

# Combining Accuracy and Prior Sensitivity for Classifier Design Under Prior Uncertainty

Thomas Landgrebe and Robert P.W. Duin

Elect. Eng., Maths and Comp. Sc., Delft University of Technology, The Netherlands  
{t.c.w.landgrebe, r.p.w.duin}@ewi.tudelft.nl

**Abstract.** Considering the classification problem in which class priors or misallocation costs are not known precisely, receiver operator characteristic (ROC) analysis has become a standard tool in pattern recognition for obtaining integrated performance measures to cope with the uncertainty. Similarly, in situations in which priors may vary in application, the ROC can be used to inspect performance over the expected range of variation. In this paper we argue that even though measures such as the area under the ROC (*AUC*) are useful in obtaining an integrated performance measure independent of the priors, it may also be important to incorporate the *sensitivity* across the expected prior-range. We show that a classifier may result in a good *AUC* score, but a poor (large) prior sensitivity, which may be undesirable. A methodology is proposed that combines both accuracy and sensitivity, providing a new model selection criterion that is relevant to certain problems. Experiments show that incorporating sensitivity is very important in some realistic scenarios, leading to better model selection in some cases.

## 1 Introduction

In pattern recognition, a typical assumption made is that class priors and misallocation costs are known precisely, and hence performance measures such as classification error-rate and classifier loss are typically used in evaluation. A topic that has received a lot of attention recently is the imprecise scenario in which these assumptions do not hold (see for example [9], [2], [1] and [10]), resulting in a number of tools and evaluations suited to this problem. In particular, receiver operator characteristic (ROC) curves [6] have become very popular due to their invariance to both class priors and costs, and are thus used as a basis for performance evaluation and classifier decision threshold optimisation in these imprecise environments. The Area Under the ROC (*AUC*) measure has thus been proposed, providing a performance evaluation that is independent of priors.

In this paper we argue (and show) that considering the integrated performance (*AUC*) alone may not be the optimal strategy for model selection in these situations. This is because the *AUC* measure discounts an important characteristic, namely the performance *sensitivity* across the prior range (we distinguish prior sensitivity from the sensitivity measure often used in medical decision making, which is equivalent to true positive rate). In fact, we show that in some cases, two

classifiers may compete in terms of *AUC*, but have significantly different sensitivities over the same prior range i.e. one of the classifiers may have a performance that varies rapidly from low to high values, whereas the other may be more stable. In some problems e.g. medical decision making, the former scenario may be unacceptable, emphasising the fact that this sensitivity should also be considered. A simple criterion is proposed that combines both *AUC* and sensitivity, called *AccSens*, allowing for a more appropriate criterion for some problems<sup>1</sup>.

The paper is organised as follows: Section 2 introduces the notation in the well-defined case, restricted to two-class problems for simplicity, and derives the ROC. In Section 3, the problem of uncertain/varying class priors is considered, discussing the *AUC* measure, which is invariant of priors. Section 4 discusses the importance of considering prior-dependent sensitivity in conjunction with integrated error, illustrated via a case study, and Section 5 subsequently introduces a new criterion, *AccSens*. A number of real experiments are presented in Section 6 that show some cases in which competing classifiers (using *AUC*) have significantly different sensitivities (and vice versa). Conclusions are presented in Section 7.

## 2 Problem Formulation and ROC Analysis

Consider a 2-class classification task between classes  $\omega_1$  and  $\omega_2$ , with prior probabilities  $P(\omega_1)$  and  $P(\omega_2)$  respectively, and class-conditional probabilities denoted  $p(\mathbf{x}|\omega_1)$  and  $p(\mathbf{x}|\omega_2)$ . Each object is represented by a feature vector  $\mathbf{x}$ , with dimensionality  $d$ . Figure 1 presents an example of a 1-dimensional, two-class example (means at  $-1.6$  and  $1.6$  respectively, and equal variances of 2), and  $\theta_d$  represents an equal prior, equal cost operating point.

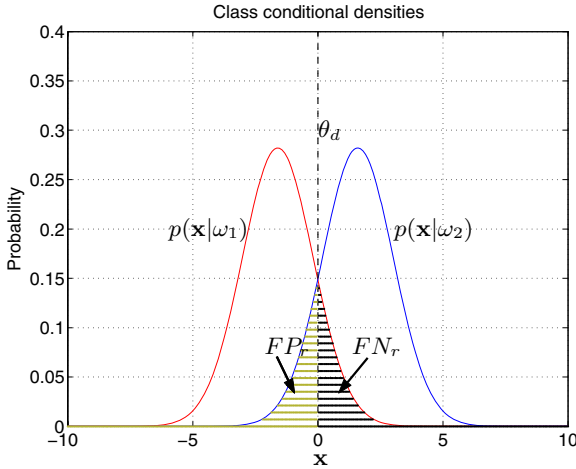
Two types of classification errors exist in the two-class case, namely the false positive rate ( $FP_r$ ), and the false negative rate ( $FN_r$ ), derived as follows, where  $\theta_w$  is the classification weight, determining the operating point:

$$\begin{aligned}
 FP_r(\theta_w) &= (1 - \theta_w)P(\omega_2) \int p(\mathbf{x}|\omega_2)I_1(\mathbf{x}|\theta_w)dx \\
 I_1(\mathbf{x}|\theta_w) &= \begin{cases} 1 & \text{if } \theta_w P(\omega_1)p(\mathbf{x}|\omega_1) > (1 - \theta_w)P(\omega_2)p(\mathbf{x}|\omega_2) \\ 0 & \text{otherwise} \end{cases} \\
 FN_r(\theta_w) &= \theta_w P(\omega_1) \int p(\mathbf{x}|\omega_1)I_2(\mathbf{x}|\theta_w)dx \\
 I_2(\mathbf{x}|\theta_w) &= \begin{cases} 1 & \text{if } (1 - \theta_w)P(\omega_2)p(\mathbf{x}|\omega_2) \geq \theta_w P(\omega_1)p(\mathbf{x}|\omega_1) \\ 0 & \text{otherwise} \end{cases}
 \end{aligned} \tag{1}$$

In the (realistic) case that distributions are not known, but are estimated from data (that is assumed representative), class conditional density estimates are denoted  $\hat{p}(\mathbf{x}|\omega_1)$  and  $\hat{p}(\mathbf{x}|\omega_2)$ , and population prior estimates are denoted  $\pi_1$  and  $\pi_2$ . These are typically estimated from an independent training set that

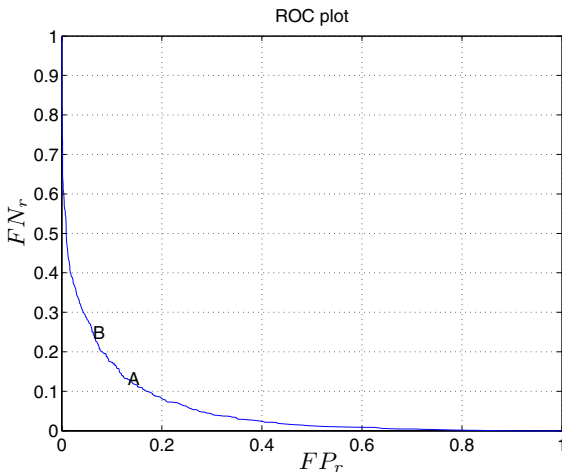
---

<sup>1</sup> Even though we emphasise a varying/uncertain class prior, the theory and analysis in this paper extends also to the related problem of varying misallocation costs [1], since these both have a similar impact from an ROC perspective in that a variation in either prior or cost results in a varying performance, strictly along the ROC [9].



**Fig. 1.** One-dimensional example illustrating two overlapping Gaussian distributions, and the two error-types associated with an equal error, equal cost operating point  $\theta_d$ .

is assumed drawn representatively from the true distribution. Equation 1 can then be extended to this case. The classifier weight  $\theta_w$  allows for  $FP_r$  to be traded off against  $FN_r$  (and vice-versa) to suit a given application. A particular setting of  $\theta_w$  results in a single operating point, with a corresponding  $FN_r$  and  $FP_r$  combination. Varying  $\theta_w$  (where  $0 \leq \theta_w \leq 1$ ) allows for specification of any desired operating point. An ROC plot [6] consists of a trade-off curve between  $FN_r$  and  $FP_r$  (as a function of  $\theta_w$ ). As such, the ROC is a useful tool in optimising and evaluating classifiers.



**Fig. 2.** ROC plot for the example in Figure 1

In the well-defined case that the priors can be estimated sufficiently well, and remain constant (e.g. estimated from training data, and generalising to an application scenario), the classification problem can be optimised (and evaluated) directly using the ROC. Strategies vary, but the most popular ones are as follows (also demonstrated on the ROC plot in Figure 2, which is the ROC plot generated from the example in Figure 1):

- **Equal error optimisation:** In this case,  $FP_r$  errors have the same consequences as  $FN_r$  errors, and the objective of the optimisation is to select a  $\theta_w$  such that  $FP_r = FN_r$ . In Figure 2, point *A* shows this operating point.
- **Cost-sensitive optimisation:** In some applications e.g. medical decision making, different errors have different misclassification costs (denoted  $c_1$  for  $FN_r$  errors, and  $c_2$  for  $FP_r$  errors). In this case  $\theta_w$  should be chosen such that the overall system loss is minimised, where the loss  $L$  can be computed as  $L = \theta_w c_1 \pi_1 FN_r + (1 - \theta_w) c_2 \pi_2 FP_r$  (profits are ignored here i.e. consequences of correct classifications). In Figure 2, point *B* illustrates an operating point for the equal prior case, with  $c_1 = 0.2$  and  $c_2 = 0.8$ .

### 3 Varying Priors, Uncertain Environments

The previous discussion assumed that the priors can be well estimated, and remain fixed in application. However, in many real applications this is not the case (see [9], [2]), confounding the problem of optimising the operating point and model selection (fairly comparing classifiers). In these cases, priors may not be known beforehand, or priors in an independent training set are not representative, or the priors may in fact vary in application. In these cases, even though an immediate optimisation and comparison is not appropriate, several techniques have been proposed for classifier design e.g. [9]. These typically use the ROC plot, since it has the desirable property of being independent of priors/costs (i.e. the same ROC results irrespectively), allowing classifier performance to be inspected for a range of priors (or costs). In particular, the Area Under the ROC ( $AUC$ ) measure [2] has been derived to give an integrated performance measure, allowing for model comparison independent of the prior. The  $AUC$  measure is defined as:

$$AUC = 1 - \int (FN_r) dFP_r \quad (2)$$

This performance measure results in a normalised score between 0 and 1, with 1 corresponding to perfect classification, 0.5 to random classification, and below 0.5 as worse than random (i.e. swap classifier labels). The  $AUC$  measure can also be computed over a range of priors/operating points, accounting for knowledge of the degree of uncertainty/variation. Thus, even though priors may be uncertain/varying, the best overall classifier can be chosen based on the most favourable integrated performance<sup>2</sup>.

<sup>2</sup> For threshold optimisation, the best strategy may be to use a  $\theta_w$  corresponding to the centre of the known range, or to apply the minimax criterion [3].

## 4 The Importance of Incorporating Sensitivity

In this paper we demonstrate that comparing classifiers in uncertain environments on the basis of integrated error ( $AUC$ ) only may not necessarily be the best strategy to take. This argument arose based on comparison of ROC plots for a number of competing classifiers (the experiments will show some realistic scenarios). It was observed that in some cases, two competing classifiers resulted in a similar  $AUC$  score, but inspection of the ROC made it clear that in one case, the performance range was small, but in another, much larger. This implies that for the problem in which priors may vary, the latter classifier may result in very poor performance at one extreme, and very good performance at the other. Depending on the problem, it may be much better to select the former model that is generally more stable over the expected prior range. Next a case study is presented to demonstrate such a scenario.

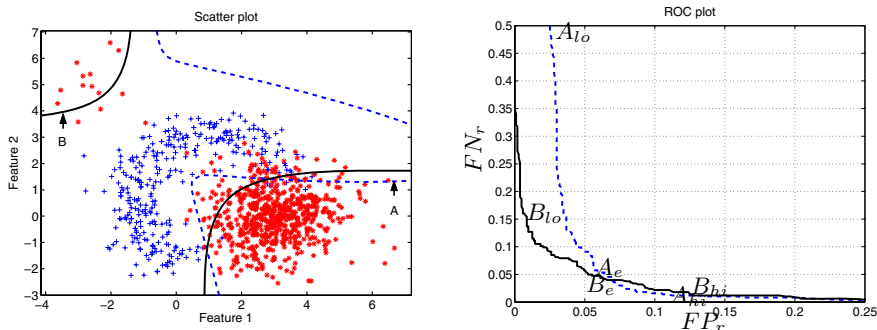
### 4.1 Case Study

Figure 3 depicts a demonstration of a model-selection scenario, comparing two different classifiers, denoted  $A$  and  $B$  respectively. Each classifier is trained on the distribution shown in the left plot, consisting of a two-class problem between  $\omega_1$  and  $\omega_2$  respectively, where  $\omega_1$  objects are drawn from  $N(\mu = 3.0, 2; \omega = 1) + \frac{1}{32}N(\mu = -2.0, 5.0; \sigma = 1)$  ( $N$  is the normal distribution with mean  $\mu$  and variance  $\sigma$ ), and  $\omega_2$  is one class from the banana distribution [4]. In this synthetic problem, 1500 objects are drawn from the true distribution to create a training set, and a further 1500 objects are drawn independently to result in an independent test set<sup>3</sup>. The two classifiers  $A$  and  $B$  are then trained on the training set, resulting in the decision boundaries at a single operating point as depicted in the left plot.  $A$  is a mixture of Gaussians classifier, with two mixtures chosen for  $\omega_1$ , and one for  $\omega_2$ . Classifier  $B$  is a support vector classifier with a second order polynomial kernel.

In this problem, it is assumed that the priors may vary (in application) such that  $0.05 \leq \pi_1 \leq 0.9$ , i.e. the abundance of  $\omega_1$  varies between 5% and 90%, and the costs are assumed equal (priors at the low and high extremes for  $\omega_1$  are denoted  $\pi_1^{lo}$  and  $\pi_1^{hi}$  respectively, computed by analysing where on the ROC the performance drifts to for the new prior, relative to the original operating point). The scatter-plot shows the resultant classifier decision boundaries of the two classifiers at the equal error point (i.e. equal priors). The ROC plot on the right depicts classifier performance for a range of operating points. For the first extreme, i.e.  $\pi_1 = 0.05$ ,  $A_{lo}$  and  $B_{lo}$  show the respective operating points for the two classifiers. For the second extreme, i.e. at  $\pi_1 = 0.9$ ,  $A_{hi}$  and  $B_{hi}$  again demonstrate how the operating point shifts.  $A_e$  and  $B_e$  show the positions of the equal-error points.

It can immediately be observed that the two classifiers have a distinct performance characteristic as a function of the prior values, even though the equal error points are rather similar. Table 1 compares some performance measures

<sup>3</sup> Cross-validation is ignored here as this example is for demonstration purposes only.



**Fig. 3.** Case study illustrating performance of two competing classifier models A and B. The left plot shows the data distribution, as well as the respective decision boundaries at a single operating point. The right plot is an ROC-plot for the two models across a range of priors.  $A_{lo}$  and  $B_{lo}$  are operating points at  $\pi_1 = 0.05$ , and similarly  $A_e$  and  $B_e$  are equal-error points, and  $A_{hi}$  and  $B_{hi}$  correspond to  $\pi_1 = 0.9$ .

between classifiers A and B. Firstly the error rate shows that both classifiers result in a similar performance for the equal prior case. The  $AUC$  measure integrates the classification error over the range of priors (between  $A_{lo}$  and  $A_{hi}$ ), and again this measure shows that both classifiers have similar performance across the prior range as a whole. However, when investigating the sensitivity with respect to the priors, it can be seen that classifier A is much more sensitive than B across the range, with the  $FN_r$  varying by up to 47.3%. Prior sensitivity (denoted  $Sens$ ) is computed as the Euclidean distance between the upper and lower prior range, from a  $\pi_1^{lo}$  situation, to  $\pi_1^{hi}$ . This is performed by considering the applicable ranges of  $FN_r$  and  $FP_r$ :

$$Sens = \frac{1}{\sqrt{2}} \sqrt{((FN_r(\pi_1^{lo}) - FN_r(\pi_1^{hi}))^2 + (FP_r(\pi_1^{hi}) - FP_r(\pi_1^{lo}))^2)} \quad (3)$$

This measure scales between 0 and 1, where a low score indicates the favourable condition of low sensitivity, whereas a high score indicates a large sensitivity to prior variation. Note that  $Sens$  is a simple measure in that it subtracts only the extreme values, justified by the fact that an ROC increases monotonically.

In this type of problem, classifier B is clearly more appropriate since it is far less sensitive to a perturbation in prior. It is also clear that the error-rate measure and  $AUC$  are not sufficient on their own in this case to choose the best models, and that the prior sensitivity across the range of interest should be included to aid in the model selection process.

## 5 Combining Accuracy and Sensitivity

The case study made it clear that in the uncertain prior situation, classifier sensitivity should be considered in conjunction with integrated error over the prior

**Table 1.** Performance measures for the synthetic example. Error-rate is denoted  $\epsilon$ ,  $AUC$  is the integrated error measure across the prior range, and the sensitivity  $Sens$  shows how much the performance varies ( $\frac{\%}{100}$ ) across the prior range.

Model	$\epsilon$	$AUC$	$Sens$
<b>A</b>	0.057	0.942	0.340
<b>B</b>	0.052	0.945	0.131

range. The next step is to develop a criterion that combines these two performance measures, that is useful for evaluation/model selection in this domain. It is conceivable that some problems may have different consequences for accuracy and sensitivity performances e.g. in some cases a low overall error (i.e. high  $AUC$ ) may be more important than a low sensitivity, in which case  $Sens$  could be weighted lower than  $AUC$ . In another case, e.g. medical decision making, a high sensitivity to priors may be more unacceptable than a slightly lower  $AUC$ . Thus, for generality, we introduce a weighting corresponding to each term, that can be used to penalise either according to the problem (analogous to misallocation costs). The  $AUC$  weight is denoted  $w_e$ , and the  $Sens$  weight is denoted  $w_s$ . We then define the combined measure, called  $AccSens$ , consisting of the geometric mean of the weighted sum of  $AUC$  and  $Sens$ , as defined in Equation 4. This is appropriate because both measures are scaled between 0 and 1. In the case that  $w_e$  and  $w_s$  are both set to unity (equal importance), the  $AccSens$  error measure also scales between 0 and 1, where a low score is favourable (the  $\frac{1}{\sqrt{2}}$  normalises the measure to this range).

$$AccSens = \frac{1}{\sqrt{2}} \sqrt{w_e((1 - AUC)^2) + w_s(Sens^2)} \quad (4)$$

For the case study example (assuming unit weighting), the  $AccSens$  errors are 0.244 for model  $A$ , and 0.100 for model  $B$ , indicating that  $B$  is superior.

## 6 Experiments

A number of experiments on realistic datasets have been undertaken. The objective is to select the most competitive model, considering the problem of varying/uncertain priors, with a known  $\pi_1$  range:  $0.1 \leq \pi_1 \leq 0.9$ . Additionally, we assume  $AUC$  and  $Sens$  are weighted equally. For each model, we investigate an integrated error over the prior range ( $AUC$ ), the  $Sens$  (sensitivity) across the range (Equation 3), the  $AccSens$  measure to combine the two, and finally the equal error rate  $\epsilon$  for comparison purposes. In each experiment, a 10-fold randomised hold-out procedure is performed, effectively resulting in 10 ROC plots upon which the aforementioned statistics are computed. Significance between models is assessed using ANOVA (99.5% significance level). The following datasets are used:

- **Road sign:** A road sign classification dataset [8] consisting of various *sign* and non-*sign* examples represented by images (793 pixels). All *signs* have

**Table 2.** Results of real experiments, comparing *AUC*, *Sens*, *AccSens*, and  $\epsilon$  (equal-error point) for a number of models per dataset. Standard deviations are shown.

Model	AUC	Sens	AccSens	$\epsilon$
<b>Road sign</b>				
1) <i>pca8 mogc 4 4</i>	0.881(0.026)	0.272(0.039)	0.211(0.029)	0.127(0.022)
2) <i>pca12 mogc 2 2</i>	0.886(0.058)	0.180(0.029)	0.154(0.028)	0.093(0.021)
3) <i>sc svc r 16</i>	0.951(0.016)	0.149(0.028)	0.111(0.021)	0.052(0.014)
4) <i>pca17 mogc 2 4</i>	0.876(0.100)	0.080(0.026)	0.112(0.056)	0.043(0.017)
5) <i>sc svc r 22</i>	0.952(0.016)	0.128(0.019)	0.100(0.015)	0.049(0.013)
6) <i>pca14 mogc 2 4</i>	0.907(0.061)	0.109(0.021)	0.106(0.033)	0.055(0.016)
<b>Phoneme</b>				
1) <i>sc knnc3</i>	0.905(0.013)	0.271(0.049)	0.204(0.028)	0.140(0.011)
2) <i>sc knnc1</i>	0.913(0.009)	0.248(0.013)	0.186(0.010)	0.107(0.008)
3) <i>sc parzenc</i>	0.891(0.014)	0.294(0.023)	0.222(0.018)	0.128(0.015)
<b>Sonar</b>				
1) <i>sc knnc3</i>	0.887(0.027)	0.310(0.107)	0.235(0.073)	0.147(0.039)
2) <i>sc knnc1</i>	0.892(0.036)	0.280(0.054)	0.213(0.043)	0.122(0.050)
3) <i>pca6 parzenc</i>	0.850(0.050)	0.405(0.069)	0.308(0.046)	0.167(0.054)
4) <i>sc svc p4</i>	0.829(0.056)	0.533(0.141)	0.398(0.100)	0.218(0.066)
<b>Ionosphere</b>				
1) <i>pca0.999 ldc</i>	0.855(0.039)	0.385(0.118)	0.292(0.084)	0.145(0.043)
2) <i>fisherm qdc</i>	0.855(0.037)	0.337(0.053)	0.260(0.041)	0.140(0.036)
3) <i>fisherm mogc 3 3</i>	0.834(0.035)	0.365(0.093)	0.285(0.063)	0.160(0.040)
4) <i>sc svc r 1.0</i>	0.853(0.171)	0.545(0.231)	0.434(0.095)	0.128(0.044)

been grouped together into a single class (381 objects), to be discriminated from non-*signs* (888 objects).

- **Phoneme:** This dataset is sourced from the ELENA project [5], in which the task is to distinguish between oral and nasal sounds, based on five coefficients (harmonics) of cochlear spectra. In this problem, the “nasal” class (3818 objects) is to be discriminated from the “oral” class (1586 objects).
- **Sonar** and **Ionosphere** are two well-known datasets from the *UCI* machine learning database [7].

Results are presented in Table 2. Various representation and classification algorithms have been used. Preprocessing/representation: *sc* denotes unit variance scaling, *pca* is a principle component mapping followed by the number of components used, or the fraction of variance retained, and *fisherm* is a Fisher mapping. Classifiers: *knnc* denotes the *k*-nearest neighbour classifier followed by the number of neighbours considered, *parzenc* is a Parzen-window classifier, *ldc* and *qdc* are Bayes linear and quadratic classifiers respectively, *mogc* is a mixture of Gaussians classifier followed by the number of mixtures per class, and *svc* is a support vector classifier, with *p* denoting a polynomial kernel followed by the order, and *r* denoting a Gaussian kernel, followed by the variance parameter.

Results show that there are many cases in which incorporation of sensitivity is important for this problem. In the *Road sign* case, an example of this is demonstrated by comparing models 1) and 2). Both show a similar *AUC* score,



but 2) is much less sensitive to prior variation. The *AccSens* measure is sensitive to this difference, showing significance (based on an ANOVA hypothesis test). Another interesting comparison is between 3) and 4), in which case model 3) has a significantly higher *AUC*, but 4) has a significantly better *Sens*. Both result in the same *AccSens* score. Models 3), 4), 5), and 6) all compete from an *AccSens* perspective (significantly better than 1) and 2)). In the *Phoneme* dataset, model 3) competes with 1) and 2) in terms of *AUC*, but 2) results in a better *Sens*, and thus results in a superior *AccSens* score (significant). This clearly illustrates the point of the paper once again - without considering sensitivity, model 3) could have been chosen instead of 1) or 2). In the *Sonar* dataset, model 2) appears superior in terms of both *AUC* and *Sens*, and thus there was no benefit of the new measure in this case. Finally, in the *Ionosphere* dataset, models 1), 2) and 4) result in similar *AUC* scores, but 2) appears less sensitive than 4) (not very significant). Using the *AccSens* measure, 1), 2) and 3) are significantly better than 4). As a final general comment on experimental results, it is apparent that there are cases in which a model selection based on *AUC* only is not the optimal procedure. Thus, we argue that in the prior uncertain/unstable environment, prior sensitivity should also be considered, using for example the *AccSens* measure.

## 7 Conclusions

In this paper the problem of varying/uncertain priors was investigated. ROC analysis has become a standard tool in this domain, with the Area Under the ROC (*AUC*) a popular model selection criterion. We argued that even though this integrated measure can be used to compare classifiers independent of priors, it may also be important to consider how *stable* a model is over the relevant range. A case study and some realistic experiments were presented that demonstrated how classifiers that compete in terms of *AUC* may differ significantly in terms of sensitivity (and vice-versa). It may thus be more sensible for the given problem to consider both. A simple measure, called *AccSens* was proposed, that combines the (weighted) geometric means of *AUC* and sensitivity, allowing for model comparison that considers both integrated accuracy (*AUC*), and prior sensitivity. A few real experiments demonstrated that this methodology is superior in some situations.

**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.

## References

- [1] N.M. Adams and D.J. Hand. Comparing classifiers when misallocation costs are uncertain. *Pattern Recognition*, 32(7):1139–1147, 1999.
- [2] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

- [3] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley - Interscience, second edition, 2001.
- [4] R.P.W. Duin. *PRTools, A Matlab Toolbox for Pattern Recognition*. Pattern Recognition Group, TUDelft, January 2000.
- [5] ELENA. European ESPRIT 5516 project. *phoneme dataset*, 2004.
- [6] C. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 3(4), 1978.
- [7] P.M. Murphy and D.W. Aha. UCI repository of machine learning databases, <ftp://ftp.ics.uci.edu/pub/machine-learning-databases>. *University of California, Department of Information and Computer Science*, 1992.
- [8] P. Paclík. Building road sign classifiers. *PhD thesis, CTU Prague, Czech Republic*, December 2004.
- [9] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.
- [10] J Yuen. Bayesian approaches to plant disease forecasting. *Plant Health Progress*, November 2003.