Maximizing the Area under the ROC Curve using Incremental Reduced Error Pruning

Henrik Boström HENKE@DSV.SU.SE
Dept. of Computer and Systems Sciences
Stockholm University and Royal Institute of Technology
Forum 100, 164 40 Kista, Sweden

Abstract

The use of incremental reduced error pruning for maximizing the area under the ROC curve (AUC) instead of accuracy is investigated. A commonly used accuracy-based exclusion criterion is shown to include rules that result in concave ROC curves as well as to exclude rules that result in convex ROC curves. A previously proposed exclusion criterion for unordered rule sets, based on the lift, is on the other hand shown to be equivalent to requiring a convex ROC curve when adding a new rule. An empirical evaluation shows that using lift for ordered rule sets leads to a significant improvement. Furthermore, the generation of unordered rule sets is shown to allow for more fine-grained rankings than ordered rule sets, which is confirmed by a significant gain in the empirical evaluation. Eliminating rules that do not have a positive effect on the estimated AUC is shown to slightly improve AUC for ordered rule sets, while no improvement is obtained for unordered rule sets.

1. Introduction

There has recently been a growing interest in using ROC curves for analyzing rule learning methods (Fürnkranz & Flach, 2003; Fürnkranz & Flach, 2004; Fürnkranz & Flach, 2005) as well as using rule learning methods for maximizing the area under the ROC curve (AUC) (Fawcett, 2001; Lavrac et al., 2004; Prati & Flach, 2004)). The main motivations for using AUC as an evaluation criterion instead of accuracy, which traditionally has been the most common criterion for comparing rule induction methods (e.g. (Cohen, 1995)), are that it is insensitive to the actual class distribution on which the model is tested and that it does not assume equal misclassification costs (Bradley, 1997; Provost et al., 1998). As noted in (Bradley, 1997), the AUC can be interpreted as the probability of ranking a true positive example higher than a false positive when ordering examples according to decreasing likelihood of being positive.

Incremental reduced error pruning, which was originally introduced in (Fürnkranz & Widmer, 1994), is a technique that has been extensively used for efficient separate-and-conquer rule learning (e.g., (Fürnkranz & Widmer, 1994; Cohen, 1995; Frank & Witten, 1998; Dain et al., 2004; Boström, 2004)). By pruning each rule immediately after its generation and removing examples covered by the pruned rule, the number of generated rules is kept relatively small compared to keeping each rule unpruned and removing the relatively few examples covered by each, more specific, rule. Since the computational cost grows as the product of the number of generated rules and the number of training examples, incremental reduced error pruning normally allows substantially larger training sets to be handled. A number of criteria for deciding how to prune generated rules and whether or not to exclude a generated rule have previously been proposed and evaluated with respect to maximizing accuracy (Fürnkranz & Widmer, 1994; Cohen, 1995; Boström, 2004).

In this work, we investigate the use of incremental reduced error pruning for maximizing AUC. This includes investigating whether previously proposed pruning and exclusion criteria for maximizing accuracy also are reasonable for maximizing AUC. It turns out that one of the most frequently used pruning criteria, precision, already has been shown to maximize AUC (Fürnkranz & Flach, 2005), and hence may be used also for this purpose. We show, however, that

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*

the most commonly used exclusion criterion, based on accuracy, is less suited, since it may lead to concave ROC curves as well as to excluding rules that would result in convex ROC curves. On the other hand, a previously proposed exclusion criterion for unordered rule sets, based on the lift, is shown to include a rule if and only if it leads to a convex ROC curve.

We also study using incremental reduced error pruning for maximizing AUC by generating both ordered and unordered rule sets. In contrast to ordered rule sets (also known as decision lists (Rivest, 1987)), which classify examples according to the first applicable rule, a prediction is formed from all applicable rules in an unordered rule set (see (Fawcett, 2001) for comparisons of a number of methods for forming the prediction to maximize AUC). Moreover, incremental reduced error pruning for ordered rule sets generate rules for all classes except one, which is used to label a default rule, while incremental reduced error pruning for unordered rule sets results in rules that characterize all classes. We will explain why these two differences may in fact be advantageous when generating unordered rule sets to maximize AUC.

In the next section, we present the two variants of incremental reduced error pruning (resulting in ordered and unordered rule sets respectively), and present a method for post-processing generated rules by eliminating rules that do not appear to improve AUC. We analyze the suitability of previously proposed exclusion criteria for maximizing AUC. We also explain why generating unordered rule sets could be expected to give a higher AUC than by generating ordered rule sets. In Section 3, an empirical comparison of the methods is given, and finally, in Section 4, we conclude by discussing made observations and outline some directions for future work.

2. Methods

2.1. Incremental Reduced Error Pruning

In Fig. 1 and Fig. 2 two variants of incremental reduced error pruning are shown. The first, called IREP-O, generates ordered rule sets and is a variant of the algorithms presented in (Fürnkranz & Widmer, 1994; Cohen, 1995), while the second, called IREP-U, generates unordered rule sets and is taken from (Boström, 2004). The main differences between the algorithms is that that the latter generates rules for all classes, while the former will form a default rule for the last class. Furthermore, the prune set is kept constant in the latter algorithm, allowing each rule to be evaluated and pruned independently of previously generated rules, while the former removes covered examples from both the grow and prune sets. It should be noted that in the original formulation of incremental reduced error pruning for ordered rule sets (Fürnkranz & Widmer, 1994), only two-class problems were handled, while this was extended to multi-class problems in (Cohen, 1995). The algorithm for ordered rule sets presented here slightly differs from the previous in that a prune set is generated initially, from which examples are removed only if they are covered by a generated rule that should not be excluded. In the original formulation, the remaining examples to be covered were repeatedly divided into a grow and prune set each time a new rule was to be generated, and the rule generation was terminated whenever a rule was found that should not be included.[1]

Two problems that need to be addressed when applying unordered rules is how to classify examples that are covered by multiple, possible conflicting rules, and how to classify examples that are not covered at all. The former problem is in this work addressed by applying naïve Bayes as in (Boström, 2004), while the latter is addressed by classifying the example according to the class distribution of those examples in the prune set that are not covered by any rules.[2]

For both algorithms, class probability distributions are formed using the covered examples in the prune set together with Laplace correction (Cestnik & Bratko, 1991).

2.2. Pruning and Exclusion Criteria for Maximizing AUC

Several commonly employed pruning criteria for incremental reduced error pruning have been shown to be equivalent to maximizing precision, i.e., the fraction $\frac{p}{p+n}$, where p and n are the number of covered positive and negative examples respectively (Fürnkranz & Flach, 2005). In the same work, it is noted that maximizing precision in fact is equivalent to attempting to maximize AUC. To see this, assume we start with a default rule assigning zero probability of being positive to all examples (i.e., the ROC curve is a straight line from $(0, 0)$ to $(0, 1)$, where the x- and y-coordinates give the fraction of covered false and true positives respectively). If we add a rule that covers p

[1] In (Cohen, 1995), an alternative stopping condition was introduced, allowing the number of bits required to encode the rules and class labels to grow up to d when adding a rule compared to the minimum encoding found so far, where d is a user-specified parameter.

[2] If this set is empty, the distribution is formed using the original prune set.

```
function IREP-O($OrderedClasses$,$Examples$)
  $Rules := \emptyset$
  Make stratified split of $Examples$ into
    $Grow$ and $Prune$
  for each $Class \in OrderedClasses$ do
    if $Last(Class, OrderedClasses)$ then
      $Rules := Rules \cup \{DefaultRule(Prune)\}$
    else
      $Pos := \{e : e \in Grow \wedge Class(e) = Class\}$
      $Neg := Grow \setminus Pos$
      while $Pos \neq \emptyset$ do
        $Rule := GrowRule(Pos, Neg)$
        $Rule := PruneRule(Rule, Prune)$
        if $Exclude(Rule, Prune)$ then
          $Grow := Grow \setminus Covers(Rule, Pos)$
          $Pos := Pos \setminus Covers(Rule, Pos)$
        else
          $Rules := Rules \cup \{Rule\}$
          $Grow := Grow \setminus Covers(Rule, Grow)$
          $Prune := Prune \setminus Covers(Rule, Prune)$
  return $Rules$
```

Figure 1. The IREP-O algorithm.

positive and n negative examples to this classifier, examples covered by this rule will be given a higher rank than those classified by the default rule alone. The ROC curve will now consist of two segments, passing through $(0, 0)$, $(n/N, p/P)$ and $(1, 1)$, where N and P are the total number of negative and positive examples respectively. In order to maximize AUC, we would like to maximize the slope of the first segment[3], which is given by $\frac{p/P}{n/N}$. Since P and N are constant for all candidate rules, maximizing the slope of the ROC curve is equivalent to maximizing precision[4].

A commonly employed exclusion criterion when generating ordered rule sets is $\frac{p}{p+n} \leq 1/2$ (Fürnkranz & Widmer, 1994; Cohen, 1995), which is natural when maximizing accuracy, since an added rule for the positive class may otherwise be allowed to make more errors than correct classifications. However, when maximizing AUC, this criterion may in fact allow a rule to be added for which the slope of the corresponding first segment of the ROC curve is less than one, i.e., the corresponding ROC curve is concave, since the slope depends on the total number of positive and negative examples as shown above. For example, if $p = 2n$ and $P = 3N$ then the rule would be included using this criterion (since $2/3 > 1/2$), but the slope will be $\frac{n}{N}\frac{2}{3} \leq 2/3$. Moreover, this criterion may also exclude rules that result in a slope greater than one. For example, if $n = 2p$ and $N = 3P$ then the rule would be excluded using this criterion (since $1/3 \leq 1/2$), but the slope will be $\frac{p}{P}\frac{3}{2}$, which is greater than one, if $p/P > 2/3$.

```
function IREP-U($Classes$,$Examples$)
  $Rules := \emptyset$
  Make stratified split of $Examples$ into
  $Grow$ and $Prune$
  for each $Class \in Classes$ do
    $Pos := \{e : e \in Grow \wedge Class(e) = Class\}$
    $Neg := Grow \setminus Pos$
    while $Pos \neq \emptyset$ do
      $Rule := GrowRule(Pos, Neg)$
      $Rule := PruneRule(Rule, Prune)$
      if not $Exclude(Rule, Prune)$ then
        $Rules := Rules \cup \{Rule\}$
      $Pos := Pos \setminus Covers(Rule, Pos)$
  return $Rules$
```

Figure 2. The IREP-U algorithm.

For unordered rule sets, lift (i.e., $\frac{\frac{p}{p+n}}{\frac{P}{P+N}}$) has been the basis for both a pruning and an exclusion criterion (Boström, 2004).[5] It should be noted that using lift as a pruning criterion is equivalent to using precision, since $\frac{P}{P+N}$ is constant. However, excluding rules with a lift less than or equal to one turns out to be equivalent to requiring a convex ROC curve for an included rule (i.e., the slope of the first segment must be greater than one), since

$$\frac{\frac{p}{p+n}}{\frac{P}{P+N}} \leq 1 \iff \frac{p}{p+n} \leq \frac{P}{P+N} \iff$$

$$p(P+N) \leq P(p+n) \iff pN \leq Pn \iff$$

$$\frac{p}{P} \leq \frac{n}{N} \iff \frac{\frac{p}{P}}{\frac{n}{N}} \leq 1$$

[3] This would not necessarily be optimal if we were allowed to add one rule only, but this strategy assumes that an arbitrary number of additional rules may be added.

[4] Maximizing $\frac{p}{p+n}$ is equivalent to minimizing $\frac{p+n}{p} = 1 + n/p$ which in turn is equivalent to maximizing p/n.

[5] The term *likelihood ratio to default* was used instead of lift in that work.

2.3. Post-processing Rule Sets w.r.t. AUC

It has been observed that significant gains can be obtained when using incremental reduced error pruning for maximizing accuracy, by post-processing generated rules through considering replacements of each rule with more general or specific versions followed by eliminating rules that increase the total description length (Cohen, 1995). A similar procedure may be used also for maximizing AUC. In this work, we consider a simplified procedure, in which each rule is either kept or completely eliminated (i.e., replacement rules are not considered), and instead of minimizing the description length, rules that do not contribute positively to the AUC (as estimated on the prune set) are removed.

It should be noted that when removing a rule from an unordered rule set, the class distributions of the remaining rules are not affected, since the coverage of each rule on the prune set is independent of the other rules. Hence, one pass through the rules suffices for finding out which rules should be removed. On the other hand, when removing a rule from an ordered rule set, the class distributions of the successive rules may be affected. Hence it matters in what order rules are removed and several passes over the rules may be required. In our study, rules are considered in the same order as they were generated, and whenever a rule is removed, the remaining rules are considered from the beginning.

2.4. Ordered vs. Unordered IREP for Maximizing AUC

As mentioned in the introduction, the fact that when using unordered rule sets, classifications may be formed from several rules and that rules are generated for each class can be beneficial when trying to maximize AUC, as explained below.

Assume that we are facing a two-class learning task, where each class requires two rules if defined separately. Assume further that attached to each rule is a class probability distribution. An ordered rule set would then typically consist of three rules $H_O = R_1, R_2, R_3$, where the two first rules would assign the same most probable class (positive) to covered examples, while the last would act as a default rule, assigning the other class (negative) to any examples that are not covered by the first two rules. From a ranking perspective, where we want to order a set of examples from the most likely positive to the least likely, the ordered rule set H_O allows for partitioning the examples in (at most) three groups, where all examples in a group are given the same score (i.e., probability of being positive).[6] In particular, all examples that would be classified as negative are placed in the same group and could hence not be differentiated.

Table 1. Employed Methods

Acronym	Algorithm	Post-Processing	Excl. crit.
DL	IREP-O	no	accuracy
DLP	IREP-O	yes	accuracy
DL-L	IREP-O	no	lift
DLP-L	IREP-O	yes	lift
RS	IREP-U	no	lift
RSP	IREP-U	yes	lift

On the other hand, an unordered rule set would typically consist of four rules $H_U = \{R_1, R_2, R_3, R_4\}$, for which the class distributions of the two first would give the positive class a higher probability than the negative and vice versa for the last two rules. Since an example that is to be ranked in principle can be covered by any subset of the four rules, we have at most 2^4 possible groups to place the example in. This means that examples (either classified as positive or negative) can be ranked according to a much more fine-grained scale. Even if no or few of the rules that would assign different classes do overlap, the possibility of differentiating examples independently of whether they are classified as positive or negative still allows for the examples to be partitioned into more groups.

3. Empirical Evaluation

3.1. Experimental Setting

3.1.1. Methods

The methods that are to be compared are variants of the IREP-O and IREP-U algorithms using two different exclusion criteria for IREP-O (accuracy and lift respectively) and with and without post-processing for both algorithms. All methods use precision as a pruning criterion, and 2/3 of the training examples are used for growing rules, while 1/3 are used for pruning. All methods are given the same grow and prune sets. The employed methods are summarized in Table 1.

[6] There will be fewer possible groups if the same probability distribution is attached to multiple rules.

Table 2. AUC for all 6 methods on the 34 data sets.

Data set	DL	DLP	DL-L	DLP-L	RS	RSP
breast-cancer (2 cl.)	59.17	59.17	61.22	62.11	**66.58**	65.85
breast-cancer-wisconsin (2 cl.)	95.31	95.07	96.42	96.40	**99.13**	98.76
crx (2 cl.)	85.70	86.64	85.76	87.39	**89.95**	89.41
cylinder-bands (2 cl.)	**73.98**	73.94	71.46	70.93	72.97	72.62
hepatitis (2 cl.)	65.98	66.35	73.34	73.42	**82.96**	82.00
house-votes (2 cl.)	97.38	97.61	97.49	97.81	**97.91**	97.19
ionosphere (2 cl.)	91.72	90.69	91.81	90.53	**95.09**	93.83
kr-vs-kp (2 cl.)	95.92	96.11	97.03	97.25	99.54	**99.67**
mushroom (2 cl.)	99.80	99.80	99.80	99.80	**99.99**	99.98
pima-indians-diabetes (2 cl.)	65.27	65.27	69.48	69.24	**76.90**	76.66
promoters (2 cl.)	71.21	72.10	73.82	76.23	85.11	**85.87**
sick-euthyroid (2 cl.)	81.75	81.46	90.57	91.11	**95.67**	95.63
spambase (2 cl.)	83.19	83.18	82.75	82.68	94.40	**94.48**
spectf (2 cl.)	57.47	57.47	62.69	62.69	**87.23**	86.00
tic-tac-toe (2 cl.)	96.37	96.40	97.56	97.63	**99.69**	99.65
balance-scale (3 cl.)	80.19	80.41	78.44	79.77	**95.83**	95.74
splice (3 cl.)	95.70	95.82	96.46	96.81	**98.05**	97.95
tae (3 cl.)	50.73	50.73	50.78	50.84	**52.77**	**52.77**
iris (3 cl.)	94.75	95.60	95.00	95.87	97.91	**98.12**
lung-cancer (3 cl.)	62.05	62.05	64.01	64.01	**70.01**	**70.01**
new-thyroid (3 cl.)	86.98	87.17	90.44	90.55	96.43	**96.50**
post-operative-patients (3 cl.)	**50.00**	**50.00**	47.83	43.08	40.67	39.46
wine (3 cl.)	89.35	89.89	90.64	90.31	**99.18**	98.84
car (4 cl.)	84.99	85.12	93.27	94.49	97.93	**98.11**
lymphography (4 cl.)	71.82	70.37	75.55	74.35	76.32	**80.14**
cleveland-heart-disease (5 cl.)	53.06	53.06	**68.33**	67.71	65.50	65.93
glass (6 cl.)	66.80	66.34	67.83	69.25	72.25	**72.79**
dermatology (6 cl.)	88.64	88.53	94.88	94.33	**97.40**	97.27
image-segmentation (7 cl.)	88.32	88.06	89.82	89.85	**92.11**	91.26
ecoli (8 cl.)	88.12	88.04	92.26	92.27	87.83	**93.09**
yeast (10 cl.)	69.25	69.54	73.46	**74.43**	70.60	73.90
soybean-large (19 cl.)	92.45	92.32	92.51	92.22	92.22	**93.94**
primary-tumor (21 cl.)	58.22	58.29	67.97	68.26	66.47	**70.66**
audiology (24 cl.)	81.37	81.46	83.59	**84.36**	80.82	81.24

3.1.2. Methodology and data sets

We have chosen to compare the methods w.r.t. AUC using ten-fold cross-validation on 34 data sets from the UCI Repository (Blake & Merz, 1998). The names of these data sets together with the number of classes are listed in Table 2. The AUC was calculated for each method on all examples according to (Fawcett, 2003) and all methods were given exactly the same training and test examples. For data sets with more than two classes, the total AUC was calculated (Fawcett, 2001).[7]

3.1.3. Test hypotheses

There are actually a number of hypotheses to be tested: does lift result in a higher AUC than using accuracy as an exclusion criterion for ordered rule sets, is the suggested post-processing method beneficial for (ordered and unordered) incremental reduced error pruning, and does unordered incremental reduced error pruning outperform the ordered variant.

3.2. Experimental Results

The AUC for all methods on all 34 data sets are shown in Table 2, where the best result for each data set is in bold-face and the rows are ordered after the number of classes in each data set.

In Table 3, the number of wins and losses for each pair of methods is shown, together with the p-value of obtaining that result if the null hypothesis holds (i.e., both methods are equally likely to win).

One can see that using lift as an exclusion criterion indeed is clearly more effective than using accuracy for ordered rule sets (independently of whether or not

[7] For two-class problems, the total AUC is equivalent to AUC.

Table 3. Summary of all wins and losses for all 6 methods (row wins/column wins).

	DL	DLP	DL-L	DLP-L	RS	RSP
DL	-	13/15 (8.51e-01)	4/29 (1.09e-05)	6/27 (3.24e-04)	5/29 (3.86e-05)	4/30 (6.16e-06)
DLP	15/13 (8.51e-01)	-	7/26 (1.32e-03)	6/27 (3.24e-04)	5/29 (3.86e-05)	4/30 (6.16e-06)
DL-L	29/4 (1.09e-05)	26/7 (1.32e-03)	-	12/20 (2.15e-01)	7/27 (8.21e-04)	4/30 (6.16e-06)
DLP-L	27/6 (3.24e-04)	27/6 (3.24e-04)	20/12 (2.15e-01)	-	7/27 (8.21e-04)	5/29 (3.86e-05)
RS	29/5 (3.86e-05)	29/5 (3.86e-05)	27/7 (8.21e-04)	27/7 (8.21e-04)	-	18/14 (5.97e-01)
RSP	30/4 (6.16e-06)	30/4 (6.16e-06)	30/4 (6.16e-06)	29/5 (3.86e-05)	14/18 (5.97e-01)	-

post-processing is employed). The p-value of obtaining the observed number of wins and losses is 1.09×10^{-05} without post-processing and 3.24×10^{-04} with post-processing, allowing the null hypothesis to be safely rejected.

One can also see that generating unordered rule sets (with or without post-processing) results in a higher AUC than when generating ordered rule sets (for both exclusion criteria). With no post-processing, the win/loss ratio between generating unordered rule sets and ordered rule sets using lift is 27/7 and the p-value is 8.21×10^{-04}, and with post-processing, the win/loss ratio is 29/5 and the p-value is 3.86×10^{-05}, in both cases allowing the null hypothesis to be rejected.

When it comes to whether or not the proposed post-processing procedure actually is beneficial w.r.t. AUC, the picture is less clear. For ordered rule sets generated from using lift as the exclusion criterion, the use of post-processing appears to be beneficial with a win/loss ratio to not using post-processing of 20/12, which however does not allow the null hypothesis to be rejected (the p-value is 0.215). For ordered rule sets generated with the accuracy-based criterion and for unordered rule sets, there seem to be no gains w.r.t. AUC from using the proposed post-processing method. However, from the point of view of interpretability, the rule sets become much smaller with post-processing as shown in Table 4. One might conclude that the rule sets can be simplified without significantly loosing performance.

It can also be observed in Table 4 that the number of rules typically increases when using lift instead of accuracy as exclusion criterion for ordered rule sets (the number of rules increases 32 times and decreases 1 time without post-processing, while the number of rules increases 29 times and decreases 2 times with post-processing). This indicates that the benefit of using lift compared to accuracy actually comes from including rules that otherwise would have been excluded (due to too low accuracy), rather than from eliminating rules that introduce concavities.

4. Concluding Remarks

We have studied the use of incremental reduced error pruning for maximizing AUC instead of accuracy. While a commonly employed pruning criterion, based on precision, has been shown to maximize AUC, we show that a commonly used exclusion criterion, based on accuracy, may include rules that result in concave ROC curves, as well as exclude rules that result in convex ROC curves. We showed that a previously proposed exclusion criterion, based on lift, includes a rule if and only if the resulting ROC curve is convex, and an empirical evaluation gave strong evidence for that the use of this criterion improves AUC compared to the accuracy-based criterion for ordered rule sets.

We have also presented arguments for why one might expect the generation of unordered rule sets to give a higher AUC than when generating ordered rule sets: the number of ways to partition the examples is increased by forming class distributions from multiple rules as well as that rules are generated for all classes. This was also confirmed to be highly beneficial by the empirical evaluation. Although we did not experiment with alternative ways of forming the predictions, this study in a way confirms the observation in (Fawcett, 2001), that a higher AUC is obtained by combining all applicable rules than using a single rule.

Finally, we studied whether eliminating rules that do

Table 4. Mean no. of rules for all 6 methods on the 34 data sets.

Data set	DL	DLP	DL-L	DLP-L	RS	RSP
breast-cancer (2 cl.)	**2.8**	**2.8**	5.0	4.4	11.3	9.8
breast-cancer-wisconsin (2 cl.)	8.8	**8.1**	9.9	**8.1**	23.2	13.7
crx (2 cl.)	11.3	**8.2**	12.7	9.1	27.2	19.3
cylinder-bands (2 cl.)	9.4	**8.2**	10.3	9.1	20.7	18.5
hepatitis (2 cl.)	2.2	**2.1**	2.8	2.5	12.2	8.9
house-votes (2 cl.)	4.1	**3.2**	4.8	3.6	13.1	8.3
ionosphere (2 cl.)	5.4	**5.1**	6.6	6.2	15.6	12.2
kr-vs-kp (2 cl.)	23.3	**16.1**	25.0	17.4	54.1	35.5
mushroom (2 cl.)	11.2	**9.8**	11.2	**9.8**	45.6	23.1
pima-indians-diabetes (2 cl.)	6.7	**5.9**	9.5	8.2	26.5	19.1
promoters (2 cl.)	4.0	**3.7**	5.4	4.6	8.0	7.4
sick-euthyroid (2 cl.)	5.6	**5.0**	7.2	5.9	31.8	21.4
spambase (2 cl.)	32.5	23.2	32.3	**22.9**	101.5	84.4
tic-tac-toe (2 cl.)	9.6	**9.0**	11.1	10.1	35.4	21.3
spectf (2 cl.)	3.0	**2.9**	3.7	3.6	23.8	17.2
balance-scale (3 cl.)	14.2	11.2	14.5	**10.9**	58.1	37.8
iris (3 cl.)	5.8	**4.3**	6.2	**4.3**	11.3	7.0
lung-cancer (3 cl.)	**1.9**	**1.9**	2.3	2.3	3.4	3.3
new-thyroid (3 cl.)	4.4	**4.2**	5.4	5.1	9.3	7.4
post-operative-patients (3 cl.)	**1.0**	**1.0**	1.6	1.4	3.7	3.4
splice (3 cl.)	15.6	**13.3**	20.8	16.3	104.0	87.8
tae (3 cl.)	**2.8**	**2.8**	6.2	5.6	8.1	8.0
wine (3 cl.)	5.9	**5.2**	6.5	5.7	11.8	8.8
car (4 cl.)	16.6	**15.2**	31.8	24.8	51.6	41.5
lymphography (4 cl.)	3.6	**3.1**	4.8	4.1	11.7	8.6
cleveland-heart-disease (5 cl.)	**1.5**	**1.5**	5.7	5.3	13.2	10.9
dermatology (6 cl.)	8.4	**6.9**	11.3	8.4	15.8	10.6
glass (6 cl.)	5.0	**4.7**	8.5	7.4	10.6	9.1
image-segmentation (7 cl.)	7.6	**6.9**	10.3	8.9	15.9	13.4
ecoli (8 cl.)	7.6	**6.5**	10.9	8.8	28.9	16.1
yeast (10 cl.)	13.5	**10.6**	25.0	17.9	39.6	19.9
soybean-large (19 cl.)	17.2	**16.3**	20.9	18.8	33.2	26.9
primary-tumor (21 cl.)	3.6	**3.4**	9.7	8.5	28.5	16.2
audiology (24 cl.)	5.7	**5.4**	10.5	8.9	19.3	14.5

not contribute positively to the AUC (as estimated on the prune set) actually lead to any improvements. A slight improvement was observed for ordered rule sets (when using lift as an exclusion criterion), while no improvement was observed for unordered rule sets, other than that the size of the resulting rule set was reduced.

There are number of open questions that deserve further study. One is why post-processing of unordered rule sets did not lead to any improvements. This includes explaining why the estimated AUC often mislead the post-processing procedure to remove rules, that actually would have been beneficial. More robust estimation methods or post-processing criteria may be helpful. Another direction for future work is to to investigate alternative ways of post-processing the generated rules, e.g. by also considering replacement rules.

Precision and lift were estimated by the relative frequencies in the prune sets, while Laplace correction was used to form the class distributions of each rule. It has not been investigated whether using some correction when estimating the former two actually would lead to any improvement, and whether some alternative to Laplace, e.g., the m-estimate (Cestnik & Bratko, 1991) would be beneficial w.r.t. AUC.

Acknowledgements

Thanks to the anonymous reviewers for useful comments and to the Ljubljana Oncology Institute, Slovenia for providing the breast cancer, lymphography and primary tumor data sets.

References

Blake, C., & Merz, C. (1998). UCI repository of machine learning databases.

Boström, H. (2004). Pruning and exclusion criteria for unordered incremental reduced error pruning. *Proceedings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning* (pp. 17–29).

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Cestnik, B., & Bratko, I. (1991). On estimating probabilities in tree pruning. *Proceedings of the Fifth European Working Session on Learning* (pp. 151–163). Springer-Verlag.

Cohen, W. (1995). Fast effective rule induction. *Proc. of the 12th International Conference on Machine Learning* (pp. 115–123). Morgan Kaufmann.

Dain, O., Cunningham, R., & Boyer, S. (2004). Irep++, a faster rule learning algorithm. *Proceedings of the Fourth SIAM International Conference on Data Mining.*

Fawcett, T. (2001). Using rule sets to maximize roc performance. *Proceedings of the IEEE International Conference on Data Mining* (pp. 131–138). IEEE Computer Society.

Fawcett, T. (2003). *Roc graphs: Notes and practical considerations for data mining researchers* (Technical Report). HP Laboratories, Palo Alto.

Frank, E., & Witten, I. (1998). Generating accurate rule sets without global optimization. *Proc. 15th International Conf. on Machine Learning* (pp. 144–151). Morgan Kaufmann, San Francisco, CA.

Fürnkranz, J., & Flach, P. (2003). An analysis of rule evaluation metrics. *Proceedings of the 20th International Machine Learning Conference*. Morgan Kaufmann.

Fürnkranz, J., & Flach, P. (2004). An analysis of stopping and filtering criteria for rule learning. *Proceedings of the 15th European Machine Learning Conference* (pp. 123–133). Springer-Verlag.

Fürnkranz, J., & Flach, P. (2005). Roc 'n' rule learning – towards a better understanding of covering algorithms. *Machine Learning*, *58*, 39–77.

Fürnkranz, J., & Widmer, G. (1994). Incremental reduced error pruning. *Proceedings of the 11th International Conference on Machine Learning* (pp. 70–77). Morgan Kaufmann.

Lavrac, N., Kavsek, B., Flach, P. A., & Todorovski, L. (2004). Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, *5*, 153–188.

Prati, R., & Flach, P. (2004). Roccer: A roc convex hull rule learning algorithm. *Proceedings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning* (pp. 144–153).

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the 15th International Conference on Machine Learning* (pp. 445–453).

Rivest, R. (1987). Learning decision lists. *Machine Learning, 2(3), 229-246.*