Cautious Classifiers

Cèsar Ferri[1] and **José Hernández-Orallo**[1]

Abstract. The evaluation and use of classifiers is based on the idea that a classifier is defined as a complete function from instances to classes. Even when probabilistic classifiers are used, these are ultimately converted into categorical classifiers that must choose one class (with more or less confidence) from a set of classes. Evaluation metrics such as accuracy/error, global cost, precision, recall, f-score, specificity, sensitivity, effectiveness, macro-average, logloss, MSE or the Area Under the ROC Curve (AUC) are usually defined for "complete" classifiers. In this paper we pursue the usefulness and evaluation of "cautious" or "partial" classifiers. A cautious classifier adds an extra class "unknown" to the set of the original classes. This "unknown" class represents the cases where the prediction is uncertain or not reliable. Now, in a cost-insensitive context, accuracy and error will not be *directly* related but indirectly, through the coverage index. We develop new measures, efficacy and capacity, which find a compromise between reducing the number of misclassified data (error) and reducing the number of unclassified data (abstention). Inspired by ROC analysis we introduce several techniques to choose from a set of cautious classifiers. For probabilistic classifiers we define a discretisation method for converting them into cautious classifiers by using a "caution window". We develop new response graphs to show the way in which different classifiers behave according to the size of the window and the class bias. In a cost-sensitive context, cost matrices and confusion matrices can be directly extended to account for this new class. Moreover, we extend ROC analysis and AUC evaluations to these classifiers, by considering the degree of abstention as an additional dimension.

Keywords: Uncertainty in decision making, probabilistic classifiers, cost-sensitive learning, ROC analysis, model representation, misclassification cost.

1 INTRODUCTION

Decision making can be supported by the predictions of a classifier. A wrong decision has usually negative consequences. Hence, we would like to use classifiers that rarely make a mistake. In a dualistic framework, this means the same to look for classifiers that are almost always correct.

However, when we use human assistance for supporting decision making, there are some cases where the expert says "I don't know" and asks for further assistance (to other experts) or just prefers to postpone the decision. Frequently, we say a person is an expert or a wise person when she prefers to be silent (and ask other experts) rather than to make a mistake. In this three-valued framework, we consider someone an expert when she is frequently correct, she is rarely wrong and sometimes she says "I don't know".

There are some application areas where this definition of expert is imperative. Medical decision making, especially diagnosis, always prefers a cautious expert that sometimes says "I'm not sure of the diagnosis. More tests are needed" rather than a reckless expert that makes a wrong diagnosis.

Although there are areas where the "unknown" prediction is excluded, such as law (in case of doubt the result should be not guilty), there are some situations where it would be convenient to work with a three-valued logic for decisions. Obviously, a three-valued logic is not new in decision theory, but it has been relatively unexplored in other areas, the learning of classifiers in particular.

Pursuing this idea, in this work we present the notion of "cautious classifier". A cautious classifier will make predictions by selecting among the possible problem classes and an additional class "unknown". This kind of classifiers open up a series of issues such as how to evaluate them taking into account the degree of errors vs. the degree of abstention, how to find optimal compromises, how cost-sensitive learning can be extended to consider this kind of classifiers, and many other concerns.

A very important issue is the relationship between probabilistic classifiers (classifiers that assign probabilities of membership rather than assigning a single class) and cautious classifiers. In some way, it can be argued that it is always better to have a probabilistic classifier than a cautious classifier, because if we have the former we can derive the latter. In fact it is so, but usually we must end up with the latter, so it is worth considering how to convert from probabilistic classifiers to cautious classifiers. A great part of this work is devoted to the analysis of how to perform this conversion.

Finally, there are sets of association rules that are indeed partial classifiers and can be considered as cautious classifiers, but they are rarely treated as such because this is deemed to be negative and are usually completed in some way. In the same way, separate-and-conquer algorithms [5] work with partial rules. Classifiers are frequently completed by the use of a default rule.

The paper is organised as follows. The following section introduces the notion of cautious classifiers and two basic measures: efficacy and capacity. Section 3 discusses the comparison and evaluation of several cautious classifiers, using similar techniques as ROC analysis. Section 4 deals with probabilistic classifiers and how they can be converted into cautious classifiers. This can be done through the concept of confidence threshold (below which the classifier abstains), a stratified threshold or the use of a class bias and a window size. Section 5 introduces the notion of cautious cost matrix, where costs are also defined for abstentions. Section 6 loops the loop by discussing the use of ROC analysis and AUC measures for cautious classifiers. Section 7 presents some simple examples to better understand the applicability and behaviour of some of the presented measures. Section 8 closes the paper with an overall discussion of the presented ideas and the future work.

[1] Universitat Politècnica de València, Spain. Email: {cferri, jorallo}@dsic.upv.es. This work has been supported by: ICT for EU-India Cross Cultural Dissemination Project ALA/95/23/2003/077-054, Generalitat Valenciana under grant GRUPOS03/25 and MetaMidas, Acción Integrada HU 2003-0003, and STREAM TIC 2001-2705-C03-01.

2 DEFINITION AND EVALUATION OF SINGLE CAUTIOUS CLASSIFIERS

Let us define I as the set of possible instances (unlabelled examples) of a given problem. Given a set of classes C, we define $C' = C \cup \{\perp\}$. While a traditional classifier is defined as a function $I \rightarrow C$, a cautious classifier is defined as a function $I \rightarrow C'$. It is important to note that the element $\perp$ does not exist in the target function or the dataset. Both target function and datasets are defined as functions $I \rightarrow C$.

2.1 Extended Confusion Matrix. Measures Derived

A Confusion Matrix of the predictions of a classifier is a very practical and intuitive way of seeing the distribution between actual classes and predictions. In our case, the confusion matrix is defined as a function $C' \times C \rightarrow Nat$. We use the notation $M(i,j)$ for referring to the cardinality of elements with predicted class i and actual class j.

Example 1:

For instance, given 100 test examples and a classifier P, an example of a traditional Confusion Matrix for three classes {a, b, c} might be as follows:

		Actual a	Actual b	Actual c
Predicted	a	20	2	3
	b	0	30	3
	c	0	2	40

This matrix is understood as follows. From the hundred examples, 20 were of class 'a' and all were correctly classified, 34 were of class 'b' from which 30 were correctly classified as 'b', 2 misclassified as 'a' and 2 misclassified as 'c'. Finally, 46 were of class 'c' from which 40 were correctly classified as 'c', 3 misclassified as 'a' and 3 misclassified as 'b'.

However, a cautious classifier R gives a different portrait:

		Actual A	Actual b	Actual c
Predicted	a	19	1	2
	b	0	30	0
	c	0	1	38
	$\perp$	1	2	6

We use the symbol $\perp$ for abstention. The last row denoted by $\perp$ means that it has abstained for one example of class 'a', two of class 'b' and six of class 'c'. This new row forces traditional measures to be rethought. From here we can re-define several measures:

$$Card = \sum_{i \in C', j \in C} M(i,j)$$

$$Coverage = \frac{\sum_{i \in C, j \in C} M(i,j)}{Card}$$

$$Abstention = \frac{\sum_{j \in C} M(\perp,j)}{Card}$$

$$Accuracy = \frac{\sum_{i \in C, j \in C, i=j} M(i,j)}{\sum_{i \in C, j \in C} M(i,j)}$$

$$Error = \frac{\sum_{i \in C, j \in C, i \neq j} M(i,j)}{Card}$$

For example 1, we have Card=100, Coverage= 0.91, Abstention= 0.09, Accuracy= 0.956, Error= 0.04.

Of course, the way in which we have defined accuracy and error may seem arbitrary. Accuracy and coverage are similar to precision and recall in information retrieval. Note that accuracy (unlike error) is only computed wrt. the cases where the classifier does not predict $\perp$. Moreover, *Accuracy* is normalised wrt. the known values whereas *Error* is normalised wrt. all the values. Consequently, we have:

$$Accuracy \times Coverage = Coverage - Error$$

And we also have, logically, that:

$$Coverage + Abstention = 1$$

We must also take into account that *Accuracy* is undefined if $Coverage = 0$.

Only in the case we have a traditional (non-cautious classifier), then we have the usual expression $Accuracy = 1 - Error$.

2.2 Evaluation of a Single Cautious Classifier. Efficacy and Capacity.

All the previous definitions seem fairly simple. Another straightforward consequence is that a good classifier will have to attain high accuracy with low abstention.

We can represent *Accuracy* vs. *Abstention* in a bidimensional graph. For instance, the previous cautious classifier R has Abstention = 0.09 and Accuracy= 0.956 and can be represented as follows:

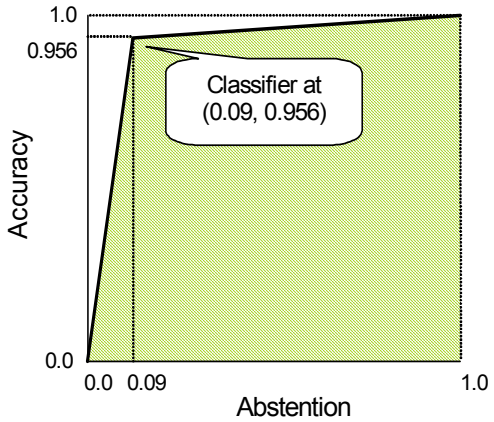

Fig. 1. Accuracy vs Abstention

It may also seem arbitrary but quite practical to consider the area formed by the segments (0,0) and (1,1) with the point given by the classifier à la ROC analysis. We call this area *Efficacy*. It aims to suggest that greater abstentions will get closer to 1 accuracy (except in the limit). However, 0 abstention usually does not give 0 accuracy, so the left bottom part of the graph is not realistic.

However, the *efficacy* if a single point can be obtained very easily as:

$$Efficacy = \frac{Accuracy - Abstention + 1}{2}$$

In this case, *Efficacy* = 0.933. It is straightforward that:

$$Efficacy = \frac{Accuracy + Coverage}{2}$$

Consequently, *Efficacy* is just the arithmetic mean between *Accuracy* and *Coverage*. Logically, if *Coverage*=1, Efficacy goes from 0.5 to 1.

It is also straightforward to see that if a cautious classifier *P* is equal to a cautious classifier *Q* except for an instance that is an error for *P* but an unknown for *Q*, then *Q* has less or equal efficacy than *P* (this is so because the denominator in Accuracy is always smaller than for Coverage).

That means that, according to efficacy, it never makes sense to convert an error to an unknown, something that is counterintuitive. Nonetheless, the efficacy measure will be still useful for probabilistic classifiers, as we will see.

An alternative could be to use some other score, such as the f-score, typical in information retrieval, which is the harmonic mean of precision and recall. In our case, we could compute:

$$\text{f-score} = \frac{2\,Accuracy \times Coverage}{Accuracy + Coverage}$$

This measure gives lower values when higher accuracy is obtained by lowering coverage.

A different alternative way (and simpler than f-score) to represent the quality of a classifier is to consider *Error* vs. *Abstention*. In this case we can consider that, given a cautious classifier *Q*, we can move some classified instances to unclassified instances and viceversa, by using random guesses, constructing a parametrised cautious classifier Q_α, by using the following formula, where *i* represents an instance.

$Q_\alpha(i)$:
- if $\alpha >$ *Abstention* then choose $Q_\alpha(i)$ between $\perp$ and $Q(i)$ with probability (α – *Abstention*) / (1 – *Abstention*).
- if $\alpha =$ *Abstention* then choose $Q_\alpha(i) = Q(i)$.
- if $\alpha <$ *Abstention* then
  - if $Q(i) \neq \perp$ then $Q_\alpha(i) = Q(i)$.
  - if $Q(i) = \perp$ then choose between a random class and $\perp$ with probability (*Abstention* – α) / *Abstention*.

Note that with this procedure the confusion matrix of $Q_\alpha(i)$ will be a function $C' \times C \rightarrow Real$ instead of $C' \times C \rightarrow Nat$.

The first condition represents when we augment the degree of abstention from the current abstention to a point where everything is abstention. The third condition has two cases. The first one, when the original classifier does not abstain, just uses the given prediction. The second case represents a random guess made when we want the classifier to have lower abstention. This random class can be chosen using a uniform distribution or the prior distribution.

For instance, we can compute the confusion matrix for $Q_{0.25}(i)$ with respect to the previous example 1. Since α is greater than abstention (0.09) we apply the first rule, having the following probability:

Prob = (0.25 – 0.09) / (1 – 0.09) = 0.1758

and the following matrix:

Predicted \ Actual	a	b	c
a	15.66	0.82	1.65
b	0	24.73	0
c	0	0.82	31.32
$\perp$	4.34	7.63	13.03

Similarly, the confusion matrix for $Q_{0.06}(i)$ is computed by using the third rule with:

Prob = (0.09 – 0.06) / (0.09) = 0.3333

Predicted \ Actual	a	b	c
a	19.33	1.22	2.67
b	0	30.22	0.67
c	0	1.22	38.67
$\perp$	0.67	1.33	4

Varying the parameter α we can represent a capacity graph for a single classifier:

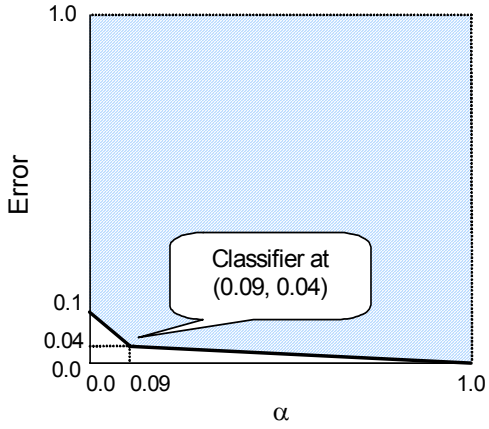

Fig. 2. Capacity graph of a classifier

The first part of the graph ([0.0, 0.09] abstention and [0.1, 0.04] error) begins at a full classifier. This full classifier is constructed by converting all the undefined values of the original cautious classifier into defined values by guessing (in this case with a uniform distribution) among the three classes. Since there are 9 unknown cases (1 of actual class *a*, 2 of actual class *b* and 6 of actual class *c*), the error will be increased by (($|C| - 1$) / $|C|$) × *Abstention*. In this case, 0.06. Since the error was originally 0.04 then with α= 0 the error would be 0.1. Note that (($|C| - 1$) / $|C|$) is the slope of the first part. The greater the number of classes the steeper this first part will be. Finally, the second part of the graph is much simpler and just ends up with no errors at *Abstention* = 1.

The first part of the graph can be slightly improved if we know the original class distribution. For instance if we know that p_a= 0.2, p_b= 0.34 and p_c= 0.46, we would have that the error at α = 0 would be increased by (1 – 0.2) × 1 + (1 – 0.34) × 2 + (1 – 0.46) × 6 = 5.36 instances. Consequently the whole error at α = 0 would be 0.0936 instead of 0.1.

Once again, it seems natural to consider the area above the function formed by Error with varying α, as a measure of quality. This area is known as *Capacity*.

$$Capacity = 1 - \left[\frac{Error \times (1 + Abstention)}{2} + \frac{\left(\frac{|C|-1}{|C|} \right) \times Abstention}{2} \right]$$

When Abstention = 0, Capacity = 1 – Error/2 = Efficacy.

3 COMPARING AND EVALUATING SETS OF CAUTIOUS CLASSIFIERS

Given a set of cautious classifiers, the first idea to evaluate them may be to compute the previous measures and choose the one with biggest efficacy or capacity. However, this choice could possibly rule out a classifier that could be useful in a context where it is preferable a greater coverage and greater error or vice versa.

The rationale is quite similar to what is argued for ROC analysis, to which we come back later on in this work.

Let us show several cautious classifiers in a capacity graph:

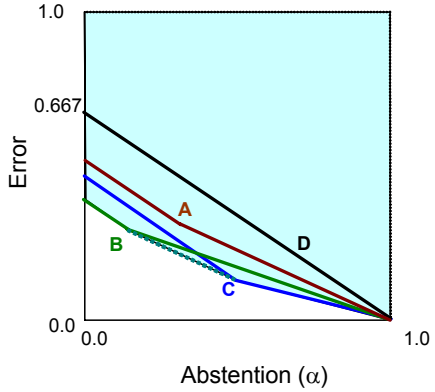

Fig. 3. Capacity graph of a set of classifiers

In the previous picture *D* represents the "no information" classifier for a 3-class problem with uniform class distribution. Nothing should be worse than this and can be considered as a baseline.

B and *C* are quite different classifiers. *C* attains a low error by sacrificing coverage. On the contrary, *B* has a greater error but a very high coverage. *A* is a classifier with an intermediate abstention degree but greater error. The rationale is similar to ROC analysis. The third classifier *A* can be ruled out safely, because it is above both classifiers.

Even more, we can connect *B* and *C* representing a fictitious (but constructible) classifier that proportionally mixes the predictions and abstentions of both classifiers.

Consequently, in this graph we would compute the convex hull of the existing classifiers (not only taking into account the points but the tails to the left of each classifier).

All this makes sense if we originally have a set of cautious classifiers. It is clear that if we have a set of categorical classical classifiers, the one with the lowest error will be optimal for the whole graph. This fact changes when we consider probabilistic classifiers.

4 FROM PROBABILISTIC CLASSIFIERS TO CAUTIOUS CLASSIFIERS

While a traditional classifier is defined as a function $I \rightarrow C$, a probabilistic classifier is defined as a function $I \rightarrow P^{|C|}$ where P is the interval [0..1] of real numbers. Consequently, a probabilistic classifier assigns to each instance a set p_i of $|C|$ probabilities, each of them corresponding to the probability of predicting class i. For this reason, a probabilistic classifier is also called a class probability estimator.

Usually, an additional condition is required or met by normalisation: $\Sigma p_i = 1$. Consequently, one of the probabilities can be obtained from the rest. In the special case of being just two classes, it is just necessary to tell one of them, and frequently it is much easier to establish thresholds for separating between both classes. Nonetheless, this is also possible for more than two classes, as we will see.

Example 2

Consider the following probability estimation tree (PET) with 7 leaves and the test results of each leaf with a test dataset.

	Estimated Probs		Test Cases	
#Leaf	p_a	p_b	Class *a*	Class *b*
1	1	0	23	1
2	0.75	0.25	10	0
3	0.7	0.3	4	2
4	0.6	0.4	0	9
5	0.35	0.65	2	3
6	0.2	0.8	1	15
7	0.1	0.9	0	30
Total :	-	-	40	60

A probability estimator, such as this PET, can be converted into a categorical classifier by establishing a way of assigning the classes according to the probability. For instance, the easiest way to do such a conversion is to assign the class with the greatest probability. In this case, nodes 1, 2, 3 and 4 would predict class *a* and 5, 6 and 7 would predict class *b*. Consequently, we would have the following confusion matrix of the classifier obtained from the previous PET:

		Actual	
		a	b
Predicted	a	37	12
	b	3	48

With Accuracy= 0.85, Error= 0.15, Coverage= 1, Efficacy= 0.925, f-score= 0.919, Capacity= 0.925.

But we could also use a different criterion and convert it into a *cautious classifier* using a confidence threshold. For instance, for 0.625 it would be:

Confidence = max (p_a, p_b)
If Confidence > 0.625 then class = argmax$_i$ (p_i)
else class = ⊥

According to this rule, we have a different confusion matrix:

		Actual a	Actual b
Predicted	a	37	3
	b	3	48
	⊥	0	9

With Accuracy= 0.934, Error= 0.06, Coverage= 0.91, Efficacy= 0.922, f-score= 0.916, Capacity= 0.9448.

The previous conversion rule can be established by a means of a threshold τ. In general, we would have:

> Confidence = max (p_i)
> If Confidence ≥ τ then class = argmax$_i$ (p_i)
> else class = ⊥

This solution, however, is not very suitable when there is an important difference, in general, between the probabilities of one class and the rest. For instance, a very imbalanced dataset would generally abstain more for one class than for others. In general, the minority class would have, proportionally, more abstentions than the majority class.

A more general and flexible conversion rule would be to consider different thresholds τ_i for each class. In this case, we would have:

> Confidence = max (p_i)
> c= argmax$_i$ (p_i)
> If Confidence ≥ τ_c then class = c
> else class = ⊥

With this rule, the prediction is always given by the class with greatest probability, but the abstention will depend on thresholds which are different for each class.

This previous rule has still some problems, since for minority classes the probabilities are usually very low and, independently of how low the threshold is, it is usually not the maximum probability. A variant of the previous rule that allows the different thresholds to be more useful is:

> If $\exists p_i, p_i \geq \tau_i$ then class = argmax$_i$ (p_i / τ_i)
> Else class = ⊥

This rule allows thresholds to be really different (e.g. $\tau_a = 0.8$ and $\tau_b = 0.4$) and, for instance, an example with probabilities $p_a = 0.55$ and $p_b = 0.45$ will be assigned class b, where for the previous rule it would be an abstention.

However, the previous rules mix up two different aspects of a threshold: the relevance given to each class (defined by the difference between the class thresholds) and the degree of abstention (defined by the high or low values of the threshold). In other words, if we want to increase the abstention degree we have to decrease all the thresholds, in a proportional way.

To tackle this problem, we generalise the previous decision rule by using a *class bias* for the probabilities and a *caution window*.

Let us define a Class Bias K= { k_i } i = 1..|C|, where $\Sigma_i\ k_i$= 1 and a window size w, $0 \leq w \leq 1$. With this, we define the decision rule as:

> DECISION RULE: (given K and w)
> For each i = 1..|C|, $\tau_i = (1 - k_i) \cdot w + k_i$
> If $\exists p_i, p_i \geq \tau_i$ then class = argmax$_i$ (p_i / τ_i)
> else class = ⊥

It is easy to see that, since $\Sigma_i\ k_i$= 1 and $\Sigma_i\ p_i$= 1, if we have w=0 then Abstention = 0. In a similar way, if we have w=1 then Abstention will be closer to 1 (not always 1 since there will be cases for which $p_i = \tau_i$).

This new decision rule just separates in a proper way two issues: with K we define the class bias and with w we define the abstention degree.

The previous definitions may seem too complex for two-class problems since one of the components of K depends on the other. However, the idea was to give a decision rule applicable to any classifier (binary or multiclass).

For instance, for the previous example if we define the class bias K = { 0.55, 0.45 } and the window w= 0.15 we would have the following confusion matrix:

		Actual a	Actual b
Predicted	a	37	3
	b	3	48
	⊥	0	9

With a fixed window we just see one "snapshot" of the situation. An interesting thing is to establish a class bias K and show the accuracy with respect to a variable window size. A refinement of this idea is to make the graph with respect to abstention, because increasing window sizes give increasing abstention. Additionally, this relationship is neither continuous nor linear and it is sometimes difficult for comparing several classifiers or different class biases for the same classifier. Consequently, it is better to use the abstention degree as a reference.

For instance, the following graph shows the evolution for increasing window size for three different class biases. As said before, we use abstention instead of window size for making the comparison easier.

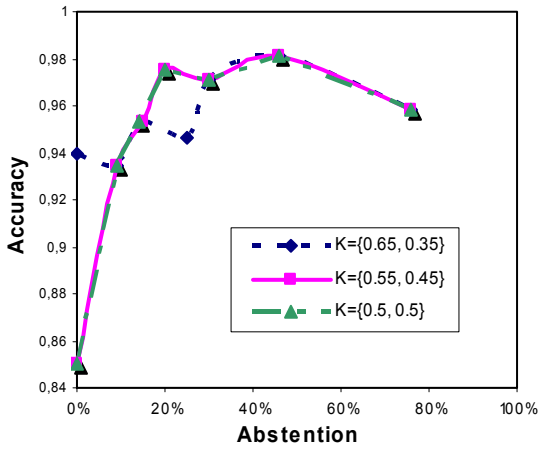

Fig. 6. Accuracy Response Graph

The measure is not shown for abstention 1 because accuracy is undefined. If we calculate the area of any of these curves (using i.e. the value 1 for abstention 1), we have a "probabilistic

capacity" measure, which can be used to evaluate several class biases on the same classifier or several classifiers. In a similar way as other graphs, we can use the *curves* to discard some class biases that are clearly superseded by other class biases for any window size (for instance, in the previous case, the class bias K= {0.55, 0.45}, can be discarded, since it is always equal or worse than the class bias K= {0.5, 0.5}.

A similar graph can be made with the graph error vs. abstention.

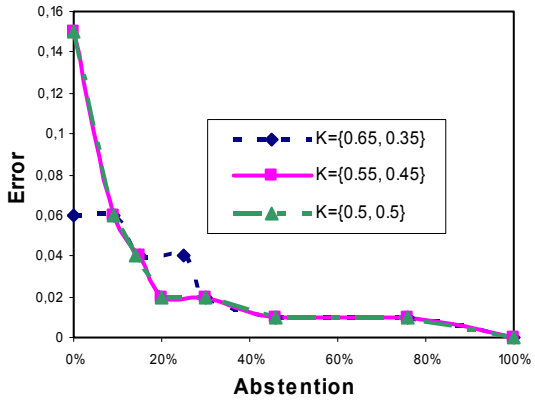

Fig. 7. Error Response Graph

Once again, the areas formed by each curve in the capacity response graph are called "probabilistic capacities", and can be considered as a measure of the optimality of each class bias. In the same way, several probabilistic classifiers can be compared by using this graph.

5 COST-SENSITIVE CAUTIOUS CLASSIFIERS

Up to now, we have considered a cost-insensitive context. Cost matrices can be extended for cautious classifiers in a very easy way.

Now, a Cost Matrix (also known as Loss Matrix), not only indicates the costs for correct and incorrect classifications, but also the cost for abstention, which can be different for each class. An example of a Cost Matrix for three classes {a, b, c} and the unknown class might be as follows:

		Actual		
		a	b	c
Predicted	a	-2.5	4	2
	b	2.1	-3.5	0
	c	1.2	1.3	-4
	⊥	0	0	0

This example shows a quite reasonable portrait, the diagonal of the matrix shows the costs for correct classification (-2.5, -3.5, -4). These values are usually negative, because a correct classification has benefits instead of costs. The other values represent different cases of misclassification. For instance, the value 2.1 in cell (b,a) means that classifying incorrectly an 'a' instance as a 'b' instance has a cost of 2.1. In this case, abstention has zero cost. Obviously, other possibilities are reasonable, provided abstention is somehow in the middle between hits and mistakes.

We use the notation L for this matrix where $L(i,j)$ refers to the element of predicted class i and actual class j.

From this matrix and the confusion matrix it is very easy to compute the cost of a classifier for a given dataset, just as the 1 by 1 matrix product, given a Resulting Matrix:

$$R(i,j) = M(i,j) \times L(i,j), \quad i \in C', j \in C$$

In the same way as before, if we have a cost matrix we can draw a cost response graph for a probabilistic classifier with different class biases, varying the window size..

6 ROC ANALYSIS OF CAUTIOUS CLASSIFIERS

ROC analysis [8][11] has been proven very useful for evaluating classifiers, especially when the cost matrix is not known a priori.

The first question that arises is how to represent a a cautious classifier in the ROC space. More precisely, we would like to obtain TPR and FPR from a 3×2 matrix.

For instance, for the previous example and the class bias K = { 0.55, 0.45 } and the window w= 0.4 we would have the following classifier:

		Actual	
		a	b
Predicted	a	33	1
	b	1	45
	⊥	6	14

We have several alternatives to obtain the TPR and the FPR:

1. Ignoring the values for ⊥

		Actual	
		a	b
Predicted	a	**0.971**	**0.0217**
	b	0.029	0.978
	⊥	-	-

This makes TPR= 0.971, FPR= 0.0177.

2. Ignoring the value for ⊥ but just for the TPR

		Actual	
		a	b
Predicted	a	**0.971**	**0.0167**
	b	0.029	0.75
	⊥	-	0.233

This makes TPR= 0.971, FPR= 0.0167.

3. Ignoring the value for ⊥ but just for the FPR

		Actual	
		a	b
Predicted	a	**0.825**	**0.0217**
	b	0.025	0.978
	⊥	0.15	-

This makes TPR= 0.825, FPR= 0.0217.

4. Not Ignoring the value for ⊥.

		Actual	
		a	b
Predicted	a	**0.825**	**0.0167**
	b	0.025	0.75
	⊥	0.15	0.233

This makes TPR= 0.825, FPR= 0.0167.

Obviously, option 2 is the most optimistic and option 3 the most pessimistic. Wrt. increasing window, option 1, 2 and 3 will end with an undetermined point in the ROC space for w=1, and option 4 will end at point (0, 0).

If we draw the ROC curve just ignoring the unknown cases (option 1), we would have a ROC curve that could be better for some degrees of abstention. Below we will show how this abstention (now ignored) can be taken into account in an evaluation metric.

We can show graphically the effect on the originally ROC curve when removing some decision instances. They are usually on the central part of the curve and the effect of the unknown window is like a secant that cuts part of the ROC curve and push it up towards the rectangle triangle that allows both remainder parts of the ROC curve to re-connect.

This is shown in the following figure:

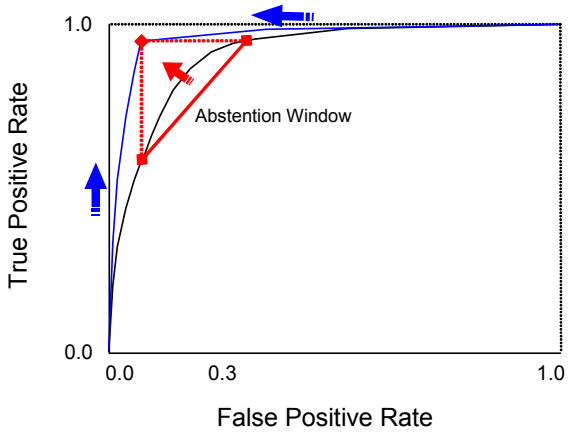

Fig. 8. ROC curve

The new ROC curve is usually more abrupt but has greater area (sacrificing coverage, logically).

Finally, we can take a probabilistic classifier with a class bias and draw the ROC curve for varying window size (TPR × FPR × Abstention). The picture would be something similar to this:

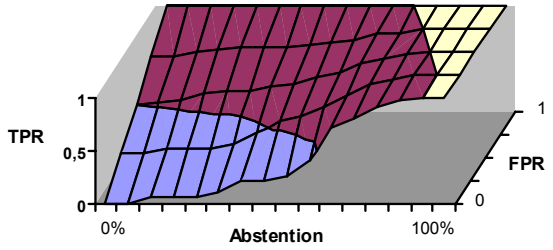

Fig. 9. ROC Response Graph

We could compute the volume of this entire surface as a measure of quality of the probabilistic classifier with a class bias, which has a clear understanding. The greater this volume the greater the AUC is for different levels of abstention.

Instead of representing the ROC curve, we can just show the AUC. In this graph, we can see that two different class biases give better regions depending on the abstention degree. This picture is similar to a ROC graph where we can compare several curves. In fact, we can include the results of several models with several class biases and compute the convex hull to determine the models and class biases that can be discarded surely.

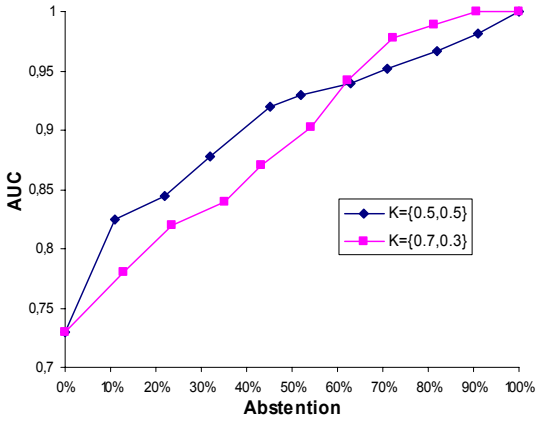

Fig. 10. AUC Response Graph

As expected, the area of each of this AUC vs Abstention graphs corresponds to the volume of the previous graphs TPR x FPR x abstention. The advantage of this latter graph is that we can use approximations of the AUC for more than two classes, such as Hand and Till's *M* function [7].

7 EXAMPLES AND DISCUSSION

In order to show how some of the previous measures can be useful, let us show two simple examples for which we compute some of the previous measures for probabilistic classifiers.

We consider two datasets from the UCI repository [1]: spam and tic-tac. For the experiments, we use the J4.8 implementation in WEKA [12] without pruning and with Laplace correction in order to obtain better probability estimation. Results are given for 20×5-fold cross validation.

For the spam dataset we can show its ROC curve:

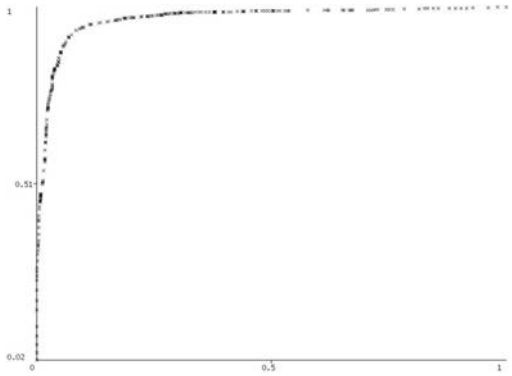

Fig. 11. ROC Curve of J4.8 model for the spam dataset

The ROC Curve shows how TPR and FPR evolve depending on the threshold chosen. The left-upper corner is very abrupt suggesting that in this case a cautious classifier is not going to increase accuracy extraordinarily for increasing degrees of abstention, since an important part of the AUC can be lost if we abstain. This is also so because, initially, the complete classifier (non-cautious) has high accuracy (0.92 for the complete classifier).

If we draw the evolution of accuracy with respect to the degree of abstention (class bias centred on 0.5), we have in the below picture, as expected, a growing "curve" which reaches 1

usually before 100% abstention (note that points are not exactly drawn each 10%, since there can be examples predicted with the same confidence). As we can see in the picture below, accuracy increases quicker for small degrees of abstention, suggesting that good results can be obtained with abstention degrees between 30% and 40%.

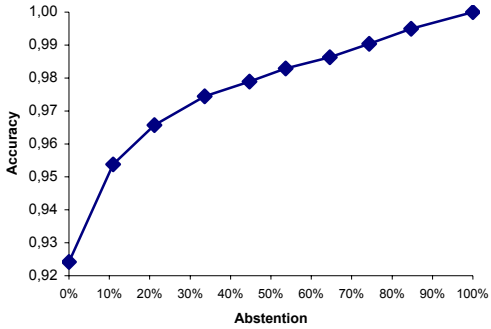

Fig. 12. Accuracy response graph of J4.8 model for the spam dataset

Similarly, the evolution of the AUC (computed as we discussed in section 6) is slightly different, as we can see in the figure below. From 0% to 10% of abstention it goes from 0.967 to 0.974, which is a significant increment and it increases slower in the middle, to reach the maximum more quickly towards the end.

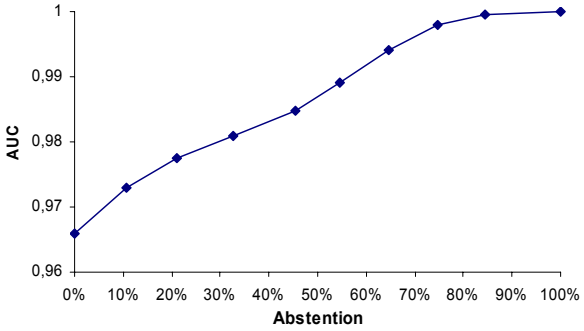

Fig. 13. Accuracy response graph of J4.8 model for the spam dataset

In order to see a context for which it is beneficial to abstain, let us consider the following cost matrix:

		Actual	
		pos	neg
Predicted	pos	0	100
	neg	20	0
	⊥	2	3

This matrix has much greater cost for errors than for abstentions. The picture below shows the evolution of the cost (mean cost of each fold of cross-validation) regarding this matrix. There is a region, between 30% and 60% of abstention, for which the cost is minimal. Even in this case where the cost of errors is much greater than abstentions, we have that the final part of the graph is increasing, showing that very high degrees of abstention are not beneficial in general, as expected. Of course, there can be cost matrices where the minimum can be reached at 0% or 100% of abstention, but the analysis of cautious classifiers is useful in the cases for which there are minimums for different degrees of abstention.

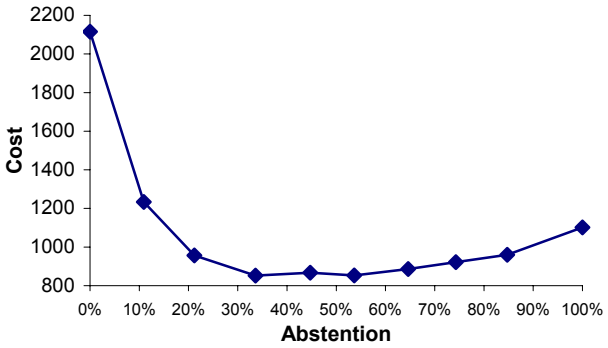

Fig. 14. Cost response graph of J4.8 model for the spam dataset

Let us now take a look to the other dataset, the tic-tac dataset. As the below figure shows, this dataset has much lower accuracy and AUC than the previous one, but, interestingly, AUC is relatively greater than accuracy is (0.8 accuracy and 0.87 AUC for the total classifier). Additionally, the ROC curve is much less abrupt in its left-upper corner, suggesting that here the benefits of a cautious classifier can be more significant.

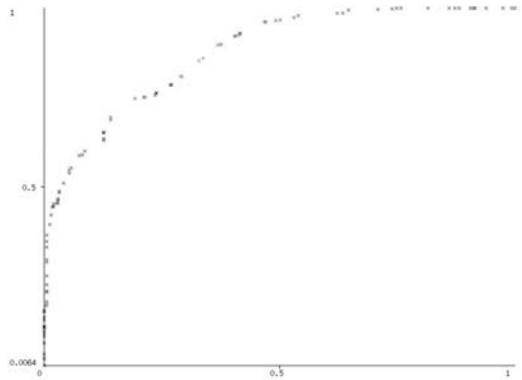

Fig. 15. ROC Curve of J4.8 model for the tic-tac dataset

As we did in the previous dataset, we also show the evolution of accuracy wrt. the abstention degree (class bias centred on 0.5):

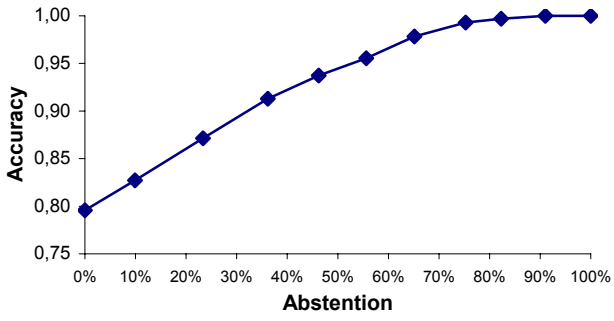

Fig. 16 Accuracy response graph of J4.8 model for the tic-tac dataset

In this case, the increase is constant and more regular from 0% to 65%. This broad region is given by the smooth left-upper corner of the ROC curve, which allows that the AUC steadily goes from 0.87 at 0% abstention to 0.89 at 11%, to 0.918 at 24%, to 0.941 at 36%, to 0.956 at 46% and much slower from that point (0.98 at 82%), as we can see in the following AUC response graph.

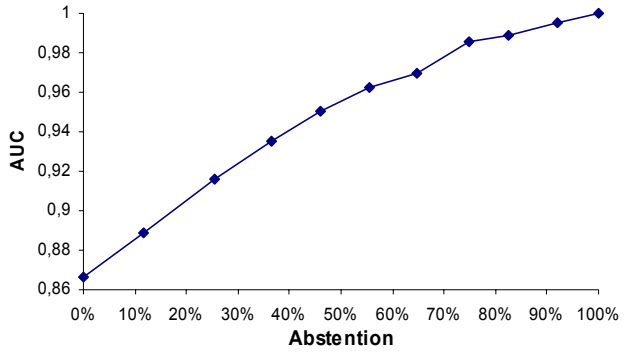

Fig. 17 AUC response graph of J4.8 model for the tic-tac dataset

Finally, let us show the cost response curve for the tic-tac problem (using the same cost matrix as before).

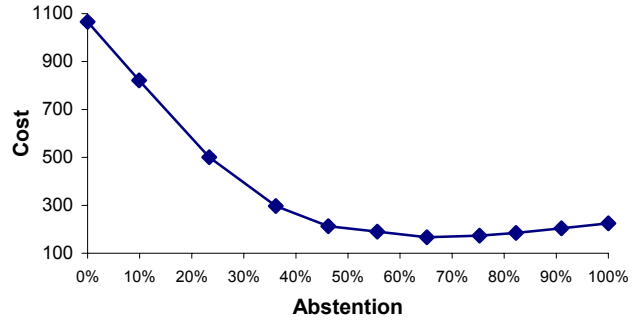

Fig. 18. Cost response graph of J4.8 model for the tic-tac dataset

In this case, the lowest cost is obtained between 65% and 75% abstention degree, showing that, for this dataset, greater levels of abstention can have more sense than in the previous dataset.

Finally, for the tic-tac dataset we compare the accuracy response graph for three different learning methods: J4.8, Naïve Bayes and Logistic Regression:

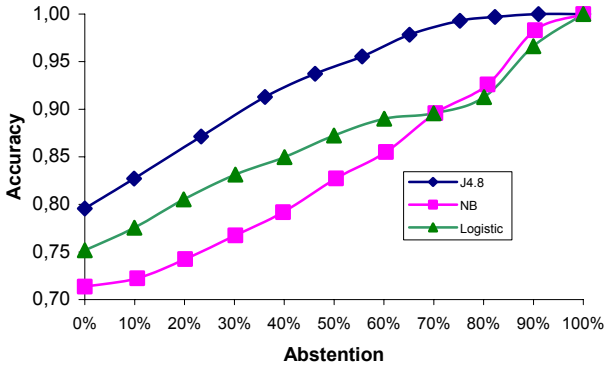

Fig. 19. Accuracy response graph for several models for the tic-tac dataset

The picture above shows very different behaviours of the three models. It is especially interesting the final part of NB and Logistic, because Logistic is better than NB for low and mid abstention degrees but NB seems better than Logistic for high abstention degrees. The modus operandi is similar to how ROC analysis is used to compare probabilistic classifiers: we can decide in which regions one classifier is better than the rest.

It is interesting to see the ROC curves for NB and Logistic.

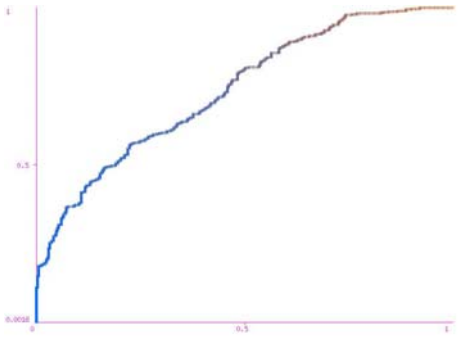

Fig. 20. ROC Curve of NB model for the tic-tac dataset

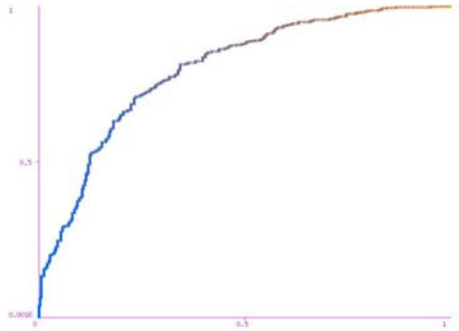

Fig. 21. ROC Curve of Logistic model for the tic-tac dataset

We can see that the start of the ROC curve of the logistic model is highly irregular, which may justify the irregular behaviour in the accuracy response graph of this classifier. Nonetheless, it is much more difficult to see this from the ROC curves than from the accuracy response graphs.

In a similar way, the picture below shows the accuracy response graph of J4.8 without and with pruning:

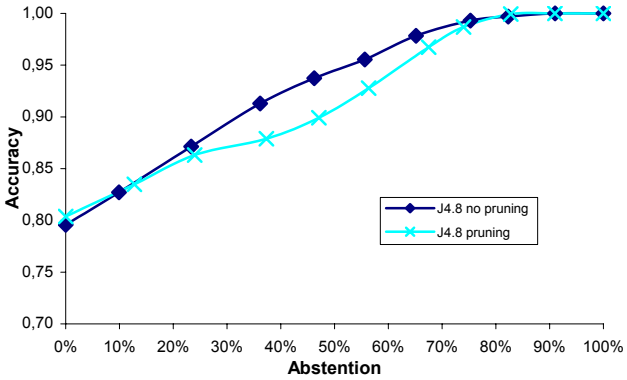

Fig. 22. Accuracy response graph for J4.8 for the tic-tac dataset

It is highly interesting to see how the same technique can construct models that behave quite differently with increasing degrees of abstention. J4.8 with pruning has initially better accuracy (0.804 vs. 0.796) but much worse AUC (0.844 vs. 0.873), as has been shown to be the case in general (see e.g. [2][3][8][9]). Since AUC is a good metric for probability estimators, we see that the unpruned decision tree does a better job selecting the examples that must be abstained, because its probability estimations are better.

This highlights the situations where good cautious classifiers can be obtained from probabilistic classifiers: the probabilistic classifier is a good probability estimator.

Summing up, from the previous sections and these simple examples, we can suggest that capacity and efficacy are useful for comparing cautious classifiers, not for comparing probabilistic classifiers or for analysing how to convert them into cautious classifiers. In fact, capacity and efficacy response curves are always decreasing, so their goal is not to show which abstention degree can improve their values, because there are not. Their purpose is just to compare static (i.e. categorical) classifiers.

Accuracy, error and AUC response graphs are useful for choosing the optimal degree of abstention and a good class bias, especially when class distribution is not balanced. From the previous examples, we have also seen that the form of the original ROC curve of a probabilistic classifier also gives important information about how its derived cautious classifiers can perform.

8 CONCLUSIONS

In the introduction, we have argued that there are many situations where a classifier can abstain, instead of predicting a class for cases where the classifier is quite uncertain. We need, however, to assess with new metrics how to perform this abstention in order to find trade-offs between abstention and accuracy/error. In this work, we have presented some new techniques for evaluating, transforming and representing cautious classifiers. Some of them are useful for comparing between original cautious classifiers, such as efficacy and capacity, others can be useful for transforming and evaluating probabilistic classifiers, such as the accuracy / AUC response curves.

This paper, nonetheless, leaves many open questions about statistical and formal interpretation of the introduced measures, which would be useful to give a better understanding of what they mean and to what they correspond. In this sense it would be interesting to analyse and define the degrees of freedom in a cautious cost matrix in order to obtain parametrised cost functions for probabilistic classifiers, depending on the class bias (which roughly corresponds to the "skew") and the abstention degree.

An open question is how to act when the original problem has an "unknown" class. In this case, there is no need to create a new class, but this "unknown" class should be treated in a different way than the other classes, in order to apply or modify the measures and techniques presented in this work.

Another interesting thing to be studied would be how to perform the combination of classifiers where we want the resulting classifier to be a cautious classifier. Several new situations appear, that can be handled with classical approaches used in ensemble methods and some of the techniques presented here. For instance, we could consider the situation of combining *n* probabilistic classifiers into one cautious classifier. In this situation, the variance of the predictions of the ensemble could be relevant to tell between a given class or the unknown class. Other situation can be presented when we want to combine *n* cautious classifiers into one cautious classifier. The proportion of "unknowns" among the *n* classifiers should be also taken into account.

The idea of cautious classifier has been used to implement a new multiclassifier scheme known as "delegating classifier" [4], where the first classifier abstains for a part of the examples and delegates them to a second classifier.

Finally, an interesting connection can be established between cautious classifiers and classifiers that take test cost into account. For instance, in medical diagnosis, it may be interesting to construct a classifier with a high abstention that performs few and very cheap tests (uses the cheapest attributes). For the cases where this classifier gives a certain result, the use of a classifier with more coverage (but further tests) is avoided. For the cases where the classifier gives an unknown result, almost nothing is lost. This can be taken forward to by considering classifiers that are able to find trade-offs (in application time) between test cost and coverage.

There are of course many relations to other approaches which consider classifiers in a non-categorical way, considering class probability estimators, fuzzy classifiers, etc.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

1. Blake, C.; Merz, C. "UCI repository of machine learning databases", 1998.
2. Ferri, C., Flach, P., Hernández-Orallo, J. Learning Decision Trees using the Area Under the ROC Curve. In C. Sammut; A. Hoffman (eds.), *Proc. Int. Conf. on Machine Learning (ICML2002)*, pp. 139-146, Morgan Kaufmann, 2002.
3. Ferri, C., Flach, P., Hernández-Orallo, J.. *Improving the AUC of Probabilistic Estimation Trees.* in *Proc. Int. Conf. on Machine Learning* (ECML 2003), LNAI, Springer 2003.
4. Ferri, C.; Flach, P.; Hernández-Orallo, J. "Delegating classifiers" conditionally accepted for ICML'04.
5. Fürnkranz, J. "Separate-and-conquer rule learning" *Artificial Intelligence Review*, 13, 3-54, 1999.
6. Hand, D.J. *Construction and assessment of classification rules*. Chichester: Wiley, 1997.
7. Hand, D.J.; Till, R.J. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems" *Machine Learning*, 45, 171-186, 2001.
8. Ling, C.X., Yan, R.J. Decision Tree with Better Ranking. In *Proc. Int. Conf. on Machine Learning (ICML2003)*, AAAI Press, 2003.
9. Provost, F., Domingos, P. Tree Induction for Probability-based Ranking. *Machine Learning* 52(3), 2003.
10. Provost, F. and Fawcett, T. "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distribution" in *Proc. of The Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp. 43-48, Menlo Park, CA: AAAI Press, 1997.
11. Swets, J., Dawes, R., and Monahan, J. "Better decisions through science" *Scientific American*, October 2000, 82-87.
12. Witten, I.H.; Frank, E. "Data Mining: Practical machine learning tools with Java implementations", Morgan Kaufmann, 2000, http://www.cs.waikato.ac.nz/ ml/weka/.