Learning Rare Class Footprints: the REFLEX Algorithm

Ray J. Hickey RJ.HICKEY@ULSTER.AC.UK
School of Computing and Information Engineering, University of Ulster, Coleraine, Co. Londonderry, N. Ireland, UK, BT52 1SA

Abstract

An *r*-contour footprint is a set of individuals each of whom has a propensity of at least *r* of belonging to a rare class. The properties of footprints are summarized. An algorithm, REFLEX, is proposed for extracting a footprint from an induced decision tree. Results of initial experiments comparing REFLEX to *m*-estimation Laplace smoothing show that both algorithms deliver broadly similar performance for different contours. Unlike Laplace, REFLEX does not require extensive tuning. When high propensity rare class disjuncts exist (> 50%), both algorithms perform better on pruned trees.

1. Introduction

Learning to classify when one or more classes is rare, known as the *class imbalance problem*, continues to be a topic of major interest. Rare classes are often associated with higher misclassification costs (Drummond & Holte, 2000). A discussion of the problem and review of recent approaches is provided by Japkowicz and Stephen (2002).

If a class is rare then either the underlying rules identifying it are subject to substantial noise or have very low support (or both). There may even be no rules for which the rare class is the majority.

Instead of building a classifier, the aim here will be to characterize a subset of a population, called a *footprint*, where the probability of an individual belonging to the rare class exceeds a given level. Often this level will be less than 50%.

2. *p*-Propensity Footprints

Assume that individuals in a population are represented using a set of description attributes and that they belong to one of two classes *common* and *rare*. Typically, it is not known for any individual which class they belong to. For each individual, based on their description, there is a propensity, *q*, to belong to the rare class and a propensity, *p*, to belong to the common class where $q + p = 1$. Given probabilities for the occurrence of descriptions, a subset of the attribute space can be assigned propensities for the common and rare classes by averaging over descriptions.

Definition 1 A (*p-propensity*) *footprint*, *F*, is a subset of the attribute description space for which the (average) propensity of the rare class is $prop(F) = p$. The probability of occurrence of the footprint is its *support*, $supp(F)$.

The term propensity is used here rather than *probability* to emphasize that tendency to be rare will be viewed as an inherent characteristic of an individual or group. Probability, on the other hand emphasizes risk assessment and decision making.

Every description space can be represented as a disjoint union of disjuncts, $\boldsymbol{D} = \{D_i\}_{i=1}^{k}$ such that, within each disjunct, the propensity of individuals is constant and no two disjuncts have the same propensity. Assume, without loss of generality, that the D_i are indexed in order of increasing propensity. Each footprint, *F*, then has a canonical representation with respect to $\boldsymbol{D}$ as

$$F = \bigcup_{i=1}^{k} E_i$$

where E_i is a *sub-disjunct*, i.e. subset, of D_i, $i = 1, \ldots, k$. Some of the E_i may be empty. If, for some i, $E_i = D_i$ then E_i is said to be *full*.

The propensity of F can be expressed as

$$prop(F) = \frac{\sum_{i=1}^{k} prop(E_i) \times supp(E_i)}{\sum_{i=1}^{k} supp(E_i)} = \sum_{i=1}^{k} s_i \times prop(E_i) \quad (1)$$

where s_i is the conditional support of E_i within F and $\sum_{i=1}^{k} s_i = 1$.

Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.

More generally, Equation 1 can be used to obtain the propensity of a union of disjoint footprints given the individual propensities.

Some properties of footprints can be obtained immediately.

Property 1 $\min_i prop(E_i) < prop(F) < \max_i prop(E_i)$

for non-empty sub-disjuncts, E_i , of footprint, F.

Property 2 For any disjoint footprints, F and G each having positive support

$$prop(F) < prop(G) \Rightarrow$$
$$prop(F) < prop(F \cup G) < prop(G).$$

Property 3 If F, F' and G are disjoint footprints with

$0 < prop(F') < prop(F) < prop(G)$ and $supp(F') \geq supp(F)$ then

$$prop(F' \cup G) < prop(F \cup G)$$

Proof Let $\alpha' = supp(F')/(supp(F') + supp(G))$ and define α correspondingly (so that $\alpha < \alpha'$). Then

$$\begin{aligned} prop(F' \cup G) &= \alpha' prop(F') + (1-\alpha') prop(G) \\ &< \alpha' prop(F) + (1-\alpha') prop(G) \\ &< \alpha prop(F) + (1-\alpha) prop(G) \\ &= prop(F \cup G) \end{aligned}$$

using Equation 1 applied to disjoint unions. □

2.1 *r*-Contour Footprints

A particularly important footprint is one in which all sub-disjuncts of at least a given propensity are present and full while all others are empty.

Definition 2 The *r-contour footprint* is

$$C_r = \bigcup_{prop(D_i) \geq r} D_i$$

Note that it is not implied in Definition 2 that there is a disjunct with the exact propensity r.

Some basics properties of contour footprints are:

Property 4 $prop(C_r) \geq r$.

Property 5 $prop(C_r)$ is non-decreasing in r.

Property 6 $supp(C_r)$ is non-increasing in r.

The following result also holds.

Theorem 1 If F is a footprint with $supp(F) > supp(C_r)$ then $prop(F) < prop(C_r)$.

Proof Suppose there exists F with $supp(F) > supp(C_r)$ and $prop(F) \geq prop(C_r)$. Amongst all such F select one, F^*, with maximum propensity. Compare sub-disjuncts in F^* to the corresponding disjuncts in C_r. In F^*, all disjuncts, E_i, with $prop(E_i) > prop(F^*)$ must be full otherwise, by Property 2, filling these would further increase $prop(F^*)$. Consider the remaining (lower) disjuncts in C_r. Some of the corresponding sub-disjuncts in F^* must not be full else F^* contains all of C_r; but $supp(F^*) > supp(C_r)$ so then F^* would contain all of C_r and additional sub-disjuncts with propensity less than r implying, from Property 2, that $prop(F^*) < prop(C_r)$ which is a contradiction. Therefore F^* must:

- include some lower sub-disjuncts, C_r^{in}, from C_r
- omit some, C_r^{out}, from the bottom of C_r
- include some, F_{lower}, from below C_r.

Thus F^* can be expressed as the disjoint union

$$F^* = F_{lower} \cup (C_r^{in} \cup C_{r'})$$

while

$$C_r = C_r^{out} \cup (C_r^{in} \cup C_{r'}).$$

where $r' = prop(F^*)$. Clearly $prop(F_{lower}) < prop(C_r^{out})$ and, by assumption, $supp(F_{lower}) > supp(C_r^{out})$. Applying Property 3 shows that $prop(F^*) < prop(C_r)$ which is a contradiction. Thus F^* does not exist and so there is no footprint, F, with $supp(F) > supp(C_r)$ and $prop(F) > prop(C_r)$. □

Amongst footprints achieving a given minimum propensity, it does not follow that the footprint that maximizes support will be a contour. The lowest r for which C_r exceeds the required propensity will, typically have a propensity well above this. It may be possible to add sub-disjuncts below the contour, which while lowering the propensity, will not take it below the minimum.

2.2 ROC Points and Footprints

Receiving operator characteristic (ROC) points (Provost, Fawcett and Kohavi, 1998) can be used to compare footprints. Given a confusion matrix for probabilities

	Actual +ve	Actual -ve
Predict +ve	True +ve (*TP*)	False +ve (*FP*)
Predict -ve	False -ve (*FN*)	True -ve (*TN*)

(with *rare* as +ve) then, if it is assumed that every individual in a footprint, F, is rare, it follows that

$$TP(F) = supp(F) \times prop(F)$$
$$FP(F) = supp(F) \times (1 - prop(F))$$

Note that

$$prop(F) = TP(F)/(TP(F) + FP(F)) \quad (2)$$

Let p_{rare} be the overall propensity of the rare class in the population. The true positive rate (*TPR*) and the false positive rate (*FPR*) are defined as $TPR(F) = TP(F)/p_{rare}$ and $FPR(F) = FP(F)/(1- p_{rare})$. A footprint is said to dominate another if its ROC point, (*TPR*, *FPR*), is more north-west, i.e. has greater *TPR* for no greater *FPR*.

Property 7 If footprints *E* and *F* are disjoint then

$$TP(E \cup F) = TP(E) + TP(F)$$
$$FP(E \cup F) = FP(E) + FP(F)$$

Property 8 If footprints *E* and *F* are disjoint and have positive support, then

$$E \cup F \text{ dominates } F \Leftrightarrow prop(E) = 1$$

In particular, no contour footprint dominates another contour footprint.

Proof If $E \cup F$ dominates F then

$$FP(E \cup F) = FP(E) + FP(F) \leq FP(F)$$

so $FP(E) = 0$ and $prop(E) = 1$ from Equation 2. Conversely, if $prop(E) = 1$, then, also from Equation 2, $TP(E) > 0$ and $FP(E) = 0$. Thus $TP(E \cup F) > TP(F)$ while $FP(E \cup F) = FP(F)$ and so $E \cup F$ dominates *F*.

For $r < r'$, either $C_r = C_{r'}$ or $C_r = E \cup C_{r'}$ where *E* and $C_{r'}$ are disjoint with positive support and $prop(E) < 1$. Thus C_r cannot dominate $C_{r'}$. If $C_r \neq C_{r'}$ then, it is clear that $TP(C_{r'}) < TP(C_r)$ and $C_{r'}$ cannot dominate C_r. □

Theorem 2 A contour footprint is not dominated by any other footprint.

Proof Suppose a footprint, *F**, dominates C_r for some *r*. It follows from Property 8 that *F* cannot be a contour footprint nor be contained within a contour footprint. Thus, as in the proof of Theorem 1,

$$F^* = F_{lower} \cup (C_r^{in} \cup C_{r'})$$

where

$$C_r = C_r^{out} \cup (C_r^{in} \cup C_{r'}).$$

By supposition, $TP(F^*) > TP(C_r)$ so from Property 7

$$TP(F_{lower}) > TP(C_r^{out})$$

that is

$$supp(F_{lower}) \times prop(F_{lower}) > supp(C_r^{out}) \times prop(C_r^{out})$$

but $prop(F_{lower}) < prop(C_r^{out})$ hence

$$supp(F_{lower}) > supp(C_r^{out})$$

and

$$supp(F^*) > supp(C_r)$$

Thus the conditions of Property 3 apply to the decomposition of *F** and so $prop(F^*) < prop(C_r)$ and

$$1 - prop(F^*) > 1 - prop(C_r)$$

Therefore

$$FP(F^*) = supp(F^*)\,(1 - prop(F^*))$$
$$> supp(C_r)\,(1 - prop(C_r)) = FP(C_r)$$

which contradicts the dominance of *F** over C_r. □

2.3 Expected Benefit

A rare class often incurs a greater misclassification cost. Elkan (2001) has shown that inappropriate specification of costs leads to inconsistency and, instead, recommends the use of benefits. An appropriate benefit matrix for footprints, corresponding to the confusion matrix in Section 2.2, is

	Actual +ve	Actual -ve
Predict +ve	B_1	$-B_2$
Predict –ve	0	0

where B_1 and B_2 are positive. The negative benefit $-B_2$ reflects the loss from wrongful identification of an individual as rare, e.g. loss of profit or wasteful use of medical treatment. In many applications, $B_1 \gg B_2$.

The (unit) expected benefit from using a footprint, *F*, to identify rare class individuals based on the confusion and benefit matrices above is

$$\begin{aligned} Expben(F) &= B_1 \times TP - B_2 \times FP \\ &= B_1 \times supp(F) \times prop(F) \\ &\quad - B_2 \times supp(F) \times (1 - prop(F)). \end{aligned} \quad (3)$$

Property 9 If *E* is dominated by *F* then

$$Expben(E) < Expben(F).$$

Property 10 $Expben(F) > 0$ if and only if

$$prop(F) > B_2 / (B_1 + B_2).$$

Property 11 Expected benefit is maximized over all footprints by C_r where

$$r = \min\{r_i = prop(D_i) : r_i > B_2/(B_1 + B_2)\}. \quad (4)$$

3. Learning Footprints

A footprint may be represented as a decision tree in which each leaf is tagged as either belonging to the footprint or not. The individual paths to the leaves tagged as being in the footprint represent its sub-disjuncts. This tree can be learned from training examples classified as *common* or *rare*.

The algorithm proposed here is called REFLEX (**R**ar**E** class **F**ootprint **L**earning from **EX**amples). It attempts to induce the *r*-contour footprint for given *r*. REFLEX is applied to a decision tree induced by an algorithm such as ID3 (Quinlan, 1986). The tree may be pruned before application of REFLEX.

In REFLEX (see table 1) the tree expansion is reversed beginning at the bottom and working upwards until footprint nodes are first encountered. A footprint node is identified by a statistical test.

3.1 The Footprint Assignment Criterion

Suppose n examples reach a node. Let the frequency distribution at the node be $(f_1 / n, f_2 / n)$ where is f_2 / n is the relative frequency of the rare class. Based on the value of f_2 / n, a decision must be made as to whether to assign the node to the footprint.

A disjunct belongs to the *r*-contour footprint if its propensity is greater than or equal to *r*. A statistical test can be applied to establish this. Such a test is really a heuristic since a multiple comparison effect applies to node frequency distributions (Jensen & Cohen, 2000).

The one-sided upper confidence bound $ub(r, \alpha)$, where α is a significance level, can be used to accept or reject the node for footprint membership: accept if $f_2 \geq ub(n, r, \alpha)$. For sufficiently large n, $ub(n, r, \alpha)$ can be approximated from the Normal distribution $N(nr, nr(1\text{-}r))$ otherwise it is obtained exactly from the Binomial distribution $B(n, r)$. The test offers some protection against sub-disjuncts with propensity less than *r* being accepted into the footprint.

If, however, there are sub-disjuncts of C_r with propensity close to the contour boundary, their chance of being assigned to the footprint can be small. For example if $r = 0.20$, $\alpha = 0.1$ and $n = 30$, then

$$ub(30, 0.2, 0.1) = 10$$

thus to be assigned to the footprint, a sub-disjunct would require $f_2 \geq 10$, that is $f_2 / n \geq 10/30 = 0.3$. There is only a 37% chance of achieving this. Disjuncts immediately above the contour boundary will, therefore, tend to be under-represented in the induced footprint until the training set size grows sufficiently.

Table 1. The REFLEX algorithm.

($FC(N)$ is the footprint criterion, for node N)

Input: an induced decision tree, T
Output: footprint, Fp

initialize $Fp = \{\}$

while there is an end node in T (i.e. all its children are leaves) for which no child, M, satisfies $FC(M)$ do

  replace the end node by a leaf

return $RefT$.

for each leaf, l, in $RefT$, do

  if l satisfies $FC(l)$
  $Fp \leftarrow Fp \cup \{\text{path to } l\}$

return Fp

Use of significance tests for the learning of small disjuncts was criticized by Holte et al. (1989) on the grounds that meaningful as well as uninformative disjuncts can be pruned away. It is also known (Provost et al., 1998) that pruning can degrade probability estimates in the leaves. The mechanism described above, however, does not trade child nodes for the parent in quite the same way. Pruning takes place only if none of the child nodes offers a footprint member. The parent, though, with a lower rare class propensity, may compensate through its greater frequency and reach significance.

As recommended by Elkan (2001), REFLEX is applied to a directly learned tree. That is, the initial induction process is not oriented towards the detection of footprint disjuncts.

4. Experimental Results

To evaluate REFLEX, an artificial universe (Hickey, 1996) was created. This specified a complete class model for the two classes *common* and *rare* in terms of six informative description attributes and three pure noise attributes. The contours for this domain are shown in Table 2 in descending order of rare class propensity. The rare class base rate is 9.57%. The attributes provide very little lift: the maximum attainable classification rate, 100 – Bayes error rate, is 90.94%.

REFLEX was tested against the Laplace *m*-estimation smoothing method (Cestnik and Bratko, 1991). Here the raw frequency score, k/n, from k rare class training examples at a leaf containing n examples, is smoothed to $(k + bm)/(n+m)$ where b is the base rate of the rare class and m is a parameter.

Table 2. C_r details for an artificial domain, Domain 1.

r	$prop(C_r)$ (%)	$supp(C_r)$	r	$prop(C_r)$ (%)	$supp(C_r)$	r	$prop(C_r)$ (%)	$supp(C_r)$	r	$prop(C_r)$ (%)	$supp(C_r)$
0.95	95.00	0.0038	0.21	43.93	0.0225	0.11	16.64	0.3581	0.05	12.18	0.7599
0.89	93.33	0.0053	0.20	36.03	0.0336	0.10	14.45	0.5338	0.04	12.12	0.7660
0.73	89.76	0.0064	0.18	23.21	0.1163	0.09	14.00	0.5771	0.03	11.46	0.8256
0.52	85.35	0.0073	0.16	22.18	0.1355	0.08	13.92	0.5855	0.01	09.57	1.0000
0.45	77.24	0.0091	0.15	19.09	0.2378	0.07	13.20	0.6539			
0.24	69.97	0.0105	0.13	18.09	0.2846	0.06	12.26	0.7521			

The smoothed value is a convex combination of k/n and b. Increasing m pulls the value more towards b. The rationale for smoothing is that small leaf frequencies tend to gravitate towards 0 or 1 as a consequence of the attribute selection competition during tree growth. The parameter m must be determined either using expert advice or from the data itself. The m-estimation technique is a generalization of simple Laplace smoothing in which k/n is smoothed to $(k+1)/(n+2)$. The latter corresponds to $b = 1/2$ and $m = 2$.

To determine the footprint, the smoothed rare class frequency was obtained for each leaf of the induced tree. Each leaf whose smoothed frequency exceeded the contour level was placed in the footprint.

Three contours 25%, 60% and 12.5% were chosen for the experiments.

A series of trials, at a range of sizes, was performed to induce footprints from training examples generated from the model. The number of replications ranged from 1000 at smaller sizes down to 10 at the largest sizes.

The significance level for REFLEX was set mostly at 0.05. For Laplace, m was selected on the basis of trials involving separate inductions on specially generated data.

From Table 2, the 25% contour is the union of the first five disjuncts. Results for fully grown ID3 trees for this contour are shown in Table 3 (a) and (b). For Laplace smoothing, $m = 10$ was found to be best.

Overall, the results for REFLEX and for Laplace smoothing were broadly similar. It required a sample of approximately 2000 to guarantee that induced footprints that would not be empty. From $n = 3000$ virtually all trials yielded a footprint attaining the required minimum propensity of 25%. REFLEX tended to produce a smaller number of disjuncts than Laplace.

From table 2, the true propensity and support for the contour are 77.24% and 0.0091. As n increased, the induced footprints moved towards these true values. Initial support levels were inflated as a result of including only non-empty footprints in the averaging.

An induced sub-disjunct is described as *hot* if belongs to C_r and as *cold* otherwise. The percentage support of hot sub-disjuncts relative to the induced footprint as a whole is shown in the *Hot Supp (%)* column. This increased with n as cold sub-disjuncts were gradually eliminated. For larger n, Laplace smoothing produced less hot support then REFLEX, which explains its lower propensities.

Benefits were set at $B_1 = 600$ and $B_2 = 200$ which, from Equation 4, correspond to a 25% contour for maximising expected benefit. Equation 3 gives the maximum as 3.8. Almost all trials produced positive expected benefit from $n = 3000$ onwards (corresponding to attainment of the required propensity). Expected benefit levels rose with n and REFLEX was slightly ahead for most of the learning curve.

To investigate the effects of pruning, the induced trees were pruned using MEP pruning (Niblett and Bratko, 1986) involving simple smoothing. The results are shown in tables 3 (c) and (d). For $n < 3000$, pruning produced a higher incidence of empty footprints (not shown). For example, at $n = 1000$, REFLEX applied to the full tree had a 10% occurrence of empty footprints which rose to 45% on pruning. Similar statistics applied to Laplace.

But for larger n, pruning had a definite advantage. The number of disjuncts decreased (which was to be expected) and Laplace was on a par with REFLEX. Propensity levels generally improved, particularly for Laplace. Hot support percentages rose and cold support was eliminated at $n = 100000$. Finally, expected benefit improved across all sample sizes and approached the maximum of 3.8.

Table 3. Statistics for inductions of the 25% contour footprint from ID3 trees using REFLEX and Laplace smoothing (m = 10) on training examples from Domain 1. SE estimates: *Prop* < 2%; *Supp* < 0.0005; *Exp Benefit* < 0.1.

n	*No. of Disj*	*Prop (%)*	*Supp*	*Hot Supp (%)*	*Exp Benefit*	*No. of Disj*	*Prop (%)*	*Supp*	*Hot Supp (%)*	*Exp Benefit*
	(a) REFLEX on full tree					(b) Laplace on full tree				
1000	2.2	42.4	0.0100	32	0.71	2.1	46.3	0.0072	44	0.88
2000	3.3	52.9	0.0090	43	1.42	3.4	49.8	0.0078	47	1.31
3000	4.6	55.9	0.0095	49	2.00	5.3	49.3	0.0103	41	1.67
5000	7.1	59.2	0.0091	57	2.26	8.2	49.5	0.0106	42	1.84
10000	10.7	68.5	0.0078	73	2.62	13.5	52.6	0.0111	46	2.20
30000	23.8	74.4	0.0079	81	3.06	32.5	54.6	0.0122	53	2.80
100000	41.6	76.8	0.0085	93	3.51	60.0	61.9	0.0118	70	3.44
	(c) REFLEX on pruned tree					(d) Laplace on pruned tree				
1000	2.0	49.3	0.0094	41	1.24	2.0	49.7	0.0085	46	1.24
2000	2.6	62.4	0.0076	59	1.95	2.9	60.4	0.0080	55	1.88
3000	3.0	71.2	0.0071	72	2.37	3.1	69.7	0.0071	70	2.32
5000	3.9	75.3	0.0073	79	2.80	4.1	75.8	0.0071	82	2.78
10000	5.3	79.4	0.0075	88	3.19	5.9	78.3	0.0076	86	3.16
30000	7.3	81.6	0.0077	99	3.50	8.8	80.1	0.0081	99	3.57
100000	10.4	79.4	0.0085	100	3.68	12.0	78.6	0.0086	100	3.70

It is generally held that pruning with a marked class imbalance and/or differing costs may not be productive and often results in the tree being pruned away. Elkan (2001) recommends no pruning when Laplace smoothing is used to estimate probabilities. Whether or not pruning is beneficial, though, may depend on the existence of high propensity disjuncts (>50%).

From Table 2, the 60% contour consists of the top three disjuncts and has propensity 89.76% and support 0.0064. For Laplace, m = 4 was found to give the best results. Thus it appears necessary to tune m to the contour as well as the domain.

The results for REFLEX and Laplace on the full trees were virtually identical across the learning curve. With fewer true disjuncts involved, a sample of about 5000 was necessary before non-empty footprints were obtained with certainty. Along the learning curve, footprints for REFLEX and Laplace contained similar numbers of disjuncts unlike for the 25% contour. At n = 100000, support reached 0.0052 (±0.0001) for REFLEX and 0.0054 (±0.0001) for Laplace. Benefits were set at B_1 = 200 and B_2 = 300 giving a maximum of 0.95. At n = 100000, REFLEX and Laplace had reached 0.84 and 0.87 respectively (±0.01).

Pruning produced qualitatively the same improvements as for the 25% contour. At n = 100000, both algorithms attained virtually full coverage of the true footprint with no cold support.

Finally, the 12.5% contour was induced (using m = 10). Table 2 shows there is a cluster of disjuncts just below and above this level. It thus provides a more demanding test. The propensity is 18.09% and support is 0.2846. REFLEX required a larger sample than Laplace to produce non-empty footprints (about n = 250 against 100). The main differences, though, were in the number of disjuncts with Laplace having more by a factor of five, and also in support where REFLEX attained only 0.112 (±0.004) at n = 100000 whereas Laplace achieved 0.277 (±0.002). With B_1 = 1400 and B_2 = 200, expected benefits (maximum is 25.46) were similar for most of the learning curve, but at n = 100000, Laplace achieved 20.07 (±0.06) against 15.99 (±0.24) for REFLEX.

Thus the stringent test applied by REFLEX for entry to the footprint has inhibited the growth of support and, consequently, of expected benefit.

4.1 A Second Domain

To investigate behaviour of the algorithms when high propensity rare class disjuncts are not available, Domain 1 was altered to produce Domain 2 while leaving the base rate approximately the same. Disjunct definitions and attribute distributions were unchanged. Rare class propensities now range from 0.32 down to 0. Details of the top six disjuncts are shown in Table 4. The base rate is 9.68%.

Table 4. Partial C_r details for an artificial domain, Domain 2.

r	$prop(C_r)$ (%)	$supp(C_r)$
0.32	32.00	0.0008
0.27	28.83	0.0023
0.25	25.66	0.0134
0.23	25.46	0.0145
0.21	21.78	0.0830
0.19	21.66	0.0868

Inductions were carried out for the 20% contour containing the top five disjuncts with propensity 21.78% and support 0.0830 (Table 4). For Laplace, m = 40 was used. For REFLEX, the 0.01 significance level was found to be better. Results are shown in Table 5.

A sample size of about 5000 was required before footprints reached the 20% level. Empty footprints, though, occurred until about n = 20000. Laplace was worse in this respect: at n = 10000, about 35% of footprints were empty against 6% for REFLEX. Support rose steadily with n and Laplace moved ahead of REFLEX (as was seen above with the12.5% contour for Domain 1). Even at n = 100000, though, the Laplace support was less than half the maximum. Induction had been made difficult by the cluster of disjuncts around the contour. Laplace gained expected benefit more quickly as n increased because of its better support.

In contrast to Domain 1, pruning produced mostly empty trees at all sample sizes (over 90% of trials). This was to be expected in a situation where there are no disjuncts with a majority for the rare class.

4.2 Relative Frequency and Simple Laplace

In a final series of experiments, footprints were induced for the four contours above using relative frequency as the criterion for the footprint, i.e. $f_2 / n \geq$ r where (f_1 / n, f_2 / n) is the frequency distribution at the leaf and f_2 / n is that of the rare class. Simple Laplace was also used.

For the full trees, performance on the 25% contour of Domain 1 was poor for both algorithms. Footprints contained a very large number of disjuncts (575 for Laplace at n = 100000). At all sample sizes, footprints contained considerable cold support resulting in low propensities and expected benefit. Laplace was much worse then frequency. Neither algorithm seems to have been able to find the high propensity disjuncts.

For the 60% contour, both algorithms started badly requiring a size of about n = 20000 to reach the contour level. Thereafter performance improved but still fell short at n = 100000.

By contrast both algorithms performed very well on the 12.5% contour. They matched REFLEX and m-estimation at larger sizes although simple Laplace produced over 1300 disjuncts at n = 100000 (compared to about 170 for REFLEX).

For the 20% contour of Domain 2, the performance of both algorithms was exceedingly poor (with Laplace the worse of the two). At no size did either reach the contour level and so all expected benefits were negative. Hot support levels, though, were better than REFLEX and m-estimation Laplace. Once again, the problem was the failure to detect the higher disjuncts.

On pruning, the 25% and 60% contour results greatly improved but were still inferior to REFLEX and m-estimation at the lower sizes. Results for 12.5% also improved slightly.

5. Discussion and Conclusion

With footprint induction, the goal is to find a characterization of individuals exceeding a given propensity for the rare class. There is no interest in individuals outside the footprint. The REFLEX algorithm applies a stringent test for membership of the footprint. Individuals excluded from it are not deemed to belong to the common class; it is just that their claim to be in the footprint was not sufficiently strong. Thus REFLEX does not seek to produce a classifier or other optimal decision procedure across the population.

The important characteristics of a footprint learner are that it should be able to produce a footprint description meeting the required propensity level with a high degree of certainty (over trials) and from as small a sample as possible. Further, the hot support for the footprint should grow effectively to the maximum (implying that expected benefit will do likewise).

The natural competitor to REFLEX is m-estimation Laplace smoothing applied, likewise, to the direct tree. Considerable effort was expended, using the large volume of artificial data available, to determine suitable values for the m parameter.

Table 5. Statistics for inductions of the 20% contour footprint from ID3 trees using REFLEX and Laplace smoothing (m = 40) on training examples from Domain 2. SE estimates: *Prop* < 0.3%; *Supp* < 0.001; *Exp Benefit* < 0.03.

n	*No. of Disj*	*Prop (%)*	*Supp*	*Hot Supp (%)*	*Exp Benefit*	*No. of Disj*	*Prop (%)*	*Supp*	*Hot Supp (%)*	*Exp Benefit*
	(a) REFLEX on full tree					(b) Laplace on full tree				
5000	2.3	19.6	0.0063	83	0.01	1.4	20.1	0.0049	84	-0.02
10000	2.9	20.1	0.0089	94	0.04	2.0	20.7	0.0059	100	0.04
20000	3.6	21.3	0.0059	90	0.09	4.1	20.4	0.0097	81	0.04
30000	4.3	22.0	0.0051	86	0.09	6.9	20.8	0.0168	87	0.13
50000	5.5	22.9	0.0069	87	0.15	11.9	20.9	0.0226	84	0.20
100000	8.4	22.6	0.0163	97	0.34	20.1	21.3	0.0324	88	0.40

The indications are that m must be re-tuned for different contours. With limited real data, tuning might not be so effective and this could impair performance. REFLEX does not require such extensive tuning. The 0.05 or 0.01 significance levels seem generally satisfactory.

In the preliminary investigation carried out here, REFLEX was found to compete well against Laplace. It many instances there were no statistical differences between the results. On some occasions REFLEX gains support less quickly than Laplace. Neither algorithm dominated the other on ROC points. Pruning was found to be effective for both algorithms when disjuncts with rare class propensity above 50% exist.

Simple Laplace and relative frequency gave highly variable performances. Their best results were on large samples where there were no high propensity disjuncts.

Future work will focus on possible refinements to the REFLEX criterion for footprint membership. Further investigations will be carried out using real data sets.

References

Cestnik, B., & Bratko, I. (1991). On estimating probabilities in tree pruning. In Y. Kodratoff (Ed.), *Machine Learning: EWSL-91* (pp. 138-150). Berlin: Springer-Verlag.

Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost: An alternative to ROC representation. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 198-207). Boston: ACM.

Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Joint Conference on Artificial* Intelligence (pp.973-978). San Francisco: Morgan Kaufmann.

Hickey, R. J. (1996). Noise modeling and evaluating learning from examples. *Artificial Intelligence*, *82*, 157-179.

Holte, R., Acker, L., & Porter, B. (1989). Concept learning and the problem of small disjuncts. *Proceedings of the Eleventh International Joint Conference on Artificial* Intelligence (pp.813-818). San Francisco: Morgan Kaufmann.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study**.** *Intelligent Data Analysis Journal, 6*, 429 - 449.

Jensen, D.D., & Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, *38*, 1-30.

Niblett, T., & Bratko, I. (1986). Learning decision rules in noisy domains. In M. A. Bramer (Ed.), *Research and Development in Expert Systems III*. Cambridge: Cambridge University Press.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445-453). San Francisco: Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81-106.