The Effect of Class Imbalance on Precision-Recall Curves

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

May 29, 2020

The Receiver Operating Characteristic (or ROC) curve and the Precision-Recall (PR) curve are two ways of summarizing the performance of a binary classifier as the threshold for deciding between the two classes is changed.

The question we address here is how the PR curve is affected by the relative abundance of the positive and negative classes in the test data. 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 $p = \mathrm{P}/(\mathrm{P}+\mathrm{N})$. The ROC curve is a plot of TPR against FPR. As is well known, the ROC is invariant to p; this is immediate from the definitions of TPR and FPR, as they are ratios within the positives and negatives respectively. TPR and FPR are properties of the classifier and the threshold chosen.

Precision is defined as

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \frac{\mathrm{P}\cdot\mathrm{TPR}}{\mathrm{P}\cdot\mathrm{TPR}+\mathrm{N}\cdot\mathrm{FPR}} = \frac{\mathrm{TPR}}{\mathrm{TPR}+\frac{\mathrm{N}}{\mathrm{P}}\mathrm{FPR}}. \tag{2}$$

Thus the precision has an explicit dependence on $\mathrm{N}/\mathrm{P} = (1-p)/p$. Note that the $\mathrm{Prec} \to 1$ as $p \to 1$, and also that $\mathrm{Prec} \to 0$ as $p \to 0$ if $\mathrm{FPR} > 0$.

The precision-recall curve plots the precision against recall, which is another name for the true positive rate. As recall is invariant to class imbalance, we can consider how the precision varies with p at fixed recall. If we start with balanced classes at $p = 1/2$ and gradually decrease p[1], we see that the corresponding precision will decrease, because the denominator increases. This does not seem to be well known, perhaps because for a given problem we do not always have the freedom to manipulate p. However, for example 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 p for each class. To enable a fairer comparison, they suggested using "normalized precison", which would use a standard value of p across classes.

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

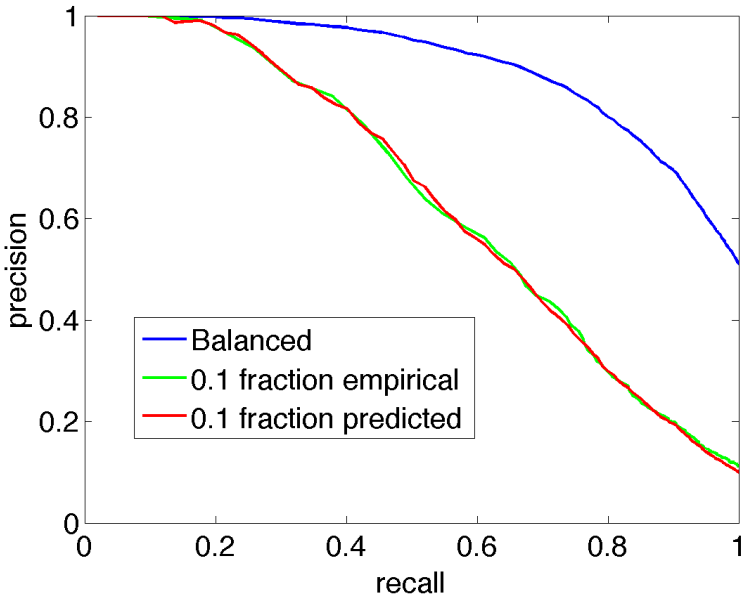

Figure 1: Precision-recall curve for varying p.

Eq. 2 allows us to *predict* how the PR curve will change with p. 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 $p = 1/2$ the blue curve was obtained, and for $p = 1/11$ the green 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).

Acknowledgements

I thank Simão Eduardo, Alfredo Nazábal and Charles Sutton for helpful discussions.

References

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*.

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