Mach Learn (2007) 68: 137–169
DOI 10.1007/s10994-007-5013-y

On the use of ROC analysis for the optimization of abstaining classifiers

Tadeusz Pietraszek

Received: 13 February 2006 / Revised: 26 March 2007 /
Accepted: 26 March 2007 / Published online: 11 July 2007
Springer Science+Business Media, LLC 2007

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 show that selecting the optimal classifier in the first model is similar to known iso-performance lines and uses only the slopes of ROC curves, whereas selecting the optimal classifier in the remaining two models is not straightforward. We investigate the properties of the convex-down ROCCH (ROC Convex Hull) and present a simple and efficient algorithm for finding the optimal classifier in these models, namely, the bounded-abstention and bounded-improvement models. We demonstrate the application of these models to effectively reduce misclassification cost in real-life classification systems. The method has been validated with an ROC building algorithm and cross-validation on 15 UCI KDD datasets.

Keywords Abstaining classifiers · ROC analysis · Cost-sensitive classification · Cautious classifiers

1 Introduction

Abstaining classifiers are classifiers that can refrain from classification in certain cases and are analogous to a human expert, who can say "I don't know". In many domains (e.g.,

An early version of this paper was published at ICML2005.

Action Editor: Johannes Fürnkranz.

T. Pietraszek (✉)
IBM Zurich Research Laboratory, Säumerstrasse 4, 8803 Rüschlikon, Switzerland
e-mail: pie@zurich.ibm.com

T. Pietraszek (✉)
e-mail: tadek@google.com

medical diagnosis) such experts are preferred to those who always make a decision but are sometimes wrong.

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

In recent years, there has been much work on ROC analysis (Fawcett 2003; Flach and Wu 2005; Provost and Fawcett 2001). 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 (Fawcett 2003).

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. In particular, we do not require that the underlying classifier gives calibrated probabilities, which is not always trivial (Cohen and Goldszmidt 2004; Zadrozny and Elkan 2001). We look at a particular type of abstaining binary classifiers, i.e., metaclassifiers constructed from two classifiers described by a single ROC curve, and show how to select such classifiers optimally according to *three different optimization criteria* that are commonly encountered in practical applications.

More specifically, in the first model, the so-called *cost-based model*, the goal is to optimize the misclassification cost calculated using an extended-cost matrix (similarly to the model discussed by Tortorella 2000, 2004). While this model can be used in situations in which misclassification costs are explicitly given, in many practical applications the exact misclassification costs are unknown and can only be estimated. Given this and the high sensitivity of the cost-based model, we proposed two other models in which abstention costs are implicit. In these models, the so-called bounded models, we calculate the cost per actually classified instance and use boundary conditions for abstention rate and the misclassification cost.

The idea is that such a setting would allow us to trade the misclassification cost for abstentions, i.e., higher abstentions result in a lower misclassification cost (obviously, given an abstention $k_{\max}$ we want to have the smallest misclassification cost rc possible and similarly, given misclassification cost $rc_{\max}$ we want to have the smallest abstention k possible), which is intuitive in many applications.

In the *bounded-abstention model*, the boundary condition is the maximum abstention rate of the classifier, whereas in the *bounded-improvement model*, the boundary condition is the maximum misclassification cost of the classifier. These models can be intuitively used in many practical applications with resource or cost constraints, in particular the ones involving a human domain expert. We will illustrate these models with two examples from the domain of computer security: a resource-bounded example and a cost-bounded example.

In the first example, suppose there is a system processing intrusion detection alerts in real time. Alerts can be either *true alerts*, which indicate an intrusion, or *false alerts*, triggered mistakenly when no intrusion took place. In case of a true alert, the system should perform an action (e.g., notify the network administrator), whereas false alerts should be quietly discarded. The system uses an imperfect automatic classifier, with a known ROC, and a human analyst, with a limited processing capability. Assuming that the system receives c events per minute and the human analyst can only analyze m events per minute, the goal of a bounded-abstention model is to decide which m/c alerts will be processed by the analyst so that the overall misclassification cost is minimized. In this case, the abstention for a fraction m/c alerts results in a lower misclassification cost of remaining alerts.

In the second example, assume a similar scenario but with cost constraints: e.g., a contract with a customer limits the maximum misclassification cost (resulting from discarded true alerts and incorrect notifications on non-intrusions) to some value $rc_{\max}$. Assuming the best automatic classifier has a misclassification cost $rc^*_{\max}$ ($rc^*_{\max} > rc_{\max}$), the goal of the bounded-improvement model is to select the smallest number of alerts to be classified by a human domain expert so that the contractual obligations are met.

We will formally define the above three models in the following sections. The point made here is to motivate our models and to show their practical relevance.

The contribution of the paper is twofold: First, 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 propose efficient algorithms to practically build an optimal abstaining classifier in each of these models using ROC analysis, and evaluate our method on a variety of UCI KDD datasets.

Parts of this paper are based on (Pietraszek 2005). In this contribution we provide a deeper investigation of the two bounded models, and provide an efficient algorithm to select the optimal classifier.

The paper is organized as follows: Sect. 2 presents the notation and introduces the ROCCH method. In Sect. 3 we introduce the concept of ROC-optimal abstaining classifiers in three models. Section 4 discusses the first model, the cost-based model, and Sect. 5 propose algorithms for efficient construction of abstaining classifiers in the other two models: bounded-abstention and bounded-improvement models. Section 6 discusses the evaluation methodology and presents the experimental results. In Sect. 7, we present related work. Finally, Sect. 8 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 *scoring classifier* $\mathcal{R}$ is a special type of binary classifier that assigns scores to instances. The value of the score denotes the likelihood that the instance is "+" and can be used to sort instances from the most likely to the least likely positive. Note that the scores do not necessarily have to be calibrated probabilities. A scoring classifier $\mathcal{R}$ 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 the parameter (in this case a threshold) 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 nonclassified instances can be classified using another (possibly more reliable, but more expensive) classifier (e.g., a multi-stage classification system as suggested by Senator 2005) or a human domain expert.

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; columns represent class labels predicted by the classifier. Element $C_{i,j}$ represents the number of instances of class i

Table 1 The 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

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 1(a). 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 modeled by a *cost matrix* Co with identical meanings of rows and columns as in the confusion matrix. The element $Co_{i,j}$ represents the cost of assigning a class j to an instance of class i.

As shown by Elkan (2001), for binary cases, all four-element cost matrices (*full cost matrices*) with identical value of $(c_{21} - c_{11})/(c_{12} - c_{22})$ form an equivalence class with respect to the decision-making process in the classification. This in particular means that adding a constant value to the rows of the cost matrix yields and equivalent cost matrix. Therefore, without loss of generality it is often assumed that a cost-matrix has the main diagonal equal to zero (i.e., the cost of correctly classifying instances is zero). In such cases, the matrix has only two nonzero values (Table 1(b)): c_{21} (cost of misclassifying a negative instance as a positive) and c_{12} (cost of misclassifying a positive instance as a negative), and from the decision-making perspective the matrix has only one degree of freedom, the so-called *cost ratio* $\mathrm{CR} = c_{21}/c_{12}$. We will come back to these assumptions when we discuss abstaining classifiers.

Classifiers in a cost-sensitive setup can be characterized by the expected cost rc, a cost-weighted sum of misclassifications divided by the number of classified instances (making it invariant of the number of classified instances):

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

2.1 ROC analysis

An ROC plane ($fp \times tp$) has axes ranging from 0 to 1 and labeled *false positive rate* ($fp = FP/(FP + TN) = FP/N$) and *true positive rate* ($tp = TP/(TP + FN) = TP/P$), as shown in Fig. 1. Note that throughout the paper we use a convention in which uppercase variables denote absolute values and the lowercase variables denote the corresponding rates. 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 scoring classifier can be used to generate a set of points on the ROC plane efficiently (Fawcett 2003).

Given a set of points on an ROC plane, the ROC Convex Hull (ROCCH) method (Provost and Fawcett 2001) constructs a piecewise-linear convex down curve (called ROCCH) $f_{\mathrm{ROC}} : fp \mapsto tp$, having the following properties: (i) $f_{\mathrm{ROC}}(0) = 0$, (ii) $f_{\mathrm{ROC}}(1) = 1$, and (iii) the

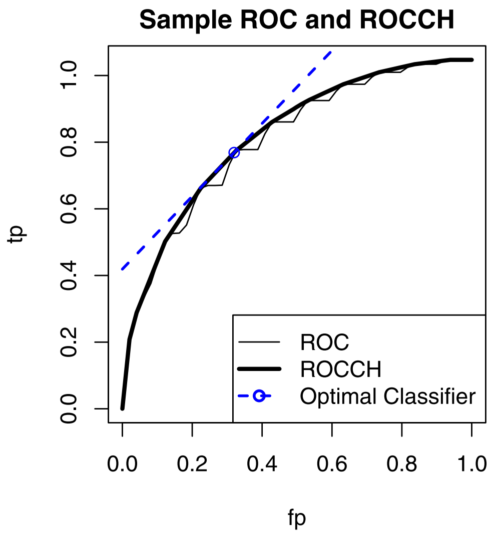


Fig. 1 Examples of ROC and ROCCH curves and the cost-optimal classifier

slope of f_{ROC} is monotonically nonincreasing. We denote the slope of a point on the ROCCH as f'_{ROC}.[1]

To find the classifier that minimizes the misclassification cost rc, we rewrite (1) as a function of one variable, FP, calculate the first derivative drc/dFP and set it equal to 0. This yields a known equation of iso-performance lines

$$f'_{\mathrm{ROC}}(fp^*) = CR\frac{N}{P}, \tag{2}$$

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 (2001), we assume that for any real $m > 0$ there exists at least one point (fp^*, tp^*) on the ROCCH having $f'_{\mathrm{ROC}}(fp^*) = m$.

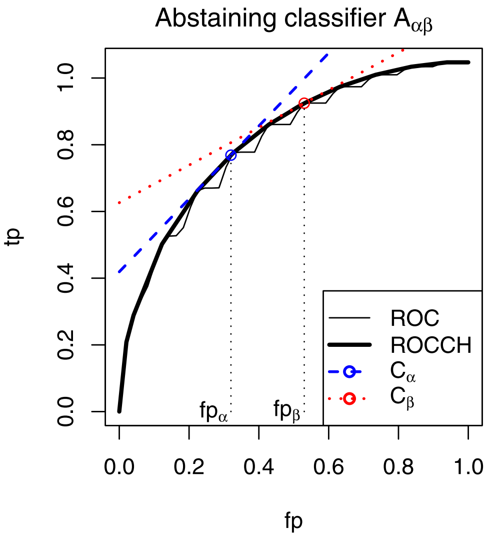
Note that the solution of this equation can be used to find a classifier that minimizes the misclassification cost for the instances used to create the ROCCH. We call such a classifier *ROC-optimal*. However, it may not be optimal for other instances. Nevertheless, if the testing instances used to build 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 an ROC curve and the binary classifiers used to construct it. An ROC-optimal classifier is defined as described in Sect. 2.1. The method constructs an abstaining metaclassifier $\mathcal{A}_{\alpha,\beta}$ using two

[1] We assume that the slope at vertices of a convex hull takes all values between the slopes of adjacent line segments.

Fig. 2 Abstaining classifier $\mathcal{A}_{\alpha,\beta}$ constructed using two classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$



binary classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ as follows:

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

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 $FP_\alpha \le FP_\beta$ (as shown in Fig. 2).

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 $fp_\alpha \le fp_\beta$, corresponding to $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$, the following conditions hold:

$$\forall i\colon \quad (\mathcal{C}_\alpha(i) = + \implies \mathcal{C}_\beta(i) = +) \wedge (\mathcal{C}_\beta(i) = - \implies \mathcal{C}_\alpha(i) = -). \tag{4}$$

Conditions (4) are the ones used by Flach and Wu (2005) in the work on repairing concavities of ROC curves. These are met in particular if the ROC curve and $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ are built from a single scoring classifier $\mathcal{R}$ (e.g., a Bayesian classifier) with two threshold values α and β ($\alpha \ge \beta$). The advantage is that for such a classifier, a simple and efficient algorithm for constructing an ROC curve exists (Fawcett 2003). For arbitrary classifiers (e.g., rule learners), (4) is generally violated. However, we observed that the fraction of instances with $\mathcal{C}_\alpha(i) = + \wedge \mathcal{C}_\beta(i) = -$ typically is small. As this is an important class of applications, this is an interesting area for future research.

Given a particular cost matrix and class distribution 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 nonclassified instances and the cost is undefined. Therefore, we proposed and investigated (Pietraszek 2005) three different criteria and models of optimization $\mathcal{E}$: the cost-based, the bounded-abstention and the bounded-improvement model, which we discuss in the following sections. Models $\mathcal{E}$

Table 2 Cost matrix Co for an abstaining classifier under Cost-Based Model. 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}

determine how nonclassified instances are accounted for in the misclassification cost and other boundary conditions. We formulate our goals as:

Given	– An 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 $\mathcal{E}$.

Cost-based model In the first evaluation model $\mathcal{E}_{CB}$, a so-called cost-based model, we use an extended 2×3 cost matrix with the third column representing the cost associated with abstaining from classifying an instance. This cost can be dependent on or independent of the true class of the instance.

Bounded models To address the shortcomings of the cost-based model and allow for situations in which the extended cost matrix is not available, we propose two models $\mathcal{E}_{BA}$ and $\mathcal{E}_{BI}$ that use a standard 2×2 cost matrix and calculate the misclassification cost per instance actually classified. The motivation is to calculate the cost only for instances the classifier attempts to classify.

In such a setup, a classifier randomly abstaining from classification would have the same misclassification cost as a normal classifier. Conversely, classifiers abstaining from classifying only for *difficult* instances may have a significantly lower misclassification cost.

However, such a system is underspecified as we do not know how to trade the misclassification cost for the number of nonclassified instances. To address this, we propose two bounded evaluation models having boundary conditions:

Bounded-abstention model $\mathcal{E}_{BA}$**,** where the system abstains for not more than a fraction $k_{\max}$ of instances and has the lowest misclassification cost,

Bounded-improvement model $\mathcal{E}_{BI}$**,** where the system has a misclassification cost not higher than $rc_{\max}$ and abstains for the lowest number of instances.

4 Cost-based model

In this model, we compare the misclassification cost, rc_{CB}, incurred by a binary and an abstaining classifier. We use an extended 2×3 cost matrix, with the third column representing the cost associated with classifying an instance as "?". Similarly, to the standard binary case we assume that the costs of correct classification is zero, however the costs of abstention can be different from different classes. This is different from a similar model studied by Tortorella (2000, 2004), who used a full cost matrix and only a single abstention cost. We will come back to these differences at the end of this section.

Given	– ROC curve generated using classifiers such that (4) holds – 2×3 cost matrix Co
Find	Classifier $\mathcal{A}_{\alpha,\beta}$ such that it minimizes misclassification cost rc_{CB}

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 $\mathcal{C}_\beta$. Similarly, false negatives for $\mathcal{C}_\beta$ imply false negatives for $\mathcal{C}_\alpha$, and we can thus formulate (5). The misclassification cost rc_{CB} is defined using a 2×3 cost matrix similar to (1), with the denominator equal to the total number of instances.

$$\begin{aligned}
rc_{CB} \cdot (N+P) &= \underbrace{(FP_\beta - FP_\alpha)c_{23}}_{\mathcal{C}_\alpha,\mathcal{C}_\beta \text{ disagree, }-} + \underbrace{(FN_\alpha - FN_\beta)c_{13}}_{\mathcal{C}_\alpha,\mathcal{C}_\beta \text{ disagree, }+} + \underbrace{FP_\alpha \cdot c_{21}}_{FP \text{ for both}} + \underbrace{FN_\beta \cdot c_{12}}_{FN \text{ for both}} \\
&= (FN_\alpha \cdot c_{13} + FP_\alpha \cdot (c_{21} - c_{23}) + FN_\beta \cdot (c_{12} - c_{13}) + FP_\beta \cdot c_{23}) \\
&= P\left(1 - f_{\mathrm{ROC}}\left(\frac{FP_\alpha}{N}\right)\right)c_{13} + FP_\alpha(c_{21} - c_{23}) \\
&\quad + P\left(1 - f_{\mathrm{ROC}}\left(\frac{FP_\beta}{N}\right)\right)(c_{12} - c_{13}) + FP_\beta \cdot c_{23}.
\end{aligned} \tag{5}$$

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

$$\begin{aligned}
\frac{\partial rc_{CB}}{\partial FP_\alpha} \cdot (N+P) &= -\frac{P}{N} f'_{\mathrm{ROC}}\left(\frac{FP_\alpha}{N}\right)c_{13} + c_{21} - c_{23}, \\
\frac{\partial rc_{CB}}{\partial FP_\beta} \cdot (N+P) &= -\frac{P}{N} f'_{\mathrm{ROC}}\left(\frac{FP_\beta}{N}\right)(c_{12} - c_{13}) + c_{23},
\end{aligned} \tag{6}$$

set the derivatives to zero and making sure that the function has a local extremum. After replacing FP_α and FP_β with the corresponding rates fp_α and fp_β, we obtain the final result:

$$\begin{aligned}
f'_{\mathrm{ROC}}(fp^*_\beta) &= \frac{c_{23}}{c_{12} - c_{13}} \frac{N}{P}, \\
f'_{\mathrm{ROC}}(fp^*_\alpha) &= \frac{c_{21} - c_{23}}{c_{13}} \frac{N}{P},
\end{aligned} \tag{7}$$

which, similarly to (2), allows us to find fp^*_α and fp^*_β, and the corresponding classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$. Note that these equations use only slopes of the ROC curve and are therefore very easy to apply.

This derivation is valid only for metaclassifiers (3) with (4), which implies $fp^*_\alpha \le fp^*_\beta$ and $f_{\mathrm{ROC}}(fp^*_\alpha) \le f_{\mathrm{ROC}}(fp^*_\beta)$. As an ROCCH is increasing and convex, its first derivative is nonnegative and nonincreasing, and we obtain $f'_{\mathrm{ROC}}(fp^*_\alpha) \ge f'_{\mathrm{ROC}}(fp^*_\beta) \ge 0$. Using the 2×3 cost matrix these conditions can be rewritten as:

$$(c_{21} \ge c_{23}) \wedge (c_{12} > c_{13}) \wedge (c_{21}c_{12} \ge c_{21}c_{13} + c_{23}c_{12}). \tag{8}$$

If condition (8) is not met, our derivation is not valid, but the solution is trivial. While it is clear that when the abstention costs are higher than the costs of incorrect classification, the optimal strategy is not to abstain (expressed by the first two terms of (8)), the interpretation of the third term is not obvious. We will prove it in the following theorem.

Theorem 1 *If (8) is not met, the classifier minimizing the misclassification cost is a binary classifier, namely, a single classifier described by (2).*

Proof Calculating (6) we obtain that if the rightmost part of (8) does not hold, $\partial rc_{CB}/\partial fp_\alpha$ is negative for all values $f'_{\mathrm{ROC}}(fp^*_\alpha) \leq f'_{\mathrm{ROC}}(fp^*) = c_{21}/c_{12} \cdot N/P$ and, similarly, $\partial rc_{CB}/\partial fp_\beta$ is positive for all values $f'_{\mathrm{ROC}}(fp^*_\beta) \geq f'_{\mathrm{ROC}}(fp^*) = c_{21}/c_{12} \cdot N/P$. This, together with the basic assumption $fp_\alpha \leq fp_\beta$ and the properties of the ROCCH, implies that $fp^*_\alpha = fp^*_\beta$, which means that the optimal abstaining classifier is a binary classifier. Such a classifier is the binary classifier described by (2). □

Equation (8) allows us to determine whether for a given 2 × 3 cost matrix Co a trivial abstaining classifier minimizing rc_{CB} exists, 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 $c_{23}/c_{13} = c_{21}/c_{12}$.

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, (8) simplifies to the harmonic mean of two misclassification costs: $c_{13} = c_{23} \leq c_{21}c_{12}/(c_{21} + c_{12})$. The second case yields the condition $c_{13} \leq c_{12}/2$ (equivalent to $c_{23} \leq c_{21}/2$). 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 (7) if (8) (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 binary classifier ($\mathcal{C}_\alpha = \mathcal{C}_\beta$).

4.1 Equivalence of the full cost matrix

Recall from Sect. 2, that in the binary case a cost matrix with the main diagonal equal to zero is equivalent to a full cost matrix with respect to classification. It is not obvious, however, that the same is true for the cost-based model. We will prove it in the following theorem:

Theorem 2 *Adding a constant value to the rows of the extended cost matrix yields an equivalent matrix with respect to the choice of the cost-optimal classifier.*

Proof Assume we have a full extended cost matrix

$$Co = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix}.$$

In the first step we will modify (5) to calculate costs for a full cost matrix to account for correct decisions made by the classifier. Classifier $\mathcal{A}_{\alpha,\beta}$ makes TP_α correct classification with class "+" and TN_β correct classifications with class "−". This means that the overall misclassification cost is increased by $Pf_{\mathrm{ROC}}(\frac{FP_\alpha}{N})c_{11} + (N - FP_\beta)c_{22}$. Hence, calculating derivatives based on (6) yields

$$\begin{aligned} \frac{\partial rc_{CB}}{\partial FP_\alpha} \cdot (N + P) &= -\frac{P}{N} f'_{\mathrm{ROC}}\left(\frac{FP_\alpha}{N}\right)(c_{13} - c_{11}) + (c_{21} - c_{23}), \\ \frac{\partial rc_{CB}}{\partial FP_\beta} \cdot (N + P) &= -\frac{P}{N} f'_{\mathrm{ROC}}\left(\frac{FP_\beta}{N}\right)(c_{12} - c_{13}) + (c_{23} - c_{22}). \end{aligned} \tag{9}$$

In both cases the derivatives use a only paired differences of two items in the same row of the matrix, which means that arbitrary constants added to the rows of the matrix cancel out and thus have no effect on the optimal classifier chosen by the algorithm. This completes the proof. □

An important implication from this proof is that the cost-based model presented by Tortorella, Tortorella (2000, 2004) with a full 2 × 2 cost-matrix and a single abstention cost is equivalent to ours (based on Theorem 2 one can subtract the values of costs of correct classification (c_{11}, c_{22}) from each row of the matrix). Note that if the costs of correct classification are not identical, converting Tortorella's model results in different values of abstention costs: c_{13} and c_{23}. Finally, this shows that from the classification perspective the extended cost matrix has only three degrees of freedom (we have four values, but any non-zero value can be fixed by multiplying the whole matrix by a constant).

5 Bounded models

In the experiments using a cost-based model, we noticed that the cost matrix and in particular cost values c_{13}, 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. Therefore, in the bounded models, we use a standard 2 × 2 cost matrix and calculate the misclassification cost only per instances classified.

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

$$\begin{aligned} rc_B &= \frac{1}{(1-k)(N+P)}(FP_\alpha \cdot c_{21} + FN_\beta \cdot c_{12}), \\ k &= \frac{1}{N+P}((FP_\beta - FP_\alpha) + (FN_\alpha - FN_\beta t)), \end{aligned} \tag{10}$$

determining the relationship between the fraction of classified instances k and the misclassification cost rc_B as a function of classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$. Similarly to the cost-based model we can rewrite these equations as functions of fp_α and fp_β:

$$\begin{aligned} rc_B &= \frac{1}{(1-k)(N+P)}(Nfp_\alpha \cdot c_{21} + P(1 - f_{\mathrm{ROC}}(fp_\beta)) \cdot c_{12}), \\ k &= \frac{1}{N+P}(N(fp_\beta - fp_\alpha) + P(f_{\mathrm{ROC}}(fp_\beta) - f_{\mathrm{ROC}}(fp_\alpha))). \end{aligned} \tag{11}$$

By putting boundary constraints on k and rc_B and trying to optimize the other variable, rc_B and k respectively, we create two interesting evaluation models we discuss in the following sections.

5.1 Bounded-abstention model

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

As stated in the introduction this model does not require an explicit 2 × 3 cost matrix, which may not be given and is particularly applicable in *resource-constrained* situations. In such cases, the bounded-abstention model yields an optimal classifier given the abstention window $k_{\max}$. For example, many classification systems involving human domain experts are typically resource constrained. Another example of such systems are real-time multi-stage classification systems in which the subsequent stages have a limited throughput and cannot process more than a fraction $k_{\max}$ of instances.

Given	– ROC curve generated using classifiers such that (4) holds – 2×2 cost matrix Co – Fraction k
Find	Classifier $\mathcal{A}_{\alpha,\beta}$ such that the classifier abstains for not more than a fraction of k instances and has the lowest cost rc_{BA}.

Unfortunately, unlike the cost-based model, the set of equations (11) for a bounded-abstention model does not have an algebraic solution in the general case, and in (Pietraszek 2005) we used general numerical optimization methods to solve it. Here we present an algorithm finding the solution that is extremely efficient for piecewise-linear ROCCH curves.

To find the solution for the bounded improvement model, we will use the constrained optimization method for the function of two variables. We will proceed in the following three steps: First, we will present the algorithm for a smooth convex down ROC curve differentiable in [0, 1] and assuming that *exactly* a fraction $k_{\max}$ of instances remains unclassified. Second, we will show under which conditions the optimal classifier can abstain for less than a fraction $k_{\max}$ of instances. Finally, we will extend the method to the piecewise linear ROCCH curves.

5.1.1 *Optimal classifier for a smooth convex down curve*

Our minimization task can be defined as follows: Find the minimum of $rc_{BA}(fp_\alpha, fp_\beta)$, subject to condition $k^*(fp_\alpha, fp_\beta) = k(fp_\alpha, fp_\beta) - k_{\max} = 0$.

For this, we will use the Lagrange method, which is a method for constrained optimization of a differentiable function under equality constrains (see e.g., Stewart 1992; Wolfram Research Inc. 2006 for a more complete coverage). Very briefly, given differentiable functions F and G the goal is to find the minimum of $F(\mathbf{X})$ given the constraint $G(\mathbf{X}) = 0$. The method calculates the so-called Lagrange multipliers λ such that

$$\nabla F(\mathbf{X}) = \lambda \nabla G(\mathbf{X}). \tag{12}$$

By solving (12) for $\mathbf{X}$ and λ and given the constraint $G(\mathbf{X}) = 0$ we obtain the desired solution.

In our case we have functions of two variables (fp_α and fp_β) and in this two-dimensional case (12) can has an interpretation that vectors ∇rc_{BA} and ∇k have the same direction. This is equivalent to $\nabla rc_{BA} \times \nabla k = \mathbf{0}$ and

$$\frac{\partial rc_{BA}}{\partial fp_\alpha} \frac{\partial k}{\partial fp_\beta} - \frac{\partial k}{\partial fp_\alpha} \frac{\partial rc_{BA}}{\partial fp_\beta} = 0. \tag{13}$$

The second condition is that $k^*(fp_\alpha, fp_\beta) = k(fp_\alpha, fp_\beta) - k_{\max} = 0$.

Calculating the derivatives, condition (13) simplifies to

$$\frac{f'_{\mathrm{ROC}}(fp_\alpha)f'_{\mathrm{ROC}}(fp_\beta)c_{12}P^2 - f'_{\mathrm{ROC}}(fp_\beta)NP(c_{21}-c_{12}) - c_{21}N^2}{(N(\underbrace{fp_\alpha - fp_\beta + 1}_{\geq 0}) + P(\underbrace{f_{\mathrm{ROC}}(fp_\alpha) - f_{\mathrm{ROC}}(fp_\beta) + 1}_{\geq 0}))^2(N+P)} = 0. \tag{14}$$

Based on the properties of the ROC curve and classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$, the denominator is always positive (with an exception for an all-abstaining classifier, $fp_\alpha = 0 \wedge fp_\beta = 1$, for which rc_{BA} is undefined), which means that (14) is equivalent to

$$f'_{\mathrm{ROC}}(fp_\beta)\left(f'_{\mathrm{ROC}}(fp_\alpha) + \frac{N}{P}\left(1 - \frac{c_{21}}{c_{12}}\right)\right) = \left(\frac{N}{P}\right)^2 \frac{c_{21}}{c_{12}}. \tag{15}$$

Theorem 3 *If f''_{ROC} is nonzero, the optimal classifier in the bounded-abstention model abstains for exactly a fraction $k_{\max}$ of instances and for a given $k \in [0,1]$ there exists exactly one classifier $\mathcal{A}_{\alpha,\beta}$.*

Proof In the first step we show that when $k = 0$, $fp_\alpha = fp_\beta = fp$, such that $f'_{\mathrm{ROC}}(fp) = \frac{N}{P}\frac{c_{21}}{c_{12}}$. The equality $fp_\alpha = fp_\beta$ results from the properties of the ROC curve and (11). Condition $f'_{\mathrm{ROC}}(fp) = \frac{N}{P}\frac{c_{21}}{c_{12}}$ results from (15).

In the second step we show that given an optimal classifier $\mathcal{A}_{\alpha,\beta}$ abstaining for exactly a fraction $k_{\max}$ of instances, we can easily generate a optimal classifier $\mathcal{A}^*_{\alpha,\beta}$ abstaining for a fraction $k^*_{\max} = k_{\max} + \delta_k$ (where $\delta_k \to 0$) of instances.

Such a classifier has coordinates $(fp_\alpha + \delta_\alpha, fp_\beta + \delta_\beta)$, in which the following condition holds:

$$\delta_k = \nabla k \cdot (\delta_\alpha, \delta_\beta). \tag{16}$$

The derivative of a smooth convex down ROC curve is positive, which means that all components of ∇k are nonzero.

Using (15) for the new point, we obtain the relationship between δ_α and δ_β:

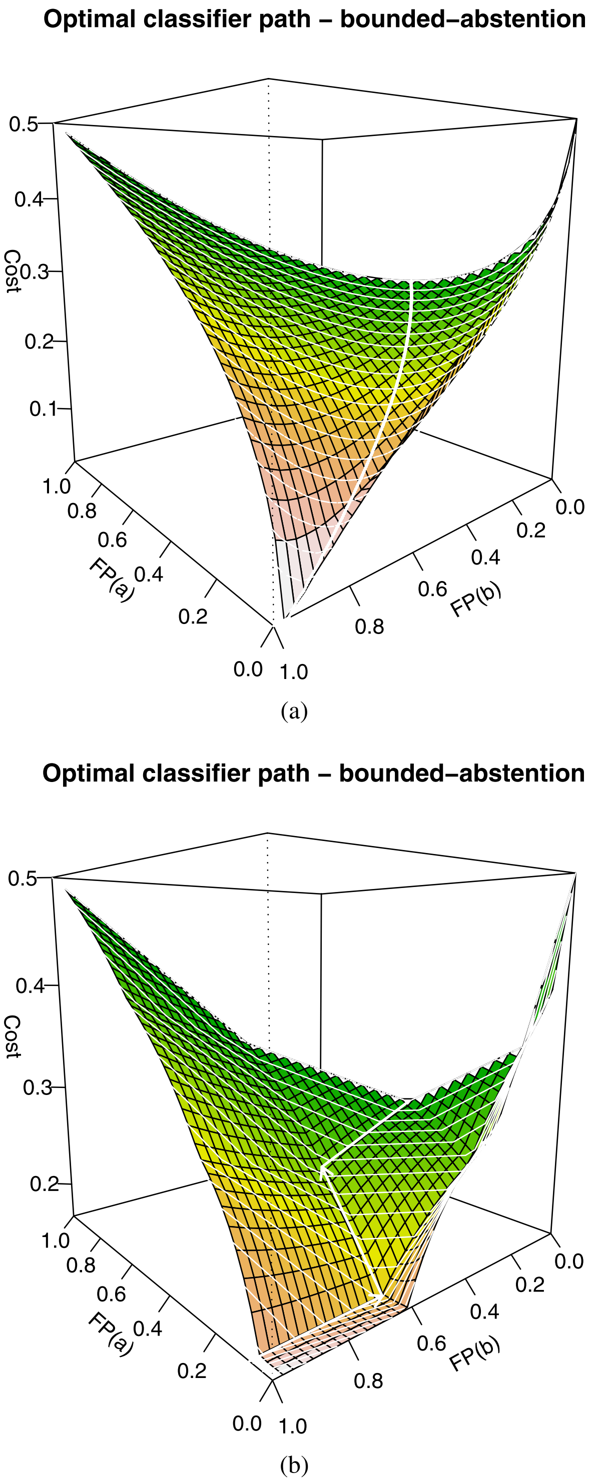
$$\begin{aligned}&(f'_{\mathrm{ROC}}(fp_\beta) + f''_{\mathrm{ROC}}(fp_\beta)\delta_\beta)\\ &\quad\cdot\left(f'_{\mathrm{ROC}}(fp_\alpha) + f''_{\mathrm{ROC}}(fp_\alpha)\delta_\alpha + \frac{N}{P}\left(1 - \frac{c_{21}}{c_{12}}\right)\right) = \left(\frac{N}{P}\right)^2 \frac{c_{21}}{c_{12}}\end{aligned} \tag{17}$$

and after simplifications we get the following result:

$$(\delta_\alpha, \delta_\beta) \cdot \left(f''_{\mathrm{ROC}}(fp_\alpha), f''_{\mathrm{ROC}}(fp_\beta)\frac{\frac{N}{P}\frac{c_{21}}{c_{12}}}{(f'_{\mathrm{ROC}}(fp_\beta))^2}\right) = 0. \tag{18}$$

Equations (16) and (18) show that for nonzero f''_{ROC} (i) there exists only one pair of $(\delta_\alpha, \delta_\beta)$ for given δ_k, (ii) $\delta_k \leq 0 \Rightarrow \delta_\alpha \leq 0 \wedge \delta_\beta \geq 0$, and (iii) $k \to 0 \Rightarrow fp_\alpha \to 0 \wedge fp_\beta \to 1$.

This completes the proof. Note that an almost similar inductive proof can be shown for classifier starting from $fp = 0 \wedge fp = 1$, with negative increments δ_k (note that the second step of the proof did not make any assumptions about the sign on δ_k). The advantage of such an approach is that there is no need to compute the value of starting point $fp_\alpha = fp_\beta = fp$ in this case. However, the derivative of rc_{BA} at $fp_\alpha = 0 \wedge fp_\beta = 1$ formally does not exist. □



Optimal classifier path – bounded–abstention

(a)

Optimal classifier path – bounded–abstention

(b)

Fig. 3 Optimal classifier paths in a bounded-abstention model. **a** Optimal classifier path for a smooth convex up curve (Algorithm 1). **b** Optimal classifier path for a piecewise ROCCH (Algorithm 2)

Algorithm 1 Algorithm for finding the optimal classifier

```
Data: ROC curve $f_{ROC}$, fraction $k_{\max}$
Result: $(fp_\alpha, fp_\beta)$ defining a classifier $\mathcal{A}_{\alpha,\beta}$, abstaining for no more than
        $k_{\max}$ instances and having the lowest misclassification cost $rc_{BA}$
$fp_\alpha \leftarrow fp_\beta \leftarrow fp$, such that $f'_{ROC}(fp) = \frac{N}{P}\frac{c_{21}}{c_{12}}.$;
while $k(fp_\alpha, fp_\beta) < k_{\max}$ do
    pick a small negative $\delta_k \to 0$ and find $\delta_\alpha, \delta_\beta$ such that
    $(\delta_\alpha, \delta_\beta) \cdot \left( f''_{ROC}(fp_\alpha), (f''_{ROC}(fp_\beta)\frac{\frac{N}{P}\frac{c_{21}}{c_{12}}}{(f'_{ROC}(fp_\beta))^2} \right) = 0$ and
    $\delta_k = \nabla k \cdot (\delta_\alpha, \delta_\beta)$;
    $fp_\alpha \leftarrow fp_\alpha + \delta_\alpha$;
    $fp_\beta \leftarrow fp_\beta + \delta_\beta$;
end
```

This proof generates an *optimal classifier path* on the hyperplane of rc_{BA} when varying $k_{\max}$ between 0 and 1, as shown in Fig. 3(a) (thick white line). Note that the thin lines show isolines of constant k (in this case with a constant increment of 0.1). The path can be constructed either by varying k from 0 to 1 or by varying k from 1 to 0. We will refer to these construction methods as "top-down" or "bottom-up".

The above derivation can be used to formulate Algorithm 1 for finding the optimal classifier in the bounded-abstention model.

5.1.2 Optimal classifier abstaining for fewer than $k_{\max}$ instances

In this section we will determine when the optimal classifier can abstain for a fraction smaller than $k_{\max}$ of instances. We will show when such a classifier exists and that when it exists it has the same misclassification cost as the optimal classifier abstaining for exactly a fraction of $k_{\max}$ instances.

Recall that the optimal abstaining classifier requires that (15) is met. In this section we will prove the following theorem:

Theorem 4 *Given an optimal classifier $\mathcal{A}_{\alpha,\beta}$ abstaining for exactly a fraction $k_{\max}$ of instances, no optimal classifier $\mathcal{A}^*_{\alpha,\beta}$ abstaining for a fraction $k^*_{\max} \leq k_{\max}$ of instances and having a misclassification cost lower than rc_{BA} exists.*

Proof Given an optimal classifier abstaining for *exactly* $k_{\max}$ instances, there exists a classifier abstaining for ($k^*_{\max} < k_{\max}$) and having the same or a lower misclassification cost iff $\partial rc_{BA}/\partial fp_\alpha \leq 0$ or $\partial rc_{BA}/\partial fp_\beta \geq 0$. In the remainder we will show when such a classifier exists.

In the calculations below we will use the following substitutions:

$$\begin{aligned}
A_\alpha &= f'_{\mathrm{ROC}}(fp_\alpha), \\
B_\alpha &= f_{\mathrm{ROC}}(fp_\alpha) - fp_\alpha f'_{\mathrm{ROC}}(fp_\alpha), \\
A_\beta &= f'_{\mathrm{ROC}}(fp_\beta), \\
B_\beta &= f_{\mathrm{ROC}}(fp_\beta) - fp_\beta f'_{\mathrm{ROC}}(fp_\beta).
\end{aligned} \tag{19}$$

Note that for a nondecreasing and convex down f_{ROC}, the following conditions hold:

$$\begin{aligned} &A_\alpha \ge A_\beta \ge 0,\\ &0 \le B_\alpha \le B_\beta \le 1,\\ &A_\alpha + B_\alpha \ge A_\beta + B_\beta \ge 1. \end{aligned} \tag{20}$$

Calculating $\partial rc_{BA}/\partial fp_\beta$. Calculating $\partial rc_{BA}/\partial fp_\beta \ge 0$, assuming that (15) holds and using substitution (19), yields the following condition:

$$\frac{\partial rc_{BA}}{\partial fp_\beta} \ge 0 \Leftrightarrow \underbrace{B_\alpha}_{\le 0} A_\beta c_{12} P^2 + \underbrace{(-A_\beta - B_\beta + 1)}_{\le 0} c_{12} NP \ge 0. \tag{21}$$

Equation (21) only holds if $B_\alpha = 0$ and $A_\beta + B_\beta = 1$ and in this case $\partial rc_{BA}/\partial fp_\alpha = 0$. From the properties of the ROC curve, this is only possible when f_{ROC} contains line segments $(0,0) - (fp_\alpha, f_{\mathrm{ROC}}(fp_\alpha))$ and $(fp_\beta, f_{\mathrm{ROC}}(fp_\beta)) - (1,1)$. In addition, condition (15) must hold.

Calculating $\partial rc_{BA}/\partial fp_\alpha$. Calculating $\frac{\partial rc_{BA}}{\partial fp_\alpha} \le 0$, assuming that (15) holds and using substitution (19), produces the following condition:

$$\frac{\partial rc_{BA}}{\partial fp_\beta} \le 0 \Leftrightarrow (A_\beta + B_\beta - 1)(A_\alpha c_{12} P^2 + (c_{12} - c_{21}) NP) \le -B_\alpha c_{21} NP. \tag{22}$$

Dividing both sides by (15) we obtain:

$$\underbrace{\frac{A_\beta + B_\beta - 1}{A_\beta}}_{\ge 0} \le \underbrace{-B_\alpha}_{\le 0} \frac{1}{NP}. \tag{23}$$

Similarly, this equation has a solution only if $B_\alpha = 0$ and $A_\beta + B_\beta = 1$.

To summarize, we proved that the classifier $\mathcal{A}_{\alpha,\beta}$ in the bounded-abstention model for $k_{\max}$ will have the lowest cost rc_{BA} when it abstains for exactly a fraction of $k_{\max}$ instances. Moreover in the special case, when $\mathcal{A}_{\alpha,\beta}$ is such that f_{ROC} contains the two line segments $(0,0) - (fp_\alpha, f_{\mathrm{ROC}}(fp_\alpha))$ and $(fp_\beta, f_{\mathrm{ROC}}(fp_\beta)) - (1,1)$, there exists an optimal classifier $\mathcal{A}^*_{\alpha,\beta}$ having the same misclassification cost and abstaining for fewer than $k_{\max}$ instances. Such a classifier will be described by the ends of following line segments: $(0,0) - (fp^*_\alpha, f_{\mathrm{ROC}}(fp^*_\alpha))$ and $(fp^*_\beta, f_{\mathrm{ROC}}(fp^*_\beta)) - (1,1)$. Such cases correspond to a flat area in Fig. 3(b). □

5.1.3 The algorithm for a convex hull f_{ROC}

Algorithm 1 is does not allow an efficient generation of a solution, as the increments δ_α, δ_β it uses are infinitely small. Moreover the property of nonzero f''_{ROC}, required by Algorithm 1, does not necessarily hold. However, our function f_{ROC} is a convex hull, a piecewise linear function, for which an efficient algorithm for finding the optimal classifier exists.

Assume the function f_{ROC} is a piecewise linear convex down curve, constructed from line n segments $S_1, S_2, \ldots, S_n$ connecting $n+1$ points $P_1, P_2, \ldots, P_{n+1}$. Each line segment S_i is described by a line segment $tp = A_i fp + B_i$, where A_i and B_i are the coefficients of a line connecting points P_i and P_{i+1}.

In this case, the value of derivatives f'_{ROC} is equal to A_i for $fp \in]fp_{P_i}; fp_{P_{i+1}}[$ and is undefined for arguments fp_{P_i} and $fp_{P_{i+1}}$. For our computations we assume that the value of f'_{ROC} for every argument fp_{P_i} takes all values between $[A_{i-1}; A_i]$. Moreover, we also assume that for fp_{P_1} the derivative takes all values $]\infty; A_1]$ and for $fp_{P_{n+1}}$ the derivative takes all values $[0; A_n]$.

Note that with a piecewise linear ROCCH, (18) cannot be used because the values of f''_{ROC} are either zero or undefined (at the vertices). However, (15) still can be used provided we allow that derivatives at vertices take all values in a range of slope values of adjacent segments.

Assuming the classifier $\mathcal{A}_{\alpha,\beta}$ optimal for a fraction $k_{\max}$ is defined by (fp_α, fp_β) where fp_α lies on the line segment S_i and fp_β lies on the line segment S_j, we construct the optimal classifier path "bottom-up" (i.e., constructing an optimal classifier $\mathcal{A}^*_{\alpha,\beta}$ for $k^*_{\max} < k_{\max}$).[2] The coordinates $(fp^*_\alpha, fp^*_\beta)$ of the classifier $\mathcal{A}^*_{\alpha,\beta}$ will depend on the value of the following expression:

$$X \leftarrow A_j\left(A_i + \frac{N}{P}\left(1 - \frac{c_{21}}{c_{12}}\right)\right) - \left(\frac{N}{P}\right)^2 \frac{c_{21}}{c_{12}}. \tag{24}$$

When $X < 0$, the classifier fp_α is located at the vertex (so that (15) holds) and the optimal classifier $\mathcal{A}^*_{\alpha,\beta}$ with $k^*_{\max} < k_{\max}$ will have $fp^*_\beta < fp_\beta$. Similarly, when $X > 0$, the classifier fp_β is located at the vertex and the optimal classifier $\mathcal{A}^*_{\alpha,\beta}$ with $k^*_{\max} < k_{\max}$ will have $fp^*_\alpha > fp_\alpha$. In both these cases the corresponding points fp^*_β and fp^*_α can be calculated from equation

$$k^*_{\max} = \frac{1}{N+P}(N(fp_\beta - fp_\alpha) + P(A_j fp_\beta + B_j - A_i fp_\alpha - B_i)), \tag{25}$$

given that the corresponding points fp_α and fp_β are fixed.

Finally, when $X = 0$, the classifier $\mathcal{A}_{\alpha,\beta}$ is located on line segments S_i, S_j outside vertices. In this case, the optimal classifier is defined ambiguously for a given $k_{\max}$ and these classifiers can be generated by finding all pairs satisfying (25) given the constraints that fp_α is within a line segment S_i and fp_β is within a line segment S_j. Specifically, it is also possible to use either of the classifiers for the two preceding cases ($X < 0$ or $X > 0$).

This leads to Algorithm 2 for finding the optimal classifier efficiently. The algorithm constructs the abstaining classifier "bottom-up" starting from points P_1 and P_n. At each step of the algorithm it calculates the value of X using (24) and depending on its sign, advances either i or j as shown in Fig. 4. If the abstention rate for the new points P_{i+1}, P_j (or P_i, P_{j-1} correspondingly) is larger than $k_{\max}$ the solution is calculated by solving a linear equation $k(fp_\alpha, fp_\beta) = k_{\max}$ with respect to fp_α (fp_β) and the algorithm terminates. Otherwise, in the next iteration the evaluation of X starts from the new point P_{i+1}, P_j (P_i, P_{j-1}).

Assuming the ROCCH consists of n line segments, the algorithm terminates in at most n steps. Therefore its complexity is $O(n)$.

5.2 Bounded-improvement model

The second bounded model is when we limit rc_B (hereafter referred to as rc_{BI}) to a threshold value $rc_{\max}$ ($rc_{BI} \le rc_{\max}$) and require that the classifier abstain for the smallest number of instances.

[2] If fp_α lies on the vertex connecting S_{i-1} and S_i, we assume the value A_i. Similarly, for fp_β lying on the vertex connecting S_j and S_{j+1}, we assume the value A_j.

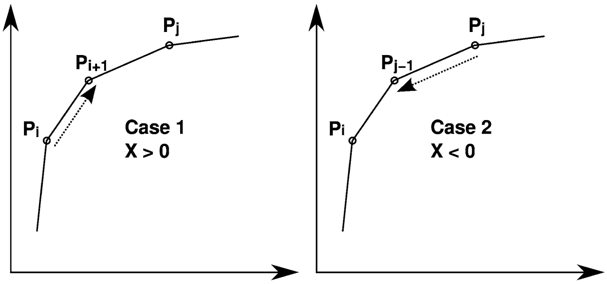


Fig. 4 Finding the optimal classifier in a bounded model: visualization of X

This model is particularly useful in a multi-stage classification systems and classification systems involving human domain experts, in which cost constraints are given. Given the maximum misclassification cost, the bounded-improvement model yields an abstaining classifier for which the misclassification cost for classified instances does not exceed a given value. Note that in this case, model performs a "local optimization" and the non-classified instances are excluded from cost calculations. Thus, if a complete coverage is needed such instances have to be classified using a more accurate but expensive classifier.

The possible real-life applications of this model include the classification of intrusion detection alerts, fraud detection systems, quality control system and medical domains. The advantage of this model is that the cost constraint is easily quantifiable and intuitive for the domain experts.

Given	– ROC curve generated using classifiers such that (4) holds – 2×2 cost matrix Co – Cost $rc_{\max}$
Find	Classifier $\mathcal{A}_{\alpha,\beta}$ such that the cost of the classifier is no greater than $rc_{\max}$ and the classifier abstains for the smallest number of instances.

This model is an inverse of the preceding model and can be solved by an algorithm similar to Algorithm 2. To show the solution for this model we use a similar approach as in the first model: First we will show the algorithm for a smooth convex down ROC curve differentiable in [0, 1] and assuming that the classifier has rc_{BI} equal to $rc_{\max}$. Second, we will show under which conditions the optimal classifier can have a misclassification cost smaller than $rc_{\max}$. Finally, we will present an efficient algorithm for piecewise linear ROCCH curves.

5.2.1 Optimal classifier for a smooth convex-down curve

To show the solution for the bounded improvement model, we will use the constrained optimization method for the function of two variables. The minimization task can be defined as follows: Find the minimum of $k(fp_\alpha, fp_\beta)$, subject to condition $rc^*_{BI}(fp_\alpha, fp_\beta) = rc_{BI}(fp_\alpha, fp_\beta) - rc_{\max} = 0$. Similarly as in Sect. 5.1, we use the Lagrange method and obtain the same condition (15). The second condition is that $rc_{BI}(fp_\alpha, fp_\beta) = rc_{\max}$.

Theorem 5 *If f''_{ROC} is nonzero, the optimal classifier in the bounded-improvement model has rc_{BI} equal to $rc_{\max}$ and for a given $rc_{BI} \in [0, rc^*_{BI}]$, where rc^* is the rc for the optimal binary classifier, there exists exactly one classifier $\mathcal{A}_{\alpha,\beta}$.*

Proof The proof is similar to Theorem 3 with an identical first condition (15) and the second condition $\delta_{rc} = \nabla rc_{BI} \cdot (\delta_\alpha, \delta_\beta)$. However, unlike in the preceding case, ∇rc_{BI} can be equal $\mathbf{0}$

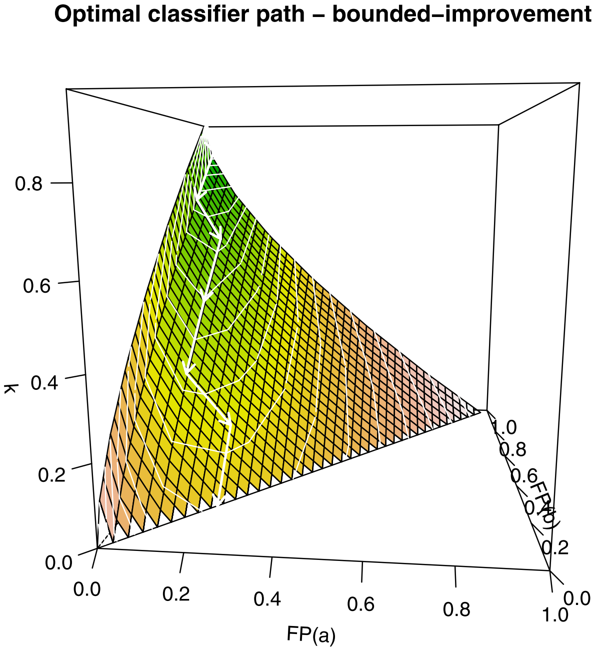
Algorithm 2 Algorithm for finding the optimal classifier in a bounded-abstention model for a piecewise-linear ROCCH curve

```
Input: ROCCH curve $f_{ROC}$, defined by $(n+1)$ points $P_1, \cdots P_{n+1}$,
       fraction $k_{\max}$
Result: $(fp_\alpha, fp_\beta)$ defining a classifier $\mathcal{A}_{\alpha,\beta}$, abstaining for no more than
        $k_{\max}$ instances and having the lowest misclassification cost $rc_{BA}$
$i_\alpha \leftarrow 1, i_\beta \leftarrow n+1$, $found \leftarrow$ FALSE ;
while (!found) do
    /* calculate coefficients for line segments $S_{i_\alpha}$ $S_{i_\beta - 1}$ */
    $A_{i_\beta-1} \leftarrow \frac{P_{i_\beta}[tp]-P_{i_\beta-1}[tp]}{P_{i_\beta}[fp]-P_{i_\beta-1}[fp]}$ ;
    $B_{i_\beta-1} \leftarrow P_{i_\beta}[tp] - A_{i_\beta-1}P_{i_\beta}[fp]$ ;
    $A_{i_\alpha} \leftarrow \frac{P_{i_\alpha+1}[tp]-P_{i_\alpha}[tp]}{P_{i_\alpha+1}[fp]-P_{i_\alpha}[fp]}$ ;
    $B_{i_\alpha} \leftarrow P_{i_\alpha}[tp] - A_{i_\alpha}P_{i_\alpha}[fp]$ ;
    /* evaluate which point to advance */
    $X \leftarrow A_{i_\beta-1}\left(A_{i_\alpha} + \frac{N}{P}(1-\frac{c_{21}}{c_{12}})\right) - \frac{N}{P}\frac{c_{21}}{c_{12}}$;
    if $X > 0$ then
        /* advance $fp_\alpha$ */
        if $k(P_{i_\alpha+1}[fp], P_{i_\beta}[fp]) \geq k_{\max}$ then
            $i_\alpha \leftarrow i_\alpha + 1$;
        else
            $fp_\beta \leftarrow P_{i_\beta}[fp]$;
            /* solve a linear eq. $k(fp_\alpha, fp_\beta) = k_{\max}$ with respect to
               $fp_\alpha$ */
            $fp_\alpha \leftarrow -\frac{k_{\max}(N+P)-P(B_{i_\beta-1}-B_{i_\alpha})-fp_\beta(N+PA_{i_\beta-1})}{N+PA_{i_\alpha}}$ ;
            $found \leftarrow$ TRUE;
        end
    else if $X < 0$ then
        /* advance $fp_\beta$ */
        if $k(P_{i_\alpha}[fp], P_{i_\beta-1}[fp]) \geq k_{\max}$ then
            $i_\beta \leftarrow i_\beta - 1$;
        else
            $fp_\alpha \leftarrow P_{i_\alpha}[fp]$;
            /* solve a linear eq. $k(fp_\alpha, fp_\beta) = k_{\max}$ with respect to
               $fp_\beta$ */
            $fp_\beta \leftarrow \frac{k_{\max}(N+P)-P(B_{i_\beta-1}-B_{i_\alpha})+fp_\alpha(N+PA_{i_\alpha})}{N+PA_{i_\beta-1}}$ ;
            $found \leftarrow$ TRUE;
        end
    else
        /* can move either $i_\alpha$ or $i_\beta$ */
        (...);
    end
end
```

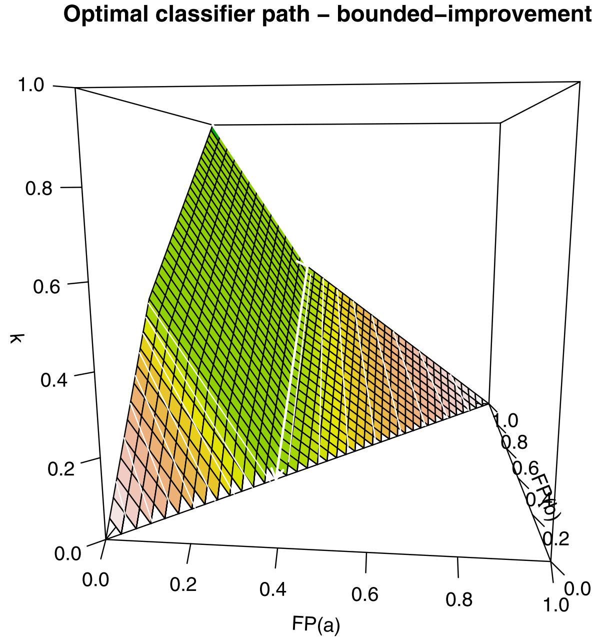
(under conditions shown in the proof of Theorem 4). In such a situation, the misclassification cost rc_{BI} will not change with the change of fp_α and fp_β. □

Similarly to the preceding case, the proof generates an optimal classifier path as shown in Fig. 5(a), where thin isolines show classifiers with identical misclassification cost rc_{BI}. The optimal classifier crosses these isolines at the points of minimal k.

Fig. 5 Optimal classifier paths in a bounded-improvement model. **a** Optimal classifier path for a piecewise ROCCH (Algorithm 2 with modifications (Sect. 5.2.3)). **b** The special case for an abstaining classifier



(a)



(b)

5.2.2 Optimal classifier with rc_{BI} lower than $rc_{\max}$

As we proved in Theorem 4, if f_{ROC} contains the line segments $(0,0)-(fp_\alpha, f_{\mathrm{ROC}}(fp_\alpha))$ and $(fp_\beta, f_{\mathrm{ROC}}(fp_\beta))-(1,1)$ and (15) holds, the classifier has the same rc_{BI} for all classifiers in these line segments.

Moreover, in this case, this rc_{BI} for this line segment is the lowest rc_{BI} an abstaining classifier can achieve with this ROC curve. Therefore for a higher $rc_{\max}$ the optimal classifier will have a lower misclassification cost. Such a situation is illustrated in Fig. 5(b).

5.2.3 The algorithm for a convex hull f_{ROC}

The algorithm is similar to Algorithm 2 with the following two modifications. Fist, it uses different conditions in lines 9 and 17, namely evaluating the misclassification cost given by

$$rc_{\max} = \frac{1}{(1-k)(N+P)}(Nfp_\alpha \cdot c_{21} + P(1-(A_j fp_\beta + B_j)) \cdot c_{12}), \tag{26}$$

where k is determined by (25). This yields

$$rc_{\max} = \frac{Nfp_\alpha \cdot c_{21} + P(1-(A_j fp_\beta + B_j)) \cdot c_{12}}{fp_\alpha(N+PA_i) - fp_\beta(N+PA_j) + P(B_j - B_i) + N + P}. \tag{27}$$

Second, in lines 13 and 21 the overall solution is calculated by solving a linear equation (27) with respect to fp_α and fp_β.

5.3 Equivalence of the full cost matrix

In the previous section we proved that in the cost-based model an extended 2×3 cost matrix has only three degrees of freedom and that adding an arbitrary constant to the rows of the cost matrix yields an equivalent matrix from a decision making perspective. This however does not hold for the 2×2 cost matrix for bounded models. We will prove the following theorem.

Theorem 6 *From a decision making perspective in bounded models, a full* 2×2 *cost matrix is not equivalent to the cost-matrix with the main diagonal equal to* 0.

Proof Similarly to Theorem 2 we modify rc_B in (10) to take correct classifications into account. This means that the nominator of (11) is increased by $Pf_{\mathrm{ROC}}(\frac{FP_\alpha}{N})c_{11} + (N - FP_\beta)c_{22}$.

Similarly to Sects. 5.1 and 5.2, using Lagrange method and calculating (13) we obtain a condition equivalent to:

$$\begin{aligned} &f'_{\mathrm{ROC}}(fp_\alpha)f'_{\mathrm{ROC}}(fp_\beta)(c_{12}-c_{11})P^2 + f'_{\mathrm{ROC}}(fp_\beta)(c_{21}-c_{22}-(c_{12}-c_{11}))NP \\ &\quad + (c_{22}-c_{21})N^2 + \underbrace{(f'_{\mathrm{ROC}}(fp_\alpha)+f'_{\mathrm{ROC}}(fp_\beta))(c_{22}-c_{11})NP}_{\neq 0} = 0. \end{aligned} \tag{28}$$

Comparing (28) with (14) we see that while the first three components are equivalent, the last one is not and, in general case, not equal to zero. This means that non-zero components on the main diagonal will change the optimal solution and thus the equivalence does not hold. This completes the proof. □

Note that condition (28) can be used in Algorithm 2 to find the optimal classifier when a full 2 × 2 cost matrix is given.

6 Experiments

To analyze the performance of our method, we tested it on 15 well-known datasets from the UCI KDD (Hettich and 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`. These datasets are all binary classification problems from 37 UCI datasets, downloaded from the Weka web-page.

We tested our method in all three models described above. In the $\mathcal{E}_{CB}$ model, the input data is a 2×3 cost matrix in the symmetric case ($c_{13} = c_{23}$). In $\mathcal{E}_{BA}$, we use a 2×2 cost matrix and $k_{\max}$ (the fraction of instances that the system does not classify). In $\mathcal{E}_{BI}$, we could not use a constant value of $rc_{\max}$ for all datasets because different datasets yield different ROC curves and misclassification costs. Instead we used a *relative cost improvement* f and calculated rc as follows: $rc = (1 - f)rc_{\text{bin}}$, where rc_{bin} is the misclassification cost of the ROC-optimal binary classifier found using (2). Hence the input data is also a 2×2 cost matrix and a fraction $f_{\min}$.

The goals of our experiments are three-fold: (i) to verify the ROC generalization and the algorithms used, (ii) to analyze the trade-off between the abstention ratio and the cost improvement and, (iii) to compare our abstaining classifiers with other known techniques allowing for abstentions. We will discuss them in the remainder of this section.

ROC generalization Our method for selecting abstaining classifiers with two thresholds is provably optimal provided that the ROC curve generalizes to an independent test set. The method presented in this paper can only be applicable in real-life situations if the generalization properties hold, which also includes the stability of algorithms used (e.g., threshold averaging for ROC curves and threshold interpolation). Therefore, the first goal of the experiments is to validate it.

Abstention ratio vs. cost improvement From the application standpoint, it is important to understand the relationship between the abstention ratio, cost parameters and the cost improvement. As showed in previous sections, depending on the constraints and the parameters given, the optimal solution can be found by using one of the three algorithms provided. Ultimately, the gain from using an abstaining classifier instead of a binary one depends on the shape of the ROC curve and is dataset specific. Therefore, the second goal of the experiments is to show the expected performance estimates on a variety of publicly available datasets (additional experiments with abstaining classifiers in the domain of computer intrusion detection can be found in Pietraszek 2007).

Direct comparison Finally, all the models, in particular the bounded models, perform "by design" better than the normal binary classifiers, therefore it is more interesting to compare them against other abstaining classifiers. From the other models (cf. Sect. 7) we chose *cautious classifiers* (Ferri and Hernández-Orallo 2004; Ferri et al. 2004) as the most appropriate for our purposes. Here we briefly discuss how cautious classifiers work and how we can make them comparable in our evaluation.

Cautious classifiers use a single multi-class probabilistic classifier and a vector K (class bias) and w (window size) the decision rule shown in Algorithm 3.

Algorithm 3 Cautious classifiers decision rule

```
Input: Classifier $\mathcal{C} : I \mapsto C$, instance $x \in I$, class bias $K$, such that
       $\sum_i k_i = 1$, window size $w$.
Result: Classification: $c \in C \cup$ “?”.
for $i \in C$ do
    $\tau_i \leftarrow (1 - k_i)w + k_i$;
end
$p_i \leftarrow \mathcal{C}(x)$ // base classifier
if $\exists p_i : p_i \geq \tau_i$ then
    $c \leftarrow \mathrm{argmax}_i(p_i/\tau_i)$;
else
    $c \leftarrow$ “?”;
end
```

This makes cautious classifiers similar to the bounded-abstention model, in which an abstention window is defined. However, although for $w = 0$ abstention is zero and the classifier abstains for almost all instances for $w = 1$, the relationship between w and the abstention is neither continuous nor linear (Ferri and Hernández-Orallo 2004). Therefore our model cannot directly compared 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.

In the evaluation we use a simple probabilistic classifier (naive Bayes) and a binary classification, which is compatible with both our algorithms and the decision rules for cautious classifiers. In fact, for a binary case cautious classifiers have only two degrees of freedom: a scalar k (combining the information about costs and the class distribution) and a window w. The interpretation of these parameters is not clear, but make a fair comparison of there methods we need to have the two classifiers have the same point of operation (i.e., the abstention window or the misclassification cost). To be able to do this we decided to: (i) compare only classifiers for $CR = 1$, thus avoiding having to encode the misclassification costs in k, (ii) perform the comparison only in the bounded abstention and bounded improvement models, (iii) use a simple algorithm based on a binary search for finding the window w, given $k_{\max}$ ($f_{\min}$) and the ROC curve in the bounded-abstention (bounded-improvement) model, respectively. In such a setting, we can directly compare rc (k) of the two types of classifiers analyzing their performance. Such a comparison is the third goal of our evaluation.

6.1 Constructing an abstaining classifier

Recall that vertices on the ROCCH can only be used to find an ROC-optimal classifier in the cost-based model (Sect. 4). In the other two models, the ROC-optimal classifier uses arbitrary points on the ROCCH, most typically one point is located at the vertex and a the other one is located on a line segment computed in Algorithm 2 and the modified version (Sect. 5.2.3).

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 did.

An ROCCH can be considered a function $f_{\mathrm{ROC}} : \tau \mapsto (fp, tp)$, 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 $f^{-1}_{\mathrm{ROC}} : (fp, tp) \mapsto \tau$ and interpolate it using splines with a function $\hat{f}^{-1}_{\mathrm{ROC}}$ defined for a continuous range of values τ. Given an arbitrary point (fp^*, tp^*) on the curve, we use the function $\hat{f}^{-1}_{\mathrm{ROC}}$ yielding τ^* to construct a classifier $\mathcal{C}_{\tau^*}$.

6.2 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. 6. We used another two-fold cross-validation ($n = 2$) to construct an 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. Although 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 scoring classifier by calculating the prediction ratio $P(+\,|\,x)/P(-\,|\,x)$.

Given the ROC curve and the input parameters (cost matrix and a value $k_{\max}$ or $f_{\min}$), the program uses the algorithms proposed to find 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 $\mathcal{A}_{\alpha,\beta}$.

Such an experiment was run for every dataset and every combination of input parameters, CR and c_{13} ($k_{\max}$ or $f_{\min}$), 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}, $k_{\max}$ or $f_{\min}$).

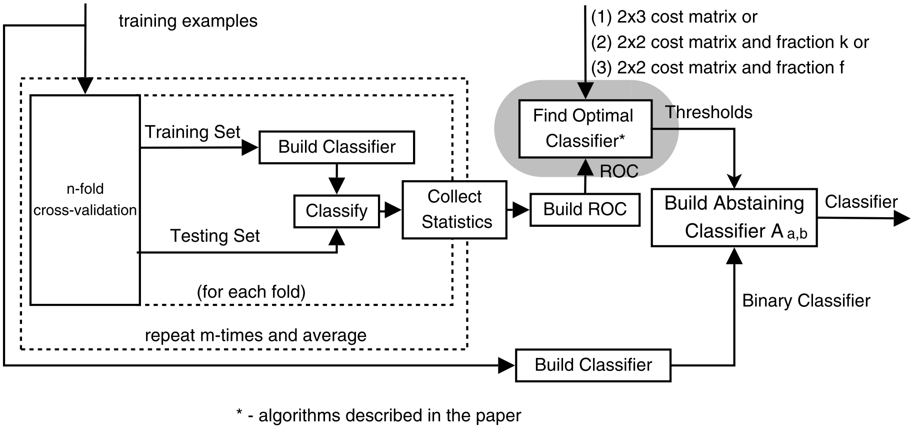


Fig. 6 Building an abstaining classifier $\mathcal{A}_{\alpha,\beta}$

We used three values of the cost ratio (CR): 0.5, 1 and 2, and four different values of c_{13} (first model), $k_{\max}$: 0.1, 0.2, 0.3 and 0.5 (second model), and $f_{\min}$: 0.1, 0.2, 0.3 and 0.5 (third model), yielding 180 experiment runs ($15 \times 3 \times 4$) for each model.

We will briefly justify this choice of parameters. For the first model, we selected values of c_{13} that are evenly spaced between 0 and the maximum value for a particular cost ratio (cf. (8)). For the other two models, we believe that, while the results will definitely be application-dependent, values of $k_{\max}$ ($f_{\min}$) that are lower than 0.1 bring too small an advantage to justify abstaining classifiers, whereas values larger than 0.5 may not be practical for real classification systems. For the CRs we tested the performance of our system for cost ratios close to 1.

We used a naive Bayes classifier from the Weka toolkit (Witten and Frank 2000) as a machine-learning method and R (R Development Core Team 2004) to perform numerical calculations.

6.3 Results—cost-based model

Out of 180 experiments (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).

Testing the stability of the generalization of ROC and the interpolation algorithms used, we calculated a relative error of the predicted (based on the ROC curve) and the actual (obtained on a validation set) misclassification cost $\Delta rc/rc = 0.15 \pm 0.01$ and the abstention ratio $\Delta k/k = 0.05 \pm 0.01$ for all datasets. The positive values mean that the classifier has on average a higher cost then expected and a marginally higher abstention ratio. This shows that the method is fairly stable.

Evaluating the relationship between the cost matrix and the abstention ratio, Fig. 7 shows the results for one representative dataset. The X-axes correspond to the cost value in a symmetric case $c_{13} = c_{23}$ (top and middle panel), and the Y-axes show the relative cost improvement (top panel) and the fraction of nonclassified instances (middle panel). The bottom 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 nonclassified instances, only indirectly determined by c_{13}. Finally, Table 3 contains tabular results for all datasets for one cost ratio and two sample costs $c_{13} \in \{0.1, 0.2\}$.

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. However for different datasets even small differences in c_{13} result in large differences of k and f. On the other hand, for many datasets, we observe an almost linear relationship between the fraction of nonclassified instances and the relative cost improvement.

6.4 Results—bounded models

6.4.1 Bounded-abstention model

Out of 180 experiments (15 datasets, four values of fractions of nonclassified instances and three cost values), 179 have a significantly lower rc_B 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.

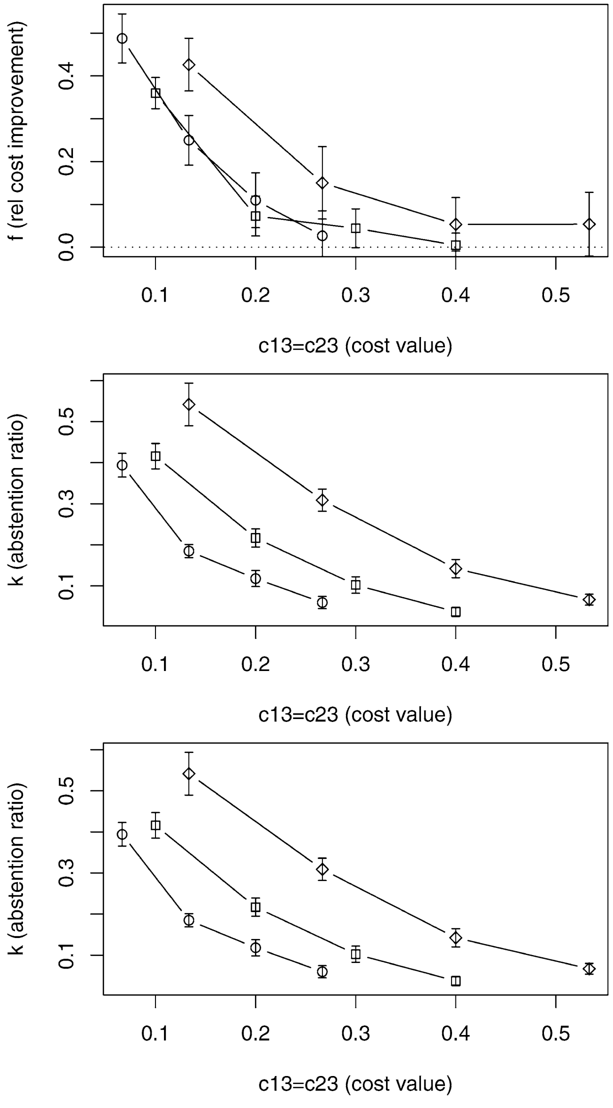


Fig. 7 Cost-based model: Relative cost improvement and fraction of nonclassified instances for ionosphere.arff, a representative dataset (○: $CR = 0.5$, □: $CR = 1$, ◇: $CR = 2$)

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 ($\Delta k/k$) for all runs is 0.078 ± 0.008. Similarly, the relative difference between the actual misclassification cost and the cost estimated based on the ROC is also very small ($\Delta rc/rc = -0.065 \pm 0.11$). This is particularly important as both k and rc are only indirectly determined by the two thresholds the algorithm calculates. This proves the stability of the algorithm on a variety of datasets.

The results for a representative dataset are shown in Fig. 8 and tabular results for all datasets for one cost ratio and two sample $k_{\max}$ are shown in Table 4. The X-axes correspond to the actual fraction of nonclassified instances and the Y-axes show the relative cost improvement (top panel) and the misclassification cost (bottom panel). The top panel shows

Table 3 Fraction of nonclassified instances (k) and relative cost improvement (f) for a cost-based model ($CR = 1$, $c_{13} = \{0.1, 0.2\}$)

Dataset	$c_{13} = 0.1$		$c_{13} = 0.2$	
	k	f	k	f
breast-cancer	0.97±0.03	0.64±0.01	0.68±0.05	0.31 ± 0.02
breast-w	0.31±0.03	0.16±0.05	0.05 ± 0	0.13 ± 0.05
colic	0.96±0.03	0.44±0.02	0.27±0.04	0.15 ± 0.03
credit-a	0.64±0.01	0.48±0.02	0.33±0.02	0.22 ± 0.02
credit-g	0.84±0.02	0.64±0.01	0.59±0.02	0.38 ± 0.01
diabetes	0.81±0.01	0.64±0.01	0.67±0.02	0.35 ± 0.01
heart-statlog	0.76±0.03	0.46±0.01	0.32±0.04	0.14 ± 0.04
hepatitis	0.46±0.06	0.51±0.03	0.29±0.03	0.29 ± 0.04
ionosphere	0.42±0.03	0.36±0.04	0.22±0.02	0.07 ± 0.05
kr-vs-kp	0.62±0.02	0.46±0.02	0.29±0.01	0.26 ± 0.01
labor	0.65±0.07	0.16±0.13	0.36±0.08	−0.09 ± 0.17
mushroom	0.23±0.02	−0.08±0.06	0.03 ± 0	0.22 ± 0.01
sick	0.13 ± 0	0.51±0.03	0.09 ± 0	0.28 ± 0.05
sonar	0.93±0.02	0.68±0.02	0.77±0.04	0.41 ± 0.03
vote	0.34±0.04	0.55±0.04	0.17±0.01	0.39 ± 0.05

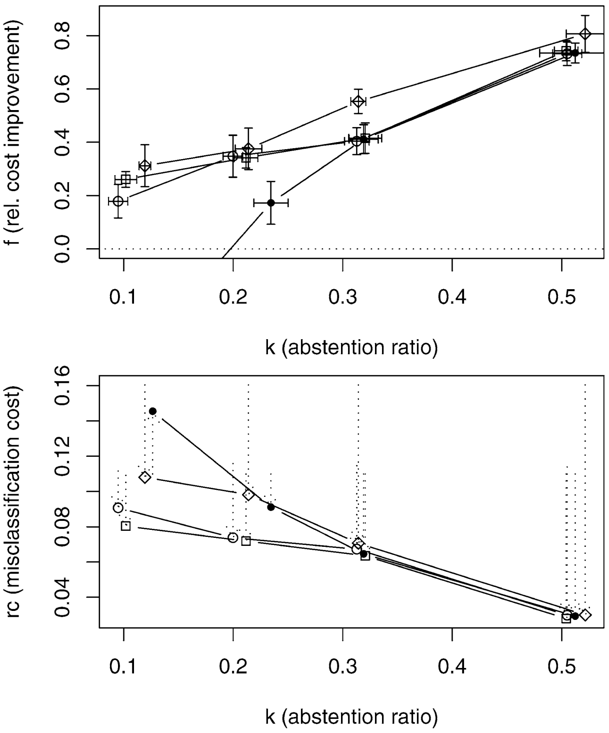


Fig. 8 Bounded-abstention model: Relative cost improvement and the absolute cost for ionosphere.arff, a representative dataset (○: $CR = 0.5$, □: $CR = 1$, ◇: $CR = 2$, •: cautious classifier for $CR = 1$)

Table 4 Relative cost improvement (f) as a function of a fraction of nonclassified instances ($k_{\max}$) for a bounded-abstention model ($CR = 1$, $k_{\max} = \{0.1, 0.5\}$)

Dataset	$k_{\max} = 0.1$		$k_{\max} = 0.5$	
	k	f	k	f
breast-cancer	0.1 ± 0.01	0.07 ± 0.01	0.53 ± 0.01	0.3 ± 0.07
breast-w	0.12 ± 0.02	0.58 ± 0.06	0.53 ± 0.04	1 ± 0
colic	0.09 ± 0.01	0.14 ± 0.02	0.48 ± 0.01	0.33 ± 0.03
credit-a	0.1 ± 0	0.17 ± 0.02	0.5 ± 0.01	0.55 ± 0.03
credit-g	0.1 ± 0	0.11 ± 0.01	0.51 ± 0.01	0.37 ± 0.07
diabetes	0.11 ± 0.01	0.11 ± 0.02	0.51 ± 0.01	0.41 ± 0.03
heart-statlog	0.11 ± 0.01	0.19 ± 0.03	0.56 ± 0.02	0.58 ± 0.09
hepatitis	0.13 ± 0.01	0.33 ± 0.04	0.53 ± 0.03	0.71 ± 0.07
ionosphere	0.1 ± 0.01	0.26 ± 0.03	0.5 ± 0.01	0.74 ± 0.04
kr-vs-kp	0.1 ± 0	0.25 ± 0.01	0.56 ± 0.02	0.89 ± 0.02
labor	0.12 ± 0.04	0.37 ± 0.15	0.58 ± 0.05	0.77 ± 0.19
mushroom	0.09 ± 0.02	0.71 ± 0.01	0.42 ± 0.02	1 ± 0
sick	0.11 ± 0	0.7 ± 0.01	0.47 ± 0	0.85 ± 0.02
sonar	0.13 ± 0.01	0.12 ± 0.03	0.56 ± 0.02	0.6 ± 0.05
vote	0.1 ± 0.01	0.46 ± 0.04	0.55 ± 0.02	0.96 ± 0.03

the relative cost improvement as a function of the fraction of instances handled by operator k. The bottom panel shows the same data with the absolute values of rc_B. The dashed arrows indicate the difference between an optimal binary classifier and an abstaining one.

In general, the higher the values of k, the higher the cost improvement; for eight datasets, namely `breast-cancer`, `credit-a`, `credit-g`, `diabetes`, `heart-statlog`, `ionosphere`, `kr-vs-kp` and `sonar`, we can observe an almost linear dependence between these variables. For four datasets (`breast-w`, `mushroom`, `sick`, `vote`) even as low an abstention as 0.1 can lead to a reduction of the misclassification cost by half (and of as much as 70% for two datasets).

Finally, as the third part of the evaluation we compared the performance of abstaining classifiers and cautious classifiers. Recall from Sect. 6 that although cautious classifiers do not use the ROC, we used it to find the threshold w to obtain the desired abstention window k. Table 5 shows that cautious classifiers are on average less stable than abstaining classifiers and yield smaller cost improvements, especially for lower values of k. In fact, cautious classifiers for $k = 0.1$ seem to perform worse than a binary classifier (negative cost improvement). For larger values of k both classifiers perform comparably. Quantitatively, comparing the cost improvement for 15 datasets and 4 fractions k (with 10 runs each) we got 31 significant wins, 24 ties and 5 losses.

6.4.2 *Bounded-improvement model*

This model is in fact the inverse of the preceding model, and thus we expected very similar results. The results for a representative dataset are shown in Fig. 9, and tabular results for all datasets for one cost ratio and two sample $f_{\min}$ are shown in Table 6. The X-axes correspond to the relative cost improvement (top panel) and the misclassification cost (bottom panel). The Y-axes show the actual fraction of nonclassified instances. The top panel shows the

Table 5 Comparison of abstaining and cautious classifiers in a bounded-abstention model (all datasets, $CR = 1$)

Abstention ratio $k_{\max}$	Abstaining classifiers		Cautious classifiers	
	k	f	k	f
0.1	0.11 ± 0.00	0.31 ± 0.02	0.16 ± 0.01	-0.32 ± 0.13
0.2	0.22 ± 0.00	0.47 ± 0.03	0.26 ± 0.01	0.13 ± 0.10
0.3	0.33 ± 0.00	0.53 ± 0.03	0.37 ± 0.01	0.48 ± 0.03
0.5	0.52 ± 0.01	0.68 ± 0.03	0.55 ± 0.01	0.68 ± 0.02

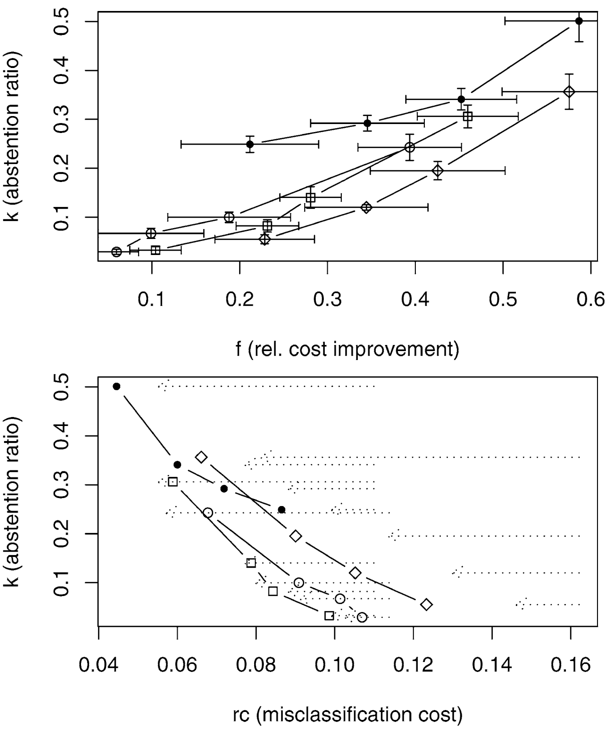


Fig. 9 Bounded-improvement model: Fraction of nonclassified instances for ionosphere.arff, a representative dataset ($\circ$: $CR = 0.5$, $\square$: $CR = 1$, $\diamond$: $CR = 2$, $\bullet$: cautious classifier for $CR = 1$)

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 set based on the performance on the training set. The mean of the relative difference of f ($\Delta f / f$) for all runs is 0.31 ± 0.15. The positive value of the mean shows that, on average, the system has a lower misclassification cost than required. Note that this value is higher than the corresponding difference in the preceding model. We conclude that this model is less stable and more sensitive to parameter changes than the preceding one. The right panel shows the same data with the X-axis giving absolute cost values. 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.

Table 6 Fraction of nonclassified instances k as a function of a relative cost improvement ($f_{\min}$) for a bounded-improvement model ($CR = 1$, $f_{\min} = \{0.1, 0.5\}$)

Dataset	$f_{\min} = 0.1$		$f_{\min} = 0.5$	
	f	k	f	k
breast-cancer	0.28 ± 0.25	0.67 ± 0.15	0.66 ± 0.24	0.93 ± 0.05
breast-w	0.15 ± 0.04	0.03 ± 0.03	0.48 ± 0.06	0.11 ± 0.03
colic	0.09 ± 0.05	0.12 ± 0.09	0.43 ± 0.11	0.86 ± 0.03
credit-a	0.13 ± 0.02	0.06 ± 0.01	0.52 ± 0.04	0.46 ± 0.02
credit-g	0.14 ± 0.05	0.39 ± 0.12	0.46 ± 0.09	0.78 ± 0.08
diabetes	0.09 ± 0.03	0.39 ± 0.17	-0.2 ± 0.91	0.89 ± 0.06
heart-statlog	0.13 ± 0.04	0.07 ± 0	0.51 ± 0.13	0.62 ± 0.08
hepatitis	0.12 ± 0.1	0.22 ± 0.19	0.54 ± 0.08	0.49 ± 0.18
ionosphere	0.1 ± 0.03	0.03 ± 0.01	0.46 ± 0.06	0.31 ± 0.02
kr-vs-kp	0.1 ± 0.01	0.05 ± 0.01	0.5 ± 0.01	0.24 ± 0.01
labor	0.36 ± 0.19	0.23 ± 0.16	0.42 ± 0.44	0.59 ± 0.16
mushroom	0.07 ± 0.01	0 ± 0	0.48 ± 0.02	0.04 ± 0.01
sick	0.27 ± 0.05	0.04 ± 0.01	0.56 ± 0.03	0.07 ± 0
sonar	0.15 ± 0.06	0.19 ± 0.01	0.52 ± 0.13	0.74 ± 0.05
vote	0.13 ± 0.04	0.03 ± 0.01	0.53 ± 0.02	0.13 ± 0.02

Table 7 Comparison of abstaining and cautious classifiers in a bounded-improvement model (all datasets, $CR = 1$)

Cost improvement $f_{\min}$	Abstaining classifiers		Cautious classifiers	
	f	k	f	k
0.1	0.17 ± 0.02	0.16 ± 0.02	-0.02 ± 0.12	0.27 ± 0.01
0.2	0.27 ± 0.02	0.24 ± 0.02	0.10 ± 0.12	0.34 ± 0.01
0.3	0.36 ± 0.02	0.31 ± 0.03	0.23 ± 0.11	0.40 ± 0.02
0.5	0.49 ± 0.03	0.45 ± 0.03	0.47 ± 0.06	0.54 ± 0.02

Similarly, to the preceding model, the four datasets can yield a 50% cost reduction while abstaining for approximately 10% of the instances. On the other hand, there are datasets in which even a 10% cost reduction is done at the cost of large abstention windows (e.g., 67% for `breast-cancer`). Considering much larger actual relative cost improvements than the desired one, we conclude that this model is more difficult to tune than the bounded-improvement model.

Finally, the comparison with cautious classifiers is shown in Table 7. Cautious classifiers are less stable and for comparable cost improvements f require higher abstentions. Quantitatively, comparing the abstention ratios for the corresponding cost improvements we obtain 31 significant wins 16 ties and 13 losses for all datasets.

7 Related work

Classifiers with reject rules were first investigated by Chow (1970) and further developed by Tortorella (2000, 2004) in the area of pattern recognition. The latter uses ROC analysis in a model corresponding to our cost-based model with a different cost matrix ($c_{13} = c_{23}$). In Sect. 4.1 we proved that these models are equivalent and thus can be used interchangeably. Furthermore, we show conditions under which a nontrivial abstaining classifier exists and also propose two bounded models with different optimization criteria.

Dubuisson and Masson (1996) and more recently Muzzolini et al. (1998) analyze statistical classifiers with a reject option in a multi-class classification setting. The authors propose two types of rejections: (i) ambiguity reject, in which an instance is classified to two or more classes with near equal probability and (ii) distance reject, in which the instance has no similarity to each of the prototypical implementation of any of the classes. Given the classification error tolerance (the probability of incorrect classification) ϵ and its confidence error C_d the Inck method (Muzzolini et al. 1998) can select the ambiguity reject threshold C_a such that the probability of ambiguity reject is minimal. This setting is the most similar to our bounded-improvement classifier model, however there is a number of differences: First, both papers make strong assumptions w.r.t. underlying classifiers and feature distributions, whereas our method can use almost arbitrary classifiers. Second, the methods are more appropriate for multi-class classifiers (Tortorella 2004) and do not take misclassification costs into account. Finally, statistical classifiers with reject rules require a feature selection algorithm and expensive integration over the pattern space to determine required probabilities. In contrast, building of the ROC curve is simple and our selection algorithms run linearly with the number of instances.

As already discussed in Sect. 6, cautious classifiers (Ferri and Hernández-Orallo 2004) are more general than ours as they support multi-class classification problems and do not use ROC curves. On the other hand, the interpretation of the abstention window w and class bias k is less intuitive and, in many cases need additional calibration (like the one performed in our evaluation). Another limitation is that cautious classifiers require calibrated probabilities from the output classifier, whereas abstaining classifiers need a scoring classifier (or in general a classifier for which condition (4) holds).

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.

Flach and Wu (2005) use a single ROC curve with identical condition (4) to "repair concavities" in the ROC curve increasing the AUC and improving the classification performance. As the method effectively uses a simple operation on scores assigned by the underlying classifier, condition (4) also holds for the "repaired" curve. This means that abstaining classifiers can be cascaded with the above method, likely improving the performance of abstaining 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 and Lavrač 2000) are another example of classifiers that may abstain from classification. They 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 (Pazzani et al. 1994), the authors do not maximize the accuracy. Moreover, confirmation rule sets are specific to the learning algorithm used.

Active learning (Lewis and Catlett 1994) minimizes the number of labeled instances by iteratively selecting a few instances to be labeled. This selection process uses an implicit abstaining classifier, where it selects instances that are lying closest to the decision boundary, however no cost-based optimization is performed.

8 Summary and conclusions

In this paper we proposed a method to build a *ROC-optimal abstaining classifier* using ROC analysis. Such a classifier minimizes the misclassification cost on instances used to build the ROC curve. Moreover, it 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, a bounded-abstention and a bounded-improvement model, which are relevant for numerous practical applications. All the models use only the base classifier and an ROC curve and do not require that the underlying has classifier calibrated output probabilities, which is not always trivial (Cohen and Goldszmidt 2004; Zadrozny and Elkan 2001).

In the first model, we used a 2 × 3 cost matrix, showed the conditions under which the abstaining classifier has a nontrivial 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 presented an efficient algorithm for finding the optimal classifier. We presented an implementation and verified our method in all three models on a variety of UCI datasets.

In our experimental validation we confirmed that the generalization of ROC is stable with the algorithms used and that the obtained improvements agree predictions based on the curve. Analyzing the relationship between abstention window and the cost improvement, we observed that in many cases even small abstentions results in significant costs improvements. This makes it a promising method for real-life applications. Finally, in comparison with cautious classifiers in the bounded models our method proved to perform better in most cases and be more stable.

The models presented in this paper, in particular bounded models, can be easily used in a variety of practical applications, due to their easy interpretation and low complexity of algorithms used. Another desirable property of abstaining classifiers is that reducing the overall number of misclassifications (even in asymmetrical classification problems) and introducing abstentions makes it easier for the human domain experts to review the classification. As argued by Axelsson (1999) with the high number of false positives, the human domain expert tends to generalize an classify everything as "negative". Abstaining classifiers reduce both false positives and false negatives, allowing the human domain expert focus on instances selected for abstention. This makes abstaining classifiers particularly appealing for problems with highly skewed class distributions and misclassification costs such as intrusion detection (Pietraszek 2007).

In this paper we showed how to select an optimal classifier given a scoring classifier $\mathcal{R}$, or in the general case, given two classifiers $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$ selected from a ROC curve with a condition (4) met. This setting can be extended to a case, in which we have a set of scoring classifiers and want to find an abstaining classifier for this set. One way of doing it would be to construct a hybrid ROC for this set of rankers and its ROCCH, however this could violate (4) and led to a sub-optimal selection. Another way is to use extended ROC analysis to create 3D surfaces $fp \times tp \times k$ and discard some parts of ROC surfaces. If the number of potential surfaces to chose from is high, such a strategy could be more efficient.

This, as well as analyzing the performance of the base algorithm for hybrid ROCCHs and applying the algorithm for multi-class classification problems are interesting areas for future research.

Acknowledgements I thank Luc De Raedt for discussions and his comments on this manuscript. Thanks also go to Johannes Fürnkranz and the anonymous reviewers for their useful comments.

References

Axelsson, S. (1999). The base-rate fallacy and its implications for the intrusion detection. In *Proceedings of the 6th ACM conference on computer and communications security* (pp. 1–7). Singapore: Kent Ridge Digital Labs.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*(1), 41–46.

Cohen, I., & Goldszmidt, M. (2004). Properties and benefits of calibrated classifiers. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Lecture notes in computer science: Vol. 3202. Proceedings of PKDD 2004: 8th European conference on principles and practice of knowledge discovery in databases* (pp. 125–136). Berlin: Springer.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Dubuisson, B., & Masson, M. (1996). A statistical decision rules with incomplete knowledge about classes. *Pattern Recognition*, *26*(1), 155–165.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the seventeenth international joint conference on artificial intelligence (IJCAI'01)* (pp. 973–978). Seattle: Kaufmann.

Fawcett, T. (2003). *ROC graphs: notes and practical considerations for researchers (HPL-2003-4)*. Tech. rep., HP Laboratories.

Ferri, C., & Hernández-Orallo, J. (2004). Cautious classifiers. In *Proceedings of ROC analysis in artificial intelligence, 1st international workshop (ROCAI-2004)* (pp. 27–36), Valencia, Spain.

Ferri, C., Flach, P., & Hernández-Orallo, J. (2004). Delegating classifiers. In *Proceedings of 21th international conference on machine leaning (ICML-2004)* (pp. 106–110). Alberta: Omnipress.

Flach, P. A., & Wu, S. (2005). Repairing concavities in ROC curves. In *Proceedings of the 19th international joint conference on artificial intelligence (IJCAI'05)* (pp. 702–707), Edinburgh, Scotland.

Gamberger, D., & Lavrač, N. (2000). Reducing misclassification costs. In *Lecture notes in artificial intelligence: Vol. 1910. Principles of data mining and knowledge discovery, 4th European conference (PKDD 2000)* (pp. 34–43), Lyon, France. Berlin: Springer.

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. In *Proceedings of ICML-94, 11th international conference on machine learning* (pp. 148–156). San Francisco: Kaufmann.

Muzzolini, R., Yang, Y.-H., & Pierson, R. (1998). Classifier design with incomplete knowledge. *Pattern Recognition*, *31*(4), 345–369.

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

Pietraszek, T. (2005). Optimizing abstaining classifiers using ROC analysis. In *Machine learning, proceedings of the twenty-second international conference (ICML 2005)* (pp. 665–672), Bonn, Germany.

Pietraszek, T. (2007). Classification of intrusion detection alerts using abstaining classifiers. *Intelligent Data Analysis Journal*, *11*(3), 293–316.

Provost, F., & Fawcett, T. (2001). Robust classification systems for imprecise environments. *Machine Learning*, *42*(3), 203–231.

R Development Core Team (2004). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-00-3.

Senator, T. E. (2005). Multi-stage classification. In *Proceedings of the 5th IEEE international conference on data mining (ICDM 2005)* (pp. 386–393). Houston: IEEE Computer Society.

Stewart, J. (1992). *Calculus*. Washington: Brooks Cole.

Tortorella, F. (2000). An optimal reject rule for binary classifiers. In *Lecture notes in computer science: Vol. 1876. Advances in pattern recognition, joint IAPR international workshops SSPR 2000 and SPR 2000* (pp. 611–620), Alicante, Spain. Berlin: Springer.

Tortorella, F. (2004). Reducing the classification cost of support vector classifiers through an ROC-based reject rule. *Pattern Analysis Applications*, *7*(2), 128–143.

Witten, I. H., & Frank, E. (2000). *Data mining: practical machine learning tools with java implementations*. San Francisco: Kaufmann.

Wolfram Research Inc. (1999–2006). Lagrange Multiplier—from Wolfram MathWorld. Web page at http://mathworld.wolfram.com/LagrangeMultiplier.html.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the eighteenth international conference on machine learning (ICML-2001)* (pp. 609–616), Williams College, Williamstown, MA. San Mateo: Kaufmann.