PRIE: A System for Generating Rulelists to Maximize ROC Performance

Tom Fawcett (tfawcett@acm.org)
Stanford Center for the Study of Language and Information

March 1, 2007

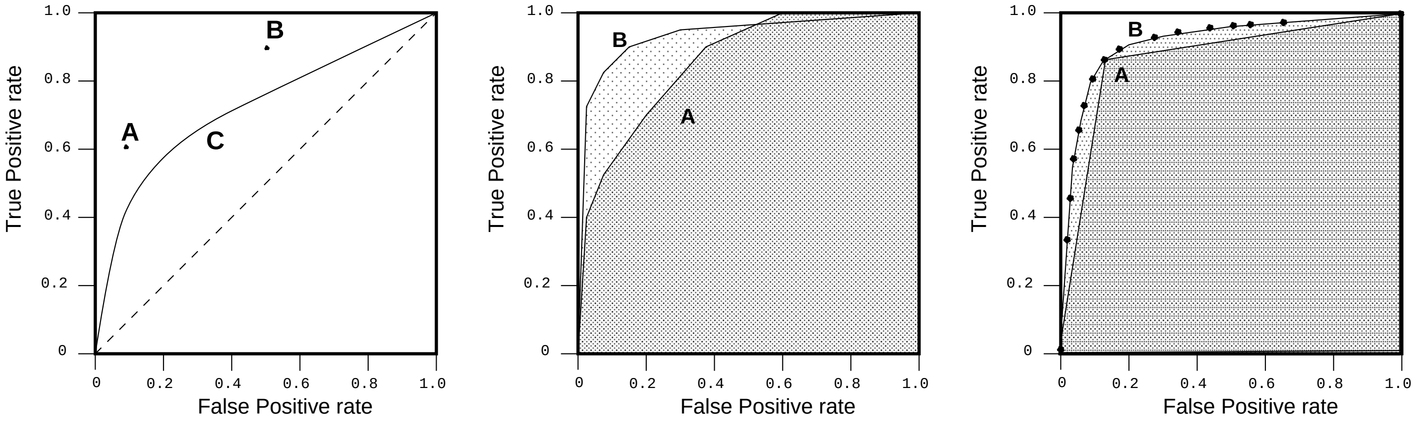
Abstract. Rules are commonly used for classification because they are modular, intelligible and easy to learn. Existing work in classification rule learning assumes the goal is to produce categorical classifications to maximize classification accuracy. Recent work in machine learning has pointed out the limitations of classification accuracy: when class distributions are skewed, or error costs are unequal, an accuracy maximizing classifier can perform poorly. This paper presents a method for learning rules directly from ROC space when the goal is to maximize the area under the ROC curve (AUC). Basic principles from rule learning and computational geometry are used to focus search for promising rule combinations. The result is a system that can learn intelligible rulelists with good ROC performance.

Keywords: Classification, ROC analysis, rule learning, cost-sensitive learning

1. Introduction

One concern of utility-based data mining is the ability to make cost-effective classification decisions. In order to make such decisions, a classifier should be able to produce, given an unseen instance, not just a hard classification (*i.e.* a class name) but also an instance score or probability that the instance belongs the class. With such an estimate, a classifier can use it, along with error costs, to determine the most cost-effective classification. Previous work has shown that methods that can produce instances scores can be used to make cost-sensitive decisions, either directly by converting them into proper probabilities using a calibration technique (Zadrozny and Elkan, 2001; Niculescu-Mizil and Caruana, 2005), or indirectly by using a technique such as the ROC Convex Hull (Provost and Fawcett, 1998a; Provost and Fawcett, 2001).

This goal calls for classification methods that are good at producing accurate instance scores; in other words, methods that maximize ROC performance. In the data mining community, researchers in classification are starting to look at the area under the ROC curve (AUC) instead of accuracy as an evaluation measure, and designing methods to maximize the AUC. Such work has been pursued in data mining, but most of it has involved methods whose concepts are difficult to interpret, such as ensembles of classifiers, support vectors machines, etc. Probabilistic classification has been investigated with other model classes such as neural networks (Santini and Bimbo, 1995) and ensembles of decision trees (Provost and Domingos, 2001; Zadrozny and



a. b. c.

Figure 1. ROC graphs and area under ROC curves.

Elkan, 2001). These models tend to be much more complex than rule sets. They attain good ROC performance but compromise intelligibility. They do not have rules' appealing properties of modularity and intelligibility[1].

Rules are commonly used in data mining because of several desirable properties: they are simple, intuitive, modular, and straightforward to generate from data. But existing methods strive to optimize classification decisions, usually by maximizing accuracy (or equivalently, minimizing error rate) on a training set. They usually try to construct small, compact rule sets while achieving high accuracy. Others have pointed out that accuracy is a poor metric to optimize (Provost et al., 1998; Ling et al., 2003), so accuracy-maximizing methods may be a poor choice for producing scoring classifiers—that is, classifiers with good ROC performance.

Therefore, an open question in data mining is how to generate rules to produce reliable instance scores. Previous work (Fawcett, 2001) showed how rules could be combined in various ways to maximize ROC performance, but the rules used in that work were generated by techniques that were not intended to maximize such performance. Most techniques attempt to maximize accuracy, and techniques that maximize accuracy do not always perform well in ROC space (Provost et al., 1998).

This paper may be seen as an extension to that prior work. We introduce a rule learning system called PRIE which is designed to maximize ROC per-

[1] *Modularity* means that each rule is local and a decision on an instance is made by only a single rule. Ensembles lack modularity because multiple classifiers may participate in deciding an instance's class, and their votes may combine in unintuitive ways. *Intelligibility* means that the classifier is reasonably understandable to those who are not proficient in data mining.

formance. PRIE exploits the structure of ROC space to guide generation of new rules. PRIE has several attractive features:

1. Because PRIE maximizes ROC performance, it naturally handles skewed datasets.

2. PRIE is able to handle multiple classes. It will attempt to optimize the combined AUC for any number of classes simultaneously.

3. PRIE's output is a single rulelist and thus is relatively intelligible and modular. To use this rulelist on a new unseen instance, the rules are evaluated sequentially and the first one matching determines the class and probability. The advantage of a rulelist over a rule set in which all rules may vote is that the latter is not modular; determining which rules were responsible for a scoring decision may be difficult and unintuitive. With a rulelist, only a single rule is responsible for deciding a score.

4. Because PRIE uses a rulelist whose rules are ordered decreasing by class likelihood, the rulelist may be used naturally with the ROCCH method (Provost and Fawcett, 2001). In use, if operating conditions (class skew and relative error costs) are known, the rulelist can be truncated to eliminate rules that will never affect a classification decision.

5. PRIE handles numerical attributes naturally, using the ROC curve implicitly to identify promising discretizations. Other classification models may discretize variables in a preprocessing pass or may use techniques unrelated to model construction. PRIE considers every discretization of a continuous attribute to comprise a separate point in ROC space, and handles these the same as any other discrete attribute.

6. PRIE can handle set-valued attributes (Cohen, 1996), in which an attribute of an instance may take on a set of discrete values instead of a single one. Such features are useful, for example, in text classification domains in which the set may represent the "bag of words" of a text document.

PRIE is unusual in that it uses basic principles from rule learning and computational geometry to focus search for promising rule combinations. The result is a system that can learn intelligible rulelists with high AUC scores.

The remainder of the paper is organized as follows. Section 2 section reviews the basics of ROC graphs and the ROC convex hull. These principles and equations form the basis of PRIE's operation. Section 3 describes PRIE and the principles behind its operation. Section 4 describes the sets of experiments, and their results, intended to validate PRIE's approach. The final section concludes and describes areas for future work.

2. ROC graphs

Prior to explaining the PRIE system we briefly review the theory of Receiver Operating Characteristics (ROC) graphs, upon which PRIE is based. ROC graphs have long been used in signal detection theory to depict the trade-off between hit rates and false alarm rates of classifiers (Egan, 1975; Swets et al., 2000). ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems (Swets, 1988)[2].

A discrete classifier applied to a test set generates two important statistics. The **True Positive rate** (also called hit rate and recall) of a classifier is:

$$\text{TP rate} \approx \frac{\text{positives correctly classified}}{\text{total positives}}$$

The **False Positive rate** (also called false alarm rate) of the classifier is:

$$\text{FP rate} \approx \frac{\text{negatives incorrectly classified}}{\text{total negatives}}$$

On an ROC graph, TP rate is plotted on the Y axis and FP rate is plotted on the X axis.

A “hard” classifier—one that outputs only a class label—produces an *(FP rate , TP rate)* pair, so it corresponds to a single point in ROC space. Classifiers A and B in Figure 1a are discrete classifiers.

Several points in ROC space are useful to note. The lower left point $(0, 0)$ represents the strategy of never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point $(1, 1)$. Any classifier that randomly guesses the class will produce performance on the diagonal line $y = x$. The point $(0, 1)$ represents perfect classification. Informally, one point in ROC space is better than another if it is to the northwest (TP rate is higher, FP rate is lower, or both) of the first.

The diagonal line $y = x$ represents the strategy of randomly guessing a class, and any classifier that appears in the lower right triangle performs worse than random guessing. This triangle is therefore usually empty.

A *ranking* or *scoring* classifier may be thresholded to produce a binary classifier: if the classifier output is above the threshold, the classifier produces a **Y**, else a **N**. Each threshold value produces a different point in ROC space, so varying the threshold from $-\infty$ to $+\infty$ produces a curve through ROC space. An ROC curve illustrates the error trade-offs available with a given classifier. Figure 1a shows the curve of a probabilistic classifier, C, in ROC

[2] Much further information on ROC graphs and analysis is available, such as Fawcett’s (2006) overview article and tutorial notes by Peter Flach (2004).

space. A more thorough discussion of ROC curves may be found elsewhere (Provost and Fawcett, 2001; Fawcett, 2006).

2.1. Area under the ROC curve

To compare classifiers we often want to reduce ROC performance to a single number representing average expected performance. A common method is to calculate the area under the ROC curve, abbreviated **AUC** (Bradley, 1997; Hanley and McNeil, 1982). Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5.

The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks (Hanley and McNeil, 1982). Figure 1b shows the areas under two ROC curves, A and B. B has greater area and therefore better average performance, though A performs better than B in the upper right region of ROC space.

ROC analysis is commonly used for two classes, but it has been extended to multiple classes (Srinivasan, 1999). Unfortunately, visualizing the result for more than two classes is non-intuitive. In practice, n classes are commonly handled by producing n different ROC graphs. Let C be the set of all classes. ROC graph i plots the classification performance using class c_i as the positive class and all other classes $c_{j \neq i} \in C$ as the negative class. Each such graph yields an AUC area.

For a single probabilistic classifier this produces n separate curves with n different AUC values. The AUC values can be combined into a single weighted sum where the weight of each class c_i is proportional to the class's prevalence in the training set:

$$AUC_{total} = \sum_{c_i \in C} AUC(c_i) \cdot p(c_i)$$

2.2. The ROC convex hull

Provost and Fawcett (1998b; 2001) have shown some important properties of the convex hull of a set of points in ROC space. The details are beyond the scope of this article, but a classifier is potentially optimal if and only if it lies on the convex hull of the set of points in ROC space. We call the convex hull of the set of points in ROC space the *ROC convex hull* (ROCCH) of the corresponding set of classifiers.

This ROCCH formulation has a number of useful implications. In use, since only the classifiers on the convex hull are potentially optimal, no others need be retained. Furthermore, in rule learning, the convex hull can be used to efficiently keep track of the best individual rules yet found. This may in turn be used to guide generation of new rules, as section 3.2 explains.

2.3. Rulelists and the ROCCH

PRIE constructs a rulelist: an ordered list of rules which, in use, will be tested one at a time against a new instance. The rules are tested in order until one of them matches or until the rulelist is exhausted. The score then emitted for the instance will be based on the rule's statistics.

Each rule is a conjunction of conditions, the satisfaction of which implies membership in a class. For simplicity, consider a two-class problem with classes **p** and **n**. An example of a simple rule and some of its performance statistics is:

$$\mathbf{x_1} \wedge \mathbf{x_2} \wedge \mathbf{x_3} \longrightarrow \mathbf{p}$$

TP=15, P=100, TPrate=.15
FP=2, N=200, FPrate=.01

The second line specifies that within the dataset, 15 **p** examples satisfy $x_1 \wedge x_2 \wedge x_3$ (True Positives). There are 100 **p** examples altogether, yielding a true positive rate (TPrate) of .15. The rule matches two **n** examples (False Positives). There are 200 **n** examples altogether, yielding a false positive rate (FPrate) of .01.

Rulelists have a natural correspondence to ROC points and to the ROCCH. Rules on a rulelist are naturally ordered by likelihood. Each rule in a rulelist corresponds to a point on an ROC hull. The probability of a given rule corresponds to the conditional probability that its conditions are true, given that all the rules preceding it are false. Figure 2 illustrates this relationship. At the bottom is a list of 11 rules learned from the UCI "car" domain with two classes, `acc` (positive class) and `unacc` (negative class). Each rule has two lines: its number and the text (antecedents and consequent) of the rule on the first line, and its local statistics on the second. The local statistics describe how many positive (P) and negative (N) examples matched against it in the training set, and from those how many were correct (TP) and incorrect (FP).

At the top of figure 2 is an ROC graph illustrating the rulelist's performance. Each point is labeled by the rule to which it corresponds. More precisely, each point corresponds to a *prefix* of the rulelist. For example, the point labeled 7 on the ROC graph represents the performance that would be derived if only the first seven rules were tested on an instance. Its lies at (.06, 0.5) in ROC space. The TP rate can be determined by adding the TPs of rules 1-7 (235) and dividing by the number of positives (P=467). The FP rate can be determined by adding the FPs of rules 1-7 (13) and dividing by the number of negatives (N=220). If an example were to be classified by this rulelist, and it failed to match rules 1-6 but matched rule 7, the example's score would be given as $235/(235+13) \approx 0.95$. If an example failed to match any rule in the rulelist, it would be given the score of $467/(220+467) \approx 0.68$, which is the prior probability of the `acc` class.

This rulelist may thus be seen as a single classifier composed of 11 pieces. It can be "thresholded" by cutting off evaluation at each of these 11 points, and thus yields an ROC curve with 11 points (plus the two endpoints at (0,0) and (1,1)). PRIE's goal is to construct such a rulelist with the best possible ROC performance. The way it goes about generating rules and assembling a rulelist from them is described next.

3. PRIE

PRIE is a separate-and-conquer rule learner (Fürnkranz, 1999). It may be thought of as comprising an outer loop and an inner loop. The outer loop iteratively chooses the best existing rule and adds it to the end of the rulelist. As rules are added to the rulelist, the instances they match are removed from consideration ("covered"). The inner loop develops new rules for the outer loop to extract. What makes PRIE different is that the outer loop covers examples based on ROC performance, and the inner loop uses ROC performance to suggest new promising rules.

Rule generation is accomplished by two interleaved processes. One process calculates the convex hull in ROC space, then repeatedly tries to generate rules that independently extend the hull. The outer process extracts the leftmost rule of the convex hull, inserts the rule at the end of a rulelist, then removes those instances matched by the rule, and re-invokes the first process. The entire rule generation method may be seen as a process that "shrinks" ROC space iteratively then attempts to find the best (most conservative) individual rule in that space.

3.1. Rule selection and instance covering (outer loop)

The inner loop is discussed in more detail below. The outer loop comprises the higher level actions of the rulelist generation process. PRIE maintains a global rulelist, initially empty, and adds successive rules to the end of it. Algorithm 1 is a basic description of the outer loop process.

PRIE maintains one ROC space for each class under consideration. Each space is a *class reference* ROC (Fawcett, 2006), *i.e.* if C is the set of classes, for every class $c_i \in C$ there is a ROC space ROC$_i$ such that the positive class $\mathbf{p} = c_i$ and the negative class $\mathbf{n} = \{c_j \in C | j \neq i\}$. Each of these ROC spaces is initialized to be simply a line connecting $(0, 0)$ and $(1, 1)$.

Algorithm 1 Rule generation algorithm

Given: Alarms: Set of alarm tuples $\langle p, \alpha, i\rangle$ where: p: probability calculated by the classifier
Output: RL: rulelist

```
$RL \leftarrow \langle\rangle$;
for $c \in$ Classes do
    Hull$_c \leftarrow \langle(0,0),(1,1)\rangle$;
end for
Create initial rules for discrete and set-valued attributes
Create initial rules for continuous attributes
while ... do
    for $c \in$ Classes do
        Develop_rules_for_ROC($c$)
    end for
    $R \leftarrow$ Extract_best_rule
    Add $R$ to end of $RL$;
end while
```

PRIE then generates singleton rules for attributes. For each discrete-valued attribute a_j with a value $v_{j,k}$ it creates a rule:

$$(a_j = v_{j,k}) \rightarrow c_i$$

For each set-valued attribute a_j with a value $v_{j,k}$, PRIE creates a rule:

$$(v_{j,k} \in a_j) \rightarrow c_i$$

For each continuous-valued attribute a_j with a value $v_{j,k}$, PRIE creates two rules:

$$\begin{array}{rcl} (a_j < v_{j,k}) & \rightarrow & c_i \\ (a_j \geq v_{j,k}) & \rightarrow & c_i \end{array}$$

For all these singleton rules, PRIE calculates the the rule statistics (TP and FP and their corresponding rates), then calculates their position in ROC$_i$ space and calculates the convex hull[3].

PRIE then enters its outer loop. Given each of the i spaces ROC$_i$ attempts to develop each space as described in section 3.2. Once this is done, it extracts the single highest likelihood rule from all of the convex hull of each class. The highest likelihood rule corresponds to the endpoint of the leftmost segment of the hull. If there are multiple vertical segments, the one with the

[3] As an implementation note, PRIE keeps bit vectors with all rules, so calculating and updating rule statistics is very fast. Calculating and updating convex hulls is also very efficient.

highest TP rate chosen. In figure 2, rule 1 would be considered the highest likelihood rule. This rule is added to the end of the global rulelist, and all of the ROC$_i$ spaces are adjusted accordingly. Finally, the convex hulls are recomputed.

Conceptually, this iterative extraction of rules may be thought of as excising successive lower left (L-shaped) portions of ROC space. Of course, extracting a rule affects all the ROC spaces so they must all be recomputed.

This process continues until all the ROC spaces are exhausted, *i.e.* they are reduced to a single line connecting $(0, 0)$ to $(1, 1)$. At this point, the global rulelist is output and PRIE terminates.

3.2. Constraining rule generation

PRIE's inner loop is responsible for generating new rules. It does this by combining existing rules, by conjoining their conditions. It operates independently over each of the ROC spaces it maintains.

One key issue for any separate-and-cover rule learning algorithm is how to generate promising new rules. Considering every possible combination of features is NP-complete so rule learning systems must use heuristics to guide their search. PRIE attempts to determine, for each of its ROC spaces, which combinations of existing rules are likely to extend the convex hull. If the new rule would be unlikely to extend the ROC convex hull, it can be eliminated from consideration. In fact, geometric properties of ROC space can be used to eliminate large numbers of rules from consideration. Before explaining how PRIE's inner loop works in section 3.3, we describe the principles underlying its process for deciding which rules to combine.

If we conjoin the conditions of two rules α and β to create a new rule γ, where will the new rule lie in ROC space? The answer depends on the intersections among the TP and FP sets of α and β; that is, upon their attribute interactions. The next two sections derive estimates for the performance of γ based on different assumptions.

3.2.1. *Independence*

Let TPrate$_\gamma$ be the expected true positive rate of γ and let x be an instance in the true positive set of γ. Then:

$$\begin{aligned} tpr_\gamma &\approx p(x \in TP_\gamma) \\ &\approx p(x \in TP_\alpha \wedge x \in TP_\beta) \end{aligned}$$

A useful simplification is to assume that the rules are conditionally independent, so the probability of an instance matching one rule is independent of the probability of it matching the other. If we assume independence of α and β,

$$\mathrm{tpr}_\gamma \approx p(x \in TP_\alpha) \cdot p(x \in TP_\beta)$$

$$\begin{aligned} &\approx \frac{\mid TP_\alpha \mid}{\mid P \mid} \cdot \frac{\mid TP_\beta \mid}{\mid P \mid} \\ &\approx \text{tpr}_\alpha \cdot \text{tpr}_\beta \end{aligned}$$

A similar derivation can be done for fpr_γ. Thus, the conjunction of two rules α and β can be expected to lie in ROC space at:

$$\gamma_{\text{I}} = \left(\text{fpr}_\alpha \cdot \text{fpr}_\beta \ , \ \text{tpr}_\alpha \cdot \text{tpr}_\beta\right) \tag{1}$$

Figure 3 shows an ROC graph of two rules, α at (0.6, 0.85) and β at (0.35, 0.70). Rule γ_{I} is plotted at (0.21, 0.595) as calculated by equation 1. We denote this point γ_{I} to emphasize that this performance estimate depends on independence of rule conditions.

3.2.2. *Optimism*

We may want to relax the independence assumption and allow for optimism. What is the *best possible* rule (*i.e.* the rule with best ROC performance) that could result from intersecting α and β? This would be one in which the tp rate was the highest possible and the fp rate was the lowest possible. Because $\gamma = \alpha \wedge \beta$, any example matching γ must match α and β, so the tp rate of γ cannot exceed either of these. The upper bound is:

$$\text{tpr}_\gamma = min(\text{tpr}_\alpha, \text{tpr}_\beta) \tag{2}$$

Calculating the lowest possible FP rate is somewhat more involved. If fpr_α and fpr_β sum to less than one, the lower bound is zero since it is possible for the intersection to be empty; otherwise their FP sets may have a non-null intersection. In general, the expression for this lower bound is:

$$\begin{aligned} \text{fpr}_\gamma &= max(\left[1 - ((1 - \text{fpr}_\alpha) + (1 - \text{fpr}_\beta))\right], 0) \\ &= max(\left[1 - 1 + \text{fpr}_\alpha - 1 + \text{fpr}_\beta)\right], 0) \\ &= max(\text{fpr}_\alpha + \text{fpr}_\beta - 1, 0) \end{aligned} \tag{3}$$

Combining equations 2 and 3, the best possible rule, denoted γ_{O}, would lie at:

$$(max(\text{fpr}_\alpha + \text{fpr}_\beta - 1, 0), min(\text{tpr}_\alpha, \text{tpr}_\beta)) \tag{4}$$

Figure 3 shows γ_{O} of α and β, as calculated by equation 4.

3.2.3. *Compromise*

We have two estimates of the performance of γ, one based on a neutral assumption of attribute independence and the other based on optimal attribute interactions. We can control how optimistic the rule learning process should be in its estimates by allowing a compromise between these two. We introduce a parameter OPTIMISM which allows us to interpolate linearly between these two points. For OPTIMISM=0, the independent estimate γ_{I} is used; for OPTIMISM=1, the most optimistic estimate γ_{O} is used; and for $0 <$ OPTIMISM < 1 the point will lie between the two estimates. We denote this final estimate γ_{E}. Figure 3 shows γ_{E} as a simple linear interpolation between γ_{I} and γ_{O} with OPTIMISM=0.5.

OPTIMISM may be seen as a parameter that controls search in rule induction with respect to the ROC convex hull. With OPTIMISM=1, only rules that could not possibly extend the convex hull will be pruned. But note that this parameter only controls generation of rule candidates—regardless of its setting, if a new rule does not extend the convex hull it will not be a candidate for inclusion in the rulelist.

We can exploit these estimates to filter unproductive rule combinations. If the estimated location of a new point in ROC space does not extend the convex hull, we need not consider creating the conjoined rule. We can thus filter from consideration any rule pair whose conjunction, when plotted in ROC space, does not lie outside the current hull.

3.2.4. *Generation*

In fact, we can do better than this. We can use these conditions to constrain rule generation by removing from consideration rule pairs that are unlikely to produce worthwhile new rules. Given a rule α and the hull segments H, we can create a corresponding set of boundary segments B. If a second rule β does not lie above B then $\alpha \wedge \beta$ will not lie above the hull line H, and this combination need not be considered.

Given a hull line segment $h_i \in H$ expressed in the line equation $y = m_{hull} \cdot x + b_{hull}$, points lying above this line must satisfy the inequality:

$$y > m_{hull} \cdot x + b_{hull}$$

Let rule α have true positive rate tp_α and false positive rate fp_α. By equation 1, for a second rule r_β to produce a rule above this hull line when conjoined with r_α, it must satisfy the inequality:

$$\begin{aligned} y \cdot tp_\alpha &> m_{hull} \cdot x \cdot fp_\alpha + b_{hull} && (5) \\ tp_\beta \cdot tp_\alpha &> m_{hull} \cdot fp_\beta \cdot fp_\alpha + b_{hull} && (6) \\ tp_\beta &> \frac{fp_\alpha \cdot m_{hull}}{tp_\alpha} \cdot fp_\beta + \frac{b_{hull}}{tp_\alpha} && (7) \end{aligned}$$

This inequality yields a "constraint line" below which no rule should be considered for combining with r_α to extend segment h_i.

This is based on the independence assumption. A similar derivation can be done for the optimism assumption, and for combining them with an OPTIMISM parameter.

Figure 4 shows an example of this. Assume there is a hull segment from (.1, .4) to (.3, .6), shown as a solid line in the figure. We have a rule labeled **a** at (.35, .6), and we want to determine which rules we should combine with **a** to extend the hull segment. Using the inequality in equation 7 we can derive the equation of a line:

$$y = 0.58 \cdot x + 0.5$$

which forms the dashed constraint line shown in figure 4. All classifiers whose ROC points lie above this line, when conjoined with rule **a**, may be expected to extend the hull segment shown. This constraint line may be used to quickly eliminate many candidates from consideration; any rule lying below this line will not extend the hull segment. Thus, rule **c** is a promising choice for conjoining with **a**, but **b** is not.

3.3. Rule generation (inner loop)

PRIE's inner loop is responsible for generating new rules by conjoining the conditions of existing rules. PRIE operates independently over each of the ROC spaces it maintains, using the techniques ideas of the last four sections to develop rules for each space.

Each hull segment is examined in turn and rule pairs are considered whose conjunction might extend the hull; that is, whose conjunction is predicted to lie to the northwest of the segment, in the portion of ROC space as yet uncovered. PRIE uses the principles in section 3.2, and specifically equation 7, to constrain its search for combinations.

Prior to generating the new rule, the antecedents of the constituent rules are checked for tests that might render the conjunction useless. For example, if one rule contains a test ($a_i = v$) and the other contains a test ($a_i = w$) for two distinct values v and w, the two would not be combined. Another such test is for conditions like $a_i < v$ and $a_i > v$ being combined in the conjoined rule. In addition, duplicate tests within the antecedents are removed.

4. Empirical Validation

In this section, we compare PRIE's performance against various other induction methods. The domains for evaluation are described and summarized in table I. In all of the experiments below, PRIE was used with an OPTIMISM value of 0.5. These results are based on 10-fold cross-validation.

4.1. PRIE vs error minimization methods

As a baseline, we compare PRIE against error minimization methods. Previous work (Fawcett, 2001) showed that rules used for scoring performed better on maximizing the AUC than the same rules with just their hard classifications. This is not a surprising result, since hard classifications produce a single ROC point, whereas scores can produce a set of points (a curve) in ROC space. Figure 1c illustrates this. However, it remains to be seen whether a system like PRIE, designed to maximize AUC performance, can perform better than error minimizing rules interpreted probabilistically.

Previous work (Fawcett, 2001) evaluated the performance of rule sets for maximizing the AUC, but there are two significant differences with that work. The goal of that work was to use standard error minimizing (*i.e.* accuracy maximizing) rule generation techniques to see how they could best be used to maximize AUC. In addition, that work experimented with a variety of rule resolution techniques—methods for combining the responses of multiple matching rules—to determine which was most effective in maximizing AUC. That paper concluded that a weighted voting scheme (called WVOTE) performed best with the error-minimizing rules.

In this section we compare the best results from that paper with PRIE's results. Table II shows the AUC results of PRIE and the AUC results of C4.5 using the WVOTE protocol. PRIE's performance is comparable to the weighted voting of accuracy-maximizing rulesets.

4.2. PRIE AND OTHER RULE-LEARNING METHODS

ROCCER (Prati and Flach, 2005) is an AUC-maximizing rule learning method. As such, it is a competitor with PRIE. ROCCER uses an association rule learning method to generate rules. It attempts to insert each rule into a rulelist, testing the resulting rulelist's ROC classification performance. If the new rule improves the performance, it is retained, else it is discarded.

Table III shows ROCCER's performance results compared to PRIE's on fifteen domains[4] reported by Prati and Flach (2005). The table shows that PRIE performs as well or better on these domains than ROCCER.

As another comparison, we consider the work of Barakat and Bradley (2006). They point out that, because support vector machines are "black box" classifiers, some work has been done on generating rulesets from SVMs. The goal is to obtain a classifier that retains much of the performance of the SVM with the intelligibility of a rule set. Barakat and Bradley specifically investigated the AUC performance of the SVMs and the rulesets derived from them. Though it requires two steps—generating an SVM from data, then generating a ruleset from the SVM—it achieves the same ends and thus is a competing approach to PRIE. It is worth comparing their results to PRIE's.

Barakat and Bradley (2006) compared their approach on only four domains. Their results are shown in table IV. After the domain name is the AUC value obtained from the original SVM trained on the data, followed by the the AUC obtained from the ruleset generated from the SVM. The fourth

[4] Satimage is not included because the contributors of that domain specified that it should not be used in cross-validation, upon which these results are based.

column shows the results of PRIE on the domain. These data show that PRIE is comparable to the rules extracted from an SVM on these domains.

5. Conclusions and Future Work

PRIE is an induction technique that is able to generate rules directly from ROC space when the goal is to maximize the area under the ROC curve. PRIE uses basic principles from rule learning and computational geometry are used to focus search for promising rule combinations. The result is a system that can learn intelligible rulelists with good ROC performance. Empirical results show that PRIE's rules compare favorably with those of other similar induction techniques.

As mentioned earlier, PRIE discretizes continuous attributes by generating discrete singleton rules from every cutpoint. It then "drops" these rules into ROC space and lets them compete with others to form new rules. Thus PRIE implicitly uses ROC space to determine effective cutpoints; the best discretizations will end up on the convex hull. To the best of our knowledge, this is the first time ROC space has been used as a discretization technique in data mining, though it is used here in conjunction with rule learning. An area of future work is to investigate this discretization method independently, against other discretization techniques, to determine the extent to which the attribute discretization is a factor in rule-learning performance.

In terms of system control, PRIE may be seen as complementary to the ROCCER system (Prati and Flach, 2005). ROCCER takes rules generated from an association rule learner and examines their ROC performance when placed into a rulelist. Thus ROCCER may be seen as generating from the rule lattice space and filtering in ROC space. On the other hand, PRIE generates rules directly from ROC space, but it must employ post-processing tests (described in the last paragraph of section 3.3) to filter obviously poor combinations. In a sense, these tests are incorporating information about the structure of rule space that an association rule learner employs directly. One interesting area of future work is to attempt a more concerted integration of these spaces, so that both rule space information and ROC performance expectations are used in parallel to maximum effect.

Acknowledgements

Ronaldo Prati provided much help with the evaluation domains of ROCCER.

Much open source software was used in this work. I wish to thank the authors and maintainers of XEmacs, TEX, LATEX, Perl and its many user-contributed packages, and the Free Software Foundation's GNU Project.

References

Barakat, N. and A. Bradley: 2006, 'Rule Extraction from Support Vector Machines: Measuring the Explanation Capability Using the Area under the ROC Curve'. In: *ICPR 2006. 18th International Conference on Pattern Recognition*, Vol. 2. pp. 812–815, IEEE Press.

Boström, H.: 2005, 'Maximizing the Area under the ROC Curve using Incremental Reduced Error Pruning'. In: *Proceedings of the ICML-2005 Workshop on ROC Analysis in Machine Learning*.

Bradley, A. P.: 1997, 'The use of the area under the ROC curve in the evaluation of machine learning algorithms'. *Pattern Recognition* **30**(7), 1145–1159.

Cohen, W. W.: 1996, 'Learning Trees and Rules with Set-Valued Features'. In: *AAAI/IAAI, Vol. 1*. pp. 709–716.

Egan, J. P.: 1975, *Signal Detection Theory and ROC Analysis*, Series in Cognitition and Perception. New York: Academic Press.

Fawcett, T.: 2001, 'Using Rule Sets to Maximize ROC Performance'. In: *Proceedings of the IEEE International Conference on Data Mining (ICDM-2001)*. pp. 131–138.

Fawcett, T.: 2006, 'An Introduction to ROC Analysis'. *Pattern Recognition Letters* **27**(8), 882–891.

Ferri, C., P. Flach, and J. Hernàndez-Orallo: 2002, 'Learning decision trees using the area under the ROC curve'. In: *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*. pp. 139–146.

Flach, P.: 2004, 'The Many Faces of ROC Analysis in Machine Learning'. ICML-04 Tutorial. Notes available from `http://www.cs.bris.ac.uk/{\~~}flach/ICML04tutorial/index.html`.

Fürnkranz, J.: 1999, 'Separate-and-Conquer Rule Learning'. *Artificial Intelligence Review* **13**(1), 3–54.

Hanley, J. A. and B. J. McNeil: 1982, 'The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve'. *Radiology* **143**, 29–36.

Ling, C. X., J. Huang, and H. Zhang: 2003, 'AUC: A Better Measure than Accuracy in Comparing Learning Algorithms'. In: *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence*. pp. 329–341, Springer.

Niculescu-Mizil, A. and R. Caruana: 2005, 'Predicting Good Probabilities With Supervised Learning'. In: *Proc. 22nd International Conference on Machine Learning (ICML'05)*.

Prati, R. and P. Flach: 2005, 'ROCCER: An Algorithm for Rule Learning Based on ROC Analysis'. In: *IJCAI 2005*. pp. 823–828.

Provost, F. and P. Domingos: 2001, 'Well-trained PETs: Improving Probability Estimation Trees'. CeDER Working Paper #IS-00-04, Stern School of Business, New York University, NY, NY 10012.

Provost, F. and T. Fawcett: 1998a, 'Robust Classification Systems for Imprecise Environments'. In: *Proceedings of AAAI-98*. pp. 706–713, Menlo Park, CA: AAAI Press.

Provost, F. and T. Fawcett: 1998b, 'Robust classification systems for imprecise environments'. In: *Proceedings of AAAI-98*. Menlo Park, CA, pp. 706–713.

Provost, F. and T. Fawcett: 2001, 'Robust Classification for Imprecise Environments'. *Machine Learning* **42**(3), 203–231.

Provost, F., T. Fawcett, and R. Kohavi: 1998, 'The Case Against Accuracy Estimation for Comparing Induction Algorithms'. In: J. Shavlik (ed.): *Proceedings of ICML-98*. San Francisco, CA, pp. 445–453. Available: `http://www.purl.org/NET/tfawcett/papers/ICML98-final.ps.gz`.

Rakotomamonjy, A.: 2004, 'Support Vector Machines and Area under ROC curve'. Technical report, University of Rouen.

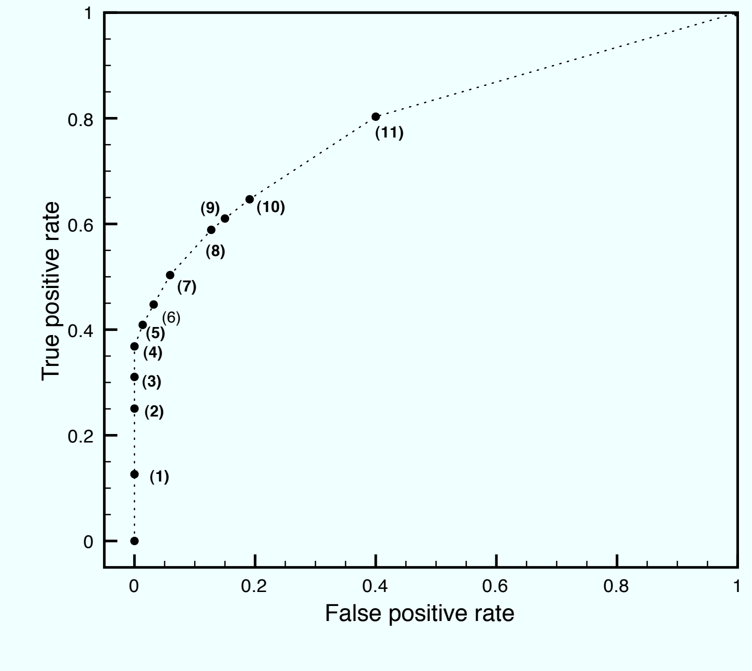
Santini, S. and D. A. Bimbo: 1995, ‘Recurrent neural networks can be trained to be maximum a posteriori probability classifiers’. *Neural Networks* **8**(1), 25–29.

Srinivasan, A.: 1999, ‘Note on the location of optimal classifiers in n-dimensional ROC space’. Technical Report PRG-TR-2-99, Oxford University Computing Laboratory, Oxford, England. Available: `citeseer.nj.nec.com/srinivasan99note.html`.

Swets, J.: 1988, ‘Measuring the accuracy of diagnostic systems’. *Science* **240**, 1285–1293.

Swets, J. A., R. M. Dawes, and J. Monahan: 2000, ‘Better Decisions through Science’. *Scientific American* **283**, 82–87.

Zadrozny, B. and C. Elkan: 2001, ‘Obtaining calibrated probability estimates from decision trees and naive bayesian classiers’. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. pp. 609–616.



N	Conditions
1	(maint = med) AND (lug_boot = big) → acc TP: 59 FP: 0 P: 467 N: 220
2	(maint = low) AND (lug_boot = big) → acc TP: 58 FP: 0 P: 408 N: 220
3	(safety = high) AND (maint = low) AND (lug_boot = med) → acc TP: 28 FP: 0 P: 350 N: 220
4	(maint = med) AND (lug_boot = med) AND (safety = high) → acc TP: 27 FP: 0 P: 322 N: 220
5	(maint = low) AND (doors = 5more) → acc TP: 19 FP: 3 P: 295 N: 220
6	(maint = low) AND (doors = 4) → acc TP: 18 FP: 4 P: 276 N: 217
7	(maint = med) AND (lug_boot = med) → a TP: 26 FP: 6 P: 258 N: 213
8	(safety = high) AND (doors = 3) → acc TP: 40 FP: 15 P: 232 N: 207
9	(safety = high) AND (lug_boot = big) AND (doors = 5more) → acc TP: 10 FP: 5 P: 192 N: 192
10	(safety = high) AND (lug_boot = big) → a TP: 17 FP: 9 P: 182 N: 187
11	(safety = high) → acc TP: 73 FP: 46 P: 165 N: 178

Figure 2. A rulelist (bottom) for the "car" domain and the corresponding ROC curve (top) with respect to the acc class.

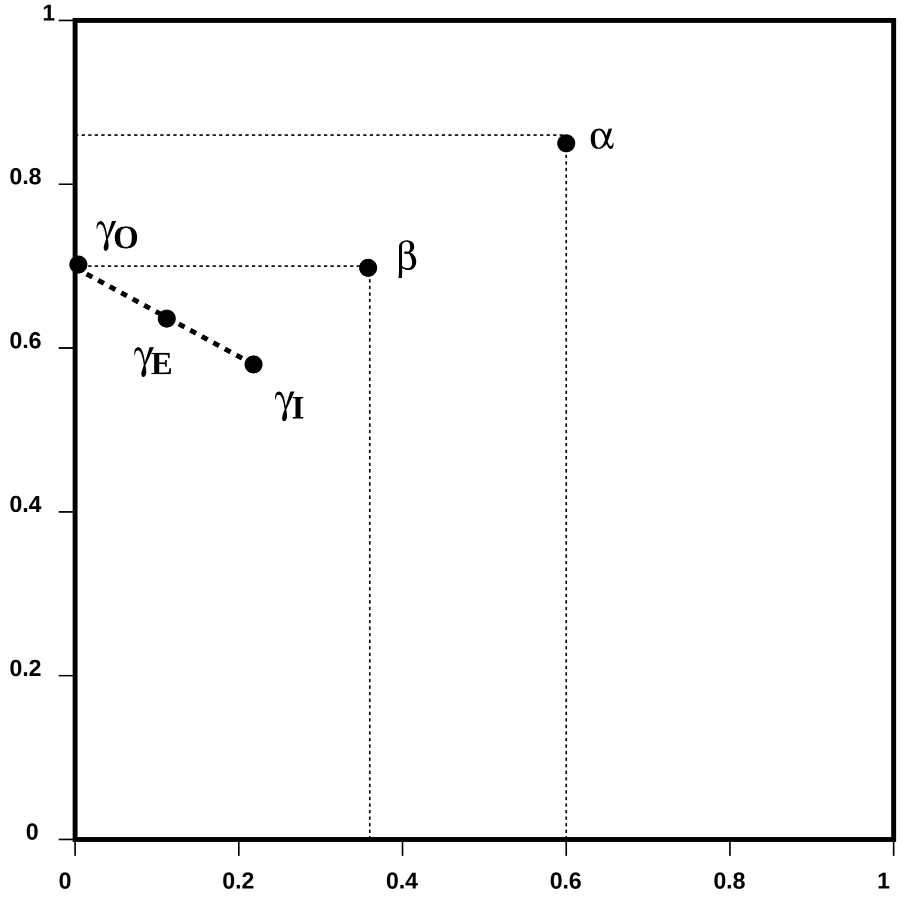


Figure 3. Estimating the performance of a new rule. α and β are two rules conjoined to form γ. γ_I is the expected new rule location assuming independence; γ_O is the best (most optimistic) rule combination we could expect; and a compromise (using OPTIMISM=0.5) produces the expected rule γ_E.

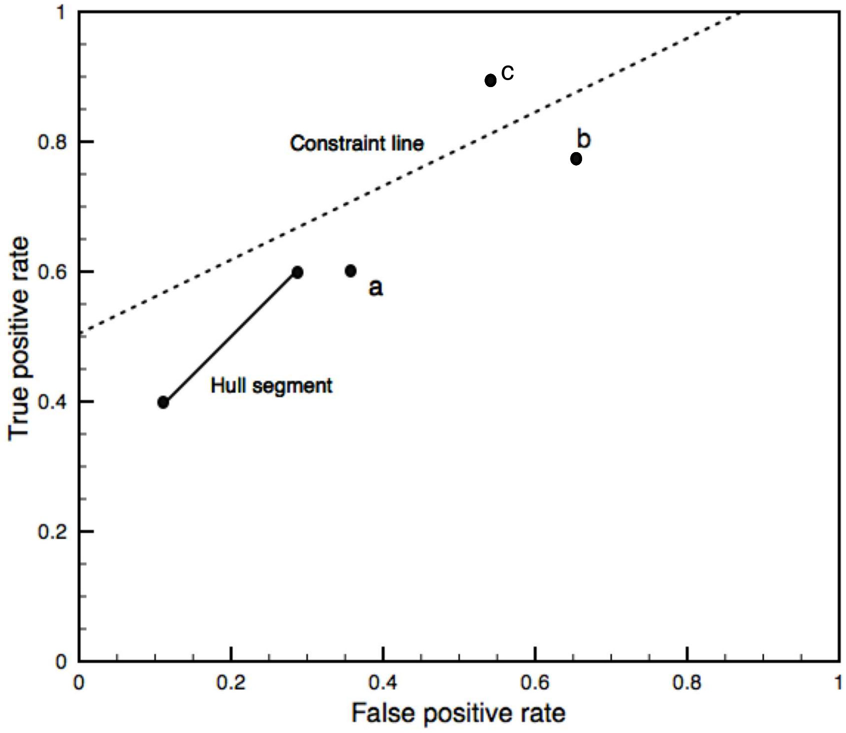


Figure 4. A constraint line used to prune rule conjunction candidates.

Table I. Domains and their characteristics

Name	**Instances**	**Classes**		**Attributes**	
		N	**Proportions**	**N**	**Type**
breast	683	2	65/34	10	0-9
breast-wisc	683	2	65/34	10	0-9
breast-wpbc	198	2	76/23	34	0-32
bupa	345	2	57/42	34	0-34
car	1728	2	70/29	34	6-28
cmc	1473	3	42/34/22	34	7-27
covtype	5000	7	24/18/18/12/11/9/3	34	4-30
crx	653	2	54/45	34	9-25
dermatology-BB[a]	366	2	69/30	34	33-1
ecoli	336	2	89/10	34	5-29
flag	194	8	35/24/20/8/4/4/1/0	34	29-4
flag-Prati	194	2	91/8	34	19-14
german	1000	2	70/30	34	20-13
glass	214	2	92/7	34	13-19
glass-RS	205	5	37/34/14/8/6	34	13-19
haberman	306	2	73/26	34	13-20
heart	270	2	55/44	34	12-21
image	2310	7	14/14/14/14/14/14/14	34	8-25
ionosphere	351	2	64/35	34	0-34
kr-vs-kp	3196	2	52/47	36	36-0
letter-a	20000	2	96/3	36	20-16
mushroom	5644	2	61/38	36	36-0
new-thyroid	215	2	86/13	36	31-5
nursery	12960	2	97/2	36	36-0
optdigits	5620	10	10/10/10/10/10/9/9/9/9/9	36	36-0
page-blocks	5473	5	89/6/2/1/0	36	26-10
pima	768	2	65/34	36	26-10
promoters	106	2	50/50	57	57-0
sonar	208	2	53/46	60	0-60
splice	3190	3	51/24/24	60	60-0
vehicle	846	2	76/23	60	42-18
yeast	1484	10	31/28/16/10/3/2/2/2/1/0	60	42-17

[a] This is the UCI dermatology domain as converted by Barakat and Bradley (2006): a two-class domain comprising the original class 1 and class 2 comprising the original classes 2 through 6.

Table II. PRIE compared against weighted voting (WVOTE) of accuracy-maximizing rules.

Dataset	**AUC using WVOTE**	**PRIE**
Breast-wisc	97.3 ± 3.6	98.3 ± 1.7
Car	98.3 ± 0.7	93.4 ± 1.5
Cmc	66.4 ± 4.9	66.3 ± 2.7
Covtype	82.2 ± 1.4	88.7 ± 1.2
Crx	90.2 ± 3.4	91.5 ± 3.4
German	68.4 ± 9.9	73.8 ± 5.1
Glass	75.5 ± 6.0	82.8 ± 16.1
Image	99.0 ± 0.5	99.1 ± 0.3
Kr-vs-kp	99.7 ± 0.2	98.8 ± 0.7
Mushroom	100.0 ± 0.0	100.0 ± 0.0
Nursery	99.8 ± .1	99.0 ± 0.4
Promoters	88.9 ± 13	84.1 ± 9.0
Sonar	76.9 ± 14	74.6 ± 13.6
Splice	97.2 ± 0.7	93.7 ± 1.6

Table III. PRIE compared to ROCCER

Dataset	**ROCCER**	**PRIE**
breast	98.63 ± 1.88	98.0 ± 1.9
bupa	65.30 ± 7.93	73.0 ± 6.7
e-coli-prati	90.31 ± 11.56	95.8 ± 4.9
flag	61.83 ± 24.14	71.1 ± 9.4
german	72.08 ± 6.02	73.8 ± 5.1
glass	79.45 ± 12.98	82.8 ± 16.1
haberman	66.41 ± 11.54	65.9 ± 6.5
heart	85.78 ± 8.43	83.0 ± 5.7
ionosphere	94.18 ± 4.49	92.1 ± 3.8
kr-vs-kp	99.35 ± 0.36	98.8 ± 0.7
letter-a	96.08 ± 0.52	99.5 ± 0.4
new-thyroid	98.40 ± 1.70	92.0 ± 12.1
nursery	97.85 ± 0.44	99.0 ± 0.4
pima	70.68 ± 5.09	80.3 ± 4.3
vehicle	96.42 ± 1.47	97.5 ± 2.1
Avg	85.13	86.84

Table IV. PRIE vs SVM rule extraction method of Barakat and Bradley

Data set	**SVM AUC ± Std**	**SVM+Rules AUC ± Std**	**PRIE AUC ± Std**
Pima	82 ± 3	94 ± 2	80.3 ± 4.3
Breast cancer	97 ± 1	96 ± 2	98 ± 2
Heart	89 ± 4	81 ± 5.1	83 ± 5.7
Dermatology	1.00 ± 0	98.4 ± 1.1	99.0 ± 1.1