2010 International Conference on Pattern Recognition

The binormal assumption on precision-recall curves

Kay H. Brodersen*†, Cheng Soon Ong*, Klaas E. Stephan† and Joachim M. Buhmann*

*Department of Computer Science, ETH Zurich, Switzerland; kay.brodersen@inf.ethz.ch
†Institute for Empirical Research in Economics, University of Zurich, Switzerland

***Abstract*—The precision-recall curve (PRC) has become a widespread conceptual basis for assessing classification performance. The curve relates the positive predictive value of a classifier to its true positive rate and often provides a useful alternative to the well-known receiver operating characteristic (ROC). The empirical PRC, however, turns out to be a highly imprecise estimate of the true curve, especially in the case of a small sample size and class imbalance in favour of negative examples. Ironically, this situation tends to occur precisely in those applications where the curve would be most useful, e.g., in anomaly detection or information retrieval. Here, we propose to estimate the PRC on the basis of a simple distributional assumption about the decision values that generalizes the established binormal model for estimating smooth ROC curves. Using simulations, we show that our approach outperforms empirical estimates, and that an account of the class imbalance is crucial for obtaining unbiased PRC estimates.**

***Keywords*-classification performance; generalizability; receiver operating characteristic; false discovery rate; information retrieval**

I. Introduction

Computing a meaningful estimate of generalizability is a key requirement for evaluating the performance of a classification algorithm [1]. In the case of binary classification, in particular, the empirical accuracy (the number of correct predictions divided by the number of test cases) frequently serves as an indicator of overall performance. However, looking exclusively at accuracies has, among others, the following two limitations [2]. First, an accuracy found to be significantly above 50% may be the result of a biased classifier tested on an imbalanced dataset [3], [4]. Second, the overall accuracy does not distinguish between the types of error that have been made. Whenever different costs are to be associated with different types of misclassification, a summary statistic is required that measures the performance of a classifier independently of how sensitivity and specificity are to be traded off.

One solution is to analyse the receiver operating characteristic (ROC) of a classifier [5], [6]. This analysis is not based on the binary predictions that the classifier made for each test case but on the ranked set of examples, as established by the underlying decision values, i.e., the internal scores that were computed for each example before applying a decision threshold. The ROC curve relates the true positive rate (TPR) to the false positive rate (FPR) obtained at every possible threshold. Thus, it provides an insight into the performance of a classifier that is independent of its detection threshold.

An increasingly popular alternative is to plot the TPR (also known as recall) against the positive predictive value (PPV), that is, the fraction of true positives in relation to all positive predictions (also known as precision). Comparing the TPR with the PPV has proven particularly useful in applications where the overall number of positive examples is small [7], e.g., in information retrieval. Replacing the FPR by the PPV turns the ROC curve into the precision-recall curve (PRC) and the area under the ROC curve (AUC) into the area under the PRC, also known as the average precision (AP).

Occasionally, the TPR is not plotted against the PPV but against the false discovery rate (FDR). However, since $\mathrm{PPV} = 1 - \mathrm{FDR}$, the two approaches are equivalent [7]. Thus, all concepts developed in this paper can be applied to both PRC and FDR curves.

In principle, the PRC could be constructed in the same way as the ROC curve: by varying the threshold and plotting the resulting empirical rate measures. In practice, however, estimating an empirical PRC is problematic as its shape is highly sensitive to idiosyncracies in the data, especially at high precisions, where the curve is most interesting [8], [9]. One way of addressing this problem is to resort to nonparametric approaches for deriving a smooth PRC estimate [9].

Here, we propose to replace the empirical PRC by a smooth estimate based on a simple distributional assumption. We begin by briefly reviewing the well-known binormal model for estimating smooth ROC curves (Section II). We then develop the main contribution of this paper. Unlike ROC curves, PRCs heavily depend on the degree of imbalance in the data. We therefore extend the model by also estimating the degree of class imbalance in the data (Section III). Based on this model, we derive an estimate of the PRC and the AP. Using simulations, we compare the two models, illustrate the impact of class-imbalance estimation on the PRC (Section IV), and briefly discuss our findings (Section V).

II. The receiver operating characteristic

In a binary classification setting, the receiver operating characteristic (ROC) provides a way of looking at the performance of a classifier that is independent of any particular

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1036

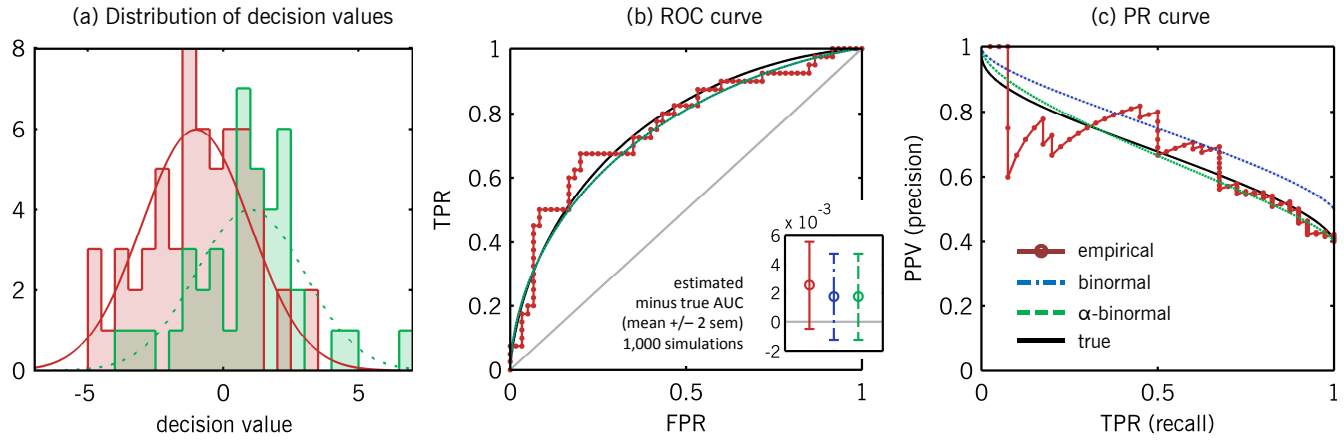

Figure 1. Simulation of empirical and model-based performance curves. Based on a set of generated decision values (a), the two diagrams in (b) and (c) show the difference between empirical curves (solid red), model-based curves based on the binormal model (dash-dotted blue), and model-based curves based on the α-binormal model (dotted green), in relation to ground truth (solid black). The inset in (b) shows the difference (mean difference $+/-$ 2 standard errors of the mean difference) between AUC estimates and ground truth across a set of 1,000 simulations.

decision threshold. Given n test cases, it is constructed by considering, at each possible threshold, the empirical confusion matrix that results from classification:

	actual $+$	actual $-$
predicted $+$	TP	FP
predicted $-$	FN	TN
total	P	N

In particular, the following threshold-dependent performance measures are commonly used:

$$\text{Accuracy} \quad ACC = \frac{TP + TN}{n} \tag{1}$$

$$\text{True positive rate} \quad TPR = \frac{TP}{P} \tag{2}$$

$$\text{False positive rate} \quad FPR = \frac{FP}{N} \tag{3}$$

$$\text{Positive predictive value} \quad PPV = \frac{TP}{TP + FP} \tag{4}$$

Derived from these quantities, the ROC curve is a plot of TPR vs. FPR across different thresholds [5], [6]. Thus, unlike the accuracy, it provides a threshold-independent way of assessing classification performance. A useful summary is provided by the Wilcoxon-Mann-Whitney statistic, more commonly known as the area under the ROC curve (AUC).

Given the decision values that were assigned to a set of test examples, we can compute empirical estimates for the TPR and the FPR, based on the confusion matrix that results from a given threshold t.

The main drawback of this approach is that the entries in the confusion matrix only change when t crosses one of the decision values. Thus, the empirical ROC curve suffers from a jagged appearance, especially when based on small datasets.

A well-known remedy is to replace empirical ROC curves by smooth estimates based on a simple parametric assumption about the underlying decision values [10], [6]. Based on a continuous decision threshold t, let $\mathcal{F}_+(t)$ and $\mathcal{F}_-(t)$ denote the resulting cumulative distribution functions of the positive and negative populations of decision values, respectively. Further, let $\alpha \in (0, 1)$ be the fraction of positive examples. The confusion matrix can then be rewritten in a parametric form:

	actual $+$	actual $-$
predicted $+$	$\alpha(1 - \mathcal{F}_+(t))$	$(1-\alpha)(1 - \mathcal{F}_-(t))$
predicted $-$	$\alpha\mathcal{F}_+(t)$	$(1-\alpha)\mathcal{F}_-(t)$
total	α	$1-\alpha$

Specifying the confusion matrix in this way yields two insights. First, neither TPR nor FPR depend on α. Thus, both the ROC curve and the AUC are independent of class imbalance, and so our model-based approach to their estimation does not need to estimate α. Second, eliminating t in the above equations leads to a simple description of the continuous ROC curve [6],

$$TPR = 1 - \mathcal{F}_+\left(\mathcal{F}_-^{-1}(1 - FPR)\right), \tag{5}$$

where, so far, all quantities are independent of a particular distributional assumption.

The binormal model

A natural way of deriving a specific instance of the parametric confusion matrix described above is to assume that decision values follow two independent Gaussian distributions: one for negative examples, and one for positive examples. This assumption leads to what is known as the *binormal model* [10]. Intriguingly, smooth ROC estimates on the basis of the binormal model are unaltered if decision values undergo a strictly increasing transformation, which

means the binormal assumption has a broad domain of application [6, Section 2.5].

III. The precision-recall curve

Although the ROC curve may be helpful in assessing the performance of a classifier independently of any given threshold, there are many situations in which additional measures are required. For example, when it comes to searching for relevant documents (information retrieval) or recognizing rare events (anomaly detection), the available data is typically heavily imbalanced in favour of the negative class. As a result, a classifier may perform poorly despite a superior AUC [3], [4]. In these cases, the precision-recall curve (PRC), which plots PPV vs. TPR across all thresholds, represents a more natural way of looking at classification performance [7], [8]. For example, it can be used to assess the recall under a given level of precision (e.g., PPV $\geq$ 90%), or to compute the area under the PRC, termed the average precision (AP). Note again that the PRC is formally equivalent to the false discovery rate (FDR) curve which is occasionally used instead and in which the AP is known as the area under the FDR curve.

We will now proceed to the main contribution of this paper, the proposition of a simple scheme for computing a smooth estimate of the PRC.

The α-binormal model

The binormal model does not include an estimate of the degree of class imbalance. This is because there is no need to consider the relative size of classes as long as we are concerned with imbalance-independent quantities such as the ROC curve or the AUC. However, when considering the PRC or the AP, this is no longer the case.

In order to derive smooth estimates of performance measures independently of whether or not they are sensitive to the degree of class imbalance, we propose to extend the binormal model by explicitly estimating not only the moments of the two Gaussians but also the mixture parameter α. We refer to this model as the *α-binormal model*. Given a set of maximum-likelihood parameter estimates ($\hat{\mu}_+$, $\hat{\sigma}_+^2$, $\hat{\mu}_-$, $\hat{\sigma}_-^2$) describing the mean and variance of the two class-conditional Gaussians, the model yields

	actual +	actual −
predicted +	$\alpha\,(1-\Phi_+(t))$	$(1-\alpha)\,(1-\Phi_-(t))$
predicted −	$\alpha\Phi_+(t)$	$(1-\alpha)\Phi_-(t)$
total	α	$1-\alpha$

where $\Phi_+(\cdot)$ and $\Phi_-(\cdot)$ are short for the Gaussian cumulative distribution functions $\Phi(\cdot;\hat{\mu}_+,\hat{\sigma}_+^2)$ and $\Phi(\cdot;\hat{\mu}_-,\hat{\sigma}_-^2)$, respectively. Thus, the α-binormal model allows us to construct a parametric confusion matrix that takes into account the degree of class imbalance observed in the given set of decision values. Based on this confusion matrix, we can write down the generic expression of the PPV (4) as a function of a particular decision threshold t,

$$PPV = \frac{\alpha(1-\Phi_+(t))}{\alpha(1-\Phi_+(t))+(1-\alpha)(1-\Phi_-(t))}, \tag{6}$$

where α is the fraction of positive examples in the data. Using $TPR = 1-\Phi_+(t)$, we can eliminate t and derive the functional form of the PRC as

$$PPV = \frac{\alpha\ TPR}{\alpha\ TPR+(1-\alpha)\big(1-\Phi_-\big(\Phi_+^{-1}(1-TPR)\big)\big)}, \tag{7}$$

corresponding to (5). Finally, as illustrated in Section IV, we can numerically approximate the integral

$$\int_0^1 PPV(TPR)\ d\ TPR, \tag{8}$$

to obtain an estimate of the average precision (AP). Note that this quantity is identical to the area under the false discovery rate (FDR) curve. MATLAB code for estimating performance measures based on the α-binormal model is available online.[1]

IV. Simulations

The utility of computing smooth estimates of the PRC could be illustrated using real-world data, but the comparison of smooth estimates with ground truth critically requires simulations. Here, we present two such simulations. First, we will look at the difference between empirical and model-based PRCs. Second, we will illustrate how the degree of class imbalance impacts on the bias in the estimate of the area under the PRC.

Empirical vs. model-based PRCs

In order to illustrate the difference between empirical and smooth estimates of the ROC curve and the PRC, we generated 100 decision values, including 60 values for a negative class (μ_-=−1, σ_-=2) and 40 decision values for a positive class ($\mu_+ = 1, \sigma_+ = 2$). The density functions as well as the resulting histograms are shown in Figure 1a. Mimicking real-world data (which may be Box-Cox transformed [11] in case of concerns about their normality), our simulated classes were neither heavily imbalanced nor completely balanced, and neither well-separated nor indiscriminable.

The empirical ROC curve and its model-based counterparts are shown in Figure 1b. While the empirical curve (dotted red line) represents but a rough approximation to ground truth (solid black line), the smooth curves provide much better estimates (blue and green lines). Note that the binormal model and the α-binormal model account for identical predictions in this case since the ROC curve is independent of the degree of class imbalance (see Section II). The inset shows the mean difference between the estimated and the true AUC, based on 1,000 simulations, indicating

[1] `http://people.inf.ethz.ch/bkay/downloads`

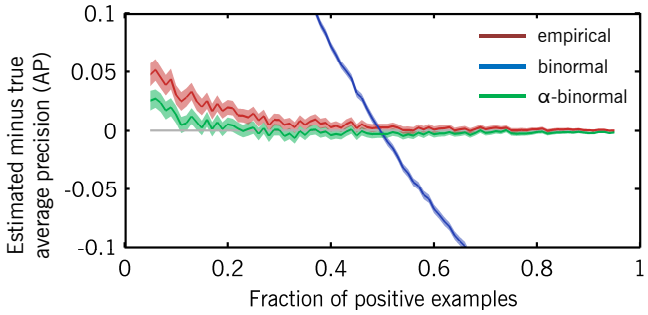

Figure 2. Effect of class imbalance on the area under the PRC. The smallest bias is exhibited by the PRC based on the α-binormal model (green) which, in particular, is significantly below the bias of the empirical curve when the data are imbalanced in favour of the negative class.

that neither the empirical (red) nor the model-based approach (blue and green) is biased.

The differences between the three types of estimation become apparent in the case of PRCs. As shown in Figure 1c, the empirical PRC is spiked and uneven, especially at low TPRs where single false discoveries elicit huge jumps. The binormal model, since it ignores class imbalance, overestimates the true curve. Only the α-binormal model provides a decent approximation to the true curve.

The impact of class imbalance on the AP

While the ROC curve is independent of class imbalance, the PRC is not. Based on 1,000 simulations for each point on the x-axis, we plotted the mean AP against the degree of class imbalance with which the underlying data was generated. As shown in Figure 2, the binormal model is only unbiased when classes are perfectly balanced ($\alpha = 0.5$). Much better predictions are afforded by the α-binormal model, whose estimates become unstable only when the number of positive examples is extremely small. Interestingly, when averaged over many simulations, the empirical AP provides an estimate almost as good as the α-binormal model, though its positive bias is more pronounced.

V. Discussion

Performance measures based on the precision-recall curve (or, equivalently, the false discovery rate curve) are helpful alternatives to the well-known ROC curve. However, empirical approaches to their estimation suffer from practical limitations, especially in the case of a small sample size and class imbalance in favour of negative examples. In this paper, we have shown that a model-based estimate can be computed on the basis of a distributional assumption about the underlying decision values and an explicit estimation of the class-mixture parameter α. Unlike more sophisticated nonparametric approaches [9], our scheme is simple and computationally inexpensive.

Using simulations, we chose to present data for the PRC and the AP. Generally, any other measure based on the comparison of two ranked sets could be investigated under the same smoothing scheme. Crucially, unlike the binormal model, the use of the α-binormal model is not restricted to quantities that are independent of α but may serve to compute smooth estimates of any quantity derived from the confusion matrix.

An important next step is to analyse in more detail the convergence rate of the scheme proposed in this paper. Further, while a simple analytical form exists for the AUC [6], to our knowledge no corresponding quantity has yet been proposed for the AP. These questions will be investigated in future studies.

Acknowledgment

This work was funded by the NEUROCHOICE project of SystemsX.ch (KES) and the University Research Priority Program 'Foundations of Human Social Behaviour' at the University of Zurich (KHB, KES).

References

[1] J. Langford, "Tutorial on practical prediction theory for classification," *Journal of Machine Learning Research*, vol. 6, pp. 273–306, 2005.

[2] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc. ICML*, 1997, pp. 445–453.

[3] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, pp. 429–449, 2002.

[4] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. ICPR*, 2010.

[5] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[6] W. J. Krzanowski and D. J. Hand, *ROC curves for continuous data*. CRC Press, 2009.

[7] K. Bleakley, G. Biau, and J.-P. Vert, "Supervised reconstruction of biological networks with local models," *Bioinformatics*, vol. 23, no. 13, pp. i57–i65, 2007.

[8] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. ICML*, 2006, pp. 233–240.

[9] S. Clémençon and N. Vayatis, "Nonparametric estimation of the precision-recall curve," in *Proc. ICML*, 2009, pp. 185–192.

[10] D. Dorfman and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory—a direct solution," *Psychometrika*, vol. 33, no. 1, pp. 117–124, 1968.

[11] K. Zou and W. Hall, "Two transformation models for estimating an ROC curve derived from continuous data." *Journal of Applied Statistics*, vol. 27, pp. 621–631, 2000.