The Effect of Class Imbalance on Precision-Recall Curves

Citation for published version:
Williams, CKI 2021, 'The Effect of Class Imbalance on Precision-Recall Curves', *Neural Computation*, vol. 33, no. 4, pp. 853-857. https://doi.org/10.1162/neco_a_01362

Digital Object Identifier (DOI):
10.1162/neco_a_01362

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Neural Computation

The Effect of Class Imbalance on Precision-Recall Curves

Christopher K I Williams
School of Informatics, University of Edinburgh, UK

March 17, 2021

Abstract

In this note I study how the precision of a binary classifier depends on the ratio r of positive to negative cases in the test set, as well as the classifier's true and false positive rates. This relationship allows prediction of how the precision-recall curve will change with r, which seems not to be well known. It also allows prediction of how F_β and the Precision Gain and Recall Gain measures of Flach and Kull (2015) vary with r.

Consider a binary classifier, where the predictions change as the threshold for deciding between the two classes is varied. The Receiver Operating Characteristic (or ROC) curve and the Precision-Recall (PR) curve are two ways of summarizing the performance of classifier in this situation. The ROC curve is invariant to the ratio r of positive to negative cases in the test set in the population limit, but the PR curve *is* affected by r. Below I show explicitly how the PR curve and derived quantities like the F_β measure (due to Van Rijsbergen 1979) are affected by r. As these are frequently used to assess the performance of classifiers, it is important that the effect of r is well understood, and adjusted for (if necessary).

The standard notation (see e.g., Witten et al. 2017, sec. 5.8) for binary classification is summarized below:

	Predicted		Sum
Actual	positive	negative	
positive	TP	FN	P
negative	FP	TN	N

There are P positive and N negative datapoints in the dataset, with the *true positive rate* (TPR) and *false positive rate* (FPR) defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = \frac{\mathrm{FP}}{\mathrm{N}}. \tag{1}$$

Let the fraction of positives in the dataset be denoted by $\pi = \mathrm{P}/(\mathrm{P}+\mathrm{N})$, and define the ratio $r = \mathrm{P}/\mathrm{N} = \pi/(1-\pi)$. If we consider the table above normalized by the sample size $n = \mathrm{P}+\mathrm{N}$, then we observe that the table's entries are fully characterized by the three quantities TPR, FPR and r, as the sum of the normalized entries must be 1. The values in the table are usually thought of as empirical counts from a sample of size n. However, one can consider the normalized table in the limit $n \to \infty$, which describes the *population* properties of the classifier at the threshold chosen.

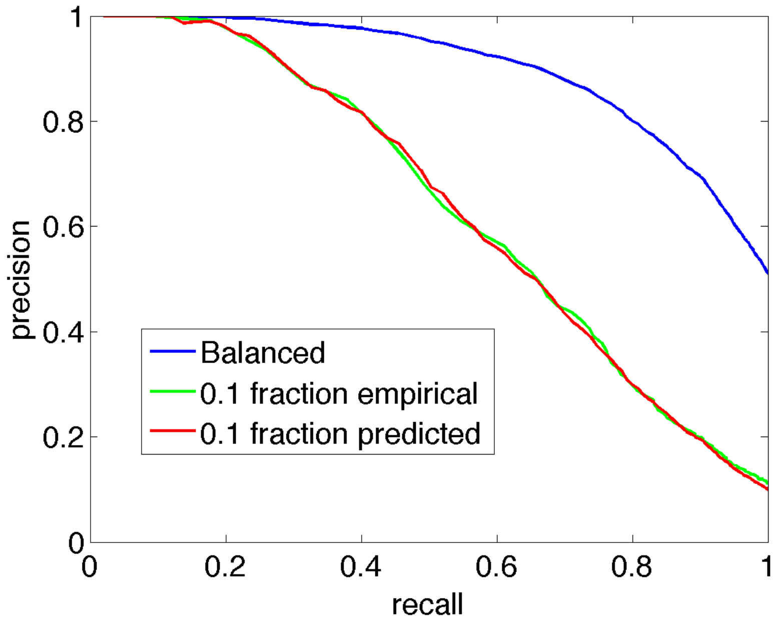

Figure 1: Precision-recall curves for varying r.

The ROC curve is a plot of TPR against FPR. As is well known (see e.g., Fawcett 2006), the population ROC is invariant to r ; this is immediate from the definitions of TPR and FPR, which are ratios within the positives and negatives respectively. Empirical ROC curves for will exhibit some variability as r varies (and indeed across different samples of the same size).

Precision is defined as

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \frac{\mathrm{P}\cdot\mathrm{TPR}}{\mathrm{P}\cdot\mathrm{TPR}+\mathrm{N}\cdot\mathrm{FPR}} = \frac{\mathrm{TPR}}{\mathrm{TPR}+\frac{1}{r}\mathrm{FPR}}. \quad (2)$$

Thus the precision has an explicit dependence on r. Note that the $\mathrm{Prec} \to 1$ as $\pi \to 1$, and also that $\mathrm{Prec} \to 0$ as $\pi \to 0$ if $\mathrm{FPR} > 0$.

The precision-recall curve plots the precision against recall Rec, which is another name for the true positive rate. As recall is invariant to class imbalance, we can consider how the precision varies with r at fixed recall. If we start with balanced classes at $r = 1$ and gradually decrease r[1], we see that the corresponding precision will decrease, because the denominator increases.

For population values of TPR, FPR and r, eq. 2 allows us to transform the precision as a function of r. For an empirical sample, it allows us to *predict* how the PR curve will change with r using the empirical values of TPR and FPR. This is illustrated in Fig 1. In this case a simple classification problem with 2d Gaussians was set up, and a logistic regression classifier trained. For a test set with $r = 1$ and $\mathrm{P} = \mathrm{N} = 5000$ the blue curve was obtained, and for $r = 0.1$ ($\mathrm{P} = 500$, $\mathrm{N} = 5000$) the green empirical curve. If at each value of recall the blue curve is scaled as per eq. 2, the red curve is obtained. Note the good agreement between the predicted and actual curves; the differences can be explained by the fact that the empirical green curve uses a smaller number of samples than the red curve (which reweights all of the balanced samples).

[1] PR curves are typically used when r is small, e.g. in information retrieval settings.

The ability to predict how the PR curve varies with r does not seem to be well known. For example, Fawcett (2006, sec 4.2) discusses "class skew" and shows PR curves for $r = 1$ and $r = 0.1$, but makes no comment on their relationship. However, Hoiem et al. (2012) have pointed out that when comparing PR curves for the detection of different visual object classes, the average precision score is sensitive to the value of r for each class. To enable a fairer comparison, they suggested using "normalized precision", which uses a standard value of r across classes[2].

Note that class imbalance r_{train} in the *training* data should not have an effect on the *test* ROC and PR curves of a probabilistic classifier[3]. To see this, consider the log odds ratio

$$\log \frac{p(C_+|\mathbf{x})}{p(C_-|\mathbf{x})} = \log \frac{p(\mathbf{x}|C_+)}{p(\mathbf{x}|C_-)} + \log r_{\mathrm{train}}, \tag{3}$$

where $r_{\mathrm{train}} = p(C_+)/p(C_-)$. For a generative classifier the LHS is obtained from the RHS and the effect of r_{train} is immediate. For a discriminative classifier eq. 3 can be used to understand the effect of r_{train} on the decision boundary. The test ROC and PR curves only depend on the sequence of confusion matrices obtained as the threshold on the classifier's log odds ratio is changed—the effect of changes in r_{train} is to shift the threshold, but not to change the sequence obtained.

The F_β measure is commonly used as a figure-of-merit that combines precision and recall. It is defined as a weighted harmonic average

$$\frac{1}{F_\beta} = \frac{1}{1+\beta^2}\frac{1}{\mathrm{Prec}} + \frac{\beta^2}{1+\beta^2}\frac{1}{\mathrm{Rec}}. \tag{4}$$

Substituting the expression for the precision from eq. 2, we obtain

$$\frac{1}{F_\beta} = \frac{1}{1+\beta^2}\frac{\mathrm{TPR} + \frac{1}{r}\mathrm{FPR}}{\mathrm{TPR}} + \frac{\beta^2}{1+\beta^2}\frac{1}{\mathrm{TPR}}, \tag{5}$$

and hence

$$F_\beta = \frac{(1+\beta^2)\mathrm{TPR}}{\mathrm{TPR} + \frac{1}{r}\mathrm{FPR} + \beta^2}, \tag{6}$$

which demonstrates the explicit dependence of F_β on r.

The performance of a classifier is often summarized by the area under the PR curve (AUPR), by analogy to the area under the ROC curve (AUROC). However, Flach and Kull (2015) argue that it is better to summarize precision-recall performance based on the F_1 score. This leads them to introduce the Precision Gain PrecG and Recall Gain RecG, defined as

$$\mathrm{PrecG} = \frac{\mathrm{Prec} - \pi}{(1-\pi)\mathrm{Prec}}, \qquad \mathrm{RecG} = \frac{\mathrm{Rec} - \pi}{(1-\pi)\mathrm{Rec}}. \tag{7}$$

Their 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). It is interesting to express PrecG and RecG in

[2] Hoiem et al. (2012) considered the PASCAL Visual Object Classes (VOC) dataset across 20 object classes, and chose their standard r based on the average proportion of positives across the classes.

[3] Or of one that provides a graded real-valued output, like a SVM.

terms of TPR, FPR and r. Using $1/(1-\pi) = 1 + r$ we obtain

$$\mathrm{PrecG} = \frac{1}{1-\pi} - \frac{r}{\mathrm{Prec}} = 1 + r - r\left(1 + \frac{1}{r}\frac{\mathrm{FPR}}{\mathrm{TPR}}\right) = 1 - \frac{\mathrm{FPR}}{\mathrm{TPR}}, \quad (8)$$

$$\mathrm{RecG} = \frac{1}{1-\pi} - \frac{r}{\mathrm{Rec}} = 1 + r\left(1 - \frac{1}{\mathrm{TPR}}\right). \quad (9)$$

Notice how PrecG is in fact independent of r, while RecG has an affine rescaling due to r. Interestingly, both PrecG and RecG each only depend on two out of the three quantities TPR, FPR and r.

The key point of the above analyses is to highlight the explicit effect of the class imbalance as expressed by r on the precision, F_β and the precision/recall gains, and to show how these quantities can be adjusted for different r if necessary. Like Hoiem et al. (2012), Siblini et al. (2020) make use of a fixed class ratio r_0, and use it to define AUPR, F-score and AUPR Gain scores that thus do not depend on r.

Acknowledgements

I thank Nick Radcliffe for a question that started this work off, Tom Dietterich for pointing out the work of Flach and Kull (2015), Peter Flach for pointing out a typo in eq. 6 in an earlier version, Wissam Siblini for alerting me to Siblini et al. (2020), the anonymous referees for comments that helped improve the paper, and Simão Eduardo, Alfredo Nazábal and Charles Sutton for helpful discussions.

References

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

Flach, P. A. and Kull, M. (2015). Precision-Recall-Gain Curves: PR Analysis Done Right. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 838–846.

Hoiem, D., Chodpathumwam, Y., and Dai, Q. (2012). Diagnosing Error in Object Detectors. In *Proceedings of the 12th European Conference on Computer Vision, ECCV 2012*.

Siblini, W., Fréry, J., He-Guelton, L., Oblé, F., and Wang, Y. Q. (2020). Master Your Metrics with Calibration. In Berthold, M., Feelders, A., and G., K., editors, *Advances in Intelligent Data Analysis XVIII. IDA 2020*. Springer, Cham. Lecture Notes in Computer Science, vol 12080.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann, second edition.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2017). *Data Mining*. Morgan Kaufmann, fourth edition.