Precision-recall operating characteristic (P-ROC) curves in imprecise environments

Thomas C.W. Landgrebe and Pavel Paclík and Robert P.W. Duin
Elect. Eng., Maths and Comp. Sc.
Delft University of Technology, The Netherlands
t.c.w.landgrebe, p.paclik, r.p.w.duin@ewi.tudelft.nl
Andrew P. Bradley
School of Info. Tech. and Elec. Eng., The University of Queensland, 4072, Australia
a.bradley@itee.uq.edu.au

Abstract

Traditionally, machine learning algorithms have been evaluated in applications where assumptions can be reliably made about class priors and/or misclassification costs. In this paper, we consider the case of imprecise environments, where little may be known about these factors and they may well vary significantly when the system is applied. Specifically, the use of precision-recall analysis is investigated and compared to the more well known performance measures such as error-rate and the receiver operating characteristic (ROC). We argue that while ROC analysis is invariant to variations in class priors, this invariance in fact hides an important factor of the evaluation in imprecise environments. Therefore, we develop a generalised precision-recall analysis methodology in which variation due to prior class probabilities is incorporated into a multi-way analysis of variance (ANOVA). The increased sensitivity and reliability of this approach is demonstrated in a remote sensing application.

1 Introduction

In pattern recognition, a common evaluation strategy is to consider classification $accuracy$ or its complement error-rate. In many empirical evaluations it is common to assume that the natural distribution (prior probabilities) of each class are known and fixed [9]. A further assumption often made is that the respective misclassification costs are known, allowing for the optimal decision threshold to be found [4]. Here, performance measures such as error-rate may be applied to compare different models as appropriate. However, in imprecise environments, misclassification costs can not be specified exactly, and class priors may not be reflected by the sampling, or even worse, the priors may in fact vary. Consequently, optimal threshold selection is ill-defined, and model selection based on a fixed threshold is unsuitable. For example, in remote sensing [8], the prior probability of various topography classes are not known a-priori, and may vary geographically. In such a situation, a performance measure should allow for an assessment that is either independent of these imprecise/ill-defined conditions or incorporates this variation.

Receiver Operator Characteristic (ROC) analysis [9], [10], has become a useful, and well-studied tool for the evaluation of classifiers in this domain. Measures such as the Area under the ROC (AUC) [10] allow for a performance evaluation independent of costs and priors by integrating performance over a range of decision thresholds. This can then be viewed as a performance measure that is integrated over a region of possible operating points.

In this paper we consider the evaluation of two-class classification problems where positive classes are to be distinguished from negative classes. In an imbalanced setting, where the prior probability of the positive class is significantly less than the negative class (the ratio of these being defined as the *skew* or λ), $accuracy$ is inadequate as a performance measure since it becomes biased towards the majority class [13]. That is, as the skew increases, $accuracy$ tends towards majority class performance, effectively ignoring the recognition capability with respect to the minority class. In these situations, other performance measures such as $precision$ (in conjunction with $recall$) may be more appropriate as they remain sensitive to the performance on each class. Figure 1 compares $accuracy$ and $precision$ as a function of skew for an example (a linear discriminant trained on the Highleyman distribution [5]), illustrating that as the skew increases, $accuracy$ tends towards TN_r (majority class performance), effectively ignoring the recognition capability with respect to the minority class.

We apply a ROC analysis methodology to the case of

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

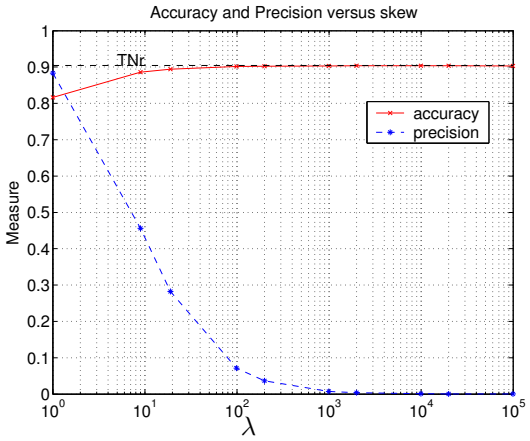

Figure 1. Comparing $accuracy$ and $precision$ for an example, as a function of skew (λ), illustrating the tendency of $accuracy$ to approach the majority class performance (TN_r) with increasing skew.

$precision$-$recall$ curves. However, we show that because $precision$ is dependent upon the degree of skewing, an additional dimension (the *skew*) must be introduced into the analysis. This effectively results in a 3-dimensional ROC *surface*. A similar approach was described in [3], where the relationships between a number of performance evaluation criteria were derived with respect to the ROC curve. In addition, we have previously presented an analysis specific to imbalanced problems, involving $precision$ operating characteristics for a number of selected operating points and priors [7]. Here however, we generalise this work so that the evaluation considers the entire operating surface, and integrated performance measures are then derived in a similar way to conventional ROC analysis. The performance of a number of models is statistically compared using a hypothesis-testing framework involving a 3-way analysis of variance (ANOVA) between classification thresholds, priors and models. We demonstrate the approach via a remote sensing application.

2 Formalisation

Consider a two-class classification problem between a positive and a negative class, ω_p and ω_n respectively, with priors π_p and π_n. An evaluation of a trained model is based on the outcomes following the application of a test set. In the 2-class case this results in a confusion matrix where test objects labelled by the trained classifier as positive fall into two categories: true positives TP and false positives FP. Correspondingly, true positive and false positive rates TP_r and FP_r, are computed by normalising TP and FP by the total number of positive (N_p) and negative (N_n) objects respectively, where N objects are involved in the test ($N = N_p + N_n$). Data samples labelled by the classifier as negative also fall in two categories, true negatives TN and false negatives FN. Also note that $TN_r = 1 - FP_r$, and $FN_r = 1 - TP_r$.

Although a confusion matrix shows all of the information about a classifier's performance, it is usual to extract measures from this matrix to illustrate specific aspects of the performance. For example:

1. Classification *accuracy*, or its complement error-rate ($error$), defined as $error = \frac{FN+FP}{N} = \pi_p FN_r + \pi_n FP_r$. This estimates the overall probability of correctly labelling a test sample, but combines results for both classes in proportion to the class priors;
2. *Recall* = TP_r. This indicates the probability of correctly detecting a positive test sample and is independent of class priors. TP_r is often utilised in medical applications where it is referred to as test *sensitivity*. In medical applications the complement to sensitivity is also used, namely *Specificity* (TN_r). Specificity indicates the probability of correctly detecting a negative test sample and is also invariant of class priors;
3. *Precision* = $\frac{TP}{TP+FP}$. This indicates the fraction of the positives detected that are actually correct. Precision effectively estimates an overall posterior probability and is therefore a meaningful performance measure when detecting rare events. Precision combines results from both positive and negative samples and so is class prior dependent. It is also often referred to as $purity$, or in medical applications as positive predictive value (PPV). Note: the complement to PPV is negative predictive value (NPV);
4. *Posfrac* = $\frac{TP+FP}{N}$. This measure is useful in applications requiring second-stage manual processing of the positive outcomes of the classifier (such as medical screening tests), and estimates the reduction in manual effort provided by the classification model.

These measures highlight different aspects of a model's classification performance and so selecting the most appropriate performance measure is clearly application dependent. In medical applications for example, sensitivity (TP_r) and specificity (TN_r) are well understood, can be related to the prior class probabilities, and so are well accepted by the end-users. Therefore, these measures are used almost exclusively in these applications. However, in applications such as database image retrieval and oil-spill detection from satellite radar images $precision$-$recall$ analysis is more appropriate [6]. In these applications $recall$ (TP_r) only really makes sense when combined with $precision$, as the prior class probabilities are unknown or highly variable. In these situations, end-users relate to $precision$-$recall$ curves as they indicate how many true positives are likely to be found in a typical search.

It is also worth noting that in a similar way in which $error$ is used as a scalar performance measure in well-defined pattern recognition problems, scalar measures such as the *F−measure* [11] are used in the well-defined $precision\text{-}recall$ case (the geometric mean of $precision$ and $recall$, in which the two measures are weighted equally), defined as $\frac{2TP_r}{TP_r+FP_r+1}$.

3 ROC analysis

The performance measures described before all relate to a single decision threshold, or operating point, for a classification model. In well defined environments, where class priors and misclassification costs are known, evaluation at a single (perhaps optimal) operating point is appropriate. However, in imprecise environments or when comparing models operating at different points, ROC analysis is more appropriate.

Given a two class problem (ω_p vs ω_n), a trained density-based classifier and a test set, the ROC curve is computed as follows[1]: the trained classifier is applied to the test set, and the aposteriori probability is estimated for each data sample. Then, a set of m thresholds ($\Theta = \theta_1, \theta_2, \ldots, \theta_m$) are applied to this probability estimate and corresponding data labellings are generated. This can be conceptualised as shifting the position of the decision boundary of a classifier across all possibilities. The confusion matrix is computed between each estimated set of labels and the true test-set labelling. The ROC curve now plots the TP_r as a function of the FP_r. This effectively results in a representation of all possible classification $accuracy$ values for a given classifier, and provided the train and test data are representative, the same ROC results irrespective of priors/costs.

It is well known that evaluation measures such as $accuracy$ vary with prior/cost [10]. Thus a classifier trained to, for example, the Bayes operating point, would report a different $accuracy$ as the priors vary. In order to maintain the Bayes error-rate, the decision threshold would have to be adjusted according to the variation in prior/cost. In cases where costs/priors are not defined well, there is a need to inspect performance for a range of different operating points and/or priors. If all operating points are used in the evaluation, the overall ROC curve will be invariant to priors [9]. Integrating performance over the whole ROC curve results in the Area Under the ROC curve (AUC) [1] [10], which is a scalar performance measure ranging from 0.5 (random classification) to 1.0 (ideal). It is also often more practical to compute the AUC over a limited range to suit the given problem.

$$AUC(\Theta) \quad = \quad \int TP_r(\Theta) dFP_r(\Theta) \tag{1}$$

[1]The true class-conditional distributions are typically not known, so the method we use to derive the ROC is an estimate of the true ROC.

This can be approximated non-parametrically via trapezoidal integration:

$$\begin{array}{c} AUC(\Theta) \approx \sum_{i=2}^{m} \Delta FP_r TP_r(\theta_i) + \frac{1}{2}\Delta TP_r \Delta FP_r \\ \Delta TP_r = TP_r(\theta_i) - TP_r(\theta_{i-1}) \\ \Delta FP_r = FP_r(\theta_i) - FP_r(\theta_{i-1}) \end{array} \tag{2}$$

The point to note here is that while the ROC curve, and therefore AUC, is invariant to priors/costs, in imprecise environments we are actually interested in the variability in performance as the priors vary (we want to select the best performing model across an expected range of priors). Therefore, the traditional ROC analysis tools are not appropriate and require extension to imprecise environments.

4 Precision-recall analysis

Whereas ROC analysis represents $TP_r(\Theta)$ against $FP_r(\Theta)$, the $precision\text{-}recall$ operating characteristics represent $TP_r(\Theta)$ against $precision(\Theta)$. As discussed in [7], we showed that $precision$ is in fact dependent on the priors, i.e., a new operating characteristic is obtained if the priors vary, as opposed to the ROC where thresholds/operating points and priors are synonymous. The consequence is that the operating characteristic constitutes a surface of operating points, with each prior resulting in a slice of this surface. The $precision$ definition can be written as:

$$precision(\Theta) = \frac{TP_r(\Theta)}{TP_r(\Theta) + \lambda FP_r(\Theta)} \tag{3}$$

This allows the performances to be obtained analytically, given an ROC (derived as in Equation 2). In Figure 2, an example of receiver (TP_r vs FP_r), and $precision\text{-}recall$ (TP_r vs $precision$) operating characteristic curves are shown for an example classifier and dataset. The $precision$ characteristics are shown for three different prior settings ($\pi_p = 0.5,\ 0.1,$ and 0.01) to demonstrate the prior dependence from a balanced to an imbalanced situation. It is clear that the $precision$ characteristic varies significantly with λ.

The AUC is computed by integrating across all classification thresholds Θ. Similarly, the $precision\text{-}recall$ characteristic can be integrated across both classification thresholds Θ and priors λ, thus obtaining an integrated performance measure, called $AUPREC$. This can again be derived using the trapezoidal approximation, resulting in Equation 4. With this formulation, the original ROC can be used, together with the given skew, to analytically compute the new performance measures.

$$\begin{array}{c} AUPREC(\lambda) = \int TP_r(\Theta) dprecision(\Theta, \lambda) \\ \approx \frac{1}{2} \sum_{i=2}^{m} \Delta TP_r [\frac{TP_r(\theta_i)}{TP_r(\theta_i) + \lambda FP_r(\theta_i)} + \frac{TP_r(\theta_{i-1})}{TP_r(\theta_{i-1}) + \lambda (FP_r \theta_{i-1})}] \end{array} \tag{4}$$

The $AUPREC$ results in a performance score for a single skew setting. However, we wish to estimate performance

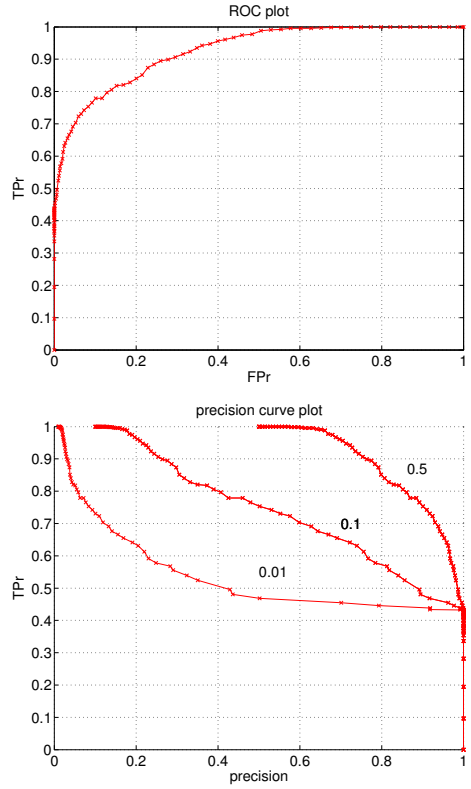

Figure 2. Demonstrating an ROC curve (top), and $precision$-$recall$ characteristics (bottom).

in problems in which the skew/costs are unknown, or only a range can be specified. In this case we wish to evaluate $precision$ across a range of priors. We therefore define an integrated $precision$ measure called $IAUPREC$. For a range of skew values (or priors) $\boldsymbol{\lambda} = \{\lambda^{lo}, \lambda^{hi}\}$, we obtain the $IAUPREC$ as shown in Equation 5.

$$IAUPREC(\lambda_{lo}, \lambda_{hi}) = \int_{\lambda_{lo}}^{\lambda_{hi}} AUPREC(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \quad (5)$$

5 Hypothesis testing by 3-way ANOVA

In this paper, we use analysis of variance (ANOVA) to test the null hypothesis that a number of models have, on the average, the same performance. If there is evidence to reject this hypothesis then we can look at the alternative hypothesis that one classifier has better performance than the others. ANOVA is simply an extension of Hypothesis tests of means (such as the t and F tests) to the case of multiple groups (in our case, > 2 classifiers) [12]. This avoids the necessity of performing multiple hypothesis tests for each pair of classifiers as we effectively test all hypotheses simultaneously.

ANOVA provides a method for splitting the variation in the data between multiple components (e.g., experimental error, classifier model, cross validation fold and prior probability). If the null hypothesis is true, then all components provide an independent estimate of the experimental error (that is, no components have a significant effect on performance). Clearly we expect that some of these components will affect performance and although we may not interested in them specifically, we use them as *blocking factors* to improve test sensitivity. Conventionally in ANOVA an F-test is used, however other non-parametric tests can also be used (e.g., rank statistics are used in the Friedman test). In this paper, we use a conventional F-test, and we specifically compare the efficacy of a 2-way ANOVA, with IAUPREC as the performance measure, to a 3-way ANOVA, with AUPREC as the performance measure, and π_p as a blocking factor. All tests are performed at the $p = 0.005$ level of significance, which gives a 1 in 200 probability of rejecting the null hypothesis *by chance*.

6 Experiments

In this section a number of experiments are undertaken in a real problem domain to demonstrate the efficacy of the proposed $precision$-$recall$ analysis. A remote sensing application is targeted, which we call *Satellite*[2]. As discussed in [8], this problem is appropriate because the prior probabilities of the various classes vary geographically. The data consists of 6435 multi-spectral values of a satellite image, with 36 dimensions (4 spectral bands in a 9 pixel neighbourhood). Six classes have been identified to characterise the topography, of which the second and fourth classes (cotton crop and damp grey soil) are considered ω_p (1329 examples), and the remaining ones ω_n (5106 examples). The goal of the experiments is to select a classifier that remains relatively robust to variations in the priors, measured in this case by $precision$.

Three classification models are compared, referred to as A, B, and C respectively, where the first uses a principal-component analysis representation (3 components), followed by a mixture of Gaussians classifier (3 mixtures per class), and the second two use the dissimilarity approach [2], using 15 and 50 randomly selected prototypes respectively, and a minimum-distance classifier. A 20-fold randomised hold-out method is used, in which 80% of the data is used in training, and the remainder for testing (cross-validation is not recommended for this dataset (image data), but we use it only for illustration of the principles).In comparing the models, we consider 3 measures:

- $AUPREC$ for $\pi_p = 0.5, 0.1, 0.01$, indicating the integrated $precision$ for various skew values.
- $IAUPREC([0.05, 0.20])$, indicating the integrated $precision$ for a range of priors $0.05 \leq \pi_p \leq 0.20$. This score is normalised by the area over the range.
- AUC, for reference purposes.

Results (with standard deviation) for the various measures are shown in Table 1. Initially, a general observation can

[2]Obtain from ftp://ftp.ics.uci.edu/pub/machine-learning-databases

Table 1. Summary of experimental results.

Model	(A)	(B)	(C)
$AUPREC(0.5)$	$0.554(0.014)$	$0.775(0.108)$	$0.803(0.084)$
$AUPREC(0.1)$	$0.554(0.013)$	$0.629(0.186)$	$0.781(0.082)$
$AUPREC(0.01)$	$0.552(0.013)$	$0.487(0.245)$	$0.734(0.075)$
$IAUPREC$	$0.554(0.013)$	$0.642(0.177)$	$0.783(0.082)$
AUC	$0.943(0.005)$	$0.825(0.046)$	$0.905(0.019)$

be made that the absolute measures indicate that the performance of C is superior to both A and B, and that B is superior to A. We note, however, that there is a large variance in these results, especially of B and C, which makes a firm conclusion hard to draw.

Considering the IAUPREC results, a 2-way ANOVA indicates that only algorithm C is statistically better than A and B (with an F-value of 21.04), and that there is no significant difference between A and B. However, the 3-way ANOVA shows a significance between all 3 models (F-value of 483.85), with C being superior to B, and B being superior to A. This result indicates that the 3-way ANOVA is more sensitive to model differences since it directly incorporates the variance due to the priors.

Performing a 3-way ANOVA on the $AUPREC$ measures for the 3 different prior values shows that model C is indeed the best, significantly better than both B and A. Similarly, B is significantly better than A over all 3 the priors. Another observation that can be made for the three $AUPREC$ measures is that models A and C remain very stable with respect to a change in the skew, whereas model B is sensitive to skew. This is a very important result, since for a balanced case, models B and C result in similar performance ($AUPREC$ scores of $0.775(0.108)$ and $0.803(0.084)$ respectively). For the case in which $\pi_p = 0.01$, the $AUPREC$ performance for B diminishes to $0.487(0.245)$, whereas C remains relatively stable at $0.734(0.075)$. The IAUPREC score indicates a lower score over $0.05 \leq \pi_p \leq 0.20$, corroborating the fact that B is sensitive to skew. Model A is extremely insensitive to skew over the range, but because of the high bias, it would probably not be considered. These observations point out the importance of the $precision$ analysis proposed here for evaluating imbalanced, imprecise problems. The ANOVA analysis also indicated that there is a significant difference between the $AUPREC(0.5)$ and $AUPREC(0.01)$ measures, but not between the $AUPREC(0.5)$ and $AUPREC(0.1)$, and the $AUPREC(0.1)$ and $AUPREC(0.01)$ measures. These experiments demonstrate practical application of the $precision$-$recall$ analysis, and and also the importance of incorporating the priors as an additional source of variance in hypothesis testing.

7 Conclusions

In this paper we have presented an extension of the traditional ROC analysis methodology in which we form a 3-dimensional $precision$-$recall$ ROC surface. Here the class priors represent the third dimension as the $precision$ measure is dependent on the class priors. This evaluation methodology was demonstrated on a remote sensing application where priors are known to vary over a fixed range. Models were compared using a 3-way ANOVA test in order to incorporate the priors as an additional source of variation. Experiments showed that the incorporation of the priors results in a more sensitive hypothesis test than the 2-way ANOVA test. This demonstrated the efficacy of this approach in highlighting classifiers that are stable over variations in the priors, and so are suitable for application in imprecise environments.

Acknowledgements: This research is/was supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs, The Netherlands.

References

[1] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[2] R. Duin, D. de Ridder, and D. Tax. Experiments with object based discriminant functions; a featureless approach to pattern recognition. *Pattern Recognition Letters*, 18(11-13):1159–1166, 1997.

[3] P. Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. *ICML-2003, Washington DC*, pages 194–201, 2003.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.

[5] W. Highleyman. Linear decision functions, with application to pattern recognition. *Proc. IRE*, 49:31–48, 1961.

[6] M. Kubat, R. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning, Special issue on applications of machine learning and the knowledge discovery process*, 30:195–215, 1998.

[7] T. Landgrebe, P. Paclík, D. Tax, S. Verzakov, and R. Duin. Cost- based classifier evaluation for imbalanced problems. *SSSPR+SPR, Lisbon, Portugal*, pages 762–770, 2004.

[8] P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. *ICML-2001*, pages 298–305, 2001.

[9] C. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 3(4), 1978.

[10] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.

[11] C. J. Van Rijsbergen. *Information Retrieval*. London, Butterworths, second edition, 1979.

[12] R. Walpole. *Introduction to Statistics*. Macmillan, New York, third edition, 1982.

[13] G. Weiss. The effect of small disjuncts and class distribution on decision tree learning. *PhD. Dissertation, Department of Computer Science, Rutgers University*, May 2003.