Cost-Based Classifier Evaluation
for Imbalanced Problems

Thomas Landgrebe, Pavel Paclík, David M.J. Tax,
Serguei Verzakov, and Robert P.W. Duin

Elect. Eng., Maths and Comp. Sc., Delft University of Technology, The Netherlands
{t.c.w.landgrebe,p.paclik,d.m.j.tax,s.verzakov,r.p.w.duin}@ewi.tudelft.nl

Abstract. A common assumption made in the field of Pattern Recognition is that the priors inherent to the class distributions in the training set are representative of the true class distributions. However this assumption does not always hold, since the true class-distributions may be different, and in fact may vary significantly. The implication of this is that the effect on cost for a given classifier may be worse than expected. In this paper we address this issue, discussing a theoretical framework and methodology to assess the effect on cost for a classifier in imbalanced conditions. The methodology can be applied to many different types of costs. Some artificial experiments show how the methodology can be used to assess and compare classifiers. It is observed that classifiers that model the underlying distributions well are more resilient to changes in the true class distribution than weaker classifiers.

1 Introduction

Many typical discrimination problems can be expressed as a *target* versus *non-target* class problem, where the emphasis of the problem is to recover *target* examples amongst outlier or *non-target* ones. ROC analysis is often used to evaluate a classifier [9], depicting the operating characteristic in terms of the fraction of *target* examples recovered (True Positive rate or TP_r), traded off against the fraction of *non-target* examples classified as *target* (False Positive rate or FP_r). The ROC curve is a useful tool to optimise the trade-off between TP_r and FP_r. A loss matrix is often applied to these types of problems in an attempt to specify decision boundaries well suited to the problem, as discussed in [2], and [1]. Both TP_r and FP_r are invariant to changes in the class distributions [10].

TP_r and FP_r are not always the only costs used in assessing classifier performance. Some applications are assessed with other cost measurements, typical ones including *accuracy*, *purity/precision*, and *recall*, discussed in [6], and [8]. An example of this could be in automatic detection of tumours in images, where a human expert is required to make a final decision on all images flagged by the classifier as *target* (called the Positive fraction or $POSfrac$). In this application we could expect the number of *target* images (images actually depicting a tumour) to be less abundant than *non-target* images, and consequently the recognition system would be expected to minimise the amount of manual inspection
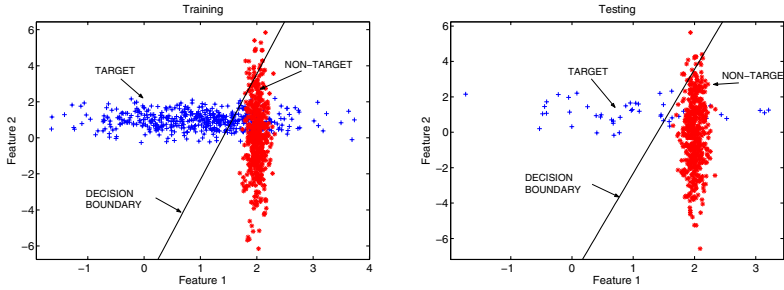
Fig. 1. Scatter diagrams to illustrate the example for balanced training conditions (left plot), but where the true class distribution is imbalanced (a prior of $\frac{1}{10}$ with respect to the *target* class simulating a real class distribution), shown in the right plot. Data is generated by the Highleyman distribution as in [4], proposed in [7].

required by the expert post-classification. A high $POSfrac$ would result in an inefficient automatic system, resulting in a high manual cost. Cost-based measurements such as $POSfrac$ and *purity* are dependent on true class distributions (as opposed to TP_r and FP_r) [5]. In some cases the actual distributions may be impossible to estimate or predict, implying that these costs may vary. It has been found that when *non-target* classes outnumber *target* classes, the effect on the costs of interest may be worse than expected [11], [8]. An example of this is shown in Figure 1 – here a balanced dataset is used for training (equal priors), but the right plot depicts the situation that arises when the true class distribution differs. Here the absolute number of *non-target* examples misclassified as *target* becomes comparable to the number of *target* examples correctly classified.

In this paper we discuss the evaluation with respect to cost of two-class discrimination problems between *target* and *non-target* classes in which the true class distribution is imbalanced and may vary (i.e. the abundance of examples for the two classes differ, called skewing). This is an important practical question that often arises, discussed and demonstrated here using synthetic examples in which the costs can easily be understood and compared. The objective is to formulate a procedure for evaluating classification problems of this nature. We show that in some situations where the true class distribution is extremely skewed in favour of the *non-target* class, the costs measurements could degrade considerably. As an example of how the proposed rationale can be applied, we choose two costs that are important to many applications, namely TP_r and $POSfrac$. Their relation is computed in conjunction with the ROC curve. These $POSfrac$ representations can be used to quickly, intuitively, and fairly, assess the outcome of the classifier for a given class distribution, or for a range of hypothetical class distributions (if it is unknown or varying). In a similar way, *Purity* or another cost measure could be used as part of the assessment procedure.

This paper is organised as follows: Section 2 introduces a theoretical framework, and shows how a classifier for the *target* versus *non-target* problem can be evaluated, discussing the construction of operating characteristics for the
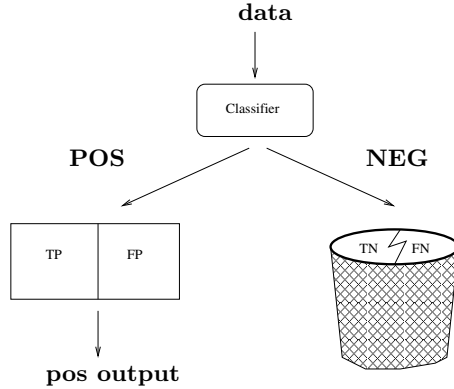
Fig. 2. A representation of the classification scheme discussed in this paper, showing a 2-class problem between a *target* and *non-target* class.

two costs emphasised here, namely TP_r and $POSfrac$. Section 3 consists of a discussion of the effect that skewed class-priors can have on the costs. A few experiments are performed in section 4 to illustrate the concepts discussed in the paper, showing direct application of the proposed methodology. Conclusions are given in section 5.

2   Classifier Evaluation between Two Classes

2.1   Notation and Problem Formulation

Consider the representation of a typical classification problem in Figure 2. Here it can be seen that a trained classifier analyses each incoming example, and labels each one as either positive (POS) or negative (NEG).

After classifying objects, four different object classifications can be distinguished (see Table 1). Data samples labeled by the tested classifier as *target* (the $POSfrac$) fall into two categories: true positives TP (true targets) and false positives FP (true non-targets). Corresponding true positive and false positive ratios TP_r and FP_r are computed by normalising TP and FP by the total amount of true targets N_t and non-targets N_n respectively.

Table 1. Defining a confusion matrix.

		estimated labels	
		target	non-target
true labels	target	TP	FN
	non-target	FP	TN

Data samples labeled by the classifier as non-target also fall in two categories, namely true negatives TN and false negatives FN. Note that $TN_r = 1 - FP_r$, and $FN_r = 1 - TP_r$. The examples labeled by the classifier as *target* are denoted POS, and those classified as *non-target* are denoted NEG. The fraction of

objects which are positively labeled is the 'Positive Fraction', $POSfrac$, defined in equation 1.

$$POSfrac = \frac{POS}{N} = \frac{POS}{POS + NEG} \tag{1}$$

where N is the total number of objects in the test set. The fraction of examples in the POS output of the classifier that are really *target* is called the *purity/precision* of the classifier, defined as $Purity = \frac{TP}{POS}$. TP_r can be written as ($TP_r = \frac{TP}{TP+FN}$), and is equivalent to *Recall*. A TP_r of 90% implies that 90% of all the target objects will be classified positive by the classifier (classified in the left branch in Figure 2). The $POSfrac$ tends to increase with TP_r for overlapping problems, indicating a fundamental trade-off between the costs. If $P(C_t)$ is very low it would be expected that the $POSfrac$ will be very small. However it will be shown in the experiments that this is not always the case – overlapping classes and weak classifiers can result in a very undesirable $POSfrac$, depending on the operating condition.

2.2   ROC Analysis and $POSfrac$ Analysis

Given a two class problem (*target* vs *non-target*), a trained density-based classifier and a test set, the ROC curve is computed as follows: the trained classifier is applied to the test set and the aposteriori probability is estimated for each data sample. Then, a set of thresholds Θ is applied to this probability estimate and corresponding data labelings 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 labeling. The ROC curve now plots the TP_r as a function of the FP_r (see the left plot in Figure 3).

Note that the ROC curve is completely insensitive to the class priors, depending only on the class conditional probabilities. When the prior of one of the classes is increased (and therefore the probability of the other class is decreased), both the TP_r and the FP_r stay exactly the same (for a fixed classifier), although the absolute number of target and non-target objects change. Costs dependent on class distribution such as $POSfrac$ are considered, the ROC curve alone is not sufficient to assess performance.

In order to compute the corresponding $POSfrac$ operating characteristic for the classifier, the same set of thresholds Θ are used. Equation 1 can be written as equation 2, which can then be posed in terms of the ROC thresholds as in equation 3.

$$POSfrac = \frac{TP + FP}{N} = \frac{TP + FP}{TP + FP + FN + TN} = \frac{TP_r N_t + FP_r N_n}{N} \tag{2}$$

$$POSfrac(\Theta) = \frac{TP_r(\Theta)N_t + FP_r(\Theta)N_n}{N} \tag{3}$$

Similarly the *Purity* cost can be derived as in Equation 4.

$$Purity(\Theta) = \frac{TP(\Theta)}{TP(\Theta) + FP(\Theta)} = \frac{TP_r(\Theta)N_t}{TP_r(\Theta)N_t + FP_r(\Theta)N_n} \tag{4}$$
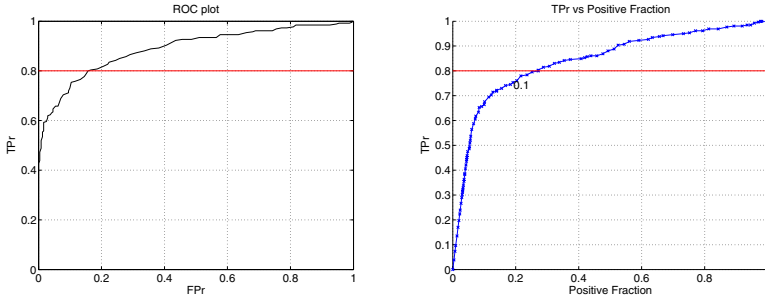
Fig. 3. ROC and $POSfrac$ plots for a linear discriminant classifier applied to High-leyman data, where $P(C_t) = 0.1$. Plots are made with respect to the *target* class.

3   The Effect of Class Imbalance on Classifier Performance

$POSfrac$ and *Purity* are two costs that are dependent on the skewness (imbalance) of the true class distribution. Define the true prior probability for the *target* class as $P(C_t)$, and for the *non-target* class as $P(C_n)$. Equations 3 and 4 can be written in terms of prior probabilities as shown in equations 5 and 6 respectively. Note that $P(C_t) = \frac{N_t}{N}$, $P(C_n) = \frac{N_n}{N}$, and $\frac{P(C_n)}{P(C_t)} = \frac{N_n}{N_t} = skewratio$.

$$POSfrac(\Theta) = \frac{TP_r(\Theta)N_t}{N} + \frac{FP_r(\Theta)N_n}{N} = P(C_t)TP_r(\Theta) + P(C_n)FP_r(\Theta) \quad (5)$$

$$Purity(\Theta) = \frac{TP_r(\Theta)}{TP_r(\Theta) + \frac{P(C_n)}{P(C_t)}FP_r(\Theta)} \quad (6)$$

Interestingly it is clear from equation 5 that the $POSfrac$ tends to FP_r as the skew ratio increases, and thus for extremely low $P(C_t)$ the ROC representation could be used alone in depicting the trade-off between costs. Now given an actual class-distribution and an operating condition, the $POSfrac$ and *Purity* can be calculated. For example, the corresponding ROC and $POSfrac$ curves for the example in Figure 1 is shown in Figure 3 for a linear discriminant classifier. Here $P(C_t) = 0.1$. It can be seen that for this condition, a TP_r of 80% would result in a $POSfrac$ of just under 30%. However, it could be that the exact class distribution is not known, and in fact sometimes the class distribution can vary from very small to extremely high levels. In these cases it can still be possible to evaluate and compare classifiers by investigating the operating characteristics across a range of priors in implementation. For example in the overlapping class problem in Figure 1, if the actual class distribution is completely unknown, it could be of value to investigate the TP_r and $POSfrac$ operating characteristic for a range of operating conditions. In Figure 4 the operating characteristic is shown for $P(C_t) = 0.5$, 0.1, and 0.001. The results for the linear classifier are shown in the left plot (the ROC plot remains constant as in the left plot of Figure 3). The right plot then shows the operating characteristic for a quadratic
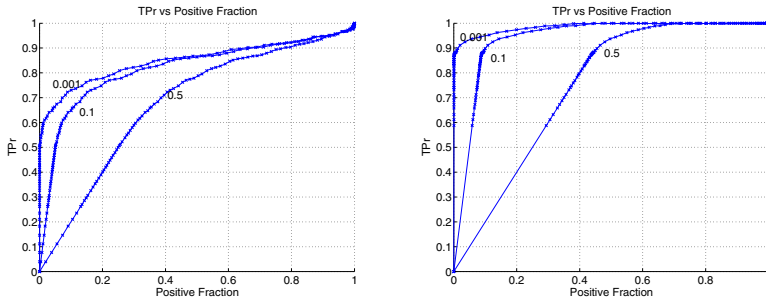
Fig. 4. $POSfrac$ plots for a linear discriminant (left plot) and quadratic discriminant (right plot) classifier applied to Highleyman data, where $P(C_t) = 0.5$, 0.1, and 0.001. The values on each curve represent different true class distributions.

classifier under the same conditions. It is clear that this classifier is much more resilient to changes in the true class distribution. Clearly as the linear classifier reaches a TP_r of 75% for a $POSfrac$ of 0.1, the quadratic classifier is substantially better at over 90%. Considering the case at which the classifiers are set to operate at a TP_r of 80%, a balanced dataset results in a $POSfrac$ of 54.9%, and a *Purity* of 77.9% for the linear classifier. These may be acceptable results. However when the class distribution changes such that $P(C_t) = 0.01$ (1 example in 100 is a *target*) the results become much worse. Whereas it could be expected that the $POSfrac$ should also drop by two orders of magnitude since the priors did so, for the linear classifier the $POSfrac$ only dropped to 24.8%. Conversely the quadratic classifier shows much better performance and resilience to changes of the true class distribution (in this case).

The simple example discussed shows that even if the true class distribution is unknown (or varies), two classifiers can still be compared to an extent. This becomes more useful when the underlying structure of the data is unknown and the choice of classifier less obvious. One example of this type of problem is in the field of geological exploration where the prior probabilities of different minerals change geographically, and often only a range of true class distributions is known.

4   Experiments

A number of experiments on artificial data are carried out in order to illustrate the effect (and indeed severity) that skewed data can have on costs in a number of situations. Four different classifiers are implemented for each data set, trained using a 30-fold cross-validation procedure. Each classifier is trained with equal training priors. For each classifier an ROC plot is generated such as in Figure 3, as well as the TP_r versus $POSfrac$ relation for the same thresholds as the ROC plot, generated for the following class distributions (as in Figure 4): $P(C_t) = 0.5$, 0.1, and 0.001. Thus each classifier is assessed from the case of balanced class distribution, to cases where the sampling is extremely skewed. In order to easily compare results a case study is performed for each classifier to estimate
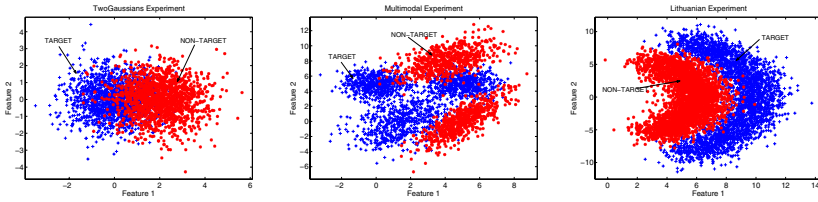
Fig. 5. Scatter diagrams of the 2-class experiments (the *Highleyman* plot is shown in Figure 1).

the effect on cost when the TP_r operating condition is fixed at 80% (in the same way $POSfrac$ or *Purity* could be used as the independent variable as the specification for the classifier). This allows each classifier to be compared easily, but still keeping cognisant that the results are for a specific operating condition only[1]. Note here that the objective of the classifier is to maintain the specified TP_r, and at the same time minimise the $POSfrac$ (and in some cases it could be more important to maximise *Purity*, but they are inversely dependent, i.e. a low $POSfrac$ results in a high *Purity*, since it can be shown that $POSfrac = \frac{TP}{N} \frac{1}{Purity}$).

The following classifiers are trained and evaluated for each data set: Linear Discriminant (LDC), Quadratic Discriminant (QDC), Mixture of Gaussian (GAUSSM), trained using the standard Expectation-Maximisation procedure, and the Parzen Classifier, as in [3] where the width parameter is optimised by maximising the log-likelihood with a leave-one-out procedure. The classifiers range from low to high complexity, the complex ones hypothetically capable of handling more difficult discrimination problems. These are all density-based classifiers, capable of utilising prior-probabilities directly, and can be compared fairly. The datasets are illustrated in Figure 5, corresponding to the following experiments, where 1500 examples are generated for each class: *TwoGaussians* with two overlapping homogenous Gaussian classes with equal covariance matrices and a high Bayes error is high at around 15.3%; *Highleyman* consisting of two overlapping Gaussian classes with different covariance matrices according to the Highleyman distribution (as in the *prtools* toolbox [4]); *Multimodal*, a multi-modal dataset with two modes corresponding to the first class, and three to the second; *Lithuanian* where two rather irregular classes overlap. All are computed under the *prtools* library [4].

4.1   Results of Experiments

For conciseness only the case study results are presented, showing the effect on $POSfrac$ cost for a single operating point, across a number of different class distributions. However the ROC and $POSfrac$ representations for the linear classifier in the *Highleyman* experiment have been shown before in the left plots

---

[1] A different operating point could for example be in favour of a different classifier.

Table 2. $POSfrac$ results (with standard deviations shown) for the four data sets, where the classifiers are fixed to operate to recover 80% of *target* examples. Results are shown for three different class distributions.

Experiment	$P(C_t)$	LDC	QDC	MOGC	PARZEN
TwoGaussians	0.5	$46.08 \pm 0.51\%$	$45.99 \pm 0.50\%$	$47.86 \pm 0.77\%$	$45.58 \pm 0.54\%$
	0.1	$18.94 \pm 0.92\%$	$18.78 \pm 0.90\%$	$22.14 \pm 1.38\%$	$18.05 \pm 0.98\%$
	0.001	$12.22 \pm 1.02\%$	$12.05 \pm 1.00\%$	$15.78 \pm 1.53\%$	$11.24 \pm 1.08\%$
Highleyman	0.5	$52.14 \pm 1.23\%$	$40.00 \pm 0.00\%$	$40.00 \pm 0.00\%$	$40.00 \pm 0.00\%$
	0.1	$29.85 \pm 2.22\%$	$8.00 \pm 0.00\%$	$8.00 \pm 0.00\%$	$8.00 \pm 0.00\%$
	0.001	$24.33 \pm 2.46\%$	$0.08 \pm 0.00\%$	$0.08 \pm 0.00\%$	$0.08 \pm 0.00\%$
Multimodal	0.5	$69.00 \pm 1.06\%$	$50.25 \pm 0.94\%$	$41.20 \pm 0.22\%$	$40.93 \pm 0.21\%$
	0.1	$60.21 \pm 1.90\%$	$26.46 \pm 1.68\%$	$10.15 \pm 0.40\%$	$9.68 \pm 0.38\%$
	0.001	$58.03 \pm 2.11\%$	$20.57 \pm 1.87\%$	$2.47 \pm 0.45\%$	$1.94 \pm 0.42\%$
Lithuanian	0.5	$46.81 \pm 0.81\%$	$42.10 \pm 0.33\%$	$40.16 \pm 0.04\%$	$40.06 \pm 0.04\%$
	0.1	$20.26 \pm 1.45\%$	$11.77 \pm 0.59\%$	$8.28 \pm 0.07\%$	$8.11 \pm 0.07\%$
	0.001	$13.68 \pm 1.61\%$	$4.27 \pm 0.65\%$	$0.39 \pm 0.08\%$	$0.21 \pm 0.08\%$

of Figures 3 and 4 respectively. Table 2 shows the $POSfrac$ results for the three different class distributions for the chosen operating point. In the *TwoGaussians* all the classifiers show an extreme effect on cost as the skewness is increased. The $POSfrac$ remains above 10% even when the distribution is such that only one example in a thousand are *target*. Here the overlap between the classes does not allow for any improvement (a high Bayes error). In the *Highleyman* experiment it can be seen that only the linear classifier is severely affected by different class distributions for the given operating condition. This experiment suggests that an inappropriate or weak classifier can be disastrous in extreme prior conditions, even though it may have seemed acceptable in training (classifiers are often trained assuming balanced conditions). The *Multimodal* experiment presented a multimodal overlapping problem, and as expected the more complex classifiers fared better. Whereas the linear and quadratic classifiers showed a $POSfrac$ of over 20% with extreme priors, the mixture-model and parzen classifiers performed a lot better. Similar results were obtained in the *Lithuanian* experiment.

5   Conclusion

This paper discussed the effect of imbalanced class distributions on cost, concentrating on 2-class problems between a class of interest called a *target* class, and a less interesting *non-target* class. A methodology was proposed in order to compare classifiers with respect to cost under conditions in which training conditions are fixed (often balanced), and true class distributions are imbalanced or varying.

It was shown that even though costs such as the true and false positive rates are independent of the true class distributions, other important costs such as $POSfrac$ and *Purity* are dependent on it. Thus for these types of problems we

proposed that in addition to traditional ROC analysis, a similar analysis of the other costs should be made simultaneously, evaluating the effect of a changing class distribution.

Following some simple experiments, it was observed that in some cases, classifiers that appeared to perform well on balanced class distribution data failed completely in imbalanced conditions. Conversely some classifiers showed resilience to the imbalance, even when extreme conditions were imposed. Thus we conclude that especially in cases in which the underlying data structure is complex or unknown, an analysis of the effect of varying and imbalanced class distributions should be included when comparing and evaluating classifiers.

Acknowledgements

References

1. C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press Inc., New York, first edition, 1995.
2. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley - Interscience, second edition, 2001.
3. R.P.W. Duin. On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Trans. Computing*, 25:1175–1179, 1976.
4. R.P.W. Duin. *PRTools Version 3.0, A Matlab Toolbox for Pattern Recognition*. Pattern Recognition Group, TUDelft, January 2000.
5. P. Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. *ICML-2003 Washington DC*, pages 194–201, 2003.
6. D. J. Hand. *Construction and Assessment of Classification Rules, ISBN 0-471-96583-9*. John Wiley and Sons, Chichester, 1997.
7. W. Highleyman. Linear decision functions, with application to pattern recognition. *Proc. IRE*, 49:31–48, 1961.
8. M. Kubat and S. Matwin. Addressing the curse of imbalanced data sets: One-sided sampling. *Proceedings, 14th ICML, Nashville*, pages 179–186, July 1997.
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. G.M. Weiss and F. Provost. The effect of class distribution on classifier learning: an empirical study. *Technical report ML-TR-44, Department of Computer Science, Rutgers University*, August 2 2001.