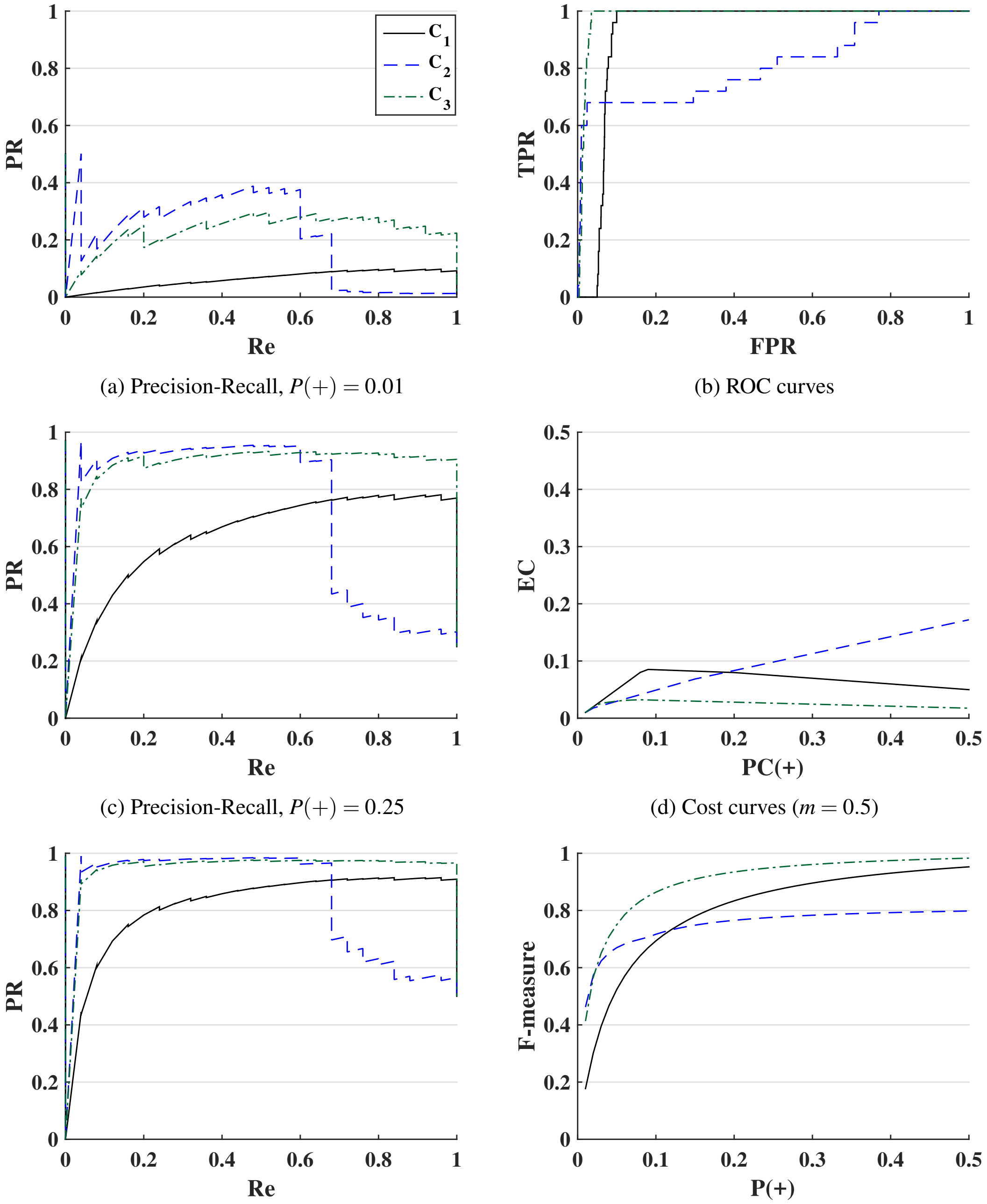
4.1. Experiments for Comparing Classifiers

In this experiment, C_1: Naive Bayes, C_2: MLP(one layer, 8 nodes) and C_3: RBF-SVM (LibSVM [39]) are designed and tested when $P_{\text{train}}(+) = P_{\text{test}}(+) = 0.01$. This experiment is done as an example of comparing:

1. How the performance of different classifiers can be compared for each considered scalar performance measure, using the corresponding (different) global measures/visualization tools.

2. How the effect of class imbalance can be observed using these global measures/visualization tools.

[2] A trajectory is defined as a set of facial ROIs that correspond to a same high quality track of a same individual across consecutive frames



(a) Precision-Recall, $P(+) = 0.01$

(b) ROC curves

(c) Precision-Recall, $P(+) = 0.25$

(d) Cost curves ($m = 0.5$)

(e) Precision-Recall, $P(+) = 0.50$

(f) F-measure curves ($\alpha = 0.5$)

Figure 9: Comparing C_1: Naive Bayes, C_2: MLP(one layer, 8 nodes) and C_3: RBF-SVM for different values of $P(+)$ in ROC, precision-recall, cost and F-measure spaces under different operating conditions.

For this experiment, one video is used for training and one video is used for testing. One individual is randomly selected as the target class and 99 individuals are selected randomly as the non-targets. The ROIs in the trajectory of the target individual are considered as the positive class and the ROIs in the trajectories of the non-target individuals are considered as the negative class Therefore, $P_{\text{train}}(+) = P_{\text{test}}(+) = 0.01$. There are 25 samples in each trajectory in this dataset. Figure 9 shows the performance of these classifiers in Precision-Recall (when $P(+) = 0.01$, 0.25 and 0.50), ROC, EC, and F-measure spaces. In Cost and F-measure spaces both m and α are set to 0.5. Note that in this case $PC(+) = P(+)$. From ROC, Cost and F-measure spaces, it is observed that C_3

outperforms the others. In cost and F-measure spaces, it is easy to compare these classifiers for any given $P(+)$. In the cost space, C_2 outperforms C_1 when $PC(+) < 0.2$ whereas in the F-measure space C_2 outperforms C_1 when $P(+) < 0.11$. When $P(+) < 0.03$, the difference between the performance of C_2 and C_3 is not easy to detect in both the cost and the F-measure spaces.

In Figure 10, both cost and F-measure spaces are used to select an optimal decision threshold for C_3 given different values of $P(+)$. Each row corresponds to one value of $P(+)$. As a reminder from sections 2.3 and 3.1, note that setting the decision threshold of a soft classifier to different values correspond to different lines and curves in cost and F-measure spaces, respectively. The optimal selection of these thresholds corresponds to the lower and upper enveloped in the cost and F-measure spaces, respectively. Therefore, In both Cost and F-measure spaces in Figure 10 the curves that correspond to all thresholds (shown as Th) are plotted with grey colour that appear as shaded areas in the figures of first and second columns.

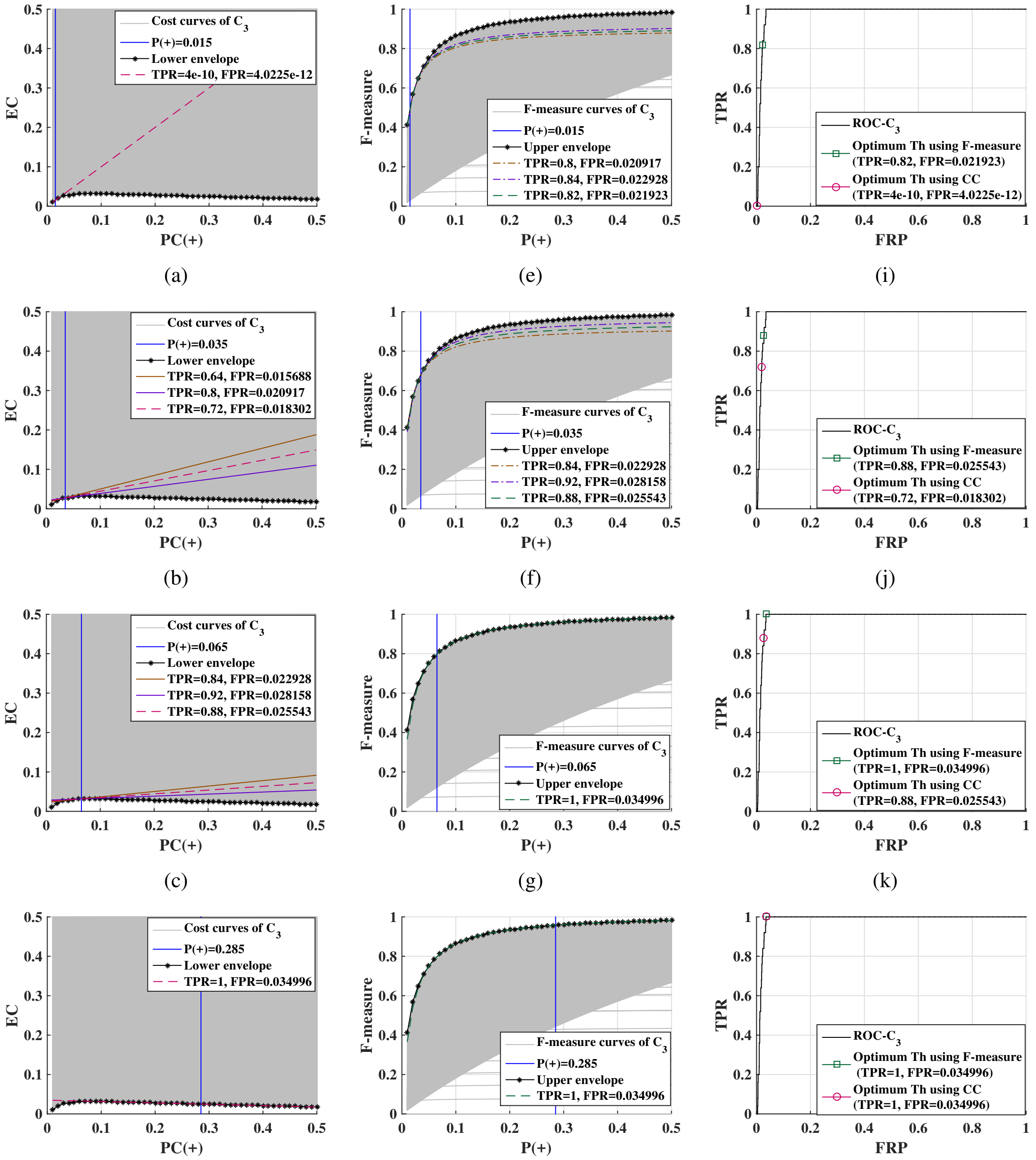
The final lower envelope curve in the Cost space and the upper envelope curve in the F-measure space is also shown in these figures with markers. In These spaces, when there is a threshold value (or (TPR, FPR) pair) that corresponds to the lower/upper envelope for a given $P(+)$ (Figures 10(a), (g), (d) and (h)), that optimal point is returned and the corresponding cost line or F-measure curve is shown. If the given $P(+)$ is not considered during finding the lower/upper envelope of the cost and F-measure plots, no threshold value (or (TPR, FPR) pair) may be found that correspond to the lower/upper envelope for that $P(+)$. For example, in Figures 10, the cost and F-measure curves are plotted with $P(+) = 0.01, 0.02, 0.03, ..., 0.5$ and $P(+) = 0.015, 0.035, 0.065, 0.285$ are not among those considered during plotting these curves. In Figures 10(e), (b), (f) and (c) there is no threshold value (or (TPR, FPR) pair) that corresponds to the lower/upper envelope for a given $P(+)$. In this case the interpolation of the adjacent threshold values is returned as the optimal threshold (the cost lines and F-measure curves corresponding to all three points are shown in the figures).

In the third column of Figures 10, the optimal threshold values (or (TPR, FPR) pairs) obtained from cost and F-measure curves are then shown and compared in the ROC space . We observed that for any $P(+) < 0.03$, the best (TPR, FPR) pair returned using the Cost space is close to $(0,0)$ in the ROC space, which is not optimal. For example in the first row, $P(+)$ is set to 0.015 which corresponds to the optimal point of $(TPR, FPR) = (0.82, 0.022)$ from the F-measure space. When $P(+)$ is set to any value between 0.03 and 0.09 (e.g. Figure 10(b), (f) and (j)), the optimal point of (TPR, FPR) obtained from the cost curve is different than wha is obtained from the F-measure space. When $P(+)$ is set to any value more than 0.09, the optimal threshold obtained from Cost and F-measure spaces are identical.

4.2. Experiments for Adaptive Ensembles

We explained in sections 3.1 and 3.4.2 that finding the upper envelope of F-measure curves that correspond to the available (TPR, FPR) points results in finding the best collection of classifiers across a range of possible P(+) during deployment. We used this idea to modify the IBC algorithm in 3.4.2. In the experiments of this section we put this algorithm in test to see if this method can result in a better performance in terms of F-measure when the classifiers that are stored for the test time are those that have been selected and combined during validation using the proposed IBC algorithm in the F-measure space.

In this experiment, a pool of 20 classifiers is trained using Bagging algorithm [38] that randomly samples



(a) (e) (i)

(b) (f) (j)

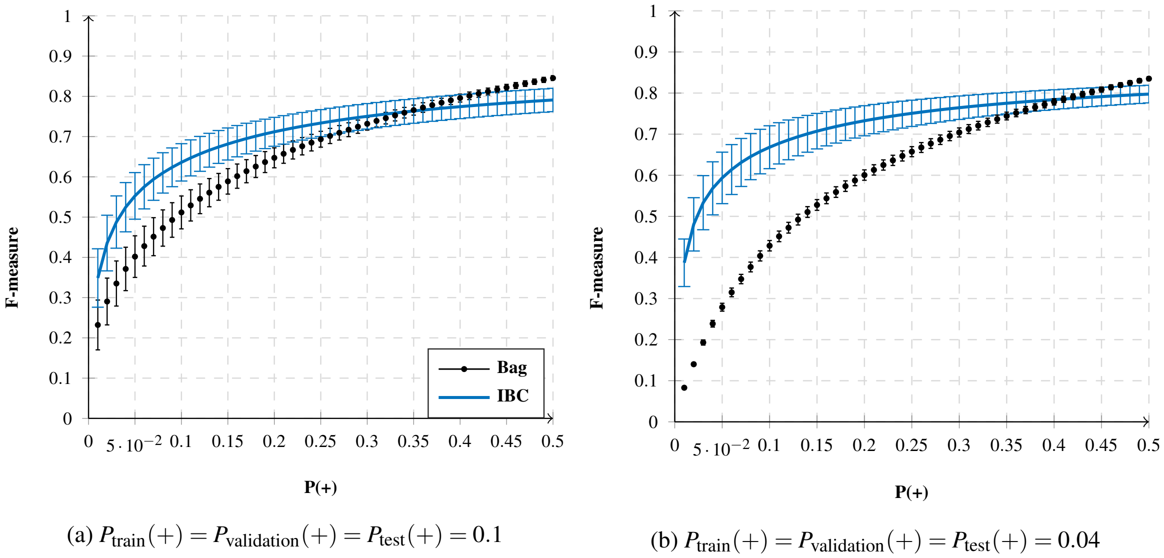
(c) (g) (k)

(d) (h) (l)

Figure 10: Results of experiments to find the optimal decision threshold (and the corresponding TPR, FPR values) in cost and F-measure spaces given a specific operating condition ($P(+)$). First column shows the cost curves, the second column shows the F-measure curves and the third column shows the ROC curves.

balanced subsets of data to train each classifier in the pool. Then, the classifiers in the pool are tested with the validation set and a subset of the classifiers is selected and combined using Algo. 1. In test time, the combination that was selected from the validation step is used.

For this experiment, face captures from one individual (that is randomly selected as target) are considered as the positive class and the face captures from a number of randomly selected individuals are considered as the negative class. In this dataset, there are 25 samples for each individual in each video. In order to consider two cases where $P_{\text{train}}(+) = P_{\text{validation}}(+) = P_{\text{test}}(+) = 0.1$ and $P_{\text{train}}(+) = P_{\text{validation}}(+) = P_{\text{test}}(+) = 0.04$, we select



(a) $P_{\text{train}}(+) = P_{\text{validation}}(+) = P_{\text{test}}(+) = 0.1$

(b) $P_{\text{train}}(+) = P_{\text{validation}}(+) = P_{\text{test}}(+) = 0.04$

Figure 11: F-measure curves of Bagging ensemble method and its modified version using IBC in the F-measure space in two settings of experiments.

9 and 24 negative class individuals, respectively.

To collect three independent sets as training, validation and testing sets, one video is used for training and one video is used for validation and testing. ROI patterns of each individual in the one video are divided into 5 folds, and for each round of experiment, one fold is considered for validation and the remaining 4 folds are considered for testing. Then the roles of the videos is reversed and the video that was used for training is used for validation and testing and vice versa. Therefore, for each individual, there are 25 samples in training data, 5 samples in validation and 20 samples in testing data, and there are 10 rounds of experiments. Then we repeat this process for another 9 randomly selected individuals as the positive class that yields $10 \times 10 = 100$ overall rounds of experiments. The F-measure curves resulting from these experiments are averaged and shown with error bars (the variance) in Figure 11.

In Figure 11, the F-measure curve of resulting combined classifier (shown as adaptive IBC) is compared to the original Bagging ensemble (shown as adaptive Bag) that combines all classifiers in the pool by score averaging. In both cases it is observed that the proposed IBC picks and combines the classifiers better and improves the performance in terms of the F-measure, over a range of imbalance levels, $P(+) < 0.25$ when $P_{\text{train}}(+) = P_{\text{validation}}(+) = P_{\text{test}}(+) = 0.1$ and $P(+) < 0.4$ when $P_{\text{train}}(+) = P_{\text{validation}}(+) = P_{\text{test}}(+) = 0.04$. This is due to the fact that when the classifiers are trained, validated and tested on higher imbalance levels they become more efficient to handle higher imbalance levels during deployment time.

5. Discussion

Here we summarize the discussion related to the results of analysis and experiments throughout the paper that support the the main contribution of the paper. We proposed a new global performance evaluation space that simply allow one to directly evaluate performance in terms of the scalar metric; the F-measure with the following properties:

- Possibility of visualizing the performance of any classifier (soft or crisp) under different imbalance levels of test data. In section 3.1 (Figures 1 and 2), we showed how a crisp classifier corresponds to a single curve in the F-measure space and a soft classifier corresponds to the upper envelope of several curves (each of which corresponds to a single threshold). We also showed that one gets different curves for a single classifier for different preference level between classes (different values of α).

- Possibility of selecting the best threshold of a classifier under the given imbalance level and preference between precision and recall. This idea was presented in section 3.4.1. In section 4.1 (Figure 10), we carried out an experiment to select the best optimal decision threshold of SVM classifier, for the given $P(+)$ values.

- Since each classifier corresponds to a curve in the proposed F-measure space, it becomes possible to compare more than two classifiers over different decision thresholds and under different imbalance levels of test data and preference between classes (see section 3.2). This also provides us the possibility of selecting the best classifier among others for the given imbalance level ($P(+)$) and preference between precision and recall (α). In section 4.1 (Figure 8), we carried out an experiment to compare the performance of three classifiers with a real-world video dataset.

- This space can also be used to select the best combination of a set of classifiers. As the second contribution of the paper in section 3.4.2, the proposed F-measure space is used to modify the Iterative Boolean Combination (IBC) method to adapt the selection and combination of classifiers in the ensemble for an optimal performance under different operating conditions (imbalance levels). In section 4.2 (Figure 11), an experiment is carried out on the video dataset to select the best combination of classifiers generated using the Bagging algorithm. The result of the experiment shows a significant improvement of performance in the F-measure space.

- The proposed F-measure space is preferred to the ROC and Precision-Recall spaces. ROC space is not focused on a specific performance measure, and is also not sensitive to class imbalance, which makes it unsuitable to classification problems with skewed class distributions. The PR space is analogous to ROC space when precision and recall are of interest, instead of TPR and FPR. Although this space is sensitive to imbalance, it does not allow to easily visualize how the F-measure behaves as a function of class skew. In Figures 2 and 9 experiments are used to show the advantage of the F-measure space to the ROC and Precision-Recall spaces.

- The F-measure space can be preferable to cost space in some applications when precision-recall is preferred to the misclassification cost like in information retrieval. In addition, the F-measure space can be preferable to cost space in some scenarios of imbalanced data classification when no specific performance measure can be defined with regard to the preference between expected cost and precision-recall. The reason is that the F-measure space is more sensitive to class imbalance and tuning the preference between classes results in a visible difference in performance in the F-measure space compared to the cost space. In section 3.3, we analyzed and compared the F-measure and cost spaces. We saw that tuning α results in a visible difference in performance while tuning m does not provide the same effect in the conventional cost space that depicts EC

against $PC(+)$. In addition in Figure 7 and the related text, the sensitivity of the F-measure to differences in TPR and FPR between two or more classifiers is more significant than the sensitivity of the EC to those. In the results of the experiments in Figure 10, we also observed that for high level of imbalance the F-measure space is preferred to the cost space for selecting the suitable decision threshold.

6. Conclusions

In this paper, the main existing global evaluation measures and visualization tools were overviewed, and a new one was proposed specifically for the scalar F-measure and for class imbalance problems. The scalar F-measure, weighs the ability of a classifier in recognizing the positive class (the minority and the class of interest) versus the misclassification rate of the negative class (the majority class). It is a suitable scalar performance measure to compare classifiers under imbalance and no visualization tool exists to depict it globally for different operational conditions. Therefore, the F-measure space is proposed as a versatile tool to visualize and compare classifiers performance under different operating conditions (i.e. skew level of data and preference between recall and precision). This space can be used to select the best decision threshold for a soft classifier as well as the best soft classifier among a group, for the given operating condition. This space can also be used to select the best classifiers based on Neyman-Pearson criterion. This space can be further used to modify learning algorithms to address imbalance. In this paper, this space is used to modify the Iterative Boolean Combination algorithm. The experiments on a real-world video dataset is carried out in order to show the use of this space to compare and select classifiers as well as the improvement of performance using the modified Iterative Boolean Combination algorithm for the Bagging ensemble learning method.

Appendix A. Appendix I: Boolean Combination of Classifiers

Diverse classifiers may be combined using the Boolean functions to achieve a more accurate and robust classification system. The Boolean combination of classifiers in ROC space have been investigated in literature for both crisp and soft classifiers [40, 41, 42, 43, 44]. The output of pairwise Boolean conjunction (AND) and disjunction (OR) of crisp classifiers may differ when they are conditionally independent or dependent. However, direct combination of responses from soft classifiers (probability estimates) considers the joint conditional probabilities of each classifier at each threshold. Therefore, no assumptions regarding the independence of the classifiers is required [11].

In this section the behaviour of curves of classifiers in ROC, EC, and F-measure spaces and their corresponding curves for pairwise combination of classifiers with Boolean functions are analysed.

Appendix A.1. Independent Classifiers

Given two conditionally independent classifiers C_i and C_j with $(\text{TPR}_i, \text{FPR}_i)$, and $(\text{TPR}_j, \text{FPR}_j)$, the TPR and FPR of $C_\wedge = C_i \wedge C_j$ (AND) and $C_\vee = C_i \vee C_j$ (OR) are obtained from:

$$\begin{aligned}
\text{TPR}_\wedge &= \text{TPR}_i \cdot \text{TPR}_j && \text{(A.1)}\\
\text{FPR}_\wedge &= \text{FPR}_i \cdot \text{FPR}_j && \text{(A.2)}\\
\text{TPR}_\vee &= \text{TPR}_i + \text{TPR}_j - \text{TPR}_i \cdot \text{TPR}_j && \text{(A.3)}\\
\text{FPR}_\vee &= \text{FPR}_i + \text{FPR}_j - \text{FPR}_i \cdot \text{FPR}_j && \text{(A.4)}
\end{aligned}$$

From these equations, $\text{TPR}_\wedge < \text{TPR}_i$, $\text{TPR}_\wedge < \text{TPR}_j$, $\text{FPR}_\wedge < \text{FPR}_i$, and $\text{FPR}_\wedge < \text{FPR}_j$. Therefore, $(\text{TPR}_\wedge, \text{FPR}_\wedge)$ is located in lower-left side of both $(\text{TPR}_i, \text{FPR}_i)$ and $(\text{TPR}_j, \text{FPR}_j)$ in ROC space. For Boolean disjunction, $\text{TPR}_\vee > \text{TPR}_i$, $\text{TPR}_\vee > \text{TPR}_j$, and $\text{FPR}_\vee > \text{FPR}_i$, $\text{FPR}_\vee > \text{FPR}_j$. Therefore, $(\text{TPR}_\vee, \text{FPR}_\vee)$ is located in upper-right side of both $(\text{TPR}_i, \text{FPR}_i)$ and $(\text{TPR}_j, \text{FPR}_j)$ in ROC space.

In cost space, the expected cost of $C_\wedge$ and $C_\vee$ could be higher or lower than the expected cost of both C_i and C_j based on the value of $PC(+)$.

$$\begin{cases} EC_\wedge < EC_i & \text{if } PC(+) < \frac{\text{FPR}_i - \text{FPR}_\wedge}{(\frac{1-m}{m})(\text{TPR}_\wedge - \text{TPR}_i) + \text{FPR}_i - \text{FPR}_\wedge} \\ EC_\wedge \geq EC_i & \text{otherwise.} \end{cases} \tag{A.5}$$

Similarly:

$$\begin{cases} EC_\vee < EC_i & \text{if } PC(+) > \frac{\text{FPR}_\vee - \text{FPR}_i}{(\frac{1-m}{m})(\text{TPR}_i - \text{TPR}_\vee) + \text{FPR}_\vee - \text{FPR}_i} \\ EC_\vee \geq EC_\vee & \text{otherwise.} \end{cases} \tag{A.6}$$

In F-space, based on the conditions explained in subsection 3.2, $F_\alpha^\wedge > F_\alpha^i$, $F_\alpha^\wedge > F_\alpha^j$, $F_\alpha^\vee > F_\alpha^i$, and $F_\alpha^\vee > F_\alpha^j$ is not true for all values of $P(+) > 0$, and the corresponding F-measure curves cross in a single point. Therefore,

$$P^*_{\wedge,\mathrm{i}}(+) = \frac{\text{FPR}_\wedge \cdot \text{TPR}_i - \text{FPR}_i \cdot \text{TPR}_\wedge}{(1 - {}^1\!/_\alpha)(\text{TPR}_i - \text{TPR}_\wedge) + \text{FPR}_\wedge \cdot \text{TPR}_i - \text{FPR}_i \cdot \text{TPR}_\wedge}$$

$$\begin{cases} F_\alpha^\wedge < F_\alpha^i & \text{if } P(+) < P^*_{\wedge,\mathrm{i}}(+) \\ F_\alpha^\wedge \geq F_\alpha^i & \text{if } P(+) \geq P^*_{\wedge,\mathrm{i}}(+) \end{cases} \tag{A.7}$$

$$P^*_{\vee,\mathrm{i}}(+) = \frac{\text{FPR}_i \cdot \text{TPR}_\vee - \text{FPR}_\vee \cdot \text{TPR}_i}{(1 - {}^1\!/_\alpha)(\text{TPR}_\vee - \text{TPR}_i) + \text{FPR}_\vee \cdot \text{TPR}_i - \text{FPR}_i \cdot \text{TPR}_\vee}$$

$$\begin{cases} F_\alpha^\vee < F_\alpha^i & \text{if } P(+) > P^*_{\vee,\mathrm{i}}(+) \\ F_\alpha^\vee \geq F_\alpha^i & \text{if } P(+) \leq P^*_{\vee,\mathrm{i}}(+) \end{cases} \tag{A.8}$$

Appendix A.2. Dependent Classifiers

In most real-world problems, the classifiers are dependent and the probability that both detectors classify positive samples correctly (positive correlation between classifiers) may take any value between $\mathrm{TPR}_i \cdot \mathrm{TPR}_j$, and $\min(\mathrm{TPR}_i, \mathrm{TPR}_j)$ [40]. Similarly the probability that both classifiers classify negative samples correctly (negative correlation between classifiers) takes a value between $(1 - \mathrm{FPR}_i)(1 - \mathrm{FPR}_j)$ and $\min((1 - \mathrm{FPR}_i), (1 - \mathrm{FPR}_j))$. Therefore,

$$\begin{aligned} \mathrm{TPR}_i \cdot \mathrm{TPR}_j &< \mathrm{TPR}_\wedge < \min(\mathrm{TPR}_i, \mathrm{TPR}_j) \qquad \text{(A.9)} \\ \mathrm{FPR}_i \cdot \mathrm{FPR}_j &< \mathrm{FPR}_\wedge < \mathrm{FPR}_i + \mathrm{FPR}_j - 1 + \min(\mathrm{FPR}_i, \mathrm{FPR}_j) \end{aligned}$$

$$\mathrm{TPR}_\vee > \mathrm{TPR}_i + \mathrm{TPR}_j - \min(\mathrm{TPR}_i, \mathrm{TPR}_j) \tag{A.10}$$

$$\mathrm{TPR}_\vee < \mathrm{TPR}_i + \mathrm{TPR}_j - \mathrm{TPR}_i \cdot \mathrm{TPR}_j \tag{A.11}$$

$$\mathrm{FPR}_\vee > 1 - \min(1 - \mathrm{FPR}_i, 1 - \mathrm{FPR}_j) \tag{A.12}$$

$$\mathrm{FPR}_\vee < \mathrm{FPR}_i + \mathrm{FPR}_j - \mathrm{FPR}_i \cdot \mathrm{FPR}_j \tag{A.13}$$

In this case $\mathrm{TPR}_\wedge < \mathrm{TPR}_i$ and $\mathrm{TPR}_\wedge < \mathrm{TPR}_j$. However, it is not easy to determine the relative position of resulting classifiers from AND and OR functions to the original classifiers in ROC, cost or F-measure spaces. Therefore, it is not possible to find an exact value neither for $(\mathrm{TPR}_\wedge, \mathrm{FPR}_\wedge)$ nor for $(\mathrm{TPR}_\vee, \mathrm{FPR}_\vee)$.

Appendix A.3. Direct Combination of Decisions

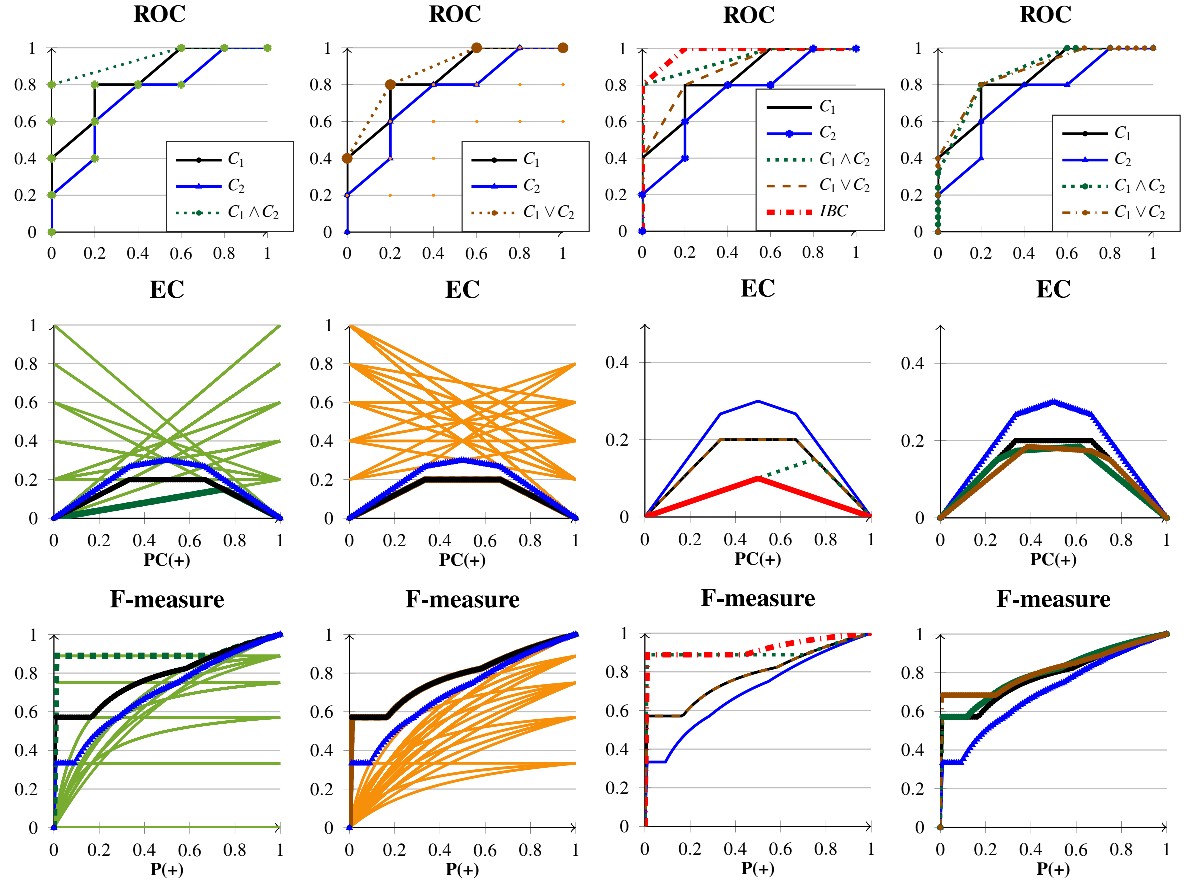
Exploiting direct combination of decisions from pairs of classifiers as done in Iterative Boolean Combination (IBC) [11] is more straightforward and implicitly accounts for dependence between classifiers. The direct combination of decisions from pairs of classifiers is carried out as follows.

The decision thresholds of each classifier are selected by sorting the scores of the classifier in ascending order. Let's consider Th^t_i $(i = 1, ..., N_i)$ as the thresholds of C_i and Th^t_j, $(j = 1, ..., N_j)$ as the thresholds of C_j. First two vectors d^k_i and d^k_j $(k = 1, ..., N_i, ..., N_i \times N_j)$ are found for each pair of (Th^k_i, Th^k_j) such that the elements of d^k_i and d^k_j are set to 1 when the decisions of C_i and C_j are correct, and to 0 when the decisions are incorrect. Then, d^k_i and d^k_j are directly combined using the Boolean functions to D_{ij}. Based on the true labels of samples, the confusion matrix is obtained from D_{ij} and ($\mathrm{TPR}_{\mathrm{ij}}$, $\mathrm{FPR}_{\mathrm{ij}}$) of the resulting classifier is calculated.

With the indirect combination method, first (TPR_i, FPR_i) and (TPR_j, FPR_j) of the classifiers are found from d^k_i and d^k_j. Then, the ($\mathrm{TPR}_{\mathrm{ij}}$, $\mathrm{FPR}_{\mathrm{ij}}$) of the combination is obtained from equations (1), (6) or (7) based on the dependency or independency of the classifiers.

Appendix A.4. Comparing Combination Methods

Given two sets of scores, the Boolean combination of two soft classifiers in Figures A.12 and A.13 is found in three ways: (1) direct combination of decisions given the scores (Figures A.12a, A.12b, A.12c) (2) assuming



(a) Direct combination of decisions (AND).

(b) Direct combination of decisions (OR).

(c) Direct combination of decisions (all).

(d) Indirect combination of independent (TPR, FPR) points.

Figure A.12: Boolean combination of two classifiers using AND, OR and all Boolean functions.

independent classifiers and using the values of TPR and FPR (Figure A.12d). (3) assuming dependent classifiers and using the values of TPR and FPR (Figure A.13). It is observed that the three methods may have different combination results. With the first two methods, AND function improves the performance for lower values of P(+) better than OR function. However, with the third method we can only identify a range for the ROC curves of AND and OR combination results. Showing this range is more complicated in cost and F-measure spaces.

In Figure A.12c, “IBC' corresponds to combination of classifiers using 10 Boolean functions ($C_i \wedge C_j$, $\neg C_i \wedge C_j$, $C_i \wedge \neg C_j$, $\neg(C_i \wedge C_j)$, $C_i \vee C_j$, $\neg C_i \vee C_j$, $C_i \vee \neg C_j$, $\neg(C_i \vee C_j)$, $C_i \bigoplus C_j$, $C_i \equiv C_j$) [11]. Comparing results of AND, OR, and IBC in Figure A.12c shows that using all Boolean functions to combine decisions of two classifiers directly results in a more accurate and robust classification system.

References

[1] T. C. Landgrebe, P. Paclik, R. P. Duin, Precision-recall operating characteristic (p-roc) curves in imprecise environments, in: 18th International Conference on Pattern Recognition (ICPR'06), Vol. 4, IEEE, 2006, pp. 123–127.

[2] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233–240.

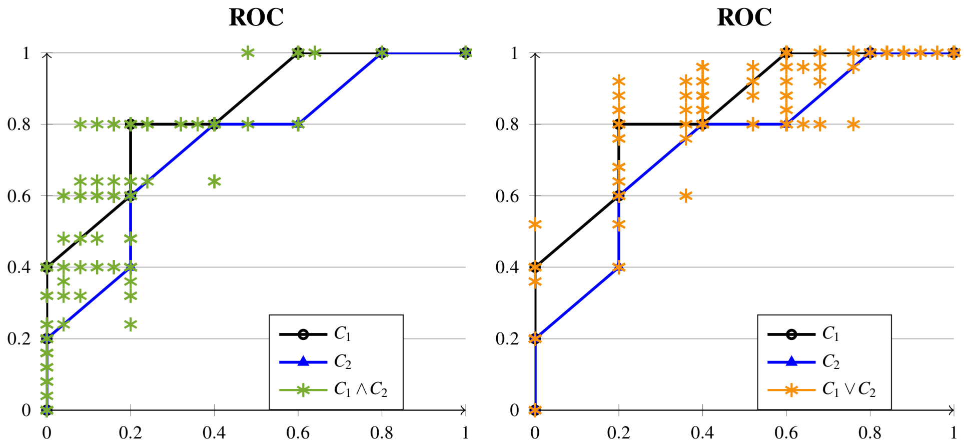


Figure A.13: Combination of two dependent classifiers given their TPR and FPR values using AND and OR functions.

[3] C. Drummond, R. C. Holte, Cost curves: An improved method for visualizing classifier performance, Machine learning 65 (1) (2006) 95–130.

[4] C. Ferri, J. Hernández-orallo, P. A. Flach, Brier curves: a new cost-based visualisation of classifier performance, in: 28th ICML), 2011.

[5] I. Pillai, G. Fumera, F. Roli, Designing multi-label classifiers that maximize f measures: State of the art, Pattern Recognition 61 (Supplement C) (2017) 394 – 404. doi:https://doi.org/10.1016/j.patcog.2016.08.008.
URL http://www.sciencedirect.com/science/article/pii/S0031320316302217

[6] K. J. Dembczynski, W. Waegeman, W. Cheng, E. Hüllermeier, An exact algorithm for f-measure maximization, in: Advances in neural information processing systems, 2011, pp. 1404–1412.

[7] Z. C. Lipton, C. Elkan, B. Naryanaswamy, Optimal thresholding of classifiers to maximize f1 measure, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 225–239.

[8] S. P. Parambath, N. Usunier, Y. Grandvalet, Optimizing f-measures by cost-sensitive classification, in: Advances in Neural Information Processing Systems, 2014, pp. 2123–2131.

[9] B. Hanczar, M. Nadif, Precision-recall space to correct external indices for biclustering, in: Proceedings of the 30th international conference on machine learning (ICML-13), 2013, pp. 136–144.

[10] P. Flach, M. Kull, Precision-recall-gain curves: Pr analysis done right, in: NIPS, 2015.

[11] W. Khreich, E. Granger, A. Miri, R. Sabourin, Iterative boolean combination of classifiers in the roc space: an application to anomaly detection with hmms, Pattern Recognition 43 (8) (2010) 2732–2752.

[12] R. Soleymani, E. Granger, G. Fumera, F-measure curves for visualizing classifier performance with imbalanced data, in: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, 2018, pp. 165–177.

[13] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (4) (2016) 221–232. doi:10.1007/s13748-016-0094-0.
URL http://dx.doi.org/10.1007/s13748-016-0094-0

[14] Y. Artan, M. A. Haider, D. L. Langer, T. H. van der Kwast, A. J. Evans, Y. Yang, M. N. Wernick, J. Trachtenberg, I. S. Yetik, Prostate cancer localization with multispectral mri using cost-sensitive support vector machines and conditional random fields, IEEE Transactions on Image Processing 19 (9) (2010) 2444–2455.

[15] C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognition Letters 30 (1) (2009) 27–38.

[16] V. Garcıa, R. Mollineda, J. Sánchez, Theoretical analysis of a performance measure for imbalanced data, in: Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10), 2010, pp. 617–620.

[17] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, Performance Measures, Springer International Publishing, Cham, 2018, pp. 47–61. doi:10.1007/978-3-319-98074-4_3.
URL https://doi.org/10.1007/978-3-319-98074-4_3

[18] R. C. Prati, G. E. Batista, M. C. Monard, A survey on graphical methods for classification predictive performance evaluation, IEEE Transactions on Knowledge and Data Engineering 23 (11) (2011) 1601–1618.

[19] C. Van Rijsbergen, Information retrieval: theory and practice, in: Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems, 1979, pp. 1–14.

[20] H. He, E. A. Garcia, Learning from imbalanced data, Knowledge and Data Engineering, IEEE Transactions on 21 (9) (2009) 1263–1284.

[21] B. W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (2) (1975) 442–451.

[22] M. Kubat, R. C. Holte, S. Matwin, Machine learning for the detection of oil spills in satellite radar images, Machine learning 30 (2-3) (1998) 195–215.

[23] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, A. Geissbuhler, Learning from imbalanced data in surveillance of nosocomial infection, Artificial Intelligence in Medicine 37 (1) (2006) 7–18.

[24] R. Ranawana, V. Palade, Optimized precision-a new measure for classifier performance evaluation, in: 2006 IEEE International Conference on Evolutionary Computation, IEEE, 2006, pp. 2254–2261.

[25] R. Batuwita, V. Palade, A new performance measure for class imbalance learning. application to bioinformatics problems, in: Machine Learning and Applications, 2009. ICMLA'09. International Conference on, IEEE, 2009, pp. 545–550.

[26] T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (8) (2006) 861–874.

[27] P. Flach, Classification in context: Adapting to changes in class and cost distribution, LMCE-2014 (2014).

[28] P. V. Radtke, E. Granger, R. Sabourin, D. O. Gorodnichy, Skew-sensitive boolean combination for adaptive ensembles–an application to face recognition in video surveillance, Information Fusion 20 (2014) 31–48.

[29] M. De-la Torre, E. Granger, R. Sabourin, D. O. Gorodnichy, Adaptive skew-sensitive ensembles for face recognition in video surveillance, Pattern Recognition 48 (11) (2015) 3385–3406.

[30] G. Edgar, Measure, topology, and fractal geometry, Springer Science & Business Media, 2007.

[31] M. De-la Torre, E. Granger, P. V. Radtke, R. Sabourin, D. O. Gorodnichy, Partially-supervised learning from facial trajectories for face recognition in video surveillance, Information Fusion 24 (2015) 31–53.

[32] C. Pagano, E. Granger, R. Sabourin, G. L. Marcialis, F. Roli, Adaptive ensembles for face recognition in changing video surveillance environments, Information Sciences 286 (2014) 75–101.

[33] R. Soleymani, E. Granger, G. Fumera, Classifier ensembles with trajectory under-sampling for face re-identification, in: Proceedings of the International Conference on Pattern Recognition Applications and Methods-Volume 1, SCITEPRESS-Science and Technology Publications, Lda, 2016, pp. 97–108.

[34] R. Soleymani, E. Granger, G. Fumera, Progressive boosting for class imbalance and its application to face re-identification, Expert Systems with Applications 101 (2018) 271–291.

[35] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, X. Chen, A benchmark and comparative study of video-based face recognition on cox face database, IEEE Transactions on Image Processing 24 (12) (2015) 5967–5981.

[36] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, Vol. 1, IEEE, 2001, pp. I–511.

[37] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (7) (2002) 971–987.

[38] R. Barandela, R. M. Valdovinos, J. S. Sánchez, New applications of ensembles of classifiers, Pattern Analysis & Applications 6 (3) (2003) 245–256.

[39] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3) (2011) 27.

[40] M. A. Black, B. A. Craig, Estimating disease prevalence in the absence of a gold standard, Statistics in medicine 21 (18) (2002) 2653–2669.

[41] S. Haker, W. M. Wells III, S. K. Warfield, I.-F. Talos, J. G. Bhagwat, D. Goldberg-Zimring, A. Mian, L. Ohno-Machado, K. H. Zou, Combining classifiers using their receiver operating characteristics and maximum likelihood estimation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2005, pp. 506–514.

[42] T. Fawcett, Roc graphs: Notes and practical considerations for researchers, Machine learning 31 (1) (2004) 1–38.

[43] M. Barreno, A. Cardenas, J. D. Tygar, Optimal roc curve for a combination of classifiers, in: Advances in Neural Information Processing Systems, 2008, pp. 57–64.

[44] Q. Tao, R. Veldhuis, Threshold-optimized decision-level fusion and its application to biometrics, Pattern Recognition 42 (5) (2009) 823–836.

Roghayeh Soleymani received M.S. in electrical engineering from Urmia University, Urmia, Iran. She is currently working towards the Ph.D. degree at École de Technologie Supérieure (Université du Québec), and is a member of LIVIA, a research laboratory focused on computer vision and artificial intelligence. Her research interests include machine learning, pattern recognition, computer vision and video surveillance with focus on multiple classifier systems, imbalanced data classification and video face recognition.

Eric Granger received Ph.D. in Electric Engineering from École Polytechnique de Montréal in 2001, and worked as a Defense Scientist at DRDC-Ottawa (1999-2001), and in R&D with Mitel Networks (2001-04). He joined the École de technologie supérieure (Université du Québec), Montreal, in 2004, where he is presently Full Professor of Systems Engineering, and director of LIVIA, a research laboratory focused on computer vision and artificial intelligence. His research interests include pattern recognition, machine learning, computer vision, domain adaptation, and incremental and weakly-supervised learning, with applications in biometrics, affective computing, video surveillance, and computer/network security.

Giorgio Fumera received Ph.D. in Electronic Eng. and Computer Science in 2002 from the University of Cagliari, and since 2010 is Associate Professor of Computer Eng. at the Dept. of Electrical and Electronic Eng. of the same University. His research interests are related to methodologies and applications of statistical pattern recognition, and include multiple classifier systems, classification with a reject option, adversarial classification, and intelligent video surveillance.