HAL open science

The bane of skew: Uncertain ranks and unrepresentative precision

Thomas A Lampert, Pierre Gançarski

▶ To cite this version:

Thomas A Lampert, Pierre Gançarski. The bane of skew: Uncertain ranks and unrepresentative precision. Machine Learning, 2014, 97 (1-2), pp.5-32. 10.1007/s10994-013-5432-x . hal-03175422

HAL Id: hal-03175422
https://hal.science/hal-03175422v1

Submitted on 20 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Machine Learning manuscript No.
(will be inserted by the editor)

The Bane of Skew:
Uncertain Ranks and Unrepresentative Precision

Thomas A. Lampert · Pierre Gançarski

Received: date / Accepted: date

Abstract While a problem's skew is often assumed to be constant, this paper discusses three settings where this assumption does not hold. Consequently, incorrectly assuming skew to be constant in these contradicting cases results in an over or under estimation of an algorithm's performance. The area under a precision-recall curve (AUCPR) is a common summary measurement used to report the performance of machine learning algorithms. It is well known that precision is dependent upon class skew, which often varies between datasets. In addition to this, it is demonstrated herein that under certain circumstances the relative ranking of an algorithm (as measured by AUCPR) is not constant and is instead also dependent upon skew. The skew at which the performance of two algorithms inverts and the relationship between precision measured at different skews are defined. This is extended to account for temporal skew characteristics and situations in which skew cannot be precisely defined. Formal proofs for these findings are presented, desirable properties are proved and their application demonstrated.

Keywords Precision · Recall · Skew · AUCPR · Evaluation · Performance · ROC

1 Introduction

A confusion matrix, or contingency table, describes a learning algorithm's performance when a specific threshold and dataset is assumed. To simplify reporting it is common to extract from the confusion matrix various performance measures. It is well known, however, that extracting the overall accuracy is inadequate when a dataset is highly skewed (He and Garcia, 2009). Alternative performance measures such as receiver operating characteristic (ROC) curves offer more detailed insight into an algorithm's performance characteristics. Nevertheless, recent literature points out that these may overestimate performance when class balances are not equal (He and Garcia, 2009) and instead precision-recall (P-R) curves are preferable (Davis and Goadrich, 2006).

T. Lampert · P. Gançarski
Département Informatique Recherche du Laboratoire ICube
Université de Strasbourg
Tel.: +33 (0)3 68 85 45 76
Fax: +33 (0)3 68 85 44 45
E-mail: {tlampert, gancarski}@unistra.fr

Precision-recall curves are being actively researched (Flach, 2003; Davis and Goadrich, 2006; Clémençon, 2009; Boyd et al, 2012) and this paper continues the trend by offering several contributions towards understanding their behaviour in skewed domains of different characteristics. For example, it is proven herein that the ranking of two (or more) algorithms (as measured by the area under precision-recall curves, or AUCPRs) can invert or re-order when different skews are taken, and the conditions under which this may take place are defined. This has direct implications when the AUCPR is employed to compare the performance of two algorithms, and also when the AUCPR is used as an optimisation function (Davis et al, 2005; Liu and Shriberg, 2007; Yue et al, 2007). To this end, another contribution is to define a function that facilitates the analytical transformation of precision between different dataset skews. This last point enables practitioners to compare performances derived from differently skewed datasets.

Furthermore, this paper addresses the question of whether class skew really is defined and determinable. For example, the proportion of buildings, trees or other features in satellite imagery will vary depending on the location of the image. Besides, the number of buildings or trees that cover an area of the planet's surface could be temporal and therefore skew is not constant. Assuming that it was, could it be quantified? And how can a detector's performance be accurately evaluated without this knowledge? Granted that in this example the skew may only change fractionally over the whole planet (but more profoundly on the local level). A more dramatic example is that of monitoring bacteria growth, the skew in this case changes exponentially with time, $\pi(t) = 2\pi(t-1)$, during the second growth phase (Zwietering et al, 1990). There are numerous other examples of domains that exhibit temporal skew from a variety of disciplines, for example road vehicle detection (Sun et al, 2006), sleep stage identification (Sieracki et al, 1985), landslide detection (Stumpf et al, 2012a) and others (Ahanotu, 1999; Fell et al, 1996; Lampert and O'Keefe, 2013; Papageorgiou, 1999; Pavlo et al, 2012; Zink and Kern Reeve, 2005), and this research should be of interest to classical and reinforcement learning paradigms alike.

With this in mind, problems can be categorised into four subsets:

1) those with known, fixed skew characteristics (most simulated and 'toy' examples fall into this category);
2) those with known, temporal skew characteristics (the bacteria example may fall into this category);
3) those with fixed skew that is indeterminable (precisely what proportion of all eye fundus images do blood vessels comprise?);
4) those with temporal skew that is indeterminable (such as the tree and building detection example above) and possibly chaotic (for example detecting red cars in road surveillance video or detecting fissures in aerial imagery, this example will be explored further in the following section).

This paper will deal with the implications of evaluating algorithms within each of these categories using real-world examples, thus emphasising the consequences of assuming incorrect skew characteristics. To clarify, the term indeterminable is used in a loose sense, that is to say that these figures could be found if all data were available, however, with current technology this doesn't seem likely (and even so, perhaps these measurements are fractal in nature and cannot be known to arbitrary precision). Furthermore, novel performance measures that give more accurate estimations of an algorithm's performance within each of the above-mentioned cases are proposed.

2 Background

As was briefly discussed in the introduction the overall accuracy, $[\text{TP}(\theta)+\text{TN}(\theta)]/N$ (where $\text{TP}(\theta)$ and $\text{TN}(\theta)$ are the number of true positive and true negative detections at threshold θ and N is the dataset's size) is known to be inadequate when a dataset is highly skewed (He and Garcia, 2009).

A common alternative is to vary the classification threshold and measure algorithm performance using receiver operating characteristic (ROC) curves. These extract the true positive rate (TPR) and false positive rate (FPR) from multiple confusion matrices that are formed as the threshold varies, and thus form curves in the resulting two-dimensional performance space. Recent literature points out that these may overestimate performance when class balances are not equal (He and Garcia, 2009) (other conditions may be encountered that reduce their effectiveness and Webb and Ting (2005) expound some of these). Instead precision-recall (P-R) curves are preferable (Davis and Goadrich, 2006), which extract precision and recall (also called true positive rate, TPR, sensitivity, and hit rate) from the confusion matrices.

Precision (also called positive predictive value) is sensitive to the ratio between the number of positive ($N_p = \text{TP}(\theta) + \text{FN}(\theta)$, where FN is the number of false negative detections) and negative instances ($N_n = \text{FP}(\theta) + \text{TN}(\theta)$, where FP is the number of false positive detections) in the dataset. This ratio is defined such that $\phi = N_p/N_n$ or alternatively $\pi = (\phi N_n)/N = N_p/N$ and is referred to as the dataset's skew. As the proportion of negative instances increases, precision decreases. Analysing its definition reveals the cause:

$$P(\theta) = \frac{\text{TP}(\theta)}{\text{TP}(\theta) + \text{FP}(\theta)}, \tag{1}$$

the precision measure is normalised by the summation of the number of TP and the number of FP detections. These values being derived from two different classes entails that the skew of the dataset affects the denominator.

Under the assumption that the positive samples are drawn from the same distribution, and that the negative samples also all originate from the same distribution, recall is however skew insensitive. It is defined such that

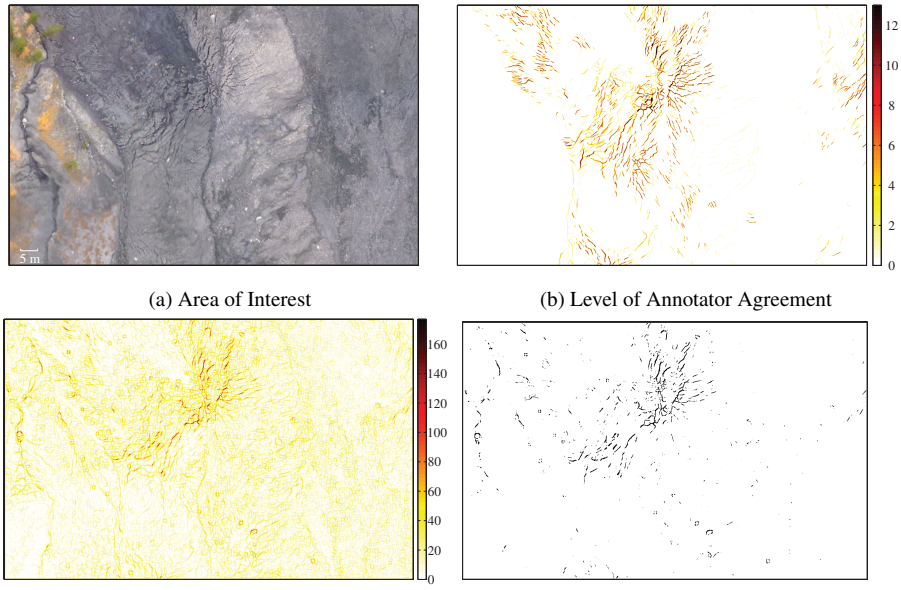
$$R(\theta) = \frac{\text{TP}(\theta)}{\text{TP}(\theta) + \text{FN}(\theta)}. \tag{2}$$

As such, recall is calculated within the positive class and is therefore not dependent upon its relative size.

An important contribution towards defining a performance measure that correctly represents skew has been made by Flach (2003), which is to analytically vary skew to produce a precision-recall surface. Taking a specific dataset skew reveals a slice of this surface. Landgrebe et al. (2006) extend this to integrate the measure with respect to skew. The work presented henceforth follows the examples set by these authors, but introduces novel solutions that remove some mathematical limitations, formally proves the measures' properties, and discusses the implications of such work.

3 Motivation

To motivate this work a typical problem of feature classification is taken, in which an algorithm is evaluated upon its ability to classify pixels of the aerial image presented in Figure



(a) Area of Interest

(b) Level of Annotator Agreement

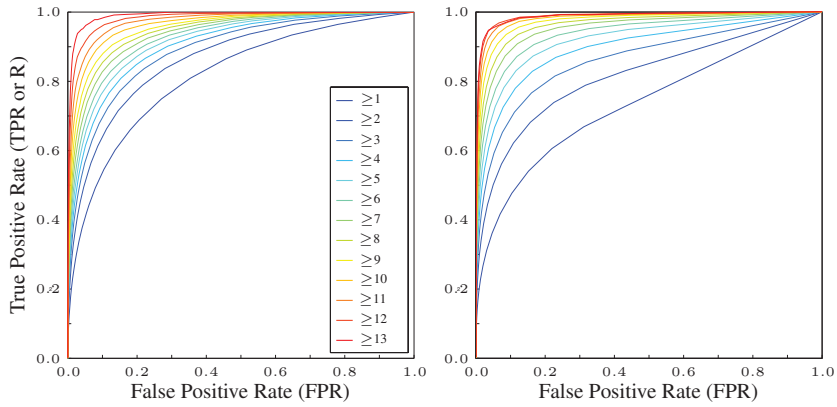
(c) Single-Scale Gaussian Level of Response

(d) Thresholded Output ($\theta = 50$)

Fig. 1: Annotator agreement of fissure locations and detector outputs. The level of colour in (b) represents the amount of agreement in that location. The level of colour in (c) represents the strength of the detector's response to the AOI presented in (a). The result of thresholding this response at threshold $\theta = 50$ is presented in (d).

1a as constituting a fissure or not (Stumpf et al, 2012b) (this problem falls into the fourth category outlined in the introduction in that the number of fissures is temporal in nature and indeterminable). The agreement of thirteen annotators in the location of these fissures is shown in Figure 1b.

The classifiers output soft class memberships, for example see Figure 1c, and therefore ROC curves can be evaluated for each of the two detection strategies used in this experiment, a Gaussian matched filter (Gauss) and Centre-Surround (C-S) filter (Stumpf et al, 2012b), for each level of agreement. This is achieved by thresholding the soft class memberships at a range of thresholds $0 \leq \theta \leq M$ (where M is the maximum of the detector's response) and then counting the number of pixels that are marked as positive by both the detector and the ground truth (GT), TP, those that are not marked by either, TN, and those marked by one and not the other, FP and FN. An example of a thresholded detector response is presented in Figure 1d, and the resulting ROC curves in Figure 2. The GTs are calculated by applying progressively increasing thresholds to the annotator agreement and thus the GTs range from the pixels that one or more annotators agreed upon, ≥ 1, to those that all annotators agreed upon, ≥ 13. It is clear that as agreement increases, so does the performance of the detector. This is congruent with the strong correlations between image features (intensity and contrast) and the annotators' agreement that are present within the data (these were measured to be -0.2245, the fissures are dark, and 0.4027 respectively). This is further exemplified in Table 1, in which the level of correlation between the detectors' outputs and annotator agreement is presented; again, high correlations are found.



(a) Single-Scale Gaussian (b) Centre-Surround

Fig. 2: Receiver Operating Curves describing the detectors' performance using different levels of agreement as the GTs.

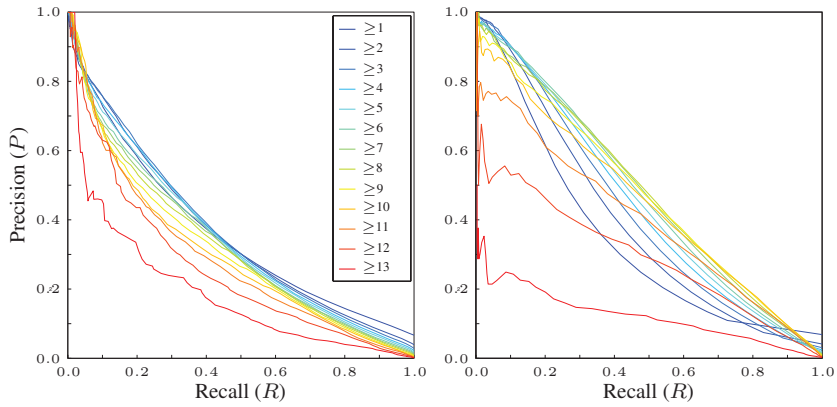
Table 1: Pearson's r Correlation coefficients between detector outputs and the annotator agreement; CCO (correlation coefficient — object) is calculated between annotator agreement and detector output within the pixels marked as a fissure by the annotators, and CCI (correlation coefficient — image) the whole image. The significance of each correlation was tested to 99% confidence and each p-value was found to be 0.0000 (to four decimal places).

Detector	CCO	CCI
Gaussian	0.5293	0.4711
C-S	0.6387	0.5259

Recent research has demonstrated that ROC curves can overestimate a detector's performance when the problem is unbalanced (He and Garcia, 2009). Instead P-R curves are preferred (Davis and Goadrich, 2006). Those derived in this problem are presented in Figure 3. Interestingly, these curves directly contradict that which has up until now been presented, as annotator agreement increases the performance (in terms of P-R curves) decreases!

As was discussed in the background section, the precision measure is dependent upon the skew of the data. In this experiment each level of agreement also exhibits a different level of skew and it is this that masks the true relationship hinted at by the correlation values. Furthermore, these performance figures are only valid for the skews from which each of the curves were determined.

Arbelaéz et al. provide an additional example within the problem of image segmentation and contour detection: increasing an image's resolution quadratically increases the prevalence of negative instances, but linearly increases the prevalence of segmentation borders or contours (being one dimensional entities) (Arbelaéz et al, 2011). To highlight the issue, suppose we apply the same algorithm to datasets having different skews. The resulting confusion matrices (for a given θ) are presented in Table 2. One can notice that in all the cases the true positive rate and false positive rate are equal, at TPR $= 0.8$ and FPR $= 0.3$,



(a) Single-Scale Gaussian (b) Centre-Surround

Fig. 3: Precision-Recall curves describing the detectors' performance using different levels of agreement as the GTs.

Table 2: Hypothetical confusion matrices simulating an algorithm's application to skewed datasets, in which the TPR = 0.8, FPR = 0.3 and $\pi = N_p/N$.

	a) $\pi = 0.09$		b) $\pi = 0.25$		c) $\pi = 0.51$		d) $\pi = 0.91$	
	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$
$+$	73	276	200	228	408	150	735	27
$-$	18	643	50	532	102	350	184	64
$P(\theta)$	0.21		0.47		0.73		0.97	

while precision varies. Although these values represent the same algorithm they report very different performances, making comparison between studies difficult. This is because they represent the algorithm's performance applied to a specific dataset. Moreover, the skew of a dataset may not be representative of the problem (the result of over or under sampling, or temporal skew, for example), resulting in an inaccurate representation of an algorithm's performance within a problem domain.

4 Measuring Performance under Different Skew Conditions

This section analyses each of the skew characteristics presented in the introduction and presents findings that help to accurately measure an algorithm's performance when applied to each.

4.1 Known Fixed Skew (First Case)

To allow for accurate performance modelling and for the comparison between P-R curves (and consequently precision values) in problems in which the skew is known and precise but has been altered, the performance differences due to skew need to be corrected. Flach (2003) defined this problem with respect to a 3D ROC space in which the third dimension

is a function of skew: "models may have been evaluated on different test sets with different class distributions, hence these points may be located in different horizontal planes [in a third dimension representing skew]." Boyd et al. (2012) describe several techniques that may result in an altered skew: aggregation for cross-validation (particularly in relational domains), aggregation amongst different tasks, and over and under sampling (He and Garcia, 2009). Several domains that are affected are: remote sensing, context recognition (Stäger et al, 2006), network intrusion, medical diagnostics, and image analysis.

We propose the following relationship to facilitate the transformation of the P-R curve's constituent precision values, which have been calculated using a dataset that has a known skew of π, into the value of precision that would have been found if the dataset had had some different skew π' (from now on a $'$ will denote a simulated variable and its absence a variable inferred using a dataset). In order for the following relationship to be valid (along with the other results presented in this paper) it is necessary to assume that all positive samples are drawn from the same distribution, and that all negative samples also originate from the same distribution. Therefore, the only difference between two differently skewed datasets is the ratio of positive and negative samples that they include.

Theorem 1 *The equation that dictates precision's transformation from one skew to another is*

$$P_{\pi'}(\theta) = \frac{\pi'}{\pi' + (1-\pi')\frac{\pi}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]}. \tag{3}$$

Proof. Taking the definition of precision, Eq. (1) and that $\text{TP}(\theta)$ is derived from recall as $\text{TP}(\theta) = R(\theta)\pi N$

$$P_\pi(\theta) = \frac{R(\theta)\pi N}{R(\theta)\pi N + \text{FP}(\theta)}.$$

Furthermore $\text{FP}(\theta)$ can be defined in terms of the FPR such that $\text{FP}(\theta) = (1-\pi)N\text{FPR}(\theta)$. Since the FPR of an algorithm is invariant to class skew under the assumption outlined previously, it can be defined in terms of precision and recall sampled at the original skew, π, such that

$$\text{FPR}(\theta) = \frac{\frac{N\pi R(\theta)}{P_\pi(\theta)} - N\pi R(\theta)}{N(1-\pi)} = \frac{\pi R(\theta)}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]$$

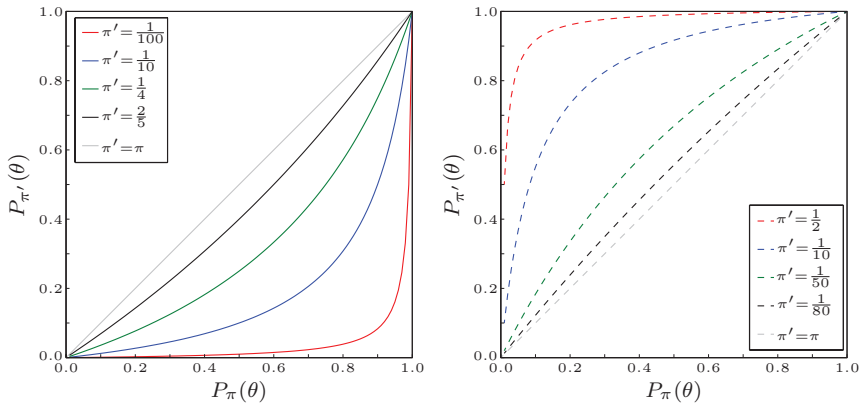
substituting this gives

$$P_{\pi'}(\theta) = \frac{\pi' R(\theta)}{\pi' R(\theta) + (1-\pi')\frac{\pi R(\theta)}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]} = \frac{\pi'}{\pi' + (1-\pi')\frac{\pi}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]}.$$

□

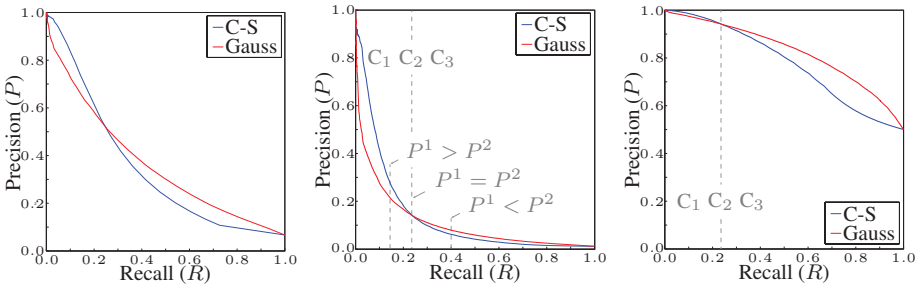
An illustration of the transformation function defined by Equation (3) is presented in Figure 4: Figure 4a represents the transformation of precision calculated at a skew of $\pi = 0.01$ to a number of other skews ($\pi' = 1/2, 1/10, 1/50, 1/80$) and Figure 4b those that were calculated at a skew of $\pi = 0.5$ to various other skews ($\pi' = 2/5, 1/4, 1/10, 1/100$). It is clear that the effect is amplified as the skew ratio π/π' deviates further from one.

To demonstrate this transformation in application, the P-R curves of the two detectors evaluated using the ≥ 1 agreement GT are extracted from Figures 3a and 3b. These are presented together in Figure 5a and the skew in this dataset is $\pi = 0.0667$ (to four decimal places). In Figures 5b and 5c the precision values have been transformed using Equation (3) to values that would have resulted had the dataset had skews of $\pi' = 0.01$ and $\pi' = 0.5$



(a) $\pi = 0.5$ (b) $\pi = 0.01$

Fig. 4: Skew transform function defined by Equation (3), applied to precision calculated upon a dataset with a skew of $\pi = 0.5$, 4b, and a skew of $\pi = 0.01$, 4b.



(a) Original curves extracted from Figure 3 (GT ≥ 1)

(b) $\pi' = 0.01$: $\text{AUCPR}_{\pi}^{\text{CS}} > \text{AUCPR}_{\pi}^{\text{GA}}$

(c) $\pi' = 0.5$: $\text{AUCPR}_{\pi}^{\text{CS}} < \text{AUCPR}_{\pi}^{\text{GA}}$

Fig. 5: Precision-Recall curves describing two algorithms' performances at different skews (and cases from the proof of Thm. 2 in App. A). The AUCs (to four decimal places) are: (a) C-S = 0.3384 and Gauss = 0.3607; (b) C-S = 0.1307 and Gauss = 0.1088; and (c) C-S = 0.7772 and Gauss = 0.8260.

and therefore each P-R curve has a different appearance (for the sake of this example the problem's skew has been assumed to be fixed, although in reality the problem belongs to the fourth skew category outlined in the introduction).

An additional property of P-R curves is also illustrated in this example: the ranking of the two detectors at any particular value of recall remains constant irrespective of skew. In Figures 5b and 5c three different rankings are observed depicted by the labels C_1, C_2 and C_3. We can make a general statement regarding this property:

Theorem 2 *Two points in P-R space that share the same recall, (R_*, P^1) and (R_*, P^2), have the same ranking (in terms of precision) irrespective of skew.*

The proof of this theorem is presented in Appendix A.

4.1.1 Area Under the Precision-Recall Curve (AUCPR)

The effects of a mismatch between skews upon aggregate functions, such as the AUCPR, are now discussed. A number of works explicitly optimise an algorithm with respect to its AUCPR as it offers a summary statistic for algorithm comparison.

Returning to Figures 5b and 5c, the AUCPR of the C-S detector (0.1307) is greater than that of the Gaussian detector (0.1088) when $\pi' = 0.01$; however, when the class balance is equal ($\pi' = 0.5$) the ranking is inverted (C-S AUCPR is 0.7772 and the Gaussian AUCPR is 0.8260). As such, if the choice had been made to prefer the C-S detector due to the test dataset having a skew of $\pi = 0.01$ but in application the skew of the problem was more balanced, a sub-optimal algorithm would have been chosen (the Gaussian detector would have been the optimal choice in this case). In the same manner, sub-optimal parameter choices may be made if the skew of a training dataset is not representative of the problem's skew (as we have seen, however, Equation (3) offers a means to circumvent this problem).

More formally, if two P-R curves cross then there may exist a dataset skew, $\pi_* \in [0, 1]$, at which the rank of the algorithms calculated on either side of this value inverts, i.e. $\mathrm{AUCPR}^1_\pi > \mathrm{AUCPR}^2_\pi$ when $\pi > \pi_*$ but $\mathrm{AUCPR}^1_\pi < \mathrm{AUCPR}^2_\pi$ when $\pi < \pi_*$. If, however, two P-R curves do not cross then the AUCPR ranking remains constant:

Corollary 1 *It directly follows from Theorem 2 that if two P-R curves do not cross, i.e. $P^1(R) \geq P^2(R)$ for all R, the ranking, measured as AUCPR, remains constant irrespective of the dataset's skew, such that* $\mathrm{AUCPR}^1 \geq \mathrm{AUCPR}^2$.

Using these findings a test can be defined to determine whether the rank of the AUC of two P-R curves inverts and, if so, Algorithm 1 can be employed to find the value of skew, π_*, at which it does. Equation (3) enables us to transform a P-R curve sampled at skew π to both extremes of the skew range, $\pi_l = \epsilon$ and $\pi_h = 1 - \epsilon$, where ϵ is a sufficiently small non-zero number (when $\pi = 0$ and $\pi = 1$ all non-trivial precisions are equal so adding and subtracting a small value avoids these cases). By calculating the AUCs of the P-R curves transformed to these extremes it can be determined whether an inversion has taken place in the range $\pi_l \leq \pi_* \leq \pi_h$. If so, the binary search defined by Algorithm 1 allows the skew π_*, at which the performance inverts, to be found. If the P-R curves cross more than once, however, the ranking may invert more than once, and therefore a linear search should instead be used (the number of inversions is equal to, or less than, the number of crossings in $0 < R < 1$).

In this manner, the skew at which the rank of the Gaussian and C-S (fissure) detectors inverts was found to be $\pi_* = 0.0329678444$ (to ten decimal places). Of course, the number of decimal places to which π_* is relevant (x in Algorithm 1) is limited by the number of TP, FP and FN detections and therefore the size of the dataset.

4.2 Known Temporal Skew (Second Case)

The previous subsection assumed that a problem's skew is fixed to some definable value, π'. Nevertheless, there are occasions in which skew may be temporal such that its behaviour over time t is defined by a function $\pi'(t)$. In this scenario Eq. (3) can be used to find the instantaneous value of precision at any point during skew's evolution by setting π' to be equal to $\pi'(t_i)$, where $0 < t_i \leq T$ is the time point associated with the desired skew.

Each of these points, however, do not reflect the algorithm's mean precision over the whole temporal evolution of skew. To achieve this Equation (3) should be extended to inte-

Algorithm 1 skew_search
Input: $P^1(\theta)$, $R^1(\theta)$, $P^2(\theta)$, $R^2(\theta)$, π'_l and π'_h
Output: π_*

$\pi'_m \leftarrow (\pi'_l + \pi'_h)/2$
if $\lfloor 10^x \pi'_l \rfloor = \lfloor 10^x \pi'_h \rfloor$ **then**
  {point of change found to x decimal places}
  return $\pi_* \leftarrow \pi'_m$
end if
if $\mathrm{sgn}(\mathrm{AUCPR}^1_{\pi'_m} - \mathrm{AUCPR}^2_{\pi'_m}) \neq \mathrm{sgn}(\mathrm{AUCPR}^1_{\pi'_h} - \mathrm{AUCPR}^2_{\pi'_h})$ **then**
  $\pi'_m \leftarrow \mathrm{skew_search}(P^1(\theta), R^1(\theta), P^2(\theta), R^2(\theta), \pi'_m, \pi'_h)$ {change in upper half}
else if $\mathrm{sgn}(\mathrm{AUCPR}^1_{\pi'_l} - \mathrm{AUCPR}^2_{\pi'_l}) \neq \mathrm{sgn}(\mathrm{AUCPR}^1_{\pi'_m} - \mathrm{AUCPR}^2_{\pi'_m})$ **then**
  $\pi'_m \leftarrow \mathrm{skew_search}(P^1(\theta), R^1(\theta), P^2(\theta), R^2(\theta), \pi'_l, \pi'_m)$ {change in lower half}
end if
return $\pi_* \leftarrow \pi'_m$ {point of change found}

grate over the function $\pi'(t)$, which collapses this added dimension into a single value, such that

$$\tilde{P}_T(\theta) = \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]}\,\mathrm{d}t \tag{4}$$

in which $\pi'(t)$ is an integrable function over the interval $[0, T]$ (however, taking a discrete summation to approximate the integration relaxes this condition) that describes skew's temporal evolution, and T is the time frame under evaluation. A number of properties of P-R space (namely: interpolation, AUCPR, unachievable P-R space, and random classifier performance) are also demonstrable within $\tilde{\text{P}}$-R space and formal proofs are given in Appendix B. As when determining performance using P-R curves, the unachievable area (Boyd et al, 2012) of $\tilde{\text{P}}$-R space should be calculated along with the $\tilde{\text{P}}$-R curve, which is possible using the results presented in Appendix B.3.

This modification allows the measure to report the mean precision value corrected with respect to any temporal characteristics of skew. Therefore avoiding the problem of measuring precision with respect to one fixed skew in a problem in which skew changes (which would result in an over- or underestimation of overall precision). It also removes the dependence upon the dataset's skew that is present in the original precision measure, which is beneficial in situations in which the dataset's skew does not reflect the problem's skew characteristics. Integrating with respect to the function $\pi'(t)$ results in a measure that is the least-squares fitting to all instantaneous precisions (as calculated using Eq. (3)) that are described by $\pi'(t)$—it therefore accurately reflects the overall value of precision.

A more detailed picture of the evolution of mean precision in a problem that is affected by temporal skew can be determined by varying T. As such, snapshots of the mean precision within the time range $0 < t \leq T$ can be obtained.

Of course, this method of measuring performance will not be desirable when precision is required at specific time points and in these situations the measure presented in the previous subsection can instead be used.

As an illustration, take the example of detecting bacteria in a petri dish during the second growth phase (this is an oversimplification for demonstration purposes, more information on

the characteristics of bacteria population growth is presented by Zwietering et al. (1990)). A hypothetical detector exists that has an FPR of 0.2 and a TPR of 0.8 and there is a hypothetical dataset such that $N = 10000$, $N_p = 4000$, and $N_n = 6000$. Each instance in the dataset is a pixel's intensity value, and each bacterium is the size of one pixel. If we were to assume that skew is fixed according to the described dataset, therefore $\pi = 0.4$, then precision would be calculated to be 0.7273. Alternatively, if the skew's temporal evolution is accounted for using Eq. (4), such that $T = 20$, $\pi'(0) = 1/N$, and $\pi'(t) = 2^t\pi'(0)$ (if $\pi'(t) > 1$ then $\pi'(t) = 1$), precision is calculated to be 0.4689—significantly lower but more accurately reflecting the algorithm's performance within the domain.

4.3 Fixed and Temporal but Indeterminable Skew (Third and Fourth Case)

The previous two subsections discussed the classical view of P-R curves in which it is assumed that a problem's skew is well defined. It was discussed in the introduction, however, that for some problems skew is not the concrete or determinable phenomenon that it is often assumed to be. An alternative view was proposed in which skew may be unknown whilst also possibly being temporal. This section proposes a measure that accounts for this uncertainty whilst preserving the property introduced in Section 4.1, specifically the ability to compare P-R curves that have been determined upon datasets with different skews.

4.3.1 Related Work

Flach (2003) and Landgrebe et al. (2006) propose modifications to the precision measure aiming to better represent a problem's underlying distribution. Flach (2003) normalises the quantities $\text{TP}(\theta)$ and $\text{FP}(\theta)$ by N_p and N_n (respectively), thus replacing them with the TPR and FPR:

$$\check{P}(\theta) = \frac{\text{TPR}(\theta)}{\text{TPR}(\theta) + \lambda'\text{FPR}(\theta)} \tag{5}$$

where $\lambda' = N_n/N_p = \pi^{-1} - 1$ is the negative to positive ratio. Therefore $\lambda' < 1$ implies that positive instances are more important and $\lambda' > 1$ the opposite (Flach, 2003). Normalising the number of TP and FP detections by the size of their respective classes removes any dependence upon the skew of a dataset. This measure only allows for the analytical variance of skew however and does not collapse this third dimension into a measure that represents performance over a range of skews.

Landgrebe et al. (2006) propose an AUC measure that integrates along this third dimension of skew and therefore an alternative form of precision can be inferred from it, such that

$$\hat{P}(\theta) = \int_{\lambda'_1}^{\lambda'_2} \frac{\text{TPR}(\theta)}{\text{TPR}(\theta) + \lambda'\text{FPR}(\theta)}\, \mathrm{d}\lambda'. \tag{6}$$

To simulate a dataset that contains only negative instances $\text{TPR}'(\theta)$ must equal zero for any threshold θ, and for this to take place the quantity λ' must tend towards infinity. This means that the range $0 \leq \lambda' < 1$ represents higher positive-to-negative skews and the range $1 < \lambda' \leq \infty$ represents higher negative-to-positive skews. Moreover, the quantity λ is undefined when $\pi = 0$.

The form of Equation (6) implicitly gives higher weight to lower values of λ. This complicates the measure's interpretation as the weighting function is not explicitly defined. Furthermore it imposes an assumption upon the distribution and relative importance of skew

values that may not be valid. In the case that a temporal evolution of skew is known (and therefore their relative importances), the measure outlined in Section 4.2 may be used—those skews that are more likely, occur more often in $\pi'(t)$ and therefore are given higher weight. There are examples however, in which the relative importance of different skews is unknown and/or equal. This subsection deals with the cases in which skew is indeterminable (in either a temporal or fixed situation). For example, in the fissure detection problem that motivated this work, the problem's skew can vary within a reasonable range of skews. Its temporal pattern, however, is unknown and therefore the weighting of specific skews cannot be inferred (and may not even be a valid assumption to make as the factors that influence the prevalence of fissures are many and complicated) and thus the weighting, w, of each skew is equal, i.e. $w(\pi') = 1$ for all π' (cf. the following subsection).

4.3.2 The Proposed Form

Instead, using Equation (3) we can derive an integrated precision measure in terms of TPR and FPR that avoids these issues, such that

Theorem 3

$$\bar{P}(\theta) = \frac{1}{\pi_2' - \pi_1'} \int_{\pi_1'}^{\pi_2'} \frac{\pi' \mathit{TPR}(\theta)}{\pi' \mathit{TPR}(\theta) + (1-\pi')\mathit{FPR}(\theta)} \, \mathrm{d}\pi'. \tag{7}$$

where $\pi_1', \pi_2' \in \mathbb{R}$, such that $0 \leq \pi_1' < \pi_2' < 1$, are the lower and upper bounds on skew. This measure is also normalised to lie within precision's original range, $\bar{P}(\theta) \in [0, 1]$.

Proof. From the proof of Theorem 1 we have

$$P_{\pi'}(\theta) = \frac{\pi'}{\pi' + (1-\pi')\frac{\pi}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]} = \frac{\pi' R(\theta)}{\pi' R(\theta) + (1-\pi')\frac{\pi R(\theta)}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]} \tag{8}$$

and using the equivalences also shown in the proof of Theorem 1 this can be expressed as

$$P_{\pi'}(\theta) = \frac{\pi' \mathrm{TPR}(\theta)}{\pi' \mathrm{TPR}(\theta) + (1-\pi')\mathrm{FPR}(\theta)}. \tag{9}$$

Finally, introducing the integration to account for skew uncertainty gives

$$\bar{P}(\theta) = \frac{1}{\pi_2' - \pi_1'} \int_{\pi_1'}^{\pi_2'} \frac{\pi' \mathrm{TPR}(\theta)}{\pi' \mathrm{TPR}(\theta) + (1-\pi')\mathrm{FPR}(\theta)} \, \mathrm{d}\pi'.$$

□

As such, Equation (7) overcomes the limitations outlined in Section 4.3.1 as follows:

- the terms TPR and FPR are values that are independent of skew (and represent a balanced dataset);
- by weighting the term TPR in addition to FPR it is possible to represent any skew analytically (provided that $\mathrm{TPR}(\theta) + \mathrm{FPR}(\theta) > 0$);
- the integration gives equal weighting to the more negative-than-positive, $0.0 \leq \pi' < 0.5$, and more positive-than-negative, $0.5 < \pi' \leq 1.0$, ranges.

These properties are formally proved in Appendix C, Theorems 6 and 7.

As a consequence, an algorithm's performance when applied to a problem with imprecise skew can be accurately represented. Furthermore, comparisons between integrated precision-recall ($\bar{\text{P}}$-R) curves determined upon different datasets, or upon the same dataset using different ground truths (each having different class skews) can be made (provided that they are determined using the same limits). As was presented in relation to $\tilde{\text{P}}$-R space the following properties are demonstrable within $\bar{\text{P}}$-R space and formal proofs are given in Appendix C: interpolation, AUCPR, unachievable P-R space, and random classifier performance. Furthermore, the unachievable area of $\bar{\text{P}}$-R space can be determined using the results presented in Appendix C.3.

As has been discussed, the proposed measure gives equal weight to each point of skew due to the fact that in cases three and four presented in the introduction it is assumed that no information exists to suggest otherwise. To increase generality however, Eq. (7) can be modified to account for situations in which the relative weighting of skews are known. By introducing a weighting function, $w(\pi')$, such that

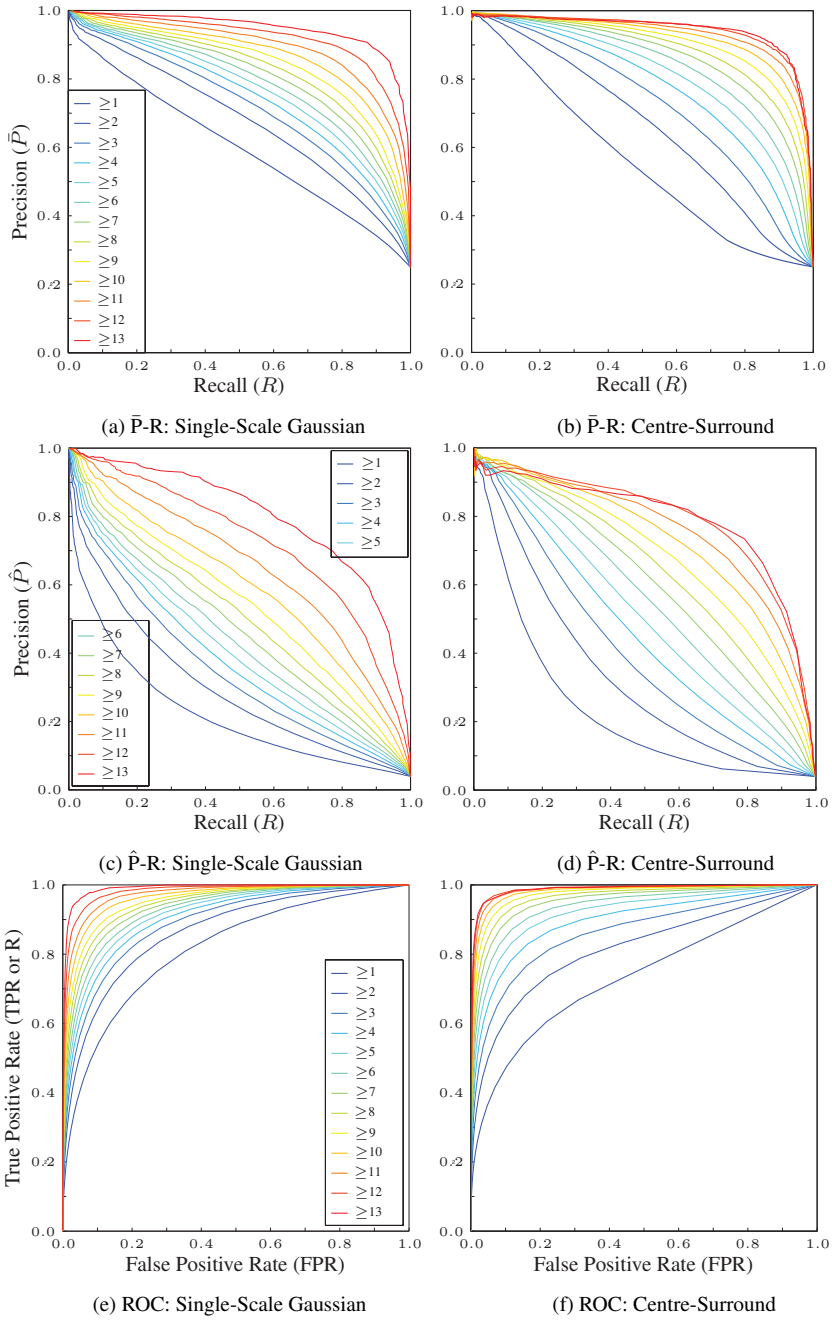
$$\bar{P}(\theta) = \frac{1}{\pi'_2 - \pi'_1} \int_{\pi'_1}^{\pi'_2} \frac{w(\pi')\pi'\text{TPR}(\theta)}{\pi'\text{TPR}(\theta) + (1-\pi')\text{FPR}(\theta)} \, \mathrm{d}\pi'$$

where $w(\pi')$ is a positive function defined on the interval $[0, 1]$ that gives weight to each skew such that $\int_{\pi_1}^{\pi_2} w(\pi') \, \mathrm{d}\pi' = 1$, this is achieved. Furthermore, except for the analytical representation of the minimum achievable AUC presented in Eq. (20) of Section C.3 and the performance of a random classifier, the findings presented in Appendix C are also valid assuming that the weighting function is appropriately introduced. Trapezoidal integration can be used to calculate the minimum achievable AUC$\bar{\text{P}}$R and the function $\bar{P}_{RC}(\theta) = 1/(\pi'_2 - \pi'_1) \int_{\pi'_1}^{\pi'_2} w(\pi')\pi' \, \mathrm{d}\pi'$ describes the performance of a random classifier when a weighting function $w(\pi')$ is employed. Expressed in this way, Langrebe et al.'s implicit weighting in Eq. (6) can be defined as $w(\pi') = 1/\pi'^2$.

In some situations it is desirable to ignore the higher positive-to-negative skew ranges ($0.5 < \pi' < 1$) and the proposed measure gives full flexibility to satisfy this requirement by leaving the selection of the skew range to the practitioner's discretion.

5 Discussion

Revisiting the example used to motivate this work (see Section 3) whilst applying the advances presented in the preceding sections produces $\bar{\text{P}}$-R curves such as those presented in Figures 6a and 6b. Each curve is calculated using a different GT and therefore each is derived from data having different skews. The bounds were determined by expert opinion due to limited data and were $\pi'_1 = 0.0$ and $\pi'_2 = 0.5$ (a method for experimentally estimating these will be discussed next), these values represent the variance of skew that arises from its temporal and uncertain nature. As no further information of the relative frequencies of these skews is available and the factors that influence skew in this problem are so many, each is given equal weighting in the evaluation (this will be discussed further in the following text). The $\bar{\text{P}}$-R curves are now comparable as the influence of skew has been normalised and as such the measured performances remain faithful to the correlations observed between the annotators' opinions and the detectors' outputs, see Table 1. The difference between the performance of each detector is smaller when compared to the P-R curves presented in Figure 3, when the variance of skew is considered. For example the difference between the



(a) $\bar{\text{P}}$-R: Single-Scale Gaussian

(b) $\bar{\text{P}}$-R: Centre-Surround

(c) $\hat{\text{P}}$-R: Single-Scale Gaussian

(d) $\hat{\text{P}}$-R: Centre-Surround

(e) ROC: Single-Scale Gaussian

(f) ROC: Centre-Surround

Fig. 6: $\bar{\text{P}}$recision-Recall, ROC, and $\hat{\text{P}}$recision-Recall curves describing the detectors' performance using GTs of different agreement levels, $\pi'_1 = 0.0$ and $\pi'_2 = 0.5$, and $\lambda'_1 = 1$ and $\lambda'_2 = 100$ (which are equivalent to $\pi'_1 = 0.0099$, to four decimal places, and $\pi'_2 = 0.5$).

Table 3: The area under curve measurements of the P-R and $\bar{\text{P}}$-R curves presented in Figures 3a and 6a respectively, in addition to their values normalised with respect to the unachievable AUC (Boyd et al, 2012).

Level of Agreement	AUCPR	$\text{AUCPR} - \text{AUCPR}_{\text{MIN}}$	$\frac{\text{AUCPR} - \text{AUCPR}_{\text{MIN}}}{1 - \text{AUCPR}_{\text{MIN}}}$	$\text{AUC}\bar{\text{P}}\text{R}$	$\text{AUC}\bar{\text{P}}\text{R} - \text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}$	$\frac{\text{AUC}\bar{\text{P}}\text{R} - \text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}}{1 - \text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}}$
≥ 1	0.3615	0.3274	0.3390	0.5996	0.4573	0.5332
≥ 2	0.3633	0.3434	0.3504	0.6760	0.5337	0.6223
≥ 3	0.3635	0.3490	0.3541	0.7159	0.5736	0.6688
≥ 4	0.3561	0.3450	0.3489	0.7422	0.5998	0.6994
≥ 5	0.3476	0.3388	0.3418	0.7639	0.6215	0.7247
≥ 6	0.3381	0.3311	0.3334	0.7834	0.6410	0.7474
≥ 7	0.3329	0.3274	0.3292	0.8043	0.6620	0.7719
≥ 8	0.3253	0.3210	0.3224	0.8239	0.6816	0.7947
≥ 9	0.3140	0.3108	0.3118	0.8437	0.7014	0.8178
≥ 10	0.3024	0.3001	0.3008	0.8675	0.7251	0.8455
≥ 11	0.2895	0.2881	0.2885	0.8920	0.7496	0.8741
≥ 12	0.2576	0.2569	0.2571	0.9187	0.7763	0.9052
≥ 13	0.1879	0.1877	0.1877	0.9454	0.8031	0.9364

AUCPRs of the C-S and Gauss algorithms when using the ≥ 5 and ≥ 8 GTs are 0.1042 and 0.1500 respectively but these are reduced to 0.0392 and 0.0569 when taking the proposed $\text{AUC}\bar{\text{P}}\text{R}$.

Boyd et al. (2012) propose to normalise the AUCPR measure with respect to the unachievable area. Table 3 presents the AUCs attributed to the P-R curves presented in Figure 3a normalised in this manner. It is clear that this method does not resolve the incorrect ordering that the P-R curves produce. The minimum achievable area of P-R space is not dependent upon the performance of the algorithm but on the skew of the dataset and therefore it cannot be used to correct such disparities. Instead, as was found in Figures 6a and 6b, the $\text{AUC}\bar{\text{P}}\text{R}$ reports the correct ordering (both with and without normalising with respect to the unachievable $\bar{\text{P}}$-R space).

To compare the measure to that inferred from Langrebe et al.'s AUC measure it is necessary to express the integration range as λ'. Since λ is undefined when $\pi = 0$ we use the range $\lambda'_1 = 1$ and $\lambda'_2 = 100$, which is equivalently expressed as $\pi'_1 = 0.0099$ (to four decimal places) and $\pi'_2 = 0.5$ (from one-hundred times as many negative samples as positive samples, to a balanced dataset). The results are presented in Figures 6c and 6d. It is noticeable that precision is lower for any given value of recall because the measure gives more weighting to lower skew values. Consequently, as λ'_2 increases (and equivalently $\pi'_1 \to 0$) the resulting curves are pushed towards the lower-left corner of $\hat{\text{P}}$-R space due to this implicit weighting (if λ'_2 is set to a sufficiently high value the curves are indistinguishable from the axes). This would be representative of the precision if it was the case that lower skews are more likely and therefore more important, however, this assumption cannot be said to be generally true.

Returning to the running example, the results of representing performance relative to the skew of the problem's domain (using $\bar{\mathrm{P}}$-R and $\tilde{\mathrm{P}}$-R curves) instead of the skew of the dataset or the normalised skew represented by ROC curves gives different insight into an algorithm's performance. This is illustrated by comparing Figures 6a and 6b with Figures 6e and 6f (for ease of reading these figures are repetitions of those presented in Figure 2), and Figure 3. For example, in Figure 6e the Gaussian detector appears to be an almost perfect detector when the highest agreement GT is under question (with an AUC of 0.9910), whereas Figure 6a reveals that there is still room for improvement (with an AUC of 0.9454). A larger spread between the curves that result from the highest and lowest agreement GTs is also observed in Figures 6a and 6b when compared to Figures 6e and 6f, giving a more detailed picture of the detector's limitations (the difference between the AUCs of the highest and lowest curves in Figure 6a is 0.3458 whereas in Figure 6e it is 0.1760, and 0.3853 in Figure 6b compared to 0.2543 in Figure 6f). Of course the extent of these differences are dependent upon two facts: the performance of the detector, and the difference in skew between the evaluation dataset and the integration limits chosen. In this example the detectors achieve state-of-the-art performance, in other situations the differences will be more profound. Figure 5 demonstrates how variable P-R curves are with respect to a dataset's skew, however, the ROC curves presented in Figures 6e and 6f would have been the same irrespective of the skew of the dataset from which they were calculated (assuming that the positive and negative samples are drawn from the same distributions, Webb and Ting 2005 discuss the behaviour of ROC curves when these assumptions are not valid).

Owing to the fact that ROC curves represent an algorithm's TPR and FPR they normalise the effects of skew and therefore report a performance that is primarily informative when the problem has a balanced skew. The work presented within this article removes this assumption, whilst admitting that a problem's skew characteristics are rarely concretely known (although offers tools for use when it is known but varies from the dataset used during evaluation) and therefore enables practitioners to obtain the expected, or average, curves for the range of skews that are identified. The reporting of this curve, however, smooths over any extreme skews that may be present in its temporal evolution, in the case of $\tilde{\mathrm{P}}$-R curves, or at either end of the range π_1' and π_2', in the case of $\bar{\mathrm{P}}$-R curves, and these may be of use to the practitioner. The work presented in Section 4.1 can be used to alleviate this limitation by allowing the practitioner to obtain an algorithm's performance curves at either extreme of the range of skews under question, in addition to its mean or expected curve, and thus model performance's variance.

As such, the findings presented in Section 4.1 and the $\bar{\mathrm{P}}$ and $\tilde{\mathrm{P}}$ measures give practitioners a more accurate summary of an algorithm's performance when applied to the different situations presented in the introduction. The term more accurate is used to imply that the measure is specific to the skew characteristics of a problem and not simply the dataset that has been selected for evaluation.

The non-linearity of the relationship between precisions calculated at different skews, Eq. (3) and Fig. 4, highlights the necessity of integrating over the range π_1' to π_2' or the function $\pi'(t)$. Simply calculating a definite precision at the middle of this range, $\pi' = (\pi_1' + \pi_2')/2$, for example, does not provide an estimate of the mean performance over the whole range (instead it results in an over- or under-estimation of precision). This also implies that imposing a fixed skew upon a problem in which the skew varies results in an incorrect performance estimate.

Table 4: Skew, $\pi = N_p/N$, found in machine learning datasets (to three decimal places). The top seven datasets relate the the domain of eye fundus images (see text).

Database	Skew
STARE (GT: VK) (Hoover et al, 2000)	0.109
STARE (GT: AH)	0.076
ARIA (GT: BDP)	0.184
ARIA (GT: BSS)	0.168
HEI-MED	0.090
DRIVE (GT: 1)	0.126
DRIVE (GT: 2)	0.123
Spectrogram (Lampert and O'Keefe, 2011)	0.006
UCI ML Adult	0.242
UCI ML Breast Cancer Wisconsin	0.373
UCI ML Poker Hand (Flush Class)	0.002
UCI ML Poker Hand (Nothing Class)	0.501
UCI ML Internet Advertisements	0.134
NASA Software Defect Dataset CM1	0.098
NASA Software Defect Dataset PC1	0.931

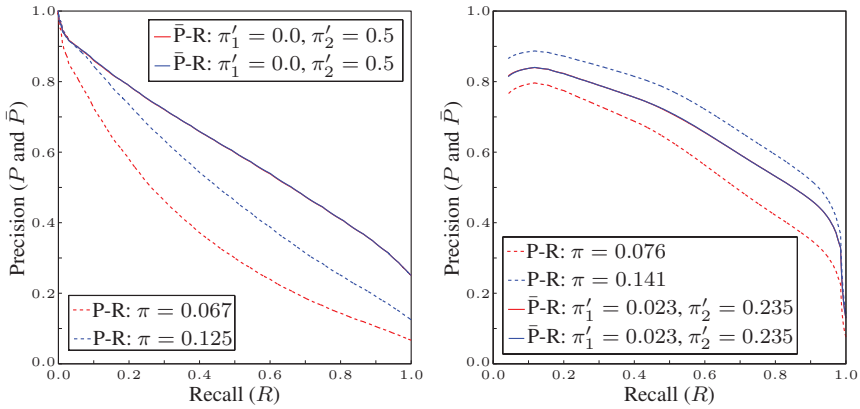
5.1 Estimating π'_1 and π'_2 from Data

It has been emphasised that the skew present in datasets not only varies between domains but also between datasets representing the same domain. It may therefore not be possible to know the problem's skew with certainty (the third category proposed in the introduction).

An empirical evaluation of the skews found in a number of datasets is presented in Table 4. Information of this sort allows us to estimate the variance of skews found within a problem when it cannot be analytically determined. It is reasonable to assume that the limits π'_1 and π'_2 of $\bar{P}$ are chosen such that $\pi'_1 = \bar{\pi} - 3\hat{\pi}$ and $\pi'_2 = \bar{\pi} + 3\hat{\pi}$ (under the condition that $0 \leq \pi'_1 < \pi'_2 < 1$), where $\bar{\pi}$ represents the mean and $\hat{\pi}$ the standard deviation, therefore making the measure representative of skew's variance (a value of three standard deviations is chosen such that approximately 99.73% of the skew's variance is represented, assuming that it is normally distributed). This enables accurate representation of an algorithm's performance applied to a domain and not a specific dataset.

To evaluate the fissure detector's performance thus far π'_1 and π'_2 have been estimated by an expert, now an additional example will be introduced to demonstrate the estimation of these limits using a large spread of data. For example, four datasets of eye fundus images (STARE, ARIA, HEI-MED, and DRIVE), which contain markings of retinal blood vessels, were analysed in such a way. The markings of each annotator were treated separately, forming seven skew estimations. The mean and standard deviation were found to be $\bar{\pi} = 0.129$ and $\hat{\pi} = 0.035$, and as such π'_1 and π'_2 would be chosen such that $\pi'_1 = 0.023$ and $\pi'_2 = 0.235$. Of course this kind of analysis may not always be possible, and in this case reasonable limits may be estimated (such as for the fissure detection problem).

This additional information may also lead to the application of an alternative skew weighting (as was discussed in Section 4.3.2). For example, assuming that the skew in eye fundus datasets is normally distributed, Eq. (7) can be modified as described in Section 4.3.2 and the weighting function $w(\pi') \sim \frac{1}{\hat{\pi}\sqrt{2\pi}} e^{-\frac{(\pi'-\bar{\pi})^2}{2\hat{\pi}^2}}$ used (the integration will then take



(a) Fissure Image (b) STARE Dataset

Fig. 7: P-R and P̄-R curves calculated upon the fissure image (using the Gaussian detector) and the STARE dataset (using the MLVessel v1.4 package) with original and an altered skews.

place within the range $\pi_1' = 0$ and $\pi_2' = 1$). For the following examples, however, we will continue to use the uniform weighting that is implicit in Eq. (7).

5.2 Further Applications

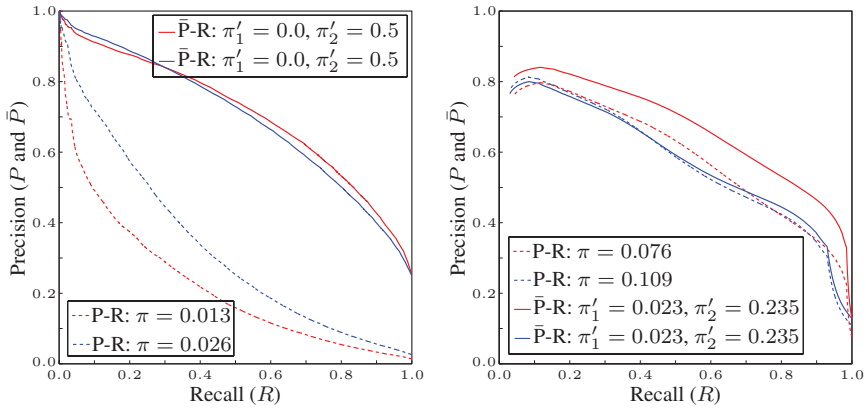
The following is now possible:

Optimisation: It was demonstrated in Section 4.1 that making use of a training set that does not represent the problem's skew can lead to suboptimal design choices. This is alleviated by explicitly modelling the skew range at which the algorithm is targeted.

Domain performance: By adjusting the skew (by removing negative instances) of the fissure image presented throughout this paper from 0.067 (using the ≥ 1 GT) to an to arbitrary value of 0.125 and of the STARE dataset from 0.076 (using AH GT) to an arbitrary value of 0.141, it is demonstrated in Figure 7 that the proposed measure allows the performance of an algorithm to be effectively modelled within its intended application domain, irrespective of the skew of the dataset upon which it was determined[1] (note that the P̄-R curves are equal and therefore appear as one).

Different ground truths: As has been demonstrated previously, skew's effect on precision prohibits comparison between different GTs. Figure 8 presents P-R curves derived from the Gaussian and MLVessel classifiers evaluated upon the fissure and STARE datasets (respectively) but using different GTs. The results from the fissure dataset (Figure 8a) demonstrate that the P-R curves indicate a large difference in performance when evaluating upon the two different GTs, however, when skew is normalised the P̄-R curves demonstrate that the detector models both annotators roughly equally. Inspecting the STARE results (Figure

[1] These results were obtained using the Gaussian detector applied to the fissure image and the MLVessel v1.4 package (Soares et al, 2006) applied to the STARE dataset. The effect of skew on precision diminishes as FPR reduces so to emphasise the effects in the STARE dataset, 10% of the output was affected by uniformly distributed random noise (within the range 0–255).



(a) Fissure Image (b) STARE Dataset

Fig. 8: P-R and P̄-R curves calculated upon the fissure image (using the Gaussian detector) taking the 8th (red) and 12th (blue) annotations as GTs and the STARE dataset (using the MLVessel v1.4 package) taking the VK (red) and AH (blue) GTs.

8b) both P-R evaluations result in similar performance, with the higher achieving varying with recall. As mentioned, however, these performances are not comparable as each ground truth has a different skew. By normalising skew, the P̄-R curves clearly show that the algorithm better models the VK ground truth. It is worth noting that any of the methods aimed at normalising skew presented in this paper, Equations (3), (4) and (7), may be used to reveal these trends.

Source code for all the functions and the documented experiments presented within this paper is available from *http://sites.google.com/site/tomalampert*.

6 Conclusions

This paper has introduced a number of interesting consequences of measuring performance with precision-recall curves. An implicit assumption in most evaluations is that skew is constant, however, this paper has given counter-examples to this assumption and has proposed that skew can have temporal variance as well. Furthermore, it is rarely possible to accurately measure a problem's skew. This has wide reaching consequences when attempting to accurately measure an algorithm's performance within a problem domain. It has been shown that accounting for temporal changes in skew results in a vastly different value of precision to that obtained when ignoring it.

Several methods to aid in the evaluation of algorithms in domains that have fixed and temporal skews have been presented. A method for transforming the precision measure into different levels of skew has been proved, for use when a dataset's skew characteristics do not match those of the problem (e.g. through the use of over or under sampling). It has been demonstrated that an algorithm's rank, as measured by the AUCPR, is not constant with respect to skew. The condition under which this variance occurs has been defined, and a test and method for finding at which skew AUCPRs become equal, proposed.

Furthermore, a method for calculating precision whilst accounting for temporal fluctuations has been defined for use in domains in which skew is not fixed. Using these findings the correct form of an integrated precision measure has been derived for use in domains in which the characteristics of the problem's skew are not known. Finally, example applications for each of these proposals have been demonstrated.

A Precision Properties

Theorem 2 Two points in P-R space that share the same recall, (R_*, P^1) and (R_*, P^2), have the same ranking (in terms of precision) irrespective of skew.

Proof. Keeping in mind the definition of precision, Equation (1), and that TP $= \pi R_* N$; there are three cases, each of which are illustrated in Figure 5, such that:

C_1) $P^1 > P^2$

$$\begin{aligned} &= \frac{\pi R_* N}{\pi R_* N + \mathrm{FP}^1} > \frac{\pi R_* N}{\pi R_* N + \mathrm{FP}^2} \\ &= \frac{\pi R_*}{\pi R_* + (1-\pi)\mathrm{FPR}^1} > \frac{\pi R_*}{\pi R_* + (1-\pi)\mathrm{FPR}^2} \end{aligned}$$

since FP $=$ FPR$(1-\pi)N$. Therefore $(1-\pi)\mathrm{FPR}^1 < (1-\pi)\mathrm{FPR}^2$ for all $\pi > 0$.

C_2) $P^1 = P^2$. Under the same skew π and the same dataset size N, $\mathrm{TP}^1 = \mathrm{TP}^2$. Therefore, from Eq. (1), $\mathrm{FP}^1 = \mathrm{FP}^2$ and since FP $=$ FPR$(1-\pi)N$, $(1-\pi)\mathrm{FPR}^1 = (1-\pi)\mathrm{FPR}^2$ for all π.

C_3) $P^1 < P^2$. As C_1 but transpose P^1 and P^2. □

B Temporally Integrated Precision Properties

B.1 $\tilde{\mathrm{P}}$-R Interpolation

Due to the discretisation of the threshold upon the algorithm's output, two points in $\tilde{\mathrm{P}}$-R space may be distant from each other. In this case, interpolation may be employed to estimate the connecting line.

By modifying Davis and Goadrich's (2006) method interpolation between two points on the $\tilde{\mathrm{P}}$-R curve, a and b, is non-linear and can be achieved such that

$$\left(\frac{\mathrm{TP}(\theta_a) + i}{N_p}, \frac{1}{T} \int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)] \frac{\pi}{1-\pi} \left[\frac{\mathrm{FP}(\theta_a) + Bi}{\mathrm{TP}(\theta_a) + i} \right]} \, \mathrm{d}t \right) \tag{10}$$

where θ_a and θ_b are thresholds such that $1 \leq i < \mathrm{TP}(\theta_b) - \mathrm{TP}(\theta_a)$, $B = \frac{\mathrm{FP}(\theta_b) - \mathrm{FP}(\theta_a)}{\mathrm{TP}(\theta_b) - \mathrm{TP}(\theta_a)}$ and $i \in \mathbb{N}$.

B.2 Area Under $\tilde{\mathrm{P}}$-R Curve

It is often advantageous to determine the AUC in ROC space. This measure is often used to optimise an algorithm, and maximising the area under a P-R curve has been shown to be more favourable than optimising the area under a ROC curve (Davis and Goadrich, 2006) (in ROC space it is possible that two AUCs are similar whereas the equivalent AUCs in P-R space are not). An equivalent measure can be defined for a $\tilde{\mathrm{P}}$-R curve under the assumption that a curve's constituent points are sufficiently close, which can be ensured by employing the interpolation outlined in Eq. (10). The AUC$\tilde{\mathrm{P}}$R measure can therefore be approximated using trapezoidal integration (Bradley, 1997), such that

$$\mathrm{AUC}\tilde{\mathrm{P}}\mathrm{R} = \int_0^1 \tilde{P}_T(\theta) \, \mathrm{d}\theta \approx \sum_{i=2}^{H} [R(\theta_i) - R(\theta_{i-1})] \, \frac{\tilde{P}_T(\theta_{i-1}) + \tilde{P}_T(\theta_i)}{2} \tag{11}$$

where H is the number of threshold levels used.

B.3 Unachievable $\tilde{\text{P}}$-R Space

Boyd et al. (2012) show that the definition of precision results in an unachievable area of P-R space that is dependent upon the skew of a dataset, $\pi = N_p/N$. The same is true for the temporally integrated precision measure.

Theorem 4 *Temporally integrated precision $\tilde{P}_T(\theta)$ satisfies*

$$\tilde{P}_T(\theta) \geq \frac{1}{T}\int_0^T \frac{\pi'(t)R(\theta)}{\pi'(t)R(\theta) + [1-\pi'(t)]}\,\mathrm{d}t. \tag{12}$$

Proof. The definition of $\tilde{P}_T(\theta)$, Eq. (4), states that

$$\tilde{P}_T(\theta) = \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]}\,\mathrm{d}t$$

from the definition of precision, Eq. (1) and the definition of recall, Eq. (2), $\text{TP}(\theta) = \pi R(\theta)N$ (since $\text{FN}(\theta) = \pi N - \text{TP}(\theta)$), we have

$$\begin{aligned}
\tilde{P}_T(\theta) &= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{\text{TP}(\theta)+\text{FP}(\theta)}{\text{TP}(\theta)} - 1\right]}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{\text{FP}(\theta)}{\pi R(\theta)N}\right]}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \frac{\pi'(t)R(\theta)}{\pi'(t)R(\theta) + [1-\pi'(t)]\frac{\text{FP}(\theta)}{(1-\pi)N}}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \frac{\pi'(t)R(\theta)}{\pi'(t)R(\theta) + [1-\pi'(t)]\frac{\text{FP}(\theta)}{N_n}}\,\mathrm{d}t
\end{aligned}$$

and since the number of FP detections cannot be greater than the number of negative instances, $\text{FP}(\theta) \leq N_n$, we have

$$\tilde{P}_T(\theta) \geq \frac{1}{T}\int_0^T \frac{\pi'(t)R(\theta)}{\pi'(t)R(\theta) + [1-\pi'(t)]}\,\mathrm{d}t.$$

□

If a point in $\tilde{\text{P}}$-R space satisfies Eq. (12) then it is achievable.

As mentioned within the main text (Section 4.2), the minimum AUC within $\tilde{\text{P}}$-R space ($\text{AUC}\tilde{\text{PR}}_{\text{MIN}}$) cannot be analytically defined because the form of the skew function $\pi'(t)$ that will be used in application is unknown. Nevertheless, from Eq. (12) we know that the minimum achievable $\tilde{\text{P}}$-R curve is

$$\tilde{P}_T(\theta) = \frac{1}{T}\int_0^T \frac{\pi'(t)R(\theta)}{\pi'(t)R(\theta) + [1-\pi'(t)]}\,\mathrm{d}t \tag{13}$$

and therefore Eq. (11) can be used to estimate $\text{AUC}\tilde{\text{PR}}_{\text{MIN}}$. An algorithm's reported AUC should therefore be $\text{AUC}\tilde{\text{PR}} - \text{AUC}\tilde{\text{PR}}_{\text{MIN}}$. In fact, a normalised AUC measure has been proposed (Boyd et al, 2012) that can be modified to suite this case

$$\text{AUC}\tilde{\text{PR}} = \frac{\text{AUC}\tilde{\text{PR}} - \text{AUC}\tilde{\text{PR}}_{\text{MIN}}}{1 - \text{AUC}\tilde{\text{PR}}_{\text{MIN}}}. \tag{14}$$

Nevertheless, as $\text{AUC}\tilde{\text{PR}}_{\text{MIN}}$ is only dependent upon $\pi'(t)$, and assuming that this function remains constant between evaluations, the relative ranking of an algorithm's performance will also remain equal.

B.4 Performance of a Random Classifier

It is often desirable to represent the performance of a random classifier in ROC space. In $\tilde{\text{P}}$-R space a random classifier's curve is entirely dependent upon the function $\pi'(t)$.

Theorem 5 *A random classifier produces a constant $\tilde{P}$-R curve equal to*

$$\tilde{P}_{RC}(\theta) = \frac{1}{T}\int_0^T \pi'(t)\,\mathrm{d}t. \tag{15}$$

Proof. From the definition of $\tilde{\text{P}}$ in Eq. (4)

$$\begin{aligned}
\tilde{P}_T(\theta) &= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{1}{P_\pi(\theta)} - 1\right]}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{\mathrm{TP}(\theta)+\mathrm{FP}(\theta)}{\mathrm{TP}(\theta)} - 1\right]}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{\mathrm{TPR}(\theta)\pi N+(1-\pi)N\mathrm{FPR}(\theta)}{\mathrm{TPR}(\theta)\pi N} - 1\right]}\,\mathrm{d}t
\end{aligned}$$

since $\mathrm{TP}(\theta) = \mathrm{TPR}(\theta)\pi N$, $\mathrm{FP}(\theta) = (1-\pi)N\mathrm{FPR}(\theta)$. For a random classifier $\mathrm{TPR}(\theta) = \mathrm{FPR}(\theta)$ (Fawcett, 2006) therefore

$$\begin{aligned}
\tilde{P}_{\mathrm{RC}}(\theta) &= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]\frac{\pi}{1-\pi}\left[\frac{\pi+(1-\pi)}{\pi} - 1\right]}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \frac{\pi'(t)}{\pi'(t) + [1-\pi'(t)]}\,\mathrm{d}t \\
&= \frac{1}{T}\int_0^T \pi'(t)\,\mathrm{d}t.
\end{aligned}$$

□

Corollary 2 *It directly follows that the AUC$\tilde{P}$R of a random classifier, within the bounds of $0 \leq R \leq 1$, is defined such that AUC$\tilde{P}R_{RC}(\theta) = \frac{1}{T}\int_0^T \pi'(t)\,\mathrm{d}t$.*

C Integrated Precision Properties

The properties of the integrated precision measure presented in Section 4.3.2 are now formally proved. For brevity the normalisation factor will henceforth be referred to as $\gamma = 1/(\pi'_2 - \pi'_1)$. We start by proving an alternative but equivalent form of Eq. (7) that will be the starting point of the remaining proofs in this section.

Theorem 6 *The integrated precision measure $\bar{P}(\theta)$, defined in Eq. (7), is equivalent to*

$$\bar{P}(\theta) = \gamma\int_{\pi'_1}^{\pi'_2} \frac{\pi' TP(\theta)}{\pi' TP(\theta) + (1-\pi')\phi FP(\theta)}\,\mathrm{d}\pi'. \tag{16}$$

Proof. Since $\text{FPR}(\theta) = \frac{\text{FP}(\theta)}{N_n}$ and $\text{TPR}(\theta) = \frac{\text{TP}(\theta)}{N_p}$. Substituting these into Eq. (7) gives

$$
\begin{aligned}
&= \gamma \int_{\pi_1'}^{\pi_2'} \frac{\pi'[\text{TP}(\theta)N_p^{-1}]}{\pi'[\text{TP}(\theta)N_p^{-1}] + (1-\pi')[\text{FP}(\theta)N_n^{-1}]}\,\mathrm{d}\pi' \\
&= \gamma \int_{\pi_1'}^{\pi_2'} \frac{\pi'\text{TP}(\theta)}{\pi'\text{TP}(\theta) + (1-\pi')N_p[\text{FP}(\theta)N_n^{-1}]}\,\mathrm{d}\pi' \\
&= \gamma \int_{\pi_1'}^{\pi_2'} \frac{\pi'\text{TP}(\theta)}{\pi'\text{TP}(\theta) + (1-\pi')\frac{N_p}{N_n}\text{FP}(\theta)}\,\mathrm{d}\pi' \\
&= \gamma \int_{\pi_1'}^{\pi_2'} \frac{\pi'\text{TP}(\theta)}{\pi'\text{TP}(\theta) + (1-\pi')\phi\text{FP}(\theta)}\,\mathrm{d}\pi'
\end{aligned}
$$

□

Equation (7) is based solely upon the true positive and false positive rates. Furthermore varying the skew of the dataset is equivalent to varying the performance of the algorithm, measured in terms of true positive and false positive rates. Such that increasing (or decreasing) $\text{TPR}(\theta)$ by a factor of π' along with decreasing (or increasing) $\text{FPR}(\theta)$ by a factor of $1-\pi'$ is equivalent (in terms of precision) to changing the skew of the dataset from being balanced to having a skew of π'.

Theorem 7 *The proposed precision measure $\bar{P}(\theta)$, defined in Equation* (7)*, can analytically represent a dataset having any skew.*

Proof. Theorem 6 shows that $\bar{P}(\theta)$ is independent of dataset skew and is instead dependent upon π'. This term balances $\text{TP}(\theta)$ and $\text{FP}(\theta)$. A dataset without positive instances ($N_p = 0$) results in zero TP detections, and similarly a dataset containing no negative instances ($N_n = 0$) results in zero FP detections. These represent the two extremes of dataset skew and can be analytically represented in Eq. (7) by $\pi' = 0$ and $\pi' = 1$ respectively. Any skew in between these extremes is therefore contained within the range $0 \leq \pi' \leq 1$. □

C.1 $\bar{\text{P}}$-R Interpolation

Due to the discretisation of the threshold upon the algorithm's output, two points in $\bar{\text{P}}$-R space may be distant from each other. In this case, interpolation may be employed to estimate the connecting line.

By modifying Davis and Goadrich's (2006) method interpolation between two points on the $\bar{\text{P}}$-R curve, a and b, is non-linear and can be achieved such that

$$
\left(\frac{\text{TP}(\theta_a) + i}{N_p}, \gamma \int_{\pi_1'}^{\pi_2'} \frac{\pi'(\text{TP}(\theta_a) + i)}{\pi'(\text{TP}(\theta_a) + i) + (1-\pi')\phi\text{FP}(\theta_a) + Bi}\,\mathrm{d}\pi' \right) \tag{17}
$$

where θ_a and θ_b are thresholds such that $1 \leq i < \text{TP}(\theta_b) - \text{TP}(\theta_a)$, $B = \frac{\text{FP}(\theta_b) - \text{FP}(\theta_a)}{\text{TP}(\theta_b) - \text{TP}(\theta_a)}$ and $i \in \mathbb{N}$.

C.2 Area Under $\bar{\text{P}}$-R Curve

It is often advantageous to determine the AUC in ROC space. This measure is often used to optimise an algorithm, and maximising the area under a P-R curve has been shown to be more favourable than optimising the area under a ROC curve (Davis and Goadrich, 2006) (in ROC space it is possible that two AUCs are similar whereas the equivalent AUCs in P-R space are not). An equivalent measure can be defined for a $\bar{\text{P}}$-R curve under the assumption that a curve's constituent points are sufficiently close, which can be ensured by employing the interpolation outlined in Eq. (17). The $\text{AUC}\bar{\text{P}}\text{R}$ measure can therefore be approximated using trapezoidal integration (Bradley, 1997), such that

$$
\text{AUC}\bar{\text{P}}\text{R} = \int_0^1 \bar{P}(\theta)\,\mathrm{d}\theta \approx \sum_{i=2}^{T} [R(\theta_i) - R(\theta_{i-1})]\,\frac{\bar{P}(\theta_{i-1}) + \bar{P}(\theta_i)}{2} \tag{18}
$$

where T is the number of threshold levels used.

C.3 Unachievable $\bar{\text{P}}$-R Space

Boyd et al. (2012) show that the definition of precision results in an unachievable area of P-R space that is dependent upon the skew of a dataset, $\pi = N_p/N$. The same is true for the integrated precision measure.

Theorem 8 *Integrated precision $\bar{P}(\theta)$ satisfies*

$$\bar{P}(\theta) \geq \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R(\theta)}{\pi' R(\theta) + 1 - \pi'} \, \mathrm{d}\pi'. \tag{19}$$

Proof. The definition of $\bar{P}(\theta)$, Eq. (7), states that

$$\bar{P}(\theta) = \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' \text{TP}(\theta)}{\pi' \text{TP}(\theta) + (1-\pi')\phi \text{FP}(\theta)} \, \mathrm{d}\pi' \geq \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' \text{TP}(\theta)}{\pi' \text{TP}(\theta) + (1-\pi')\phi(1-\pi)N} \, \mathrm{d}\pi'$$

since the number of FP detections cannot be greater than the number of negative instances, $\text{FP}(\theta) \leq N_n$. From the definition of recall, Eq. (2), $\text{TP}(\theta) = \pi R(\theta) N$ (since $\text{FN}(\theta) = \pi N - \text{TP}(\theta)$). Therefore,

$$\begin{aligned}
\bar{P}(\theta) &\geq \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R(\theta) \pi N}{\pi' R(\theta) \pi N + (1-\pi')\phi(1-\pi)N} \, \mathrm{d}\pi' \\
&= \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R(\theta) \frac{N_p}{N} N}{\pi' R(\theta) \frac{N_p}{N} N + (1-\pi')\phi(1-\frac{N_p}{N})N} \, \mathrm{d}\pi' \\
&= \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R(\theta) N_p}{\pi' R(\theta) N_p + (1-\pi') \frac{N_p}{N_n} \frac{N_n}{N} N} \, \mathrm{d}\pi' \\
&= \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R(\theta) N_p}{\pi' R(\theta) N_p + (1-\pi') N_p} \, \mathrm{d}\pi' \\
&= \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R(\theta)}{\pi' R(\theta) + 1 - \pi'} \, \mathrm{d}\pi'.
\end{aligned}$$

□

If a point in $\bar{\text{P}}$-R space satisfies Eq. (19) then it is achievable. *Nota bene*, a point's achievability is only dependent upon π'_1 and π'_2—the analytical skew—and is therefore independent of a dataset.

The curve that represents the boundary between the unachievable and achievable regions of $\bar{\text{P}}$-R space is defined by the worst performing algorithm (Boyd et al, 2012), i.e. every negative instance is classified as positive. This $\bar{\text{P}}$-R curve therefore passes through $(0,0)$ and $\left(1, \frac{\pi'_2 + \pi'_1}{2}\right)$ (this can be easily verified by substituting $\text{TPR} = 1$ and $\text{FPR} = 1$ into Eq. (16)). Interpolation between two points on a $\bar{\text{P}}$-R line is given by Eq. (17), and is therefore $\left(R, \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R}{\pi' R + 1 - \pi'} \, \mathrm{d}\pi'\right)$, where $R = \frac{i}{N_p}$, for $0 \leq i \leq N_p$. This can now be used to calculate the area of the unachievable region, which is implicitly included in AUC$\bar{\text{P}}$R (thus overestimating performance).

Theorem 9 *The area of the unachievable region in $\bar{\text{P}}$-R space, and analogously the minimum achievable* AUC$\bar{\text{P}}$R*, for limits π'_1 and π'_2 is given such that*

$$\begin{aligned}
\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(\pi'_1, \pi'_2) = \gamma \Big[&-2\pi'_1 + 2\pi'_2 \\
&+ (-1+\pi'_1)\ln(1-\pi'_1) - (-1+\pi'_2)\ln(1-\pi'_2) \\
&+ Li_2(\pi'_1) - Li_2(\pi'_2) \Big].
\end{aligned} \tag{20}$$

where $Li_2(\pi') = \sum_{k=1}^{\infty} \frac{\pi'^k}{k^2}$ is a polylogarithmic function.

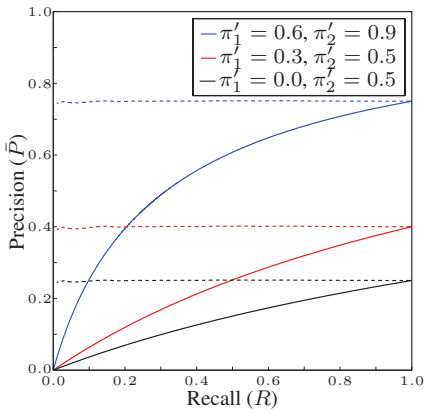


Fig. 9: Minimum achievable $\bar{\text{P}}$-R curves (solid lines) and those of random classification (dashed lines) as π'_1 and π'_2 vary. The minimum AUCs (to four decimal places) are: $\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(0.0, 0.5) = 0.1424$; $\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(0.3, 0.5) = 0.2349$; and $\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(0.6, 0.9) = 0.5471$.

Proof. Equation (19) gives the lower bound on precision at a particular recall, therefore the unachievable area is the area under the curve $\bar{P}_{\text{MIN}}(R) = \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R}{\pi' R + 1 - \pi'} \, \mathrm{d}\pi'$, such that

$$\begin{aligned}\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(\pi'_1, \pi'_2) = &\int_0^1 \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' R}{\pi' R + 1 - \pi'} \, \mathrm{d}\pi' \, \mathrm{d}R \\ =&\gamma \Big[-2\pi'_1 + 2\pi'_2 + (-1 + \pi'_1) \ln(1 - \pi'_1) \\ &- (-1 + \pi'_2) \ln(1 - \pi'_2) + \text{Li}_2(\pi'_1) - \text{Li}_2(\pi'_2) \Big]\end{aligned}$$

and $\text{Li}_2(\pi')$ can be efficiently computed to an accuracy of nineteen decimal places (Osácar et al, 1995). □

Therefore an algorithm's reported AUC should be $\text{AUC}\bar{\text{P}}\text{R} - \text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(\pi'_1, \pi'_2)$. In fact, a normalised AUC measure has been proposed (Boyd et al, 2012) that can be modified to suite this case

$$\text{AUC}\bar{\text{P}}\text{R}(\pi'_1, \pi'_2) = \frac{\text{AUC}\bar{\text{P}}\text{R} - \text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(\pi'_1, \pi'_2)}{1 - \text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(\pi'_1, \pi'_2)}. \tag{21}$$

Nevertheless, as $\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}$ is only dependent upon π'_1 and π'_2, and assuming that they remain constant, the relative ranking of algorithms' performances also remain equal. Figure 9 illustrates $\text{AUC}\bar{\text{P}}\text{R}_{\text{MIN}}(\pi'_1, \pi'_2)$ for a number of different skew limits (solid lines).

C.4 Performance of a Random Classifier

It is often desirable to represent the performance of a random classifier in ROC space. In $\bar{\text{P}}$-R space a random classifier's curve is dependent upon the integration limits, π'_1 and π'_2.

Theorem 10 *A random classifier produces a constant $\bar{P}$-R curve equal to*

$$\bar{P}_{RC}(\theta) = \frac{\pi'_2 + \pi'_1}{2}, \qquad 0 \leq R(\theta) \leq 1. \tag{22}$$

Proof. Since a random classifier gives $\mathrm{TPR}(\theta) = \mathrm{FPR}(\theta)$ (Fawcett, 2006), substituting into Eq. (16) gives

$$\bar{P}(\theta) = \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi' \mathrm{TPR}(\theta)}{\pi' \mathrm{TPR}(\theta) + (1-\pi')\mathrm{TPR}(\theta)} \, \mathrm{d}\pi' = \gamma \int_{\pi'_1}^{\pi'_2} \frac{\pi'}{\pi' + (1-\pi')} \, \mathrm{d}\pi' = \frac{\pi'_2 + \pi'_1}{2}.$$

□

Corollary 3 *It directly follows that the AUC$\bar{P}$R of a random classifier, within the bounds of* $0 \leq R \leq 1$*, is defined such that AUC$\bar{P}$R$_{RC}(\theta) = (\pi'_2 + \pi'_1)/2$.*

Acknowledgements This work is part of the FOSTER project, which is funded by the French Research Agency (Contract ANR Cosinus, ANR-10-COSI-012-03-FOSTER, 2011–2014). The participating annotators from LIVE, IPGS, and ICube (University of Strasbourg) are gratefully acknowledged.

References

Ahanotu D (1999) Heavy-duty vehicle weight and horespower distributions: measurement of class-specific temporal and spatial variability. PhD thesis, Civil and Environmental Engineering, Georgia Institute of Technology

Arbelaéz P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans PAMI 33(5):898–916

Boyd K, Page D, Santos Costa V, Davis J (2012) Unachievable region in precision-recall space and its effect on empirical evaluation. In: ICML

Bradley A (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159

Clémençon S (2009) Nonparametric estimation of the precision-recall curve. In: ICML, pp 185–192

Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: ICML, pp 233–240

Davis J, Burnside E, Dutra I, Page C, Costa V (2005) An integrated approach to learning bayesian networks of rules. In: ECML, pp 84–95

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874

Fell J, Röschke J, Mann K, Schäffner C (1996) Discrimination of sleep stages: a comparison between spectral and nonlinear eeg measures. Electroencephalography and Clinical Neurophysiology 98(5):401–410

Flach P (2003) The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: ICML, pp 194–201

He H, Garcia E (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284

Hoover A, Kouznetsova V, Goldbaum M (2000) Locating blood vessels in retinal images by piece-wise threhsold probing of a matched filter response. IEEE Trans Med Imag 19(3):203–210

Lampert T, O'Keefe S (2011) A detailed investigation into low-level feature detection in spectrogram images. Pattern Recognition 44(9):2076–2092

Lampert T, O'Keefe S (2013) On the detection of tracks in spectrogram images. Pattern Recognition 46(5):1396–1408

Landgrebe T, Paclík P, Duin R, Bradley A (2006) Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In: ICPR, pp 123–127

Liu Y, Shriberg E (2007) Comparing evaluation metrics for sentence boundary detection. In: ICASSP, pp 451–458

Osácar C, Palacián J, Palacios M (1995) Numerical evaluation of the dilogarithm of complex argument. Celestial Mechanics and Dynamical Astronomy 62(1):93–98

Papageorgiou C (1999) Trainable pedestrian detection. In: Int. Conf. on Image Processing, vol 4, pp 35–39

Pavlo A, Curino C, Zdonik S (2012) Skew-aware automatic database partitioning in shared-nothing, parallel OLTP systems. In: SIGMOD, pp 61–72

Sieracki M, Johnson P, Sieburth J (1985) Detection, enumeration, and sizing of planktonic bacteria by image-analyzed epifluorescence microscopy. Applied and Environmental Microbiology 49(4):799–810

Soares J, Leandro J, Cesar-Jr R, Jelinek H, Cree M (2006) Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Trans Med Imag 25(9):1214–1222

Stäger M, Lukowicz P, Tröster G (2006) Dealing with class skew in context recognition. In: ICDCS Workshops, p 58

Stumpf A, Kerle N, Puissant A, Lachiche N, Malet JP (2012a) Adaptive spatial sampling with active random forest for object-oriented landslide mapping. In: IGARSS, IEEE, pp 87–90

Stumpf A, Lampert T, Malet JP, Kerle N (2012b) Multi-scale line detection for landslide fissure mapping. In: IGARSS, IEEE, pp 5450–5453

Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: A review. IEEE Trans Pattern Anal Mach Intell 28(5):694–711

Webb G, Ting KM (2005) On the application of ROC analysis to predict classification performance under varying class distributions. Machine Learning 58(1):25–32

Yue Y, Finley T, Radlinski F, Joachims T (2007) A support vector method for optimizing average precision. In: SIGIR, pp 271–278

Zink A, Kern Reeve H (2005) Predicting the temporal dynamics of reproductive skew and group membership in communal breeders. Behavioral Ecology 16(5):880–888

Zwietering M, Jongenburger I, Rombouts F, Riet KV (1990) Modeling of the bacterial growth curve. Applied and Environmental Microbiology 56(6):1875–1881