Optimizing Abstaining Classifiers using ROC Analysis

Tadeusz Pietraszek PIE@ZURICH.IBM.COM
IBM Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland

Abstract

Classifiers that refrain from classification in certain cases can significantly reduce the misclassification cost. However, the parameters for such abstaining classifiers are often set in a rather ad-hoc manner. We propose a method to optimally build a specific type of abstaining binary classifiers using ROC analysis. These classifiers are built based on optimization criteria in the following three models: cost-based, bounded-abstention and bounded-improvement. We demonstrate the usage and applications of these models to effectively reduce misclassification cost in real classification systems. The method has been validated with a ROC building algorithm and cross-validation on 15 UCI KDD datasets.

1. Introduction

In recent years, there has been much work on ROC analysis (Fawcett, 2003; Flach & Wu, 2003; Provost & Fawcett, 1998). An advantage of ROC analysis in machine learning is that it offers a flexible and robust framework for evaluating classifier performance with varying class distributions or misclassification costs.

Abstaining classifiers are classifiers that can refrain from classification in certain cases and are analogous to a human expert, who in certain cases can say "I don't know". In many domains (e.g., medical diagnosis) such experts are preferred to those who always make a decision and are sometimes are wrong.

Machine learning has frequently used abstaining classifiers (Chow, 1970; Ferri & Hernández-Orallo, 2004; Pazzani et al., 1994; Tortorella, 2000) and also as parts of other techniques (Ferri et al., 2004; Gamberger & Lavrač, 2000; Lewis & Catlett, 1994). Similarly to the human expert analogy, the motivation is that such a classifier, when it makes a decision, will perform better than a normal classifier. However, as these classifiers are not directly comparable, the comparison is often limited to coverage–accuracy graphs (Ferri & Hernández-Orallo, 2004; Pazzani et al., 1994).

In our paper, we apply ROC analysis to build an abstaining classifier that minimizes the misclassification cost. Our method is based solely on ROC curves and is independent of the classifiers used. We look at a particular type of abstaining binary classifiers—metaclassifiers constructed from two classifiers described by a single ROC curve—and show how to select such classifiers optimally.

The contribution of the paper is twofold: We define an abstaining binary classifier built as a metaclassifier and propose three models of practical relevance: the cost-based model (an extension of (Tortorella, 2000)), the bounded-abstention model, and the bounded-improvement model. These models define the optimization criteria and allow us to compare binary and abstaining classifiers. Second, we show how to practically build an optimal abstaining classifier in each of these models using ROC analysis.

The paper is organized as follows: Section 2 presents the notation and introduces the ROCCH method. In Section 3 we introduce the concept of ROC-optimal abstaining classifiers in three models. In Section 4 we discuss their construction. Section 5 discusses the evaluation methodology and presents the experimental results. In Section 6 we present related work. Finally, Section 7 contains the conclusions and future work.

2. Background and Notation

A *binary classifier* $\mathcal{C}$ is a function that assigns a binary class label to an instance, usually testing an instance with respect to a particular property. We will denote the class labels of a binary classifier as "+" and "−".

A *ranker* $\mathcal{R}$ (also known as scoring classifier) is a special type of binary classifier that assigns ranks to instances. The value of the rank denotes the likelihood

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005. Copyright 2005 by the author(s)/owner(s).

that the instance is "+" and can be used to sort instances from the most likely to the least likely positive. A ranker $\mathcal{R}$ can be converted to a binary classifier $\mathcal{C}_\tau$ as follows: $\forall i : \mathcal{C}_\tau(i) = + \iff \mathcal{R}(i) > \tau$. Variable τ in $\mathcal{C}_\tau$ denotes parameter (in this case a threshold) that was used to construct the classifier.

Abstaining binary classifiers $\mathcal{A}$ (or abstaining classifiers for short) are classifiers that in certain situations abstain from classification. We denote this as assigning a third class "?". Such non-classified instances can be classified using another (possibly more reliable, but more expensive) classifier or a human domain expert. This classification exceeds the scope of this paper.

The performance of a binary classifier is described by means of a 2×2-dimensional *confusion matrix* C. Rows in C represent actual class labels, and columns represent class labels predicted by the classifier. Element $C_{i,j}$ represents the number of instances of class i classified as class j by the system. For a binary classifier the elements are called true positives (*TP*), false negatives (*FN*), false positives (*FP*), and true negatives (*TN*) as shown in Table 1a. The sum of *TP* and *FN* is equal to the number of positive instances (P). Similarly the number of negative instances (N) equals $FP + TN$.

Asymmetrical classification problems can be modelled by a *cost matrix* Co with identical meanings of rows and columns as in the confusion matrix. Element $Co_{i,j}$ represents the cost of assigning a class j to an instance of class i. Most often the cost of correct classification is zero, i.e., $Co_{i,i} = 0$. In such cases, the matrix has only two non-zero values for binary classifiers (Table 1b): c_{21} (cost of misclassifying a negative instance as a positive) and c_{12} (cost of misclassifying a positive instance as a negative). In fact, such a cost matrix has only one degree of freedom, the *cost ratio* $CR = \frac{c_{21}}{c_{12}}$.

Classifiers in a cost-sensitive setup can be characterized by the cost rc—a cost-weighted sum of misclassifications divided by the number of classified instances:

$$rc = \frac{FN \cdot c_{12} + FP \cdot c_{21}}{TP + FN + FP + TN} \quad . \tag{1}$$

2.1. ROC Analysis

Very briefly, a ROC plane has axes ranging from 0 to 1 and labeled *false positive rate* ($fp = \frac{FP}{FP+TN} = \frac{FP}{N}$) and *true positive rate* ($tp = \frac{TP}{TP+FN} = \frac{TP}{P}$). Evaluating a binary classifier $\mathcal{C}_\tau$ on a dataset produces exactly one point (fp_τ, tp_τ) on the ROC plane. Many classifiers (e.g., Bayesian classifiers) or methods for building classifiers have parameters τ that can be varied to produce different points on the ROC plane. In particular, a single ranker can be used to efficiently generate a set of points on the ROC plane (Fawcett, 2003).

Table 1: Confusion and cost matrices for binary classification. The columns (C) represent classes assigned by the classifier; the rows (A) represent actual classes.

(a) Confusion matrix C

A \ C	+	−	
+	*TP*	*FN*	P
−	*FP*	*TN*	N

(b) Cost matrix Co

A \ C	+	−
+	0	c_{12}
−	c_{21}	0

Given a set of points on a ROC plane, the ROC Convex Hull (ROCCH) method (Provost & Fawcett, 1998) constructs a piecewise-linear convex down curve (called *ROCCH*) $f_R : fp \mapsto tp$, having the following properties: (i) $f_R(0) = 0$, (ii) $f_R(1) = 1$, and (iii) the slope of f_R is monotonically non-increasing. We denote the slope of a point on the ROCCH as f'_R[†].

To find the classifier that minimizes the misclassification cost rc, we rewrite Equation (1) as a function of one variable, *FP*, calculate the first derivative $\frac{d\,rc}{d\,FP}$ and set it equal to 0. This yields a known equation of iso-performance lines

$$f'_R(fp^*) = CR\,\frac{N}{P} \quad , \tag{2}$$

which shows the optimal slope of the curve given a certain cost ratio (CR), N negative, and P positive instances. Similarly to Provost and Fawcett (1998), we assume that for any real $m > 0$ there exists at least one point (fp^*, tp^*) on the ROCCH having $f'_R(fp^*) = m$.

Note that the solution of this equation can be used to find a classifier that minimizes the misclassification cost for the instances used to create the ROCCH. We call such a classifier *ROC-optimal*. Note that it may not be optimal on other instances. However, if the testing instances used to build the ROCCH and the other instances are representative, such a ROC-optimal classifier will also perform well on other testing sets.

3. ROC-Optimal Abstaining Classifier

Our method builds an *ROC-optimal* abstaining classifier as a metaclassifier using a ROC curve and the binary classifiers used to construct it. A ROC-optimal classifier is defined as described in Sect. 2.1. The method constructs an abstaining metaclassifier $\mathcal{A}_{\alpha,\beta}$ using two binary classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ as follows:

$$\mathcal{A}_{\alpha,\beta}(x) = \begin{cases} + & \mathcal{C}_\alpha(x) = + \wedge \mathcal{C}_\beta(x) = + \\ ? & \mathcal{C}_\alpha(x) = - \wedge \mathcal{C}_\beta(x) = + \\ - & \mathcal{C}_\beta(x) = - \wedge \mathcal{C}_\alpha(x) = - \end{cases} \quad . \tag{3}$$

[†] For this paper we assume that the slope at vertices of a convex hull takes all values between the slopes of adjacent line segments.

Each classifier has a corresponding confusion matrix, $(TP_\alpha, FN_\alpha, FP_\alpha, TN_\alpha)$ and $(TP_\beta, FN_\beta, FP_\beta, TN_\beta)$, which will be used in the next sections. Classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ belong to a family of classifiers $\mathcal{C}_\tau$, described by a single ROC curve with $FP_\alpha \leq FP_\beta$.

Our method is independent of the machine-learning technique used. However, we require that for any two points (fp_α, tp_α), (fp_β, tp_β) on the ROC curve, with $fp_\alpha \leq fp_\beta$, corresponding to $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$, the following conditions hold:

$$\begin{aligned} \forall i : (\mathcal{C}_\alpha(i) = + \implies \mathcal{C}_\beta(i) = +) \wedge \\ (\mathcal{C}_\beta(i) = - \implies \mathcal{C}_\alpha(i) = -) \ . \end{aligned} \tag{4}$$

Conditions (4) are the ones used in (Flach & Wu, 2003). These are met in particular if the ROC curve and $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ are built from a single ranker $\mathcal{R}$ (e.g., a Bayesian classifier) with two threshold values α and β ($\alpha \geq \beta$). The advantage is that for such a classifier, a simple and efficient algorithm exists for constructing a ROC curve (Fawcett, 2003). For arbitrary classifiers (e.g., rule learners), (4) is generally violated. However, we observed that the fraction of instances with $\mathcal{C}_\alpha(i) = + \wedge \mathcal{C}_\beta(i) = -$ is typically small, and that applying our method is such cases still yields good results. As this is an interesting class of applications, we plan to elaborate on it as a future work item.

Given a particular cost matrix and class distribution $\frac{N}{P}$, the optimal binary classifier can easily be chosen as a one that minimizes the misclassification cost (1). However, no such notion exists for abstaining classifiers, as the tradeoff between non-classified instances and the cost is undefined. Therefore, we propose and investigate three different criteria and models of optimization: the cost-based, the bounded-abstention and the bounded-improvement model, which we discuss in the following sections. We formulate our goals as:

Given	– A ROC curve generated using classifiers $\mathcal{C}_\tau$, such that (4) holds. – A Cost matrix Co. – Evaluation model $\mathcal{E}$.
Find	A classifier $\mathcal{A}_{\alpha,\beta}$ such that $\mathcal{A}_{\alpha,\beta}$ is optimal in model $\mathcal{E}$.

3.1. Cost-Based Model

In this model, we compare the misclassification cost, rc, incurred by a binary and an abstaining classifier. We use an extended 2×3 cost matrix, with the the third column representing the cost associated with classifying an instance as "?". Note that this cost can be different for instances belonging to different classes, which extends the cost matrix introduced in (Tortorella, 2000).

Table 2: Cost matrix Co for an abstaining classifier. Columns and rows are the same as in Table 1. The third column denotes the abstention class.

A \ C	+	−	?
+	0	c_{12}	c_{13}
−	c_{21}	0	c_{23}

Having defined the cost matrix, we use a similar approach as in Sect. 2.1 for finding the optimal classifier. Note that the classifications made by $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ are not independent. Equation (4) implies that false positives for $\mathcal{C}_\alpha$ imply false positives for $\mathcal{C}_\beta$. The similar holds for false negatives, and we can thus formulate (5). The misclassification cost $\tilde{rc}$ is defined using a 2×3 cost matrix similarly to (1), with the denominator equal to the total number of instances.

$$\tilde{rc} = \frac{1}{N+P} \Big(\underbrace{(FP_\beta - FP_\alpha)\, c_{23}}_{\mathcal{C}_\alpha,\ \mathcal{C}_\beta \text{ disagree, } -} + \underbrace{(FN_\alpha - FN_\beta)\, c_{13}}_{\mathcal{C}_\alpha,\ \mathcal{C}_\beta \text{ disagree, } +} + \underbrace{FP_\alpha \cdot c_{21}}_{FP \text{ for both}} + \underbrace{FN_\beta \cdot c_{12}}_{FN \text{ for both}} \Big) \tag{5}$$

We rewrite (5) as a function of only two variables: FP_α and FP_β, so that to find the local minimum we calculate partial derivatives for these variables. After calculations, setting the derivatives to zero, making sure that the function has a local extremum, and replacing FP_α and FP_β with the corresponding rates fp_α and fp_β, we obtain the final result:

$$\begin{aligned} f'_R(fp^*_\beta) &= \frac{c_{23}}{c_{12} - c_{13}} \frac{N}{P} \\ f'_R(fp^*_\alpha) &= \frac{c_{21} - c_{23}}{c_{13}} \frac{N}{P} \ , \end{aligned} \tag{6}$$

which, similarly to (2), allows us to find fp^*_α and fp^*_β, and corresponding classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$.

This derivation is valid only for metaclassifiers (3) with (4), which implies $fp^*_\alpha \leq fp^*_\beta$ and $f_R(fp^*_\alpha) \leq f_R(fp^*_\beta)$. As a ROCCH is increasing and convex, its first derivative is non-negative and non-increasing, and we obtain $f'_R(fp^*_\alpha) \geq f'_R(fp^*_\beta) \geq 0$. Using the 2×3 cost matrix these conditions can be rewritten as:

$$(c_{21} \geq c_{23}) \wedge (c_{12} > c_{13}) \wedge (c_{21}c_{12} \geq c_{21}c_{13} + c_{23}c_{12}) \ . \tag{7}$$

If condition (7) is not met, our derivation is not valid; however the solution is trivial in this case.

Theorem 1. *If (7) is not met, the classifier minimizing the misclassification cost is a trivial binary classifier—a single classifier described by (2).*

Proof. We omit the complete proof for space reasons and only briefly outline it. First, we have to show that for a ROC-optimal abstaining classifier $f'_R(fp^*_\alpha) \geq f'_R(fp^*) \geq f'_R(fp^*_\beta) \geq 0$, where fp^* describes the ROC-optimal binary classifier. Second, we have to show that if (7) is not met, partial derivatives $\frac{\partial \tilde{rc}}{\partial fp_\alpha}$ and $\frac{\partial \tilde{rc}}{\partial fp_\beta}$ are positive for $fp^*_\alpha < fp^*$ and $fp^*_\beta > fp^*$. Therefore we conclude that the ROC-optimal classifier is a binary classifier in this case. □

Equation (7) allows us to determine whether for a given 2×3 cost matrix Co exists a trivial abstaining classifier minimizing rc, but gives little guidance to setting parameters in this matrix. For this we consider two interesting cases: (i) a symmetric case $c_{13} = c_{23}$, and (ii) a proportional case $\frac{c_{23}}{c_{13}} = \frac{c_{21}}{c_{12}}$.

The first case has some misclassification cost CR with identical costs of classifying instances as "?". This case typically occurs when, for example, the cost incurred by the human expert to investigate such instances is irrespective of their true class. In this case, (7) simplifies to the harmonic mean of two misclassification costs: $c_{13} = c_{23} \leq \frac{c_{21}c_{12}}{c_{21}+c_{12}}$. The second case gives us the condition $c_{13} \leq \frac{c_{12}}{2}$ (equivalent to $c_{23} \leq \frac{c_{21}}{2}$). This case occurs if the cost of classifying an event as the third class is proportional to the misclassification cost. These simplified equations allow a meaningful adjustment of parameters c_{13} and c_{23} for abstaining classifiers.

To summarize, the ROC-optimal abstaining classifier in a cost-based model can be found using (6) if (7) (or the special cases discussed below) holds on a given cost matrix. In the opposite case, our derivation is not valid; however the ROC-optimal classifier is a trivial binary classifier ($\mathcal{C}_\alpha = \mathcal{C}_\beta$).

3.2. Bounded Models

In the simulations using a cost-based model (see Sect. 5.3.1) we noticed that the cost matrix and in particular cost values c_{13} and c_{23} have a large impact on the number of instances classified as"?". Therefore we think that, while the cost-based model can be used in domains where the 2×3 cost matrix is *explicitly given*, it may be *difficult to apply in other domains*, where parameters c_{13}, c_{23} would have to be estimated.

To address this shortcoming we propose a model that uses a standard 2×2 cost matrix. In such a setup, we calculate the misclassification cost per instance actually classified. The motivation is to calculate the cost only for instances the classifier attempts to classify.

Using a standard cost equation, (1), with the denominator $TP + FP + FN + TN = (1-k)(N+P)$, where k is the fraction of non-classified instances, we obtain:

$$\begin{aligned} rc &= \frac{1}{(1-k)(N+P)} \left(FP_\alpha \cdot c_{21} + FN_\beta \cdot c_{12} \right) \\ k &= \frac{1}{N+P} \left((FP_\beta - FP_\alpha) + (FN_\alpha - FN_\beta) \right) \quad , \end{aligned} \tag{8}$$

which determine the relationship between the fraction of classified instances k and the misclassification cost rc as a function of $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$. By putting boundary constraints on k and rc and trying to optimize rc and k, respectively, we create two interesting evaluation models, which we discuss below.

3.2.1. Bounded-Abstention Model

By limiting k to some threshold value $k_{\max}$ ($k \leq k_{\max}$) we obtain a model in which the classifier can abstain for at most a fraction k of instances. In this case the optimization criterion is that the classifier should have the lowest misclassification cost rc.

This has several real-life applications, e.g., in situations where non-classified instances will be handled by a classifier with limited processing speed (e.g., a human expert). In such cases, assuming a constant flow of instances with speed c and a constant manual processing speed m, $m < c$, we obtain $k_{\max} = \frac{m}{c}$.

We rewrite Equations (8) as functions of two variables fp_α and fp_β and introduce two auxiliary functions $rc(fp_\alpha, fp_\beta)$, expressing the relative misclassification cost, and $k(fp_\alpha, fp_\beta)$, denoting the number of non-classified instances. The minimization goal can be expressed as follows: Among all pairs $(fp^*_\alpha, fp^*_\beta)$ that satisfy $k(fp^*_\alpha, fp^*_\beta) \leq k_{\max}(N+P)$, find the ones that minimize $rc(fp^*_\alpha, fp^*_\beta)$.

$$\begin{aligned} rc(fp_\alpha, fp_\beta) &= \frac{N \cdot fp_\alpha \cdot c_{21} + P\,(1 - f_R(fp_\beta)) \cdot c_{12}}{N + P - k(fp_\alpha, fp_\beta)} \\ k(fp_\alpha, fp_\beta) &= P\ (f_R(fp_\beta) - f_R(fp_\alpha)) \\ &\quad + N(fp_\beta - fp_\alpha) \end{aligned} \tag{9}$$

Unfortunately, unlike (6), the Equations (9) for a bounded-abstention model have no algebraic solution in the general case. Therefore we minimize it using numerical methods.

3.2.2. Bounded-Improvement Model

The second bounded model is when we limit rc to a threshold value $rc_{\max}$ ($rc \leq rc_{\max}$) and require that the classifier abstain for the smallest number of instances. Similarly to the previous model, optimizing this model requires the use of numerical methods. Using the definitions of $k(fp_\alpha, fp_\beta)$ and $rc(fp_\alpha, fp_\beta)$ in (9), we express the minimization goal as follows:

Among all pairs $(fp^*_\alpha, fp^*_\beta)$ such that $rc(fp^*_\alpha, fp^*_\beta) \leq rc_{\max}$, find the ones that minimize $k(fp^*_\alpha, fp^*_\beta)$.

This model has several real-life applications, e.g., in a medical domain, where given a certain test and its characteristics (ROC curve) the goal is to reduce the misclassification cost to a user-defined value $rc_{\max}$ allowing for the smallest number of abstentions.

For the evaluation of this model the following remark is in place. As different datasets yield different ROC curves and misclassification costs, we could not use a constant value of $rc_{\max}$ for all datasets. Instead we used a fraction cost improvement f and calculated rc' as follows: $rc_{\max} = (1 - f)\, rc$, where rc is the misclassification cost of the ROC-optimal binary classifier found using (2).

4. Constructing Abstaining Classifiers

In this section we discuss how to construct an optimal abstaining classifier in the three models. Based on (Provost & Fawcett, 1998; Tortorella, 2000), in the cost-based model, the ROC-optimal classifier is always located on the vertices of the ROCCH. This is intuitive as classifiers corresponding to two adjacent vertices of the ROCCH have the same slopes and the same misclassification costs as classifiers corresponding to the line segment joining these vertices. However, this is not always the case in the two bounded models we introduced in Sections 3.2.1 and 3.2.2.

Theorem 2. *In the (i) bounded-abstention and (ii) bounded-improvement models, the optimal classifier is not* always *located on the vertices of the ROCCH.*

Proof. (by counterexample)
(i) Assume the optimal classifier $\mathcal{A}_{\alpha,\beta}$ has its classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ located on the vertices of the particular convex hull $(0,0)$, $(0.5,1)$, $(1,1)$ with $c_{12} = c_{21} = 1$, $N = P$ and $k_{\max} < 0.25$. In this case $\mathcal{C}_\alpha$ must be equal to $\mathcal{C}_\beta$ (otherwise $k \geq 0.25$). Therefore, $fp_\alpha = fp_\beta = 0.5$ and from (9) $rc(fp_\alpha, fp_\beta) = 0.25$.

Assume a classifier $\tilde{\mathcal{A}_{\alpha,\beta}}$ has $\tilde{fp}_\alpha = 0.5 - \delta$ and $\tilde{fp}_\beta = 0.5$, with a small positive δ so that $k(\tilde{fp}_\alpha, \tilde{fp}_\beta) < 0.25$). In this case (9) simplifies to $rc(\tilde{fp}_\alpha, \tilde{fp}_\beta) = 0.25 - \frac{2\delta}{4-6\delta} < 0.25$. This contradicts the assumption that $\mathcal{A}_{\alpha,\beta}$ is an optimal classifier in a bounded-abstention model and completes the proof.

(ii) A similar proof can be shown for a bounded-improvement model. We omit it for space reasons. □

To conclude, vertices on the ROCCH can be used to find a ROC-optimal classifier only in the cost-based model. In the remaining two models, the ROC-optimal classifier uses arbitrary points on the ROCCH. Such classifiers, corresponding to points lying on the line segment can be constructed using a weighted random selection of votes of classifiers corresponding to two adjacent vertices (Fawcett, 2003). However, our prototype uses another method, which was more stable and produced less variance than the random selection.

A ROCCH can be considered a function $f : \tau \mapsto (fp, tp)$, where $\tau \in T$ is a set of discrete parameters, varying which one constructs classifiers $\mathcal{C}_\tau$ corresponding to different points on the ROCCH. In our algorithm we compute an inverse function $f^{-1} : (fp, tp) \mapsto \tau$ and interpolate it using splines with a function $\hat{f^{-1}}$, defined for a continuous range of values τ. Given an arbitrary point $(fp^*,\ tp^*)$ on the curve, we use the function $\hat{f^{-1}}$ yielding τ^* to construct a classifier $\mathcal{C}_{\tau^*}$

5. Experiments

To analyze the performance of our method we tested it on 15 well-known datasets from the UCI KDD (Hettich & Bay, 1999) database: `breast-cancer`, `breast-w`, `colic`, `credit-a`, `credit-g`, `diabetes`, `heart-statlog`, `hepatitis`, `ionosphere`, `kr-vs-kp`, `labor`, `mushroom`, `sick`, `sonar`, and `vote`.

We tested our method in all three models described above. In the model 1, the input data is a 2 × 3 cost matrix in the symmetric case ($c_{13} = c_{23}$). In the model 2, we use a 2 × 2 cost matrix and k (a fraction of instances that the system does not classify). In the model 3, the input data is also a 2 × 2 cost matrix and a fraction f, i.e., a relative cost improvement over the optimal binary classifier (defined as $\frac{rc_{\text{binary}} - rc_{\text{tri-state}}}{rc_{\text{binary}}}$).

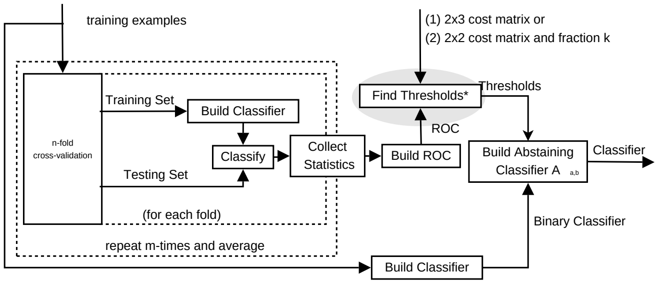
5.1. Testing Methodology

The experiment for each dataset was a two-fold cross-validation repeated five times with different seed values for the pseudo-random generator (we used 5 × 2 cv, as it has a low-level Type-I error for significance testing (Dietterich, 1998)). We averaged the results for these runs and calculated 95% confidence intervals, shown as error bars on each plot. In the cross-validation, we used a training set to build an abstaining classifier, which was subsequently evaluated on the testing set.

The process of building an abstaining classifier is shown in Fig. 1. We used another two-fold cross-validation ($n = 2$) to construct a ROC curve. The cross-validation was executed five times ($m = 5$), and the resulting ROC curves were averaged (threshold averaging (Fawcett, 2003)) to generate a smooth curve.

While the method is applicable for any machine learning algorithm that satisfies (4), we used a simple Naive Bayes classifier as a base classifier, converting it to a ranker by calculating the prediction ratio $\frac{P(x,+)}{P(x,-)}$.

Given the ROC curve and the input parameters (cost matrix and a value k), the program numerically finds values α and β describing $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ and the ROC-optimal classifier (in each model). These values were used to set the thresholds in a Naive Bayes classifier built using the entire training set to create $\mathcal{A}_{\alpha,\beta}$.



* - algorithm described in the paper

Figure 1: Building an abstaining classifier $\mathcal{A}_{\alpha,\beta}$.

Such an experiment was run for every dataset and every combination of input parameters, *CR*, c_{13} (respectively k or f), thus producing multiple plots (one for each dataset), multiple series (one for each cost ratio), and multiple points (one for each value of c_{13}, k or f).

We used three values of the cost ratio (*CR*): 0.5, 1 and 2, and four different values of c_{13} (first model), k: 0.1, 0.2, 0.3 and 0.5 (second model), and f: 0.1, 0.2, 0.3 and 0.5 (third model), yielding 180 experiment runs ($15 \times 3 \times 4$) for each model. The values of CR, c_{13}, k and f were chosen from the range of values the models are expected to be used with.

We used Bayesian classifier from Weka toolkit (Witten & Frank, 2000) as a machine-learning method. For the numerical optimization for bounded models we used the Nelder-Mead optimization algorithm (Nedler & Mead, 1965).

5.2. Results—Cost-Based Model

Out of 180 simulations (15 datasets, four values of c_{13}, and three cost values) 152 are significantly better (lower rc) than the corresponding optimal binary classifier (one-sided paired t-test with a significance level of 0.95). The optimal binary classifier was the same Bayesian classifier with a single threshold set using (2).

The results for a representative dataset are shown in Fig. 2. The complete results are shown in a technical report (Pietraszek, 2004). The X-axes correspond to the cost value in a symmetric case $c_{13} = c_{23}$ (left and center panel), and the Y-axes show the relative cost improvement (left panel) and the fraction of non-classified instances (center panel). The right panel displays the relationship between the fraction of skipped instances and the overall cost improvement. Horizontal error bars show 95% confidence intervals for the fraction of non-classified instances, only indirectly determined by c_{13}.

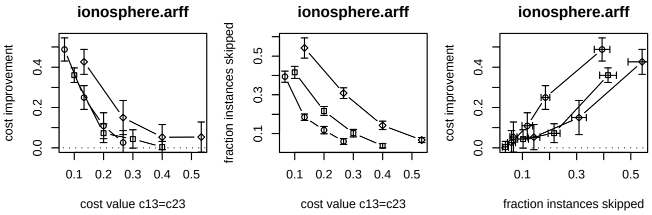


Figure 2: Cost-based model: Relative cost improvement and fraction of non-classified instances for a representative dataset ($\circ : CR = 0.5$, $\Box : CR = 1$, $\Diamond : CR = 2$).

We clearly observe that lower misclassification costs $c_{13} = c_{23}$ result in a higher number of instances being classified as "?" and higher relative cost improvement. Moreover, an almost linear relationship exists between the fraction of non-classified instances and the relative cost improvement (right panel).

5.3. Results—Bounded Models

5.3.1. Bounded-Abstention Model

Out of 180 simulations (15 datasets, four values of fractions of non-classified instances and three cost values) 179 have significantly lower rc than the corresponding optimal binary classifier (one-sided paired t-test with a significance level of 0.95). The optimal binary classifier is a Bayesian classifier with a single threshold.

We also observed that in most cases the resulting classifier classified the desired fraction of instances as the third class; the mean of the relative difference of k ($\frac{\Delta k}{k}$) for all runs is 0.078 ($\sigma = 0.19$). This is particularly important as it is only indirectly determined by the two thresholds the algorithm calculates.

The results for a representative dataset are shown in Fig. 3. The X-axes correspond to the actual fraction of non-classified instances and the Y-axes show the relative cost improvement (left panel) and the misclassification cost (right panel). The left panel shows the relative cost improvement as a function of the fraction of instances handled by operator k. In general, the higher the values of k, the higher the cost improvement; for 8 datasets, namely: `breast-w`, `credit-a`, `credit-g`, `diabetes`, `heart-statlog`, `ionosphere`, `kr-vs-kp` and `sonar`, we can observe an almost linear dependence between these variables. The right panel shows the same data with the absolute values of rc. The dashed arrows show the difference between an optimal binary classifier and an abstaining one.

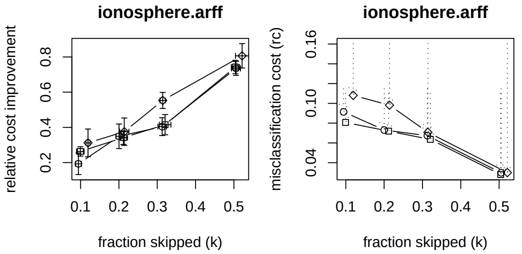


Figure 3: Bounded-abstention model: Relative cost improvement and the absolute cost for one representative dataset ($\circ : CR = 0.5$, $\Box : CR = 1$, $\Diamond : CR = 2$).

5.3.2. Bounded-Improvement Model

This model is in fact the inverse of the previous model, and thus we expected very similar results. The results for a representative dataset are shown in Fig. 4. The X-axes correspond to relative cost improvement (left panel) and the misclassification cost (right panel). The Y-axes show the actual fraction of non-classified instances.

The left panel shows the fraction of instances handled by the operator as a function of the actual misclassification cost. It is interesting to compare the actual relative cost improvement f and the assumed one (0.1, 0.2, 0.3, 0.5), as the former is only indirectly determined through two thresholds determined by the performance on the training set. The mean of the relative difference of f ($\frac{\Delta f}{f}$) for all runs is 0.31 ($\sigma = 1.18$). The positive value of the mean shows that the system has, on average, a lower misclassification cost than required. Note that this value is higher than the corresponding difference in the previous model. We conclude that this model is more sensitive to parameter changes than the previous one. The right panel shows the same data with the X-axis viewed as absolute values of costs. In addition the horizontal arrows (dashed) indicate the absolute values for the optimal binary classifier and the desired cost at the head of an arrow.

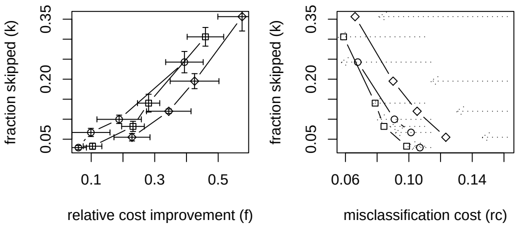


Figure 4: Bounded-improvement model: Fraction of non-classified instances for a representative dataset ($\circ : CR = 0.5$, $\Box : CR = 1$, $\Diamond : CR = 2$).

6. Related Work

Classifiers with reject rules were first investigated by Chow (1970) and further developed by Tortorella (2000). The latter uses ROC analysis in a model corresponding to our cost-based model in a more restrictive setup ($c_{13} = c_{23}$). Our work extends this model further and shows conditions, under which a non-trivial abstaining classifier exists. We also propose two bounded models with other optimization criteria.

Cautious classifiers (Ferri & Hernández-Orallo, 2004) propose abstaining classifiers with a class bias K and an abstention window w, which make them similar to our second evaluation model, where an abstention window is defined. However, although for $w = 0$ abstention is zero and the classifier abstains for approximately all instances for $w = 1$, the relationship between w and the abstention is neither continuous nor linear (Ferri & Hernández-Orallo, 2004). Therefore our model cannot be compared easily with cautious classifiers. Similarly, cautious classifiers require calibrated probabilities assigned to instances (otherwise the class bias might be difficult to interpret). In contrast our model, if used with a scoring classifier, uses only information about the ordering of instances, not the absolute values of probabilities. This makes our model more general. On the other hand, cautious classifiers are more general in the sense that they can be used with a multi-class classification, whereas our model is based on ROC analysis and is only applicable to two-class classification problems.

Delegating classifiers (Ferri et al., 2004) use a cascading model, in which classifiers at every level classify only a certain percentage of the instances. In this way every classifier, except for the last one is a cautious classifier. The authors present their results with an iterative system, using up to $n - 1$ cautious classifiers.

Pazzani et al. (1994) showed how different learning algorithms can be modified to increase accuracy at the cost of not classifying some of the instances, thus creating an abstaining classifier. However, this approach does not select the optimal classifier, is cost-insensitive and specific to the algorithms used.

Confirmation rule sets (Gamberger & Lavrač, 2000) are another example of classifiers that may abstain from classification. Confirmation rule sets use a special set of highly specific classification rules. The results of the classification (and whether the classifier makes the classification at all) depend on the number of rules that fired. Similarly to the previous approach, the authors do not maximize the accuracy.

7. Conclusions and Future Work

We proposed a method to build the *ROC-optimal abstaining classifier* using ROC analysis. Such a classifier minimizes the misclassification cost on instances used to build the ROC curve. It also has a low misclassification cost on other datasets from the same population as the one used to build the curve.

We defined the misclassification cost in three models: A cost-based model, a bounded-abstention and bounded-improvement models, which are relevant for numerous practical applications. In the first model, we used a 2×3 cost matrix, showed the conditions under which the abstaining classifier has a non-trivial minimum cost, and presented a simple analytical solution. In the bounded model, we showed how to build the abstaining classifier assuming that no more than a fraction $k_{\max}$ of instances is classified as the third class. Finally, in the third model, we showed how to build an abstaining classifier having a misclassification cost that is no greater than a user-defined value. In the latter two models, we redefined the problem as a numerical optimization problem. We presented an implementation and verified our method in all three models on a variety of UCI datasets.

As future work, we intend to extend our experiments to include other machine-learning algorithms. We will also analyze the performance of our method for algorithms for which (4) does not hold. We plan to investigate the convexity of ROC curves and how to apply our method efficiently in real-world applications, also with multi-class classification.

Acknowledgements

The author would like to thank Klaus Julisch and Axel Tanner for their comments and stimulating discussions. Thanks also go to the anonymous reviewers for the useful feedback.

References

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*, 41–46.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1923.

Fawcett, T. (2003). *ROC graphs: Note and practical considerations for researchers (HPL-2003-4)* (Technical Report). HP Laboratories.

Ferri, C., Flach, P., & Hernández-Orallo, J. (2004). Delegating classifiers. *Proceedings of 21th International Conference on Machine Leaning (ICML'04)* (pp. 106–110). Alberta, Canada: Omnipress.

Ferri, C., & Hernández-Orallo, J. (2004). Cautious classifiers. *Proceedings of ROC Analysis in Artificial Intelligence, 1st International Workshop (ROCAI-2004)* (pp. 27–36). Valencia Spain.

Flach, P. A., & Wu, S. (2003). Repairing concavities in ROC curves. *Proc. 2003 UK Workshop on Computational Intelligence* (pp. 38–44). Bristol, UK.

Gamberger, D., & Lavrač, N. (2000). Reducing misclassification costs. *Principles of Data Mining and Knowledge Discovery, 4th European Conference (PKDD 2000)* (pp. 34–43). Lyon, France: Springer Verlag.

Hettich, S., & Bay, S. D. (1999). The UCI KDD Archive. Web page at `http://kdd.ics.uci.edu`.

Lewis, D. D., & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. *Proceedings of ICML-94, 11th International Conference on Machine Learning* (pp. 148–156). Morgan Kaufmann Publishers, San Francisco, US.

Nedler, J., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308–313.

Pazzani, M. J., Murphy, P., Ali, K., & Schulenburg, D. (1994). Trading off coverage for accuracy in forecasts: Applications to clinical data analysis. *Proceedings of AAAI Symposium on AI in Medicine* (pp. 106–110). Stanford, CA.

Pietraszek, T. (2004). *Optimizing abstaining classifiers using ROC analysis (RZ3571)* (Technical Report). IBM Zurich Research Laboratory.

Provost, F., & Fawcett, T. (1998). Robust classification systems for imprecise environemnts. *Proceedings of the Fifteenth National Conference on Artificial Intellignence (AAAI-98)* (pp. 706–713). AAAI Press.

Tortorella, F. (2000). An optimal reject rule for binary classifiers. *Advances in Pattern Recognition, Joint IAPR International Workshops SSPR 2000 and SPR 2000* (pp. 611–620). Alicante, Spain: Springer-Verlag.

Witten, I. H., & Frank, E. (2000). *Data Mining: Practical machine learning tools with Java implementations.* San Francisco: Morgan Kaufmann.