The Effect of Class Imbalance on Precision-Recall Curves

Christopher K. I. Williams
ckiw@inf.ed.ac.uk
School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, U.K.

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 $F_\beta$ 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 (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 $r$. I show explicitly how the PR curve and derived quantities like the $F_\beta$ 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, Frank, Hall, & Pal, 2017, sec. 5.8) for binary classification is summarized in Table 1. There are $P$ positive and $N$ negative data points in the data set, with the true-positive rate (TPR) and false-positive rate (FPR) defined as

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}, \quad FPR = \frac{FP}{FP + TN} = \frac{FP}{N}. \quad (1)$$

Let the fraction of positives in the data set be denoted by $\pi = P/(P + N)$, and define the ratio $r = P/N = \pi/(1 - \pi)$. If we consider Table 1 normalized by the sample size $n = P + N$, 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 $n \rightarrow \infty$, which describes the population properties of the classifier at the threshold chosen.

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
Table 1: Standard Notation for Binary Classification

| Predicted | Actual |
|-----------|--------|
|           | Positive | Negative | Sum |
| Positive  | TP      | FN       | P   |
| Negative  | FP      | TN       | N   |

the definitions of TPR and FPR, which are ratios within the positives and negatives, respectively. Empirical ROC curves will exhibit some variability as \( r \) varies (and indeed across different samples of the same size).

Precision is defined as

\[
\text{Prec} = \frac{TP}{TP + FP} = \frac{P \cdot TPR}{P \cdot TPR + N \cdot FPR} = \frac{TPR}{TPR + \frac{1}{r} FPR}.
\]

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

The precision-recall curve plots the precision against recall \( \text{Rec} \), 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 \( r \), we see that the corresponding precision will decrease because the denominator increases.

For population values of TPR, FPR, and \( r \), equation 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 Figure 1. In this case, a simple classification problem with two-dimensional gaussians was set up and a logistic regression classifier trained. For a test set with \( r = 1 \) and \( P = N = 5000 \), the blue curve was obtained, and for \( r = 0.1 \) (\( P = 500 \), \( N = 5000 \)), the green empirical curve. If at each value of recall, the blue curve is scaled as per equation 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).

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, Chodpathumwam, and Dai (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

\[1^{\text{PR curves are typically used when } r \text{ is small, for example, in information retrieval settings.}}\]
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 \( r_{\text{train}} \) 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,

\[
\log \frac{p(C_+|x)}{p(C_-|x)} = \log \frac{p(x|C_+)}{p(x|C_-)} + \log r_{\text{train}},
\]

where \( r_{\text{train}} = p(C_+)/p(C_-) \). For a generative classifier, the left-hand side is obtained from the right-hand side, and the effect of \( r_{\text{train}} \) is immediate. For a discriminative classifier, equation 3 can be used to understand the effect of \( r_{\text{train}} \) 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 \( r_{\text{train}} \) is to shift the threshold but not to change the sequence obtained.

The \( F_\beta \) measure is commonly used as a figure-of-merit that combines precision and recall. It is defined as a weighted harmonic average:

\[
\frac{1}{F_\beta} = \frac{1}{1 + \beta^2} \frac{1}{\text{Prec}} + \frac{\beta^2}{1 + \beta^2} \frac{1}{\text{Rec}}.
\]

\(^2\)Hoiem et al. (2012) considered the PASCAL visual object classes (VOC) data set 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.
Substituting the expression for the precision from equation 2, we obtain

\[
\frac{1}{F_\beta} = \frac{1}{1 + \beta^2} \frac{TPR + \frac{1}{r}FPR}{TPR} + \frac{\beta^2}{1 + \beta^2} \frac{1}{TPR},
\]

and hence

\[
F_\beta = \frac{(1 + \beta^2)TPR}{TPR + \frac{1}{r}FPR + \beta^2},
\]

which demonstrates the explicit dependence of \(F_\beta\) 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 \(\text{PrecG}\) and recall gain \(\text{RecG}\), defined as

\[
\begin{align*}
\text{PrecG} &= \frac{\text{Prec} - \pi}{(1 - \pi)\text{Prec}}, \\
\text{RecG} &= \frac{\text{Rec} - \pi}{(1 - \pi)\text{Rec}}.
\end{align*}
\]

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 \(\text{PrecG}\) and \(\text{RecG}\) in terms of TPR, FPR, and \(r\). Using \(1/(1 - \pi) = 1 + r\), we obtain

\[
\begin{align*}
\text{PrecG} &= \frac{1}{1 - \pi} - \frac{r}{\text{Prec}} = 1 + r - r \left(1 + \frac{1}{r} \frac{FPR}{TPR}\right) = 1 - \frac{\text{FPR}}{TPR}, \\
\text{RecG} &= \frac{1}{1 - \pi} - \frac{r}{\text{Rec}} = 1 + r \left(1 - \frac{1}{TPR}\right).
\end{align*}
\]

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

The key point of the analyses is to highlight the explicit effect of the class imbalance as expressed by \(r\) on the precision, \(F_\beta\), 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, Fréry, He-Guelton, Oblé, and Wang (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 \(r\).

Acknowledgments

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 equation 6 in an earlier version, Wissam Siblini for alerting me to Siblini et al. (2020), the anonymous referees for comments that helped improve the note, 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., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in neural information processing systems, 28 (pp. 838–846). Red Hook, NY: Curran.

Hoiem, D., Chodpathumwam, Y., & Dai, Q. (2012). Diagnosing error in object detectors. In Proceedings of the 12th European Conference on Computer Vision. Berlin: Springer.

Siblini, W., Fréry, J., He-Guelton, L., Oblé, F., & Wang, Y. Q. (2020). Master your metrics with calibration. In M. Berthold, A. Feelders, & G. Krempl (Eds.), Advances in Intelligent Data Analysis XVIII, IDA 2020 (pp. 457–469). Cham, Switzerland: Springer.

Van Rijsbergen, C. J. (1979). Information retrieval (2nd ed.). Waltham, MA: Butterworth-Heinemann.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data mining (4th ed.). San Francisco: Morgan Kaufmann.

Received June 26, 2020; accepted October 12, 2020.