Cost Curves for Abstaining Classifiers

Caroline C. Friedel CAROLINE.FRIEDEL@BIO.IFI.LMU.DE
Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstr. 17, 80333 München, Germany

Ulrich Rückert RUECKERT@IN.TUM.DE
Stefan Kramer KRAMER@IN.TUM.DE
Institut für Informatik/I12, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München, Germany

Abstract

We present abstention cost curves, a new three-dimensional visualization technique to illustrate the strengths and weaknesses of abstaining classifiers over a broad range of cost settings. The three-dimensional plot shows the minimum expected cost over all ratios of false-positive costs, false-negative costs and abstention costs. Generalizing Drummond and Holte's cost curves, the technique allows to visualize optimal abstention settings and to compare two classifiers in varying cost scenarios. Abstention cost curves can be used to answer questions different from those addressed by ROC-based analysis. Moreover, it is possible to compute the volume under the abstention cost curve (VACC) as an indicator of the classifier's performance across all cost scenarios. In experiments on UCI datasets we found that learning algorithms exhibit different "patterns of behavior" when it comes to abstention, which is not shown by other common performance measures or visualization techniques.

1. Introduction

In many application areas of machine learning it is not sensible to predict the class for each and every instance, no matter how uncertain the prediction is. Instead, classifiers should have the opportunity to abstain from risky predictions under certain conditions. Our interest in abstaining classifiers is motivated by specific applications, for instance in chemical risk assessment, where it is considered harmful to predict the toxicity or non-toxicity of a chemical compound if the prediction is weak and not backed up by sufficient training material.

Abstaining classifiers can easily be derived from non-abstaining probabilistic or margin-based classifiers by defining appropriate thresholds which determine when to classify and when to refrain from a prediction. The lower and upper thresholds, within which no classifications are made, constitute a so-called abstention window (Ferri et al., 2004). Making use of abstention windows, a recent approach based on ROC analysis (Pietraszek, 2005) derives an optimal abstaining classifier from binary classifiers. In this approach the thresholds can be determined independently of each other from the convex hull of ROC curves. However, ROC-based approaches assume at least known misclassification costs. Moreover, classifiers and optimal abstention thresholds cannot be compared directly for a range of possible cost matrices, as it is usually done in cost curves (Drummond & Holte, 2000).

In this paper, we propose an alternative approach to ROC-based analysis of abstaining classifiers based on cost curves. The advantage of cost curves is that cost-related questions can be answered more directly, and that the performance over a range of cost scenarios can be visualized simultaneously. The proposed generalization of cost curves plots the optimal expected costs (the z-axis) against the ratio of false positive costs to false negative costs (the x-axis) and the ratio of abstention costs to false negative costs (the y-axis). The fundamental assumption is that abstention costs can be related to misclassification costs. As pointed out by other authors (Pietraszek, 2005), unclassified instances might take the time or effort of other classifiers (Ferri et al., 2004), or even human experts. Another scenario is that a new measurement has to be made for

Appearing in *Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning*, Pittsburgh, PA, 2006.

the instance to be classified. Thus, abstention costs link misclassification costs with attribute costs. Consequently, the setting is in a sense related to active learning (Greiner et al., 2002). Along those lines, we also assume that abstention costs are the same independently of the true class: Not knowing the class, the instances are handled in the very same way.

We devised a non-trivial, efficient algorithm for computing the three-dimensional plot in time linear in the examples and in the number of grid points (Friedel, 2005). The algorithm takes advantage of dependencies among optimal abstention windows for different cost scenarios to achieve its efficiency. However, the focus of this paper is not on the algorithm, but on actual abstention cost curves of diverse classifiers on standard UCI datasets. We present abstention cost curves as well as "by-products", showing the abstention rates and the location of the abstention window (the lower and upper interval endpoints). Moreover, a new aggregate measure, the volume under the abstention cost curve (VACC), is presented. VACC is related to the expected abstention costs, if all cost scenarios are equally likely.

2. Abstaining in a Cost-Sensitive Context

Before going into detail, we need to specify some basic concepts and introduce the overall setting. First of all, we assume that a classifier Cl has been induced by some machine learning algorithm. Given an instance x taken from an instance space $\mathcal{X}$, this classifier assigns a class label $y(x)$ taken from the target class $\mathcal{Y} = \{P, N\}$, where P denotes the *positive* class and N the *negative* class. To avoid confusion, we use capital letters for the actual class and lowercase letters for the labels assigned by the classifier. We would now like to analyze this classifier on a validation set $S = \{s_1, s_2, \ldots, s_r\}$ containing r instances with classes $\{y_1, y_2, \ldots, y_r\}$. As argued in the work on ROC curves (e.g. in (Provost & Fawcett, 1998)), it can make sense to use a different sampling bias for the training set than for the validation set. In this case, the class probabilities in the validation set might differ from the class probabilities of the training set or the true class probabilities. Thus, we do not explicitly assume, that the validation set shows the same class distribution as the training set, even though this is the case in many practical applications. However, we demand that the classifier outputs the predicted class label as well as some confidence score for each instance in the validation set. For simplicity we model class label and confidence score as one variable, the *margin*. The margin $m(s)$ of an instance s is positive, if the predicted class is p and negative otherwise. The absolute value of the margin is between zero and one and gives some estimate of the confidence in the prediction. Thus, the margin $m(s)$ of an instance s ranges from -1 (clearly negative) over 0 (equivocal) to +1 (clearly positive).

Applying the classifier to the validation set, yields a sequence of r (not necessarily distinct) margin values $M = (m(s_1), m(s_2), \ldots, m(s_r))$. Sorting this sequence in ascending order yields a characterization of the uncertainty in the predictions. The certain predictions are located at the left and right end of the sequence and the uncertain ones somewhere in between. Based on the information in this sequence one can then allow the classifier Cl to abstain for instances with margin values between a lower threshold l and an upper threshold u. Any such ordered pair of thresholds constitutes an *abstention window* $a := (l, u)$. A specific *abstaining classifier* is defined by an abstention window a and its prediction on an instance x is given as

$$\pi(a, x) = \begin{cases} p & \text{if } m(x) \geq u \\ \perp & \text{if } l < m(x) < u \\ n & \text{if } m(x) \leq l \end{cases} \quad (1)$$

where $\perp$ denotes "don't know".

As both the upper and lower threshold of an abstention window are real numbers, the set of possible abstention windows is uncountably infinite. Therefore, we have to restrict the abstention windows considered in some way. If we are given the margin values as a sorted vector $(m_1, \ldots, m_k)$ of distinct values – i.e., $m_1 < \cdots < m_k$ – it is sensible to choose the thresholds just in between two adjacent margin values. To model this, we define a function $v : \{0, \ldots, k\} \to \mathcal{R}$ which returns the center of the margin with index i and the next margin to the right. We extend the definition of v to the case where $i < 1$ or $i = k$ to allow for abstention windows that are unbounded on the left or on the right:

$$v(i) = \begin{cases} \frac{m_i + m_{i+1}}{2} & \text{if } 1 \leq i < k \\ -\infty & \text{if } i = 0 \\ +\infty & \text{if } i = k. \end{cases} \quad (2)$$

Note that the original margin sequence may contain the same margin value more than once, but v is defined only on the $k \leq n$ distinct margin values. The set of abstention windows $\mathcal{A}(Cl)$ for a classifier Cl is then $\mathcal{A}(Cl) := \{(v(i), v(j)) | 0 \leq i \leq j \leq k\}$. Where the classifier is clear from the context, we omit it and denote the set just by $\mathcal{A}$.

The performance of an abstention window is assessed in terms of expected cost on the validation set. To calculate this, we need information about the costs

associated with each combination of true target class and predicted target class. For our purposes, the costs are given in a *cost matrix* C such that $C(\theta,\pi)$ is the cost of labeling an instance of true class $\theta \in \{P,N\}$ with $\pi \in \{p,n,\perp\}$:

$$C := \begin{pmatrix} C(P,p) & C(P,n) & C(P,\perp) \\ C(N,p) & C(N,n) & C(N,\perp) \end{pmatrix} \quad (3)$$

As the relative frequency on the validation set can be considered as a probability measure, we use conditional probabilities to denote the classification/misclassification rates of an abstention window $a = (l,u)$ on the validation set S. For example, the *false positive rate* of the abstention window a on S is denoted by

$$P_{S,a}(p|N) := \frac{\left|\{s \in S | y(s) = N \wedge \pi(a,s) = p\}\right|}{\left|\{s \in S | y(s) = N\}\right|} \quad (4)$$

Similarly, we have the *true positive rate* $P_{S,a}(p|P)$, the *false negative rate* $P_{S,a}(n|P)$, the *positive abstention rate* $P_{S,a}(\perp|P)$, the *true negative rate* $P_{S,a}(n|N)$, and the *negative abstention rate* $P_{S,a}(\perp|N)$. With this we can calculate the *expected cost* of an abstention window a on S for cost matrix C as the sum of the products of cost and probability over all events:

$$\mathbf{EC}(C,a,S) := \sum_{\theta\in\{N,P\}} \sum_{\pi\in\{n,p,\perp\}} C(\theta,\pi) P_{S,a}(\pi|\theta) P(\theta). \quad (5)$$

In this equation $P(\theta)$ denotes the probability of an example belonging to class $\theta \in \{N,P\}$. In most applications this is just the fraction of positive and negative examples in the validation set. Sometimes, one might want to use other values for those quantities, for example to accommodate for a resampling bias.

For a given cost matrix C, we are primarily interested in the *optimal abstention window* $a_{opt} := \text{argmin}_{a\in\mathcal{A}} \mathbf{EC}(C,a,S)$, that is, the abstention window with the lowest expected cost on the validation set. We observe that the optimal abstention window does not depend on the absolute values of the costs, but only on the relation of the individual costs to each other and the class probabilities $P(P)$ and $P(N)$. For example, multiplying all values in the cost matrix by a constant factor c_m does not change the optimal window. Similarly, adding a constant c_P to the upper row and a constant c_N to the lower row of the cost matrix also has no effect on the optimal abstention window. Let C' denote C with c_P added to the upper row and c_N added to the lower row. Then:

$$\begin{aligned}\mathbf{EC}(C',a) = {} & P(P) \sum_{\pi\in\{n,p,\perp\}} (C(P,\pi) + c_P) P(\pi|P) \\ & + P(N) \sum_{\pi\in\{n,p,\perp\}} (C(N,\pi) + c_N) P(\pi|N) \\ & = \mathbf{EC}(C,a) + P(P)c_P + P(N)c_N\end{aligned}$$

Thus, $\text{argmin}_{a\in\mathcal{A}} \mathbf{EC}(C',a,S) = \text{argmin}_{a\in\mathcal{A}} \mathbf{EC}(C,a,S)$ and the optimal abstention window remains the same. Consequently, we can transform any cost matrix in a normal form C' by adding $c_P = -C(P,p)$ and $c_N = -C(N,n)$ to the upper and lower rows respectively and then multiplying with $c_m = 1/(C(P,n) - C(P,p))$. This "normalization" operation does not change the optimal abstention window, but it ensures that $C'(P,p) = C'(N,n) = 0$ and that $C'(P,n) = 1$. In the following we always assume a normalized cost matrix C' such that the optimal abstention window depends only on the relative false positive costs $C'(N,p)$ and abstention costs $C'(P,\perp)$ and $C'(N,\perp)$:

$$C' := \begin{pmatrix} 0 & 1 & C'(P,\perp) \\ C'(N,p) & 0 & C'(N,\perp) \end{pmatrix} \quad (6)$$

In many applications abstaining on an instance results in additional tests. As the true class of an instance is not known at that point, the cost of such a test is the same for both types of instances, i.e. the cost of abstention is independent of the true class of an instance. In the following we will therefore focus on cases where $C'(P,\perp) = C'(N,\perp) := C'(\perp)$[1]. This means that the optimal window of a given cost matrix in normal form is uniquely determined by just two parameters $\mu := C'(N,p)$ and $\nu := C'(\perp)$. The *normalized expected cost* of an abstention window a can then be written as a function of μ and ν:

$$\begin{aligned}& c(a,\mu,\nu) := \\ & P_{S,a}(n|P)P(P) + \mu P_{S,a}(p|N)P(N) + \nu P_{S,a}(\perp) \quad (7)\end{aligned}$$

In this problem formulation μ represents the false positive costs relative to the false negative costs, while ν controls the abstention costs relative to the false negative costs. As it turns out, abstention does not make sense for all possible settings of μ and ν. For instance, if ν is greater than μ, we can do better by classifying an instance as positive instead of abstaining. The

[1] If this condition is not fulfilled, it is is still possible to compute optimal abstention windows. However, the computational efficiency suffers from more complicated cost settings.

following lemma quantifies this phenomenon. For the sake of simplicity, we use the fractions of positive and negative instances in the validation set for $P(P)$ and $P(N)$. Therefore, we can determine $P_{S,a}(n|P)P(P)$, $P_{S,a}(p|N)P(N)$ and $P_{S,a}(\perp)$ by counting the occurrences of each event and then dividing by the number of instances r.

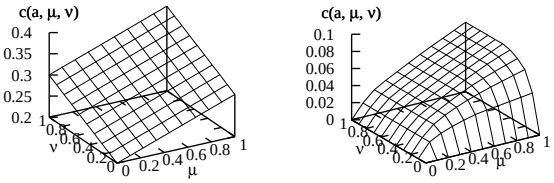
Lemma 1. *Let S, μ and ν be defined as before. If $\nu > \frac{\mu}{1+\mu}$, the optimal abstention window a_{opt} is empty, i.e. $l_{opt} = u_{opt}$ (proof omitted).*

3. Cost Curves for Abstaining Classifiers

If the cost matrix and the class probabilities in a learning setting are known exactly, one can determine the optimal abstention window a_{opt} simply by calculating the expected costs for all windows. However, for most applications costs and class distributions are uncertain and cannot be determined exactly. In such a setting one would like to assess the performance of an abstaining classifier for a broad range of cost settings. Even in the case of non-abstaining classifiers one might want to illustrate a classifier's behavior for varying cost matrices or class distributions. The two most prominent visualization techniques to do so are ROC curves (Provost & Fawcett, 1998) and cost curves (Drummond & Holte, 2000). In the following we present a novel method that allows to visualize the performance of abstaining classifiers. In principle, one could extend ROC curves or cost curves with a third dimension to accomodate for abstention. However, the meaning of the new axis in such an "extended" cost curve is not very intuitive, making it rather hard to interpret. Since visualization tools rely on easy interpretability, we follow a different approach[2].

The presented cost curve simply plots the normalized expected cost as given in equation (7). It is created by setting the x-axis to μ, the y-axis to ν and the z-axis to the normalized expected cost. Without loss of generality, we assume that the positive class is always the one with highest misclassification costs, so that $\mu \leq 1$ (if this is not the case, just flip the class labels). Furthermore, we can safely assume that $\nu \leq 1$, because otherwise the optimal abstention window is empty (as stated by lemma 1).

[2]Technically, the presented cost curve assumes a fixed class distribution to allow for easier interpretation. We feel that the gain in interpretability outweighs the need for this additional assumption. In some settings cost curves that extend (Drummond & Holte, 2000) might be more suited; see (Friedel, 2005, section 3.4) for an elaborate comparison with the cost curves presented in this paper.



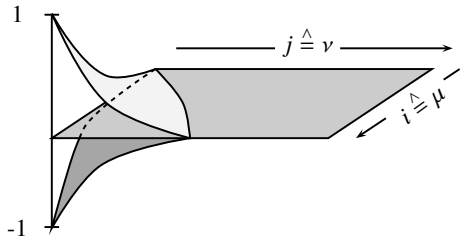
(a) (b)

Figure 1: Example cost curves for uncertain costs, but fixed class distributions. (a) shows a cost curve for a specific abstention window, (b) a cost curve for an example classifier.

We can apply the cost curves in two ways. In the first case, we plot the normalized expected cost against the false positive and abstention costs for one fixed abstention window a. Then the resulting cost curve is just a plane, because $z = c(a, x, y)$ is linear in its parameters (see Figure 1(a)). This illustrates the performance of a classifier for one particular abstention window. In the second case, the cost curve is the lower envelope of all abstention windows, i.e. $z = \min_{a \in \mathcal{A}(Cl)} c(a, x, y)$ (see Figure 1(b)). This scenario is well suited for comparing two classifiers independently of the choice of a particular abstention window. For easier analysis, the curves can be reduced to two dimensions by color coding of the expected cost (see Section 4).

Using the information from cost curves, several questions can be addressed. First, we can determine for which cost scenarios one abstaining classifier Cl_s outperforms another classifier Cl_t. This can be done by examining a so-called differential cost curve $D(s,t)$, which is defined by $d_{i,j}(s,t) := k_{i,j}(s) - k_{i,j}(t)$. $d_{i,j}(s,t)$ is negative for cost scenarios for which Cl_s outperforms Cl_t and positive otherwise. Obviously, we can also compare a non-trivial classifier with a trivial one, which either always abstains or always predicts one of the two classes. Second, we can determine which abstention window should be chosen for certain cost scenarios by plotting the lower and the upper threshold of the optimal window for each cost scenario. Third, we can plot the abstention rate instead of expected costs in order to determine where abstaining is of help at all.

Although cost curves are continuous in theory, the visualization on a computer is generally done by calculating the z-values for a grid of specific values of x and y. The number of values chosen for x and y determines the resolution of the grid and is denoted as Δ. For computational considerations, we can thus define a cost curve for a classifier Cl as a $\Delta \times \Delta$ matrix $K(p)$ with $k_{i,j}(p) := min_{a \in \mathcal{A}(Cl)} c(a, i/\Delta, j/\Delta)$



(a)

Figure 2: Schematic illustration of optimal abstention windows (upper threshold above the plane, lower threshold below) for various μ and ν. For the same μ and $\nu_1 < \nu_2$, the optimal window for ν_2 is contained in the window for ν_1.

for $0 \leq i, j \leq \Delta$. Calculating such a cost curve for moderately high values of Δ can be computationally demanding, as we have to determine the optimal abstention window for a large number of cost settings.

A naive algorithm would compute the cost curve by calculating the expected cost for each possible abstention window for each cost scenario. As the number of abstention windows is quadratic in the number of instances, this results in an algorithm in $O(\Delta^2 n^2)$. Our more efficient algorithm (Friedel, 2005) for computing cost curves largely relies on two observations:

1. The optimal abstention window a_{opt} can be computed in linear time by first determining the optimal threshold for zero abstention for the respective μ, and then finding the best abstention window located around this threshold.

2. for fixed μ and $\nu_1 < \nu_2$, the optimal abstention window for ν_2 is contained in the optimal abstention window for ν_1.

Thus, the optimal thresholds and abstention windows are arranged as illustrated by the schematic drawing in Figure 2: The plane in the center gives the optimal threshold between positive and negative classification; above we have the upper threshold of the optimal abstention window, and below the lower threshold. Based on these observations, it possible to design an efficient algorithm linear in the number of examples: In the first step, the optimal thresholds for non-abstention and the various values of μ are computed. Subsequently, the precise upper and lower thresholds around the optimal threshold found in the first step are determined.

4. Experiments

To analyze and visualize the abstention costs, we chose six two-class problems from the UCI repository: breast-w, bupa, credit-a, diabetes, haberman and vote. Five different machine learning algorithms, as implemented in the WEKA workbench (Witten & Frank, 2005), were applied to those datasets: J48, Naive Bayes (NB), PART, Random Forests (RF) and Support Vector Machines (SVM).

Alg.	Acc. (%)	AUC	VACC	Nrm. Acc.	Nrm. AUC	Nrm. VACC
			breast-w			
J48	95	0.96	0.032	0.98	0.96	1.00
NB	96	0.98	0.018	0.99	0.99	0.58
PART	95	0.97	0.030	0.98	0.98	0.95
RF	95	0.99	0.016	0.98	0.99	0.50
SVM	97	0.99	0.014	1.00	1.00	0.44
			bupa			
J48	65	0.67	0.16	0.97	0.90	0.93
NB	55	0.64	0.18	0.82	0.87	1.00
PART	62	0.67	0.17	0.93	0.91	0.97
RF	67	0.74	0.15	1.00	1.00	0.84
SVM	64	0.70	0.17	0.95	0.95	0.96
			credit-a			
J48	87	0.89	0.082	1.00	0.97	0.88
NB	78	0.90	0.093	0.90	0.98	1.00
PART	85	0.89	0.089	0.98	0.98	0.95
RF	85	0.91	0.088	0.99	1.00	0.94
SVM	85	0.86	0.081	0.98	0.95	0.87
			diabetes			
J48	73	0.75	0.15	0.96	0.90	1.00
NB	76	0.82	0.14	0.99	0.98	0.92
PART	74	0.79	0.14	0.96	0.95	0.94
RF	75	0.78	0.15	0.98	0.94	0.98
SVM	76	0.83	0.13	1.00	1.00	0.87
			haberman			
J48	69	0.61	0.12	0.93	0.87	1.00
NB	75	0.65	0.11	1.00	0.93	0.95
PART	71	0.59	0.11	0.96	0.84	0.95
RF	68	0.65	0.12	0.91	0.93	1.00
SVM	74	0.70	0.11	1.00	1.00	0.96
			vote			
J48	97	0.97	0.021	1.00	0.98	0.44
NB	90	0.97	0.046	0.93	0.98	1.00
PART	97	0.95	0.022	1.00	0.96	0.48
RF	96	0.98	0.021	1.00	0.99	0.44
SVM	96	0.99	0.022	0.99	1.00	0.47

Table 1: Summary of quantitative results of five learning algorithms applied to six UCI datasets

Our starting point is a summary of all quantitative results from ten-fold cross-validation on the datasets (see Table 1).[3] In the table, the predictive accuracy, the area under the (ROC) curve (AUC) and the volume under the abstention cost curve (VACC) are shown. The volume under the abstention cost curve can be

[3] In the experiments, we assume that the class distribution observed in the data resembles the true class distribution. Experiments assuming a uniform distribution (50:50) changed the absolute VACC numbers, but not their ordering.

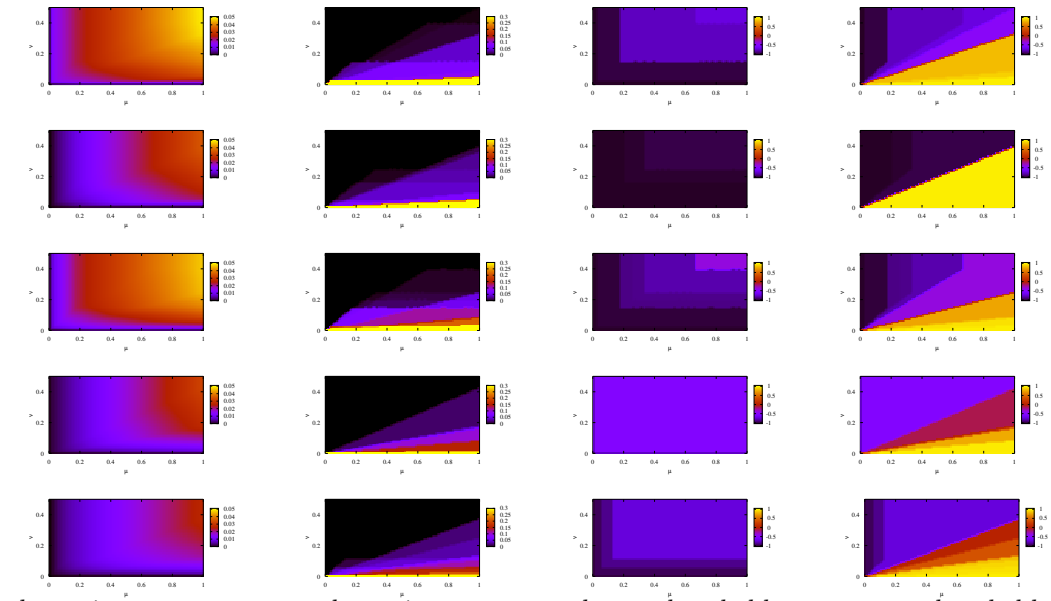
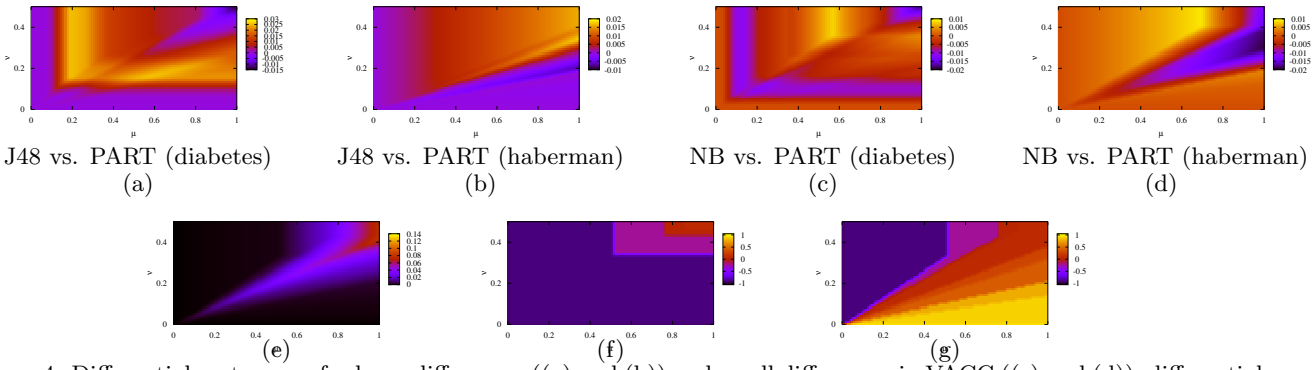


Figure 3: Abstention cost curves, optimal abstention rates and lower/upper thresholds of the optimal abstention window on the breast-w dataset. From top to bottom: J48, NB, PART, RF, and SVM (as in Table 1)

defined as the double integral over μ and ν. VACC is related to the expected value of the abstention costs if all cost scenarios are equally likely. Moreover, the normalized values of those measures are given, that is, the value of the measure divided by the maximum over the classifiers' performance for the particular dataset. The normalized values are given to facilitate an easier comparison between the measures.

Overall, one can see that VACC in fact captures a different aspect than accuracy or AUC. In the following, we discuss the quantitative results from the table one by one. On breast-w, the VACC measure indicates significant differences in terms of abstention costs, which is neither reflected in predictive accuracy nor in AUC. For instance, we can see that there is a clear order over the classifiers from different learning algorithms: SVMs perform best, followed by RF and NB, whereas PART and J48 lag behind. This is also illustrated by the plots in Figure 3, which are discussed below. On the bupa dataset, NB performs worst and RF performs best according to all measures. However, the differences are not equally visible in all measures (see, e.g., RF vs. SVM or, vice versa, NB vs. PART). On credit-a, the comparison between J48 and NB hints at a marked difference in accuracy and VACC, not shown by AUC. PART vs. SVM is a different case: Comparable values for accuracy and AUC, but a considerable gap in VACC. For the diabetes data, a distinct difference is detected for RF vs. SVM in AUC/VACC, but not in terms of accuracy. On the haberman dataset, the variation in the quantitative results is negligible (for details, see below). Finally, the results on vote reveal that NB performs dramatically worse than all other approaches, perhaps due to the violated independence assumption on this particular dataset. This drop in performance is particularly visible in the VACC results.

Next, we have a closer look at the abstention cost curves and derived plots for all five learning algorithms on the breast cancer data (see Figure 3). In the leftmost column, the optimal abstention costs over all cost scenarios are visualized. Note that all plots are cut at $\nu = 0.5$, because for greater values of abstention costs, the abstention window is already degenerate, with $l = u$. The plots reflect the numbers from Table 1 adequately, but additionally show in which regions of the space the majority of costs occur. The second column from the left visualizes the abstention rate, that is, the fraction of instances the classifiers leaves unclassified. For instance, we can infer that PART should refrain from 10% to 15% of the predictions if the abstention costs are about one tenth of the false negative costs. The two right-most colums visualize the lower and the upper interval endpoints of the abstention window. To enable a visual comparison, all curves are plotted on the same scale. Considerable dif-



J48 vs. PART (diabetes) (a) J48 vs. PART (haberman) (b) NB vs. PART (diabetes) (c) NB vs. PART (haberman) (d)

(e) (f) (g)

Figure 4: Differential cost curves for large differences ((a) and (b)) and small differences in VACC ((c) and (d)), differential cost curve SVMs vs. trivial classifier(s) on bupa (e), lower (f) and upper (g) thresholds of abstention window

ferences in the classifiers' abstention behavior become apparent.

In the plots, the isolines of l and u have a remarkably different shape. This can be explained as follows: First, both the upper and lower thresholds increase not continuously with ν or μ, but in steps. This is due to the fact that a critical value has to be reached for the cost of abstaining or classifying the instances between different threshold values, before thresholds are adjusted. Second, we observe that for values of ν for which abstaining is too expensive, the upper and the lower threshold are equal, as shown before.

The threshold shows a different behaviour only for those values of ν and μ that allow abstaining. In this range, the lower threshold depends only on the ratio between false negative costs (which are constant) and abstaining costs, and is thus independent of the false positive costs. The upper threshold on the other hand depends on both the abstaining costs ν and the false positive costs μ. In the same way as the lower threshold is effectively not affected by changes in μ in the range for which abstaining is reasonable, the upper threshold is not affected by changes in the false negative costs, which can easily be confirmed by switching the positive and negative labels.

Next, we take a look at *differential cost curves*. Differential cost curves are a tool for the practitioner to see in which regions of the cost space one classifier is to be preferred over another. In Figure 4, differential cost curves with large differences in VACC (upper row, (a) and (b)) and small differences in VACC (upper row, (c) and (d)) are shown. In Figure 4(a) and (b), J48 decision trees have smaller abstention costs than PART rules only in the bluish areas of the space. Differential cost curves also shed light on differences that do not appear in VACC, if a classifier is dominating in one region as it is dominated in another (Figure 4 (c) and (d)). The regions can be separated and quite distant in cost space, as illustrated by Figure 4 (c). The differential cost curve of NB vs. PART on haberman (Figure 4 (d)) demonstrates that even for datasets with no clear tendencies in accuracy, AUC or VACC, the plot over the cost space clearly identifies different regions of preference not shown otherwise.

Another interesting possibility is the comparison with the trivial classifier that always predicts positive, negative, or always abstains. In Figure 4 (e), we compare SVMs with trivial classifiers on the bupa dataset. In the black areas near the left upper and the right lower corner, the trivial classifer performs better than the SVM classifier. To explain this, we take a look at the lower and upper thresholds of the abstention window in Figure 4 (f) and (g). Strikingly, we find that in the upper left part $l = u = -1$, that is, everything is classified as positive, because false positives are very inexpensive compared to false negatives. However, in the lower right part $l = -1$ and $u = 1$, i.e., not a single prediction is made there, because abstention is inexpensive.

It is clear that the discussion of the above results remains largely on a descriptive level. However, ideally we would like to explain or even better, predict the behavior of classifiers on particular datasets. Unfortunately, this is hardly ever achieved in practice: In the majority of cases it is not possible to explain the error rate or AUC for a particular machine learning algorithm on a particular dataset at the current state of the art. To learn more about the behavior of the abstention cost curve and the VACC measure, we performed preliminary experiments with J48 trees, varying the confidence level for pruning, and SVMs, varying the penalty/regularization parameter C. Over all datasets, we observed only small, gradual shifts in VACC and in the shape of the curves. While it is hard to detect a general pattern, it is clear that no

abrupt changes occur. It was also striking to see that the changes over varying parameter values were consistent for both learning schemes. It seems that the VACC depends, to some extent, on the noise level of a dataset.

5. Related Work

The trade-off between coverage and accuracy has been addressed several times before, such as in articles by (Chow, 1970), who described an optimum rejection rule based on the Bayes optimal classifier, or (Pazzani et al., 1994), who showed that a number of machine learning algorithms can be modified to increase accuracy at the expense of abstention. Tortorella (Tortorella, 2005) and Pietraszek (Pietraszek, 2005) use ROC analysis to derive an optimal abstaining classifier from binary classifiers. Pietraszek extends the cost-based framework of Tortorella, for which a simple analytical solution can be derived, and proposes two models in which either the abstention rate or the error rate is bounded in order to deal with unknown abstention costs. Nevertheless, all of these ROC-based approaches assume at least known misclassification costs. In contrast, abstention cost curves, as shown in this paper, visualize optimal costs over a range of possible cost matrices. Ferri and Hernández-Orallo (Ferri & Hernández-Orallo, 2004) introduce additional measures of performance for, as they call it, cautious classifiers, based on the confusion matrix. Our definition of an abstention window can be considered as a special case of Ferri and Hernández-Orallo's model for the two-class case. However, no optimization is performed when creating cautious classifiers and only the trade-off between abstention rate and other performance measures such as accuracy is analyzed. Cautious classifiers can be combined in a nested cascade to create so-called delegating classifiers (Ferri et al., 2004). Cost-sensitive active classifiers (Greiner et al., 2002) are related to abstaining classifiers as they are allowed to demand values of yet unspecified attributes, before committing themselves to a class label based on costs of misclassifications and additional tests.

6. Conclusion

In this paper, we adopted a cost-based framework to analyze and visualize classifier performance when refraining from prediction is allowed. We presented a novel type of cost curves that makes it possible to compare classifiers as well as to determine the cost scenarios which favor abstention if costs are uncertain or the benefits of abstaining are unclear. In comprehensive experiments, we showed that adding abstention as another dimension, the performance of classifiers varies highly depending on datasets and costs. Viewing the optimal abstention behavior of various classifiers, we are entering largely unexplored territory. We performed preliminary experiments to shed some light on the dependency of VACC on other quantities, such as the noise level in a dataset. However, more work remains to be done to interpret the phenomena shown by the curves. Finally, we would like to note that another, more qualitative look at abstention is possible. In particular on structured data, refraining from classification is advisable if the instance to be classified is not like any other instance from the training set.

References

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

Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost: An alternative to ROC representation. *Proc. of the 6^{th} International Conf. on Knowledge Discovery and Data Mining* (pp. 198–207).

Ferri, C., Flach, P., & Hernández-Orallo, J. (2004). Delegating classifiers. *Proc. of the 21^{st} International Conf. on Machine Learning*.

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

Friedel, C. C. (2005). On abstaining classifiers. Master's thesis, Ludwig-Maximilians-Universität, Technische Universität München.

Greiner, R., Grove, A. J., & Roth, D. (2002). Learning cost-sensitive active classifiers. *Artificial Intelligence*, *139*, 137–174.

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

Pietraszek, T. (2005). Optimizing abstaining classifiers using ROC analysis. *Proceedings of the 22^{nd} International Conference on Machine Learning*.

Provost, F. J., & Fawcett, T. (1998). Robust classification systems for imprecise environments. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 706–713).

Tortorella, F. (2005). A ROC-based reject rule for dichotomizers. *Pattern Recognition Letters*, *26*, 167–180.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools with java implementations*. Morgan Kaufmann, San Francisco.