Precision-Recall-Gain Curves: PR Analysis Done Right

A paper to be presented at NIPS 2015

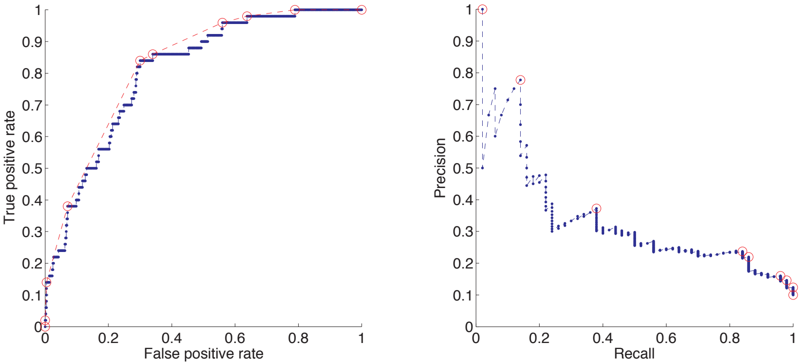
Peter A. Flach (joint work with Meelis Kull)

Intelligent Systems Laboratory, University of Bristol, United Kingdom

December 3, 2015

Talk outline

What's next?



(left) ROC curve with non-dominated points (**red** circles) and convex hull (**red** dotted line). **(right)** Corresponding Precision-Recall curve with non-dominated points (**red** circles).

Properties of ROC curves I

ROC curves are widely used in machine learning and their main properties are well understood . These properties can be summarised as follows.

Universal baselines: the major diagonal of an ROC plot depicts the line of random performance which can be achieved without training; it is universal in the sense that it doesn't depend on the class distribution.

Linear interpolation: any point on a straight line between two points representing the performance of two classifiers (or thresholds) A and B can be achieved by making a suitably biased random choice between A and B . The slope of the connecting line determines the trade-off between the classes under which any linear combination of A and B would yield equivalent performance. In particular, test set accuracy assuming uniform misclassification costs is represented by accuracy isometrics with slope $(1-\pi)/\pi$, where π is the proportion of positives .

Properties of ROC curves II

Optimality: a point D dominates another point E if D's tpr and fpr are not worse than E's and at least one of them is strictly better. The set of non-dominated points – the Pareto front – establishes the set of classifiers or thresholds that are optimal under some trade-off between the classes. Due to linearity any interpolation between non-dominated points is both achievable and non-dominated, giving rise to the *convex hull* (ROCCH).

Area: the proportion of the unit square which falls under an ROC curve ($AUROC$) estimates the probability that a randomly chosen positive is ranked higher by the model than a randomly chosen negative .
There is a linear relationship between $AUROC = \int_0^1 tpr \, d\, fpr$ and the expected accuracy $acc = \pi\, tpr + (1-\pi)(1-fpr)$ averaged over all possible predicted positive rates $rate = \pi\, tpr + (1-\pi) fpr$:

$$\mathbb{E}[acc] = \int_0^1 acc \, d\, rate = \pi(1-\pi)(2AUROC-1) + 1/2$$

For uniform class distributions this reduces to $\mathbb{E}[acc] = AUROC/2 + 1/4$.

Properties of ROC curves III

Calibration: slopes of convex hull segments can be interpreted as empirical likelihood ratios associated with a particular interval of raw classifier scores. This gives rise to a non-parametric calibration procedure which is also called isotonic regression or pool adjacent violators and results in a calibration map which maps each segment of ROCCH with slope s to a calibrated score

$$c = \frac{\pi s}{\pi s + (1-\pi)} = \frac{1}{1 + \frac{1-\pi}{\pi}\frac{1}{s}}$$

Define a skew-sensitive version of accuracy as

$$acc_c \triangleq 2c\pi\, tpr + 2(1-c)(1-\pi)(1-fpr)$$

(i.e., standard accuracy is $acc_{c=1/2}$) then a perfectly calibrated classifier outputs, for every instance, the value of c for which the instance is on the acc_c decision boundary.

Contributions of this work

(*i*) We identify the problems with current practice in Precision-Recall curves by demonstrating that they fail to satisfy each of the above properties in some respect.

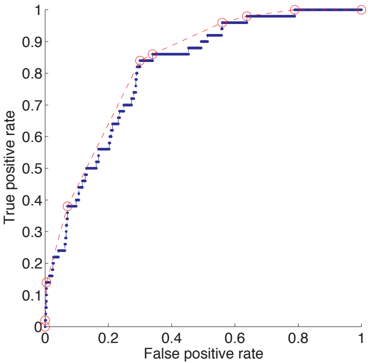
(*ii*) We propose a principled way to remedy **all** these problems by means of a change of coordinates.

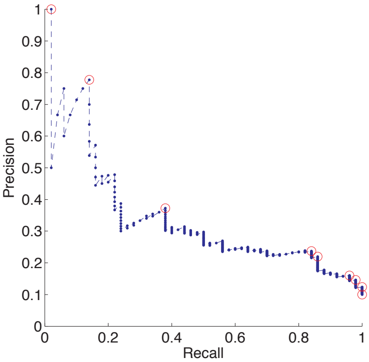
(*iii*) Our improved Precision-Recall-Gain curves enclose an area that is directly related to expected F_1 score – on a harmonic scale – in a similar way as $AUROC$ is related to expected accuracy.

(*iv*) With Precision-Recall-Gain curves it is possible to calibrate a model for F_β in the sense that the predicted score for any instance determines the value of β for which the instance is on the F_β decision boundary.

(*v*) We give experimental evidence that this matters by demonstrating that the area under traditional Precision-Recall curves can easily favour models with lower expected F_1 score than others.

What's next?





(left) ROC curve with non-dominated points (**red** circles) and convex hull (**red** dotted line). **(right)** Corresponding Precision-Recall curve with non-dominated points (**red** circles).

PR plots are not like ROC plots I

Non-universal baselines: a random classifier has precision π and hence baseline performance is a horizontal line which depends on the class distribution.

Non-linear interpolation: precision in a linearly interpolated contingency table is only a linear combination of the original precision values if the two classifiers have the same predicted positive rate (which is impossible if the two contingency tables arise from different decision thresholds on the same model). More generally, *it isn't meaningful to take the arithmetic average of precision values*.

Non-convex Pareto front: the set of non-dominated operating points continues to be well-defined but in the absence of linear interpolation this set isn't convex for PR curves, nor is it straightforward to determine by visual inspection.

PR plots are not like ROC plots II

Uninterpretable area: although many authors report the area under the PR curve ($AUPR$) it doesn't have a meaningful interpretation beyond the geometric one of expected precision when uniformly varying the recall (and even then the use of the arithmetic average cannot be justified). Furthermore, PR plots have unachievable regions at the lower right-hand side, the size of which depends on the class distribution .

No calibration: although some results exist regarding the relationship between calibrated scores and F_1 score these are unrelated to the PR curve. To the best of our knowledge there is no published procedure to output scores that are calibrated for F_β – that is, which give the value of β for which the instance is on the F_β decision boundary.

The F_β score

The F_1 score is defined as the harmonic mean of precision and recall:

$$F_1 \triangleq \frac{2}{1/prec + 1/rec} = \frac{2\,prec \cdot rec}{prec + rec} = \frac{TP}{TP + (FP + FN)/2} \tag{1}$$

This corresponds to accuracy in a modified contingency table:

	Predicted ⊕	*Predicted* ⊖	
Actual ⊕	TP	FN	Pos
Actual ⊖	FP	TP	$Neg - (TN - TP)$
	$TP + FP$	Pos	$2TP + FP + FN$

The F_β score is a weighted harmonic mean:

$$F_\beta \triangleq \frac{1}{\frac{1}{1+\beta^2}/prec + \frac{\beta^2}{1+\beta^2}/rec} = \frac{(1+\beta^2)\,TP}{(1+\beta^2)\,TP + FP + \beta^2 FN} \tag{2}$$

Related work

There is a range of recent results regarding the F-score:
(*i*) non-decomposability of the F_β score, meaning it is not an average over instances ;
(*ii*) estimators exist that are consistent: i.e., they are unbiased in the limit ;
(*iii*) given a model, operating points that are optimal for F_β can be achieved by thresholding the model's scores ;
(*iv*) a classifier yielding perfectly calibrated posterior probabilities has the property that the optimal threshold for F_1 is half the optimal F_1 at that point. and later by

The latter results tell us that optimal thresholds for F_β are lower than optimal thresholds for accuracy (or equal only in the case of the perfect model).
They don't, however, tell us how to find such thresholds other than by tuning.
We demonstrate how to identify all F_β-optimal thresholds for any β in a single calibration procedure.

What's next?

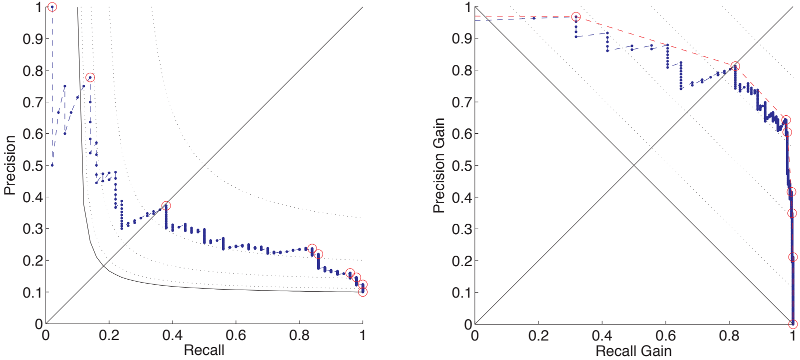
Baseline

A random classifier that predicts positive with probability p has F_β score $(1+\beta^2)p\pi/(p+\beta^2\pi)$. Hence *the baseline to beat is the always-positive classifier* rather than any random classifier. Any model with $prec < \pi$ or $rec < \pi$ loses against this baseline, hence it makes sense to consider only precision and recall values in the interval $[\pi, 1]$. Any real-valued variable $x \in [min, max]$ on a harmonic scale can be linearised by the mapping $\frac{1/x - 1/min}{1/max - 1/min} = \frac{max\cdot(x-min)}{(max-min)\cdot x}$.

Definition (Precision Gain and Recall Gain)

$$precG = \frac{prec - \pi}{(1-\pi)prec} = 1 - \frac{\pi}{1-\pi}\frac{FP}{TP} \qquad recG = \frac{rec - \pi}{(1-\pi)rec} = 1 - \frac{\pi}{1-\pi}\frac{FN}{TP} \tag{3}$$

A *Precision-Recall-Gain curve* plots Precision Gain on the y-axis against Recall Gain on the x-axis in the unit square (i.e., negative gains are ignored).



(left) Conventional PR curve with hyperbolic F_1 isometrics (dotted lines) and the baseline performance by the always-positive classifier (solid hyperbole). **(right)** Precision-Recall-Gain curve with minor diagonal as baseline, parallel F_1 isometrics and a convex Pareto front.

Linearity and optimality

Theorem

Let $P_1 = (precG_1, recG_1)$ *and* $P_2 = (precG_2, recG_2)$ *be points in the Precision-Recall-Gain space representing the performance of Models 1 and 2 with contingency tables* C_1 *and* C_2*. Then a model with an interpolated contingency table* $C_* = \lambda C_1 + (1-\lambda)C_2$ *has precision gain* $precG_* = \mu precG_1 + (1-\mu)precG_2$ *and recall gain* $recG_* = \mu recG_1 + (1-\mu)recG_2$*, where* $\mu = \lambda TP_1/(\lambda TP_1 + (1-\lambda)TP_2)$*.*

Theorem

$precG + \beta^2 recG = (1+\beta^2)FG_\beta$*, with* $FG_\beta = \frac{F_\beta - \pi}{(1-\pi)F_\beta} = 1 - \frac{\pi}{1-\pi}\frac{FP+\beta^2 FN}{(1+\beta^2)TP}$.

FG_β measures the gain in performance (on a linear scale) relative to a classifier with both precision and recall – and hence F_β – equal to π.

Area

Define $AUPRG = \int_0^1 precG \, d \, recG$ and $\Delta = recG/\pi - precG/(1-\pi)$. Hence, $-y_0/(1-\pi) \leq \Delta \leq 1/\pi$, where y_0 denotes the precision gain at the operating point where recall gain is zero.

Theorem

Let the operating points of a model with area under the Precision-Recall-Gain curve $AUPRG$ be chosen such that Δ is uniformly distributed within $[-y_0/(1-\pi), 1/\pi]$. Then the expected FG_1 score is equal to

$$\mathbb{E}[FG_1] = \frac{AUPRG/2 + 1/4 - \pi(1-{y_0}^2)/4}{1-\pi(1-y_0)} \tag{4}$$

In the special case where $y_0 = 1$ the expected FG_1 score is $AUPRG/2 + 1/4$. The expected reciprocal F_1 score can be calculated from the relationship $\mathbb{E}[1/F_1] = (1-(1-\pi)\mathbb{E}[FG_1])/\pi$ which follows from the definition of FG_β.

Calibration

Theorem

Let two classifiers be such that $prec_1 > prec_2$ and $rec_1 < rec_2$, then these two classifiers have the same F_β score if and only if

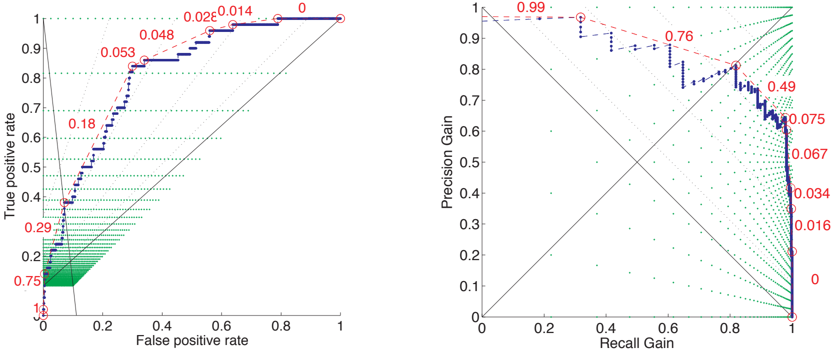
$$\beta^2 = -\frac{1/prec_1 - 1/prec_2}{1/rec_1 - 1/rec_2} = -s_{\mathrm{PRG}} \qquad (5)$$

where s_{PRG} is the slope of the connecting segment in the PRG plot.

We convert this slope to an F-calibrated score as follows:

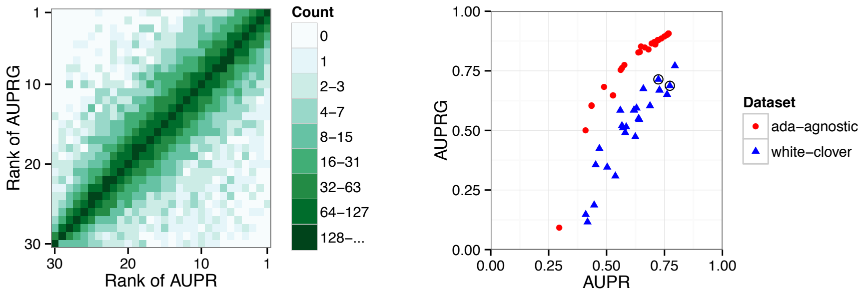
$$c_F = \frac{1}{1 - s_{\mathrm{PRG}}}$$

Notice that this cannot be obtained from the accuracy-calibrated score $\frac{1}{1+\frac{1-\pi}{\pi}\frac{1}{s_{\mathrm{ROC}}}}$.

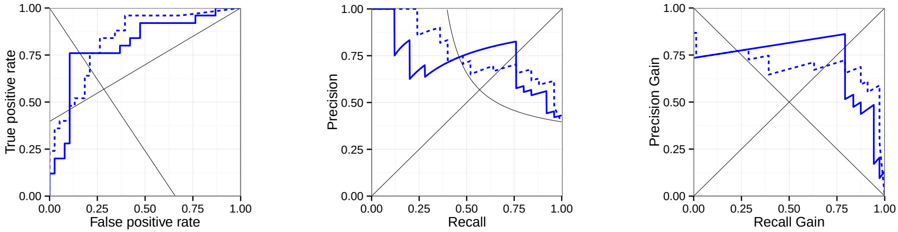


(left) ROC curve with scores empirically calibrated for accuracy. The green dots correspond to a regular grid in Precision-Recall-Gain space. **(right)** Precision-Recall-Gain curve with scores calibrated for F_β. The green dots correspond to a regular grid in ROC space, clearly indicating that ROC analysis over-emphasises the high-recall region.

What's next?



(left) Comparison of AUPRG-ranks vs AUPR-ranks. Each cell shows how many models across 886 OpenML tasks have these ranks among the 30 models in the same task. **(right)** Comparison of AUPRG vs AUPR in OpenML tasks with IDs 3872 (white-clover) and 3896 (ada-agnostic), with 30 models in each task. Some models perform worse than random ($AUPRG < 0$) and are not plotted. The models represented by the two encircled triangles are shown in detail in the next figure.



(left) ROC curves for AdaBoost (solid line) and Logistic Regression (dashed line) on the white-clover dataset (OpenML run IDs 145651 and 267741, respectively). **(middle)** Corresponding PR curves. The solid curve is on average lower with $AUPR = 0.724$ whereas the dashed curve has $AUPR = 0.773$. **(right)** Corresponding PRG curves, where the situation has reversed: the solid curve has $AUPRG = 0.714$ while the dashed curve has a lower $AUPRG$ of 0.687.

What's next?

Methodological recommendations

We recommend practitioners use the F-Gain score instead of the F-score to make sure baselines are taken into account properly and averaging is done on the appropriate scale. If required the FG_β score can be converted back to an F_β score at the end.

The second recommendation is to use Precision-Recall-Gain curves instead of PR curves, and the third to use $AUPRG$ which is easier to calculate than $AUPR$ due to linear interpolation, has a proper interpretation as an expected F-Gain score and allows performance assessment over a range of operating points.

To assist practitioners we are making R, Matlab and Java code to calculate $AUPRG$ and PRG curves available at `http://www.cs.bris.ac.uk/~flach/PRGcurves/`. We are also working on closer integration of $AUPRG$ as an evaluation metric in OpenML and performance visualisation platforms such as ViperCharts .

Closing comments

As future work we mention the interpretation of $AUPRG$ as a measure of ranking performance: we are working on an interpretation which gives non-uniform weights to the positives and as such is related to Discounted Cumulative Gain. A second line of research involves the use of cost curves for the FG_β score and associated threshold choice methods.

Acknowledgments

This work was supported by the REFRAME project granted by the European Coordinated Research on Long-Term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Engineering and Physical Sciences Research Council in the UK under grant EP/K018728/1. Discussions with Hendrik Blockeel helped to clarify the intuitions underlying this work.

References I

Kendrick Boyd, Vitor Santos Costa, Jesse Davis, and C David Page.
Unachievable region in precision-recall space and its effect on empirical evaluation.
In *International Conference on Machine Learning*, page 349, 2012.

T. Fawcett.
An introduction to ROC analysis.
Pattern Recognition Letters, 27(8):861–874, 2006.

T. Fawcett and A. Niculescu-Mizil.
PAV and the ROC convex hull.
Machine Learning, 68(1):97–106, July 2007.

References II

P. A. Flach.
The geometry of ROC space: understanding machine learning metrics through ROC isometrics.
In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, pages 194–201, 2003.

Peter A. Flach.
ROC analysis.
In Claude Sammut and GeoffreyI. Webb, editors, *Encyclopedia of Machine Learning*, pages 869–875. Springer US, 2010.

D. J. Hand and R. J. Till.
A simple generalisation of the area under the ROC curve for multiple class classification problems.
Machine Learning, 45(2):171–186, 2001.

References III

José Hernández-Orallo, Peter Flach, and Cesar Ferri.
A unified view of performance metrics: Translating threshold choice into expected classification loss.
Journal of Machine Learning Research, 13:2813–2869, 2012.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon.
Consistent binary classification with generalized performance metrics.
In *Advances in Neural Information Processing Systems*, pages 2744–2752, 2014.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy.
Optimal thresholding of classifiers to maximize F1 measure.
In *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 225–239. Springer Berlin Heidelberg, 2014.

References IV

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal.
On the statistical consistency of plug-in classifiers for non-decomposable performance measures.
In *Advances in Neural Information Processing Systems 27*, pages 1493–1501. 2014.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet.
Optimizing F-measures by cost-sensitive classification.
In *Advances in Neural Information Processing Systems*, pages 2123–2131, 2014.

F. Provost and T. Fawcett.
Robust classification for imprecise environments.
Machine Learning, 42(3):203–231, 2001.

References V

Borut Sluban and Nada Lavrač.
Vipercharts: Visual performance evaluation platform.
In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 650–653. Springer Berlin Heidelberg, 2013.

C. J. Van Rijsbergen.
Information Retrieval.
Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

Nan Ye, Kian Ming A Chai, Wee Sun Lee, and Hai Leong Chieu.
Optimizing F-measures: A tale of two approaches.
In *Proceedings of the 29th International Conference on Machine Learning*, pages 289–296, 2012.

References VI

B. Zadrozny and C. Elkan.
Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers.
In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 609–616, 2001.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown.
Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications.
The Journal of Machine Learning Research, 14(1):1033–1090, 2013.