Proceedings of the 1996 Pacific Rim International Conference on Artificial Intelligence,
Cairns, Springer-Verlag, pp. 23-34.

Cost-Sensitive Specialization

Geoffrey I. Webb

Deakin University,
School of Computing and Mathematics,
Geelong, Vic, 3217, Australia.

Abstract. Cost-sensitive specialization is a generic technique for misclassification cost sensitive induction. This technique involves specializing aspects of a classifier associated with high misclassification costs and generalizing those associated with low misclassification costs. It is widely applicable and simple to implement. It could be used to augment the effect of standard cost-sensitive induction techniques. It should directly extend to test application cost sensitive induction tasks. Experimental evaluation demonstrates consistent positive effects over a range of misclassification cost sensitive learning tasks.

1 Introduction

Most research into machine learning has considered all misclassifications to have equivalent cost. However, for many applications this assumption will not be justified. For example, when diagnosing diseases, the cost of failing to diagnose some diseases will be low, because the symptoms will eventually become more pronounced, enabling suitable diagnosis. In contrast, failure to diagnose other diseases will have high cost, as irreparable damage will occur before adequate diagnosis is eventually obtained. Similarly, mis-diagnosis of a disease will in some cases have low cost—the patient receives unnecessary treatment with few side-effects; but in others will have high cost, such as undesirable side-effects.

Previous approaches to misclassification-cost sensitive induction can be considered to fall into four main categories. The first of these divides the training data into subsets on which inductive experiments are performed in order to infer a learning bias that will minimize misclassification costs [11, 14]. The selected bias is then employed to learn a classifier from the full set of training data.

The *better safe than sorry* strategy [10] considers for high misclassification cost classes only rules with high empirical support (that cover large numbers of training examples) while rules with lower empirical support are considered for low misclassification cost classes. The classification rules are learnt independently from one another. However, on application they are considered in order from lowest to highest misclassification cost.

A number of approaches alter the empirical bias of the learning system [1, 2, 5, 8, 9]. An empirical bias is a learning bias that selects between hypotheses on the basis of how they perform on the training data. This is modified so as to provide different weights to different types of misclassification.

The final category employs background knowledge to provide biases toward suitable hypotheses [3].

This paper presents the cost-sensitive specialization strategy. This strategy was inspired by the theorem of decreasing inductive power [15]. This theorem predicts increases in the proportion of false positives to true positives on previously unseen cases when a classifier is generalized without altering empirical support. (The empirical support for a classifier is evidence based on its performance on the training data.) In the context of learning with variable misclassification costs, this theorem suggests that elements of a classifier associated with high misclassification costs should be specialized (so as to minimize the proportion of false positives to true positives). In the context of classifiers that cover all of the instance space (classifiers that never respond *I don't know*), specializing one element of the classifier requires generalization of another. As the generalized elements are expected to have higher proportions of false positive to true positives, they should be selected from those with low misclassification cost. (The misclassification cost of a class is taken herein, except where specifically indicated otherwise, to mean false positive misclassification cost—the cost of incorrectly assigning the nominated class to an object.)

In the context of learning decision trees, this translates into a strategy of generalizing leaves for classes with low misclassification costs and specializing leaves for classes with high misclassification costs. In the context of learning decision rules, this translates into generalizing rules for classes with low misclassification costs and specializing rules for classes with high misclassification costs.

However, this general strategy can be justified without recourse to the theorem of decreasing inductive power. If there are cases where, without altering the expected error rate, the leaves or rules associated with high misclassification costs can be specialized in favor of leaves or rules with lower misclassification costs then such a change will decrease expected total misclassification costs as there can be expected to be a transfer from errors with high cost to those with low cost without any change in the numbers of those errors.

This paper presents a theoretical analysis of the cost-sensitive specialization strategy and evaluates a modification to the C4.5 [12] decision tree induction system that supports cost-sensitive specialization.

2 Cost sensitive specialization

Consider a region of the instance space to which one is considering assigning a single class in an n class classification learning task. This will be the case with the region associated with a single classification rule or decision tree leaf. Let $p_i =$ the expected proportion of objects belonging to class i that will be encountered in this region of the instance space. Let $c_{ij} =$ the expected cost of classifying an object belonging to class i as belonging to class j. We assume that $\forall i : c_{ii} = 0$ (correct classifications have no cost).

Let t_i be the total expected misclassification cost if the region is assigned class i. $t_i = \sum_{j=1}^{n} p_j.c_{ij}$. The empirical bias approaches to misclassification cost sensitive induction [1, 2, 5, 8, 9] seek to minimize t_i but do not consider areas of the instance space that are occupied by no training objects. These are treated as if they make no contribution to the total expected misclassification cost.

Consider the case where there is no evidence relating to the distribution of classes within a region. In such a case one should not distinguish between classes with respect to the expected proportion of objects for that class. It follows that $\forall i, j : p_i = p_j$. In this case p_i can be considered to be a constant. From this it follows that t_i will be minimized for the class for which $\sum_{j=1}^{n} c_{ij}$ is minimized. In other words, when one has no prior knowledge about the distribution of classes within a region of the instance space, expected misclassification costs will be minimized by assigning to that region the class for which the average of misclassifications to that class is minimized.

So far we have considered the case of a complex misclassification cost function, where the misclassification cost is a function of the correct class for an object and of the class assigned to that object. Two less complex forms of cost function are worth considering. A false positive misclassification cost function is one where misclassification costs are a function of the class assigned to an object—$\forall i, j, k : i \neq j \wedge i \neq k \rightarrow c_{ji} = c_{ki}$.. A false negative misclassification cost function is one where misclassification costs are a function of the class of an object—$\forall i, j, k : i \neq j \wedge i \neq k \rightarrow c_{ij} = c_{ik}$..

For a false positive misclassification cost function, $\sum_{j=1}^{n} c_{ij}$ is ordered on the relative false positive misclassification cost for a class. For a false negative misclassification cost function $\sum_{j=1}^{n} c_{ij}$ is ordered in the reverse of the order of the false negative misclassification cost for a class.

To summarize, in general one should seek to minimize the expected misclassification cost function. Where there is no evidence as to the relative frequencies of alternative classes for a region of the instance space, expected misclassification costs are minimized by selecting the class for which the mean expected cost of misclassifications is lowest. For false positive misclassification cost functions this is the class with the lowest false positive misclassification cost. For false negative misclassification cost functions this is the class with the highest false negative misclassification cost.

Where one has available an empirical bias that is able to minimize expected errors, the above analysis suggests that between alternatives that maximize that empirical bias, one should favor classifiers for which classes with high mean expected misclassification cost are as specific as possible and classes with low mean expected misclassification costs are as general as possible. This is because if the empirical bias gives equal weighting to the two classifiers, on the assumption that it agrees with respect to the expected error rates for regions of the instance space for which the two classifiers agree, then it does not in general distinguish between the expected error rates for the classes involved in regions of the instance space which are associated with different classes by the two classifiers. As the above analysis shows, associating such regions with classes with low mean expected misclassification costs minimizes total expected misclassification costs. Cost-sensitive specialization is a bias toward specializing aspects of classifiers associated with high expected misclassification cost where such specialization is neutral with respect to the other learning biases. Such specialization has the associated effect of generalizing aspects of classifiers that are associated with low expected misclassification cost.

3 C4.5CS

C4.5CS is a decision tree post-processor that is used in conjunction with C4.5. This post-processor implements cost-sensitive specialization by seeking to specialize leaves for high misclassification cost classes in favor of leaves for low misclassification cost classes. The current implementation of C4.5CS assumes a false positive misclassification cost function.

The decision trees learnt by C4.5 have different forms of decision nodes for discrete and continuous attributes. For continuous attributes C4.5 selects a cut value and generates two branches. A test is generated of the form $v \leq cut$. Objects for which this test succeeds pass down one branch while those for which it fails pass down the other.

With respect to branches on continuous attributes that lead directly to leaves, it is straight forward to specialize the leaf for the higher misclassification cost class. This is achieved by moving the cut to the most extreme value that specializes the branch for the higher misclassification cost class. C4.5 sets the cut at the greatest value for an object from the training set that passes down the $\leq$ branch. It is therefore not possible to further specialize the $\leq$ branch without affecting the empirical support for a classifier. The method employed herein to specialize the $>$ branch is to set the cut to the least value of a training object that passes down that branch and to alter the test to $<$. This ensures the greatest possible specialization without affecting the empirical support for the classifier.

However, not all branches lead directly to leaves. Relaxing the condition attached to a branch will generalize all leaves below that branch. This will in general only be desirable when all leaves below that branch are associated with classes with costs no greater than the lowest misclassification cost of a class for a leaf below the alternative branch. To this end, a split on a continuous attribute with $\leq$ branch l and $>$ branch g is changed to a $<$ test if and only if $max\,(cost(x) : leaf(x) \wedge below(l, x)) \quad \leq \quad min\,(cost(y) : leaf(y) \wedge below(g, y))$ where $leaf(x)$ is true if and only if node x is a leaf; $below(x, y)$ is true if and only if node y is below branch x; and $cost(x)$ = the false positive misclassification cost associated with the class for leaf node x.

For discrete attributes C4.5, by default, generates a branch for each value. (An optional technique for grouping multiple values to a single branch is not considered herein but could be treated by separating out values that apply to no training objects.) It is not possible to directly generalize or specialize any of these branches by altering the test. However, when no objects from the training set follow a branch, C4.5 constructs a leaf node for the class that dominates the node from which the branch descends. As such an assignment has only weak empirical support, it might be possible to change the class for such a leaf with

Table 1. UCI data sets used for experimentation

Name	No. of Attrs.	% continuous	No. of objects	No. of classes
audiology	69	0	226	24
autos	25	44	205	7
breast cancer Slovenia	9	4	286	2
breast cancer Wisconsin	9	100	699	2
Cleveland heart disease	13	46	303	2
credit rating	15	40	690	2
echocardiogram	6	83	74	2
glass	9	100	214	3
hepatitis	19	32	155	2
house votes 84	16	0	435	2
Hungarian heart disease	13	46	295	2
hypothyroid	29	24	3772	4
iris	4	100	150	3
lymphography	18	38	148	4
new thyroid	5	100	215	3
Pima indians diabetes	8	100	768	2
primary tumor	17	12	339	22
promoters	57	0	106	2
soybean large	35	0	307	19
tic-tac-toe	9	0	958	2

little risk of increasing expected errors. Changing the class to that with the lowest misclassification cost leads to generalizing the proportion of the instance space associated with that class and specializing that associated with higher misclassification cost classes.

C4.5 develops two types of decision tree—pruned and unpruned trees. C4.5CS post-processes both types of tree. It identifies and performs all of the types of generalization described above for each tree to which it is applied.

4 Experimental evaluation

To evaluate the efficacy of this approach, C4.5CS was applied to the twenty data sets from the UCI repository of machine learning databases described in Table 1. For each data set this table lists the number of attributes by which each object is described, the proportion of these that are continuous, the number of objects in the data set and the number of classes into which these objects are divided. These data sets were selected with the intention of exploring as wide a cross-section of attribute-value machine learning tasks as possible. In the absence of true cost functions for a wide range of learning tasks and in the interest of exploring as wide a range of different types of misclassification cost sensitive task as possible, a range of different false positive misclassification cost functions were randomly generated for each data set.

For each data set, 100 runs were performed. For each run—

Table 2. Errors

	Unpruned trees				Pruned trees			
Data Set	C4.5	C4.5CS	p	Ratio	C4.5	C4.5CS	p	Ratio
audiology	11.3±0.3	11.8±0.3	0.000	1.06	10.7±0.3	11.1±0.3	0.000	1.04
autos	10.6±0.4	10.8±0.4	0.032	1.02	10.8±0.4	11.0±0.4	0.022	1.02
breast cancer Slov.	21.6±0.4	21.6±0.4	0.708	1.00	17.1±0.4	17.1±0.4	0.530	1.00
breast cancer Wisc.	8.3±0.2	8.4±0.2	0.132	1.01	7.2±0.2	7.2±0.2	1.000	1.00
Cleveland heart dis.	16.6±0.3	16.6±0.3	0.744	1.00	16.1±0.3	16.1±0.3	0.770	1.00
credit rating	25.0±0.5	25.3±0.5	0.011	1.02	21.4±0.4	21.5±0.4	0.193	1.01
echocardiogram	4.0±0.2	3.8±0.2	0.013	0.97	3.8±0.1	3.6±0.1	0.000	0.96
glass	13.7±0.3	13.8±0.3	0.734	1.00	13.8±0.3	13.8±0.3	1.000	1.00
hepatitis	6.9±0.2	6.8±0.2	0.028	0.99	6.5±0.2	6.4±0.2	0.032	0.98
Hung. heart dis.	14.0±0.3	13.9±0.3	0.140	0.99	12.9±0.3	12.8±0.3	0.195	0.99
house votes 84	5.0±0.2	5.1±0.2	0.074	1.04	4.5±0.2	4.6±0.2	0.198	1.01
hypothyroid	4.1±0.2	4.1±0.2	1.000	1.01	4.1±0.2	4.1±0.2	0.083	1.01
iris	1.5±0.1	1.5±0.1	0.368	1.00	1.5±0.1	1.5±0.1	0.250	1.00
lymphography	8.1±0.3	8.1±0.3	1.000	1.00	8.0±0.3	8.0±0.3	0.158	1.00
new thyroid	4.0±0.2	4.0±0.2	0.549	1.03	4.1±0.2	4.1±0.2	0.558	1.03
Pima indians diab.	46.2±0.5	46.0±0.5	0.107	1.00	43.0±0.5	42.9±0.5	0.235	1.00
promoters	5.4±0.2	5.5±0.2	0.566	1.03	5.4±0.2	5.4±0.2	0.664	1.03
primary tumor	40.8±0.5	40.8±0.5	0.045	1.00	40.7±0.4	40.7±0.4	0.083	1.00
soybean large	15.5±0.5	16.0±0.5	0.000	1.03	15.3±0.6	15.3±0.6	0.259	1.03
tic-tac-toe	29.6±0.7	29.5±0.7	0.101	0.99	31.9±0.7	31.7±0.7	0.030	0.99
Mean ratio				1.01				1.00

1. the data was randomly divided into training (80%) and evaluation (remaining 20%) sets.
2. misclassification costs were assigned to classes as follows.
   (a) The classes were randomly ordered from 0 to $n-1$.
   (b) Each class was assigned the misclassification cost $i.\frac{99}{n-1}+1$ (truncated to the closest integer) where i is the rank order of the class.

   This ensured that the minimum cost was 1, the maximum cost was 100 and that the remaining costs were evenly spaced between those extremes.
3. Both C4.5 and C4.5CS were applied to learn decision trees from the training set. Both pruned and unpruned decision trees were learnt by C4.5 and post-processed by C4.5CS.
4. All decision trees were applied to the evaluation set and the total number of errors and total cost of misclassifications calculated.

Table 2 presents the means and standard errors for the numbers of errors per run in these experiments. For each of unpruned and pruned trees this table presents the mean and standard error of each treatment followed by the result of a two-tailed t-test comparing these means and the ratio of total numbers of errors (C4.5CS/C4.5). The bottom row presents the mean ratio of numbers of errors for C4.5CS against C4.5. This is the mean of the ratio for each run.

Post-processing had a variable effect on the total error rate. While C4.5CS is seeking to perform specializations that will not affect total error rate, for

Table 3. Costs

	Unpruned trees				Pruned trees			
Data Set	C4.5	C4.5CS	p	Ratio	C4.5	C4.5CS	p	Ratio
audiology	560±18.7	507±18.8	0.000	0.90	536±17.6	489±17.4	0.000	0.91
autos	543±31.3	519±31.6	0.000	0.95	552±29.9	539±30.1	0.001	0.97
breast c. Slov.	1139±46.5	1124±45.6	0.001	0.99	925±58.8	923±58.3	0.330	1.00
breast c. Wisc.	408±21.1	407±21.3	0.350	0.99	329±18.1	329±18.1	1.000	1.00
Cleve. heart dis.	836±30.1	808±29.0	0.000	0.97	800±28.3	781±27.7	0.000	0.98
credit rating	1210±33.6	1155±33.2	0.000	0.96	1041±32.9	1014±32.1	0.000	0.98
echocardiogram	205±13.9	184±13.9	0.000	0.91	188±11.4	169±11.3	0.000	0.91
glass	665±28.2	641±27.5	0.000	0.96	658±28.4	633±27.2	0.000	0.96
hepatitis	330±17.7	317±17.6	0.000	0.95	320±18.3	310±19.0	0.004	0.96
Hung. heart dis.	727±31.9	706±31.8	0.000	0.97	668±33.4	652±33.0	0.000	0.98
house votes 84	261±18.6	251±17.6	0.013	0.98	227±19.9	222±20.0	0.028	0.99
hypothyroid	191±12.8	189±12.9	0.265	1.00	183±12.3	183±12.5	0.975	1.00
iris	79±7.8	78±7.7	0.338	1.00	78±7.6	77±7.4	0.341	1.00
lymphography	359±21.6	346±21.0	0.000	0.97	350±21.6	343±21.4	0.005	0.98
new thyroid	186±16.4	176±16.5	0.011	1.00	190±16.9	179±16.7	0.010	1.00
Pima ind. diab.	2353±48.6	2266±48.4	0.000	0.96	2221±50.1	2154±50.7	0.000	0.97
promoters	274±17.0	236±15.3	0.000	0.90	264±17.3	236±16.1	0.000	0.91
primary tumor	1971±38.0	1969±38.0	0.112	1.00	1956±39.6	1956±39.6	0.083	1.00
soybean large	798±34.2	741±31.4	0.000	0.94	784±41.8	757±39.2	0.037	0.97
tic-tac-toe	1618±80.4	1558±80.5	0.000	0.96	1713±98.9	1673±99.4	0.000	0.97
Mean ratio				0.96				0.97

unpruned trees it is averaging 1% more errors than C4.5. For pruned trees the overall difference in error rates is negligible. On a treatment by treatment basis the effect varied from an increase in total errors of 6% for the audio data set with unpruned trees to a decrease in total errors of 4% for the echocardiogram data set with pruned trees.

It appears that the effect of C4.5CS is smaller for pruned trees than unpruned trees. Of the 592 occasions on which the number of errors for a C4.5 unpruned tree differed from the C4.5CS unpruned tree, in 219 cases there was no difference between the corresponding pruned trees. In comparison, of the 406 occasions on which the number of errors for a C4.5 pruned tree differed from the C4.5CS pruned tree, in only 33 did the unpruned trees not also differ. A binomial sign test indicates that this difference in numbers of unique effects is significant (p=0.000). It is hardly surprising that C4.5CS has more frequent effect for unpruned than for pruned trees as pruned trees have fewer nodes and hence fewer opportunities for C4.5CS to make a change. In particular, pruning frequently deletes the leaves associated with no training items. These leaves provide the only mechanism relating to discrete attributes that C4.5CS can apply.

Table 3 presents the means and standard errors for the misclassification costs per run. This table follows the format of Table 2.

Despite the slight increase in errors apparent in Table 2, Table 3 shows that there is a marked decrease in misclassification costs. For 37 of the 40 treatments there is a decrease in mean misclassification costs. The only exceptions are the

Wisconsin breast cancer data set with pruned trees for which both systems had identical costs for all 100 runs; the hypothyroid data set for pruned trees for which the mean costs to 2 decimal places are C4.5: 183.58 and C4.5CS: 183.61 and the primary tumor data set for pruned trees for which the mean costs to 2 decimal places are C4.5: 1956.43 and C4.5CS: 1956.46. These small differences are not statistically significant. On average, misclassification costs were reduced by 4% for unpruned trees and by 3% for pruned trees. Decreases of 8% or more occurred for the audio, echocardiogram and promoters data sets for both pruned and unpruned trees. In the case of unpruned trees for the promoters data set, the decrease was 10%.

The effect appears smaller for pruned trees than unpruned. Of the 725 occasions on which the total costs differ for the unpruned C4.5 and C4.5CS trees, on 236 occasions there is no difference between the corresponding pruned trees. Of the 511 occasions on which the pruned C4.5 and C4.5CS trees differ, the corresponding unpruned trees fail to differ in only 22. A binomial sign test indicates that this difference in numbers of unique effects is significant (p=0.000). This can be explained by the decrease in opportunities for post-processing with pruned trees.

Table 4 presents the results of binomial sign tests comparing the numbers of times each system obtained a higher value than the other for each measure. With respect to unpruned decision trees, C4.5 had fewer errors significantly more often than did C4.5CS. With respect to pruned decision trees, C4.5 had fewer errors more often than did C4.5CS, but this difference was not statistically significant. For both pruned and unpruned trees, C4.5CS had lower misclassification costs significantly more often than did C4.5.

Table 4. Summary comparison

Measure	C4.5 > C4.5CS	C4.5 < C4.5CS	p
Unpruned Errors	251	341	0.0001
Pruned Errors	196	210	0.2594
Unpruned Costs	472	253	0.0000
Pruned Costs	332	179	0.0000

These experimental results demonstrate that C4.5CS can significantly reduce misclassification costs for a wide range of learning tasks. The numbers of errors that were observed demonstrate that this effect cannot be attributed to a reduction in the numbers of errors.

5 Extension to complex cost functions

C4.5CS is restricted in application to situations in which the relative misclassification costs of the classes can be ordered. Note that it is not dependent upon the assignment of accurate misclassification costs or ratios between costs. The only information that it utilizes is the order of these costs.

It can be utilized with orderings either by the costs of false positives or of false negatives. While the above work has considered misclassification costs to relate to false positives (the cost is a function of the class that is incorrectly assigned to the object) the techniques are equally applicable to situations where the misclassification costs relate to the costs of false negatives (the cost is a function of the class to which the misclassified object belongs). In the latter context, classes with high misclassification costs should be generalized (to minimize a chance of a false negative) and those with low misclassification costs should be specialized.

C4.5CS could be extended to more complex cost functions by considering at each branch on a continuous attribute the greatest cost of misclassifying an object of any class as belonging to a class represented by a leaf on the $\leq$ branch and comparing this to the minimum cost of misclassifying an object of any class as belonging to a class represented by a leaf on the $>$ branch.

A default class i for leaves that cover no objects in the training set could be selected on the basis of minimizing the mean cost of incorrectly classifying an object of another class as belonging to i.

While the modifications made by C4.5CS depend upon the relative ordering of misclassification costs but not their relative magnitude, the size of the effect of its application will depend upon the magnitude. The greater the difference in magnitude of misclassification costs for different classes the greater the expected reduction in total misclassification costs resulting from its application.

6 Increasing the degree of specialization

The cost-sensitive specialization strategy espoused by this paper involves the specialization of aspects of a classifier relating to high false positive or low false negative misclassifications and generalization of those relating to low false negative or high false positive misclassifications where those specialization and generalization actions have low impact on expected accuracy. This has been evaluated in the context of two specialization/generalization operators for C4.5 decision trees, one relating to tests on continuous attributes and the other to tests on discrete attributes.

For these operators the degree of reduction in misclassification costs is small but consistent. Where greater reductions are sought, more powerful specialization/generalization operators should be employed. Candidate operators for decision trees include the generation of characteristic leaves [4] and cuts based on evidence for neighboring regions of the instance space [16]. A variety of potentially useful specialization and generalization operators for classification rules have been described elsewhere [15].

This research deliberately avoided the use of these more powerful operators as they are likely to also improve predictive accuracy. Operators that did not improve predictive accuracy were used in order to provide clear cut support for the predicted effect without the possibility that costs were reduced simply by a reduction in errors.

7 Relationship to previous approaches

Unlike approaches based on altering the empirical bias, the cost-sensitive specialization approach does not require accurate misclassification costs. All that is required is a relative ordering of the misclassification costs. Nor does cost-sensitive specialization require additional background knowledge.

It is distinguished from approaches that use induction to select learning biases by avoiding the problem of induction and the subsequent risk that the inferred bias will turn out to be inappropriate.

Cost-sensitive specialization can be seen as a more specific statement of the *better safe than sorry* policy. In this light, the two types of mechanism that Provost and Buchanan [10] propose can be seen as techniques for specializing high misclassification cost rules and generalizing low misclassification cost rules. Relative ordering of rules does this by implicit conjunction of the conditions for subsequent rules with negations of the conditions for preceding rules. Allowing rules with lower empirical support to be developed for low misclassification cost classes has the effect of generalizing the total classifier with respect to those classes.

Cost-sensitive specialization could be applied in conjunction with alternative approaches to misclassification cost sensitive learning in the expectation of further boosting the effect of those approaches.

8 Other types of classification cost

While cost sensitive specialization has been presented as a means of minimizing misclassification costs, the same technique could also be employed to minimize other costs such as the costs of applying tests [6, 7, 13, 14]. In the context of a decision tree, specialization of branches leading to high cost tests will reduce the average cost of applying the tree as those tests will be applied less frequently. Where such specialization is neutral with respect to expected misclassification rate, there will be a reduction in expected application costs with no effect to expected accuracy.

9 Conclusion

Cost-sensitive specialization is a generic technique for cost sensitive induction. This technique involves specializing aspects of a classifier associated with high costs and generalizing those associated with low costs. It

- is widely applicable;
- is simple to implement;

- requires only relative ordering of costs rather than precise ratios between costs;
- is based on a simple intuitive principle; and
- has demonstrated consistent positive effect on a wide range of learning tasks.

While it has been implemented herein as a post-processor for an existing machine learning system, cost-sensitive specialization could also be employed directly during initial tree induction. The technique should extend in a straight forward manner to sensitivity to costs of test application. It should be emphasised that the implementation of the technique herein has been intended solely as proof-of-concept. To this end, specialisations that may be expected to increase predictive accuracy, such as has been investigated elsewhere [**?**] have been avoided. The magnitude of gains could be expected to rise if such specialisations were included.

It should also be noted that these techniques would be best applied to augment other cost sensitive induction techniques [1, 2, 3, 5, 8, 9, 10, 11], with which they are fully compatible. Cost sensitive specialisation could be applied to augment any induction technique that relies primarily upon empirical support for selecting between alternative hypotheses.

Acknowledgements

This research has been supported by the Australian Research Council.

I am grateful to Foster Provost for comments on an earlier draft of this paper.

The following data sets were compiled and donated by the corresponding donors—breast cancer Slovenia, lymphography and primary tumor: Ljubljana Oncology Institute, Slovenia; breast cancer Wisconsin: Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison; audiology: Prof. Jergen, Baylor College of Medicine. Thanks to the UCI Repository, its maintainers, Patrick Murphy and David Aha, and its donors, for providing access to the data sets used herein.

References

1. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International, Belmont, CA, 1984.
2. B. A. Draper, C. E. Brodley, and P. E. Utgoff. Goal-directed classification using linear machine decision trees. *IEEE Transactions on Pattern Recognition and Artificial Intelligence*, 16:888–893, 1994.
3. D. Gordon and D. Perlis. Explicitly biased generalization. *Computational Intelligence*, 5(2):67–81, May 1989.
4. R. C. Holte, L. E. Acker, and B. W. Porter. Concept learning and the problem of small disjuncts. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 813–818, Detroit, 1989. Morgan Kaufmann.
5. U. Knoll, G. Nakhaeizadeh, and B. Tausend. Cost-sensitive pruning of decision trees. In F. Bergadano and L. De Raedt, editors, *Proceedings of the Eighth European Conference on Machine Learning, ECML-94*, pages 383–386, Catania, Italy, 1994. Springer-Verlag.

6. S. W. Norton. Generating better decision trees. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, IJCAI-89*, pages 800–805, Detroit, MI, 1989. Morgan Kaufmann.
7. M. Nùñez. The use of background knowledge in decision tree induction. *Machine Learning*, 6:231–250, 1991.
8. M. J. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Macine Learning*, pages 217–225, New Jersey, 1994. Morgan Kaufmann.
9. F. J. Provost. Goal-directed inductive learning: Trading off accuracy for reduced error cost. In *Proceedings of the AAAI Spring Symposium on Goal Directed Learning*, pages 94–101, 1994.
10. F. J. Provost and B. G. Buchanan. Inductive policy. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pages 255–261. AAAI Press, 1992.
11. F. J. Provost and B. G. Buchanan. Inductive policy: The pragmatics of bias selection. *Machine Learning*, 20(1/2):35–61, 1995.
12. J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.
13. M. Tan. Cost-sensitive learning of classification knowledge and its applications in robotics. *Machine Learning*, 13:7–33, 1993.
14. P. D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.
15. G. I. Webb. Generality is more significant than complexity: Toward alternatives to Occam's razor. In C. Zhang, J. Debenham, and D. Lukose, editors, *AI'94 – Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence*, pages 60–67, Armidale, 1994. World Scientific.
16. G. I. Webb. Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, 4:397–417, 1996.

This article was processed using the LaTeX macro package with LLNCS style