Pre-publication draft of a paper which appeared in the Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence (AI'94) - Artificial Intelligence: Sowing the Seeds for the Future, pp 60-67

GENERALITY IS MORE SIGNIFICANT THAN COMPLEXITY: TOWARD AN ALTERNATIVE TO OCCAM'S RAZOR

GEOFFREY I WEBB
School of Computing and Mathematics, Deakin University
Geelong Vic 3217 Australia

ABSTRACT

Occam's Razor is widely employed in machine learning to select between classifiers with equal empirical support. This paper presents the theorem of decreasing inductive power: that, all other things being equal, if two classifiers *a* and *b* cover identical cases from the training set and *a* is a generalisation of *b*, *a* has higher probability than *b* of misclassifying a previously unsighted case. This theorem suggests that, to the contrary of Occam's Razor, generality, not complexity, should be used to select between classifiers with equal empirical support. Two studies are presented. The first study demonstrates that the theorem of decreasing inductive power holds for a number of commonly studied learning problems and for a number of different means of manipulating classifier generality. The second study demonstrates that generality provides a more consistent indicator of predictive accuracy in the context of a default rule than does complexity. These results suggest that the theorem of decreasing predictive power provides a suitable theoretical framework for the development of learning biases for use in selecting between classifiers with identical empirical support.

Introduction

One of the most important aspects of a computational learning system is the learning bias[1] that it embodies. The learning bias is the set of the factors that influence the system's selection of a classifier given a training set of data. Many factors may enter this bias, such as, the type of classifier that the system is capable of expressing.

Most machine learning systems perform heuristic search through the space of classifiers that they are capable of expressing, seeking a classifier that maximises some preference function. A major factor evaluated by any such preference function is how well a given classifier fits the training data. Examples of such functions include the entropy function[2], various information measures[3,4], the Laplace error estimate[5] or a preference for a classifier that correctly classifies the most positive cases while misclassifying no negative cases[6]. Such a preference function can be considered an explicit formulation of one of the learning system's primary learning biases.

However, for most learning problems, there are large numbers of competing classifiers all of which equally maximise any such function. To select between such classifiers it is necessary to invoke a secondary learning bias. Most machine learning systems explicitly or implicitly employ a bias toward the least complex of any two classifiers that equally well explain the training data. This bias is called *Occam's Razor*.

In addition to its almost universal use in machine learning, the principle of Occam's Razor is widely accepted in general scientific practice.

However, Occam's Razor has been subjected to strong philosophical attack. To summarise Quine[7], the complexity of a theory (classifier) depends entirely upon the language in which it is encoded. To claim that the acceptability of a theory depends upon the language in which it happens to be expressed appears nothing short of ludicrous.

Several attempts have been made to provide theoretical support for the principle of Occam's Razor in the machine learning context[8,9]. However, these *proofs* apply equally to any bias that favours a small random subset of the available classifiers[10].

A further issue is that the objectives of machine learning may vary greatly[11]. It seems surprising that a single secondary learning bias should always be appropriate irrespective of whether one is seeking to maximise sensitivity, specificity, positive predictive value, negative predictive value or accuracy. It would appear more likely that a plausible framework for developing secondary learning biases would alter the bias depending upon the objective of a particular learning task.

In this context, it seems hard to understand a) why the principle of Occam's Razor is so wide spread in machine learning; and b) how so many successful systems could be based on a principle with such poor theoretical foundations.

This paper provides a theoretical framework that suggests alternative secondary learning biases based on classifier generality and which explains why the principle of Occam's Razor provides an appropriate secondary learning bias in some contexts. Experimental support is provided for this framework.

The theorem of decreasing inductive power

It is common to conceptualise classification learning problems as tasks requiring the division of a geometric space, called the instance space, into a number of discrete regions each of which is labelled with a single class[12].

A common implicit underlying assumption in machine learning is the axiom of local uniformity: that objects that are close to one another in an instance space have high probability of belonging to the same class[12]. This provides justification for developing classifiers that partition the instance space into regions that are each predominantly occupied by objects from the training set belonging to a single class.

However, while this principle is simple to express, it is extremely difficult to operationalise due to obstacles to the definition of a domain independent measure of distance within an instance space. One obstacle arises from the need to standardise measurements for different attributes that may be based on incommensurable scales. Standardising values based upon the observed range of values is wide spread, but is subject to vagaries introduced by the adequacy of a sample. It is also very difficult to adequately measure distances between cases on the basis of differing values for non-ordinal attributes. Of even greater importance, even within a single ordinal attribute, there is no guarantee that the scale employed should be linear. For example, it is arbitrary whether an attribute should be recorded as an unfiltered value n or should be recorded as $\log n$ or n^x or indeed any other transformation. Clearly, the distances obtained by simple Euclidean geometry at different points in each of these different types of scale will differ substantially. For example, the distance between the unfiltered variable at values 0 and 1 will be identical to the distance at values 10 and 11. However, if the variable happens to be recorded as n^2 the distance will appear to be twenty-one times as great (the recorded values for 0 and 1 will be 0 and 1, an apparent distance of 1, but the recorded values for 10 and 11 will be 100 and 121, an apparent distance of 21).

These difficulties undermine attempts to compare classifiers by precise measurement of distances between points in the regions of an instance space that they identify.

Nonetheless, despite these difficulties, it is possible to make certain general observations about particular types of classifiers. In particular, it is possible to defend the axiom of increasing distance: that if two classifiers a and b cover identical cases from the training set and a is a generalisation of b, then, in the absence of evidence to the contrary, this provides evidence that the average distance of a point in a from an instance in the training set that belongs to the dominating class will be greater than the average distance of a point in b from an instance of the dominating class.

This principle is illustrated in Figure 1. In this figure, the outer heavy line and inner light line represent the boundaries of the region covered by the more general classifier (a)

and the more specific classifier (*b*), respectively. Positive instances are indicated by an 'x'. If Euclidean geometry is applicable and the scales are linear then it is clear that the average distance to a positive instance from points in the region of the instance space covered by the more general classifier is higher than the average distance to a positive instance from points in the area of the instance space covered by the more specific classifier. However, it should be noted that while it is plausible that the average distance is greater for the more general classifier, it is not necessarily so. For example, if the space is distorted so that distances to points at the centre of the space are greatly amplified in comparison to distances at the margins of the space, the average distance to a positive instance from points covered by the more general classifier may be less than that for the more specific classifier.

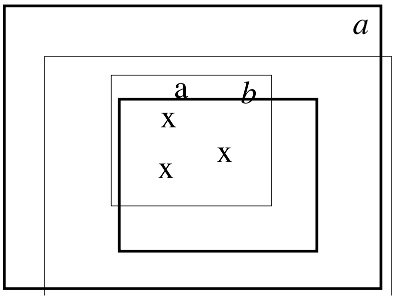

Figure 1: Regions covered by two classifiers covering identical instances from the training set.

From the axioms of local uniformity and increasing distance it is possible to derive the theorem of decreasing inductive power: that, all other things being equal, if two classifiers *a* and *b* cover identical cases from the training set and *a* is a generalisation of *b*, the probability of *a* misclassifying previously unsighted cases is higher than that of *b*. In other words, in general, the predictive accuracy of a more general classifier will be lower than that of a more specific classifier, as it will cover cases located further from previously sighted positive cases. This will be balanced, however, by the necessary outcome that the number of cases for which a more general classifier forms a classification will be greater than for a more specific classifier.

Note that while the two axioms are based upon a geometric model of classification learning, the theorem derived therefrom makes non-trivial predictions that are independent of such a geometric model.

The remainder of this paper evaluates the theorem of decreasing inductive power and examines its implications for classification learning.

Experimental conditions

To evaluate the theorem of decreasing inductive power, classifiers of varying generality that covered exactly the same cases from the training set were examined. Two alternative methods of generalisation were employed. Conjunct deletion increased generality while decreasing classifier complexity. In contrast, disjunct addition increased both generality and classifier complexity. These manipulations permitted experimental comparison of learning biases based on each of generality and Occam's Razor.

Both methods for developing alternative rule sets started by developing a set of classification rules using the DLG algorithm with further generalisation[13]. The antecedent of each of these rules took the form of a conjunction of attribute-value tests. For categorical attributes these tests took the form of a condition $a \in \{v_1, v_2, ..., v_n\}$ where *a* is an attribute and v_i is a value for that attribute. Missing values for categorical attributes were treated as distinct values. For ordinal attributes, five forms of condition were allowed: $a \leq v$, $a \geq v$, *a is missing*, $a \leq v \vee$ *a is missing* and $a \geq v \vee$ *a is missing*. The consequent of each rule was a simple classification statement.

The use of DLG with least generalisation ensured that:

a) it was not possible to specialise the antecedent of any rule without decreasing the number of cases from the training set that it covered; and

b) it was not possible to generalise the antecedent of any rule so as to increase the number of positive cases that it covered without also increasing the number of negative cases that it covered.

Rules were generated using the maximum consistent preference function. This preference function calculates the value of a classification rule as follows:

$$\text{if } \mathit{neg\_cover} > 0 \text{ then } \mathit{value} = -\mathit{neg\_cover} \text{ else } \mathit{value} = \mathit{pos\_cover}$$

where *pos_cover* represents the number of positive cases covered by a rule and *neg_cover* represents the number of negative cases covered by that rule.

DLG employs heuristic search that attempts to find rules that maximise the primary preference function. DLG with the maximum consistent preference function develops classification rules that are consistent with the training set. Evaluation was also performed using the Laplace preference function. This allows the development of rules that are inconsistent with the training set when the positive examples are numerous and the counter-examples few. The results of experiments with the Laplace preference function correspond with those presented below. Those experiments are not presented herein, due to space constraints, but are available in technical report form.

The alternative classifiers developed through conjunct deletion were generated as follows. First the most specific rules (S) were developed. These were a set of highly specific classification rules generated as described above. Next, the most general rules (G) were developed by replacing each of the most specific rules by one of its greatest identical cover generalisations. A classification rule g is a greatest identical cover generalisation of classification rule s iff g is a generalisation of s; g covers exactly the same cases from the training set as s; and there is no rule x that is a generalisation of g and of which g is not a generalisation that also covers exactly the same cases from the training set as s. For each most specific rule the OPUS$^{\mathrm{S}}$ systematic search algorithm[14] was used to find all greatest identical cover generalisations from which one was selected randomly. Note that only generalisations created by deleting conjuncts were considered during this process. The two rule sets created by these means covered exactly the same cases from the training set. In other words, the empirical evidence supporting both rule sets was identical.

Alternative rules were generated using disjunct addition as follows. Most specific rules were generated as above. Next, a variant of DLG was employed to develop an alternative set of rules employing the same preference function but guaranteeing that no rule from the first rule set was included in the new rule set. These rules were added to the initial rule set to create the general rule set (D). D and S both classified all cases from the training set identically. In other words, both rule sets had identical empirical support. It should be noted, however, that the process used to create the additional rules affected the type of rules created. For example, if the most specific rules contained the best rules, as measured by the preference function, the additional rules would necessarily have lower values as measured by this function unless rules of equivalent value also existed.

If complexity is measured as the sum of all conjuncts in all rules in a rule set, as it is throughout the rest of this paper, conjunct deletion decreases complexity while disjunct addition increases complexity. It is difficult to conceive of alternative complexity metrics that alter this general pattern.

Experimental methods

The techniques were evaluated by application to the following ten machine learning data sets from the UCI repository of machine learning data sets[15]: breast cancer, echocardiogram, glass type, hepatitis, house votes 84, hypothyroid, iris, lymphography, primary tumor, and soybean large. For all of these data sets, the cases are divided into a number of mutually exclusive classes. The induction task is to develop an expert system that can classify a object by reference to the values of its attributes.

All experiments involved one hundred repetitions of the following:

a) the data was divided into a training (80%) and evaluation (20%) set.

b) all alternative induction strategies were applied to the training set.

c) each classifier so developed was evaluated by application to the evaluation set.

Study One

The first study provided experimental evaluation of the theorem of decreasing inductive power by developing classifiers of varying generality and comparing their predictive utility. Two experiments were performed. In the first experiment the conjunct deletion generalisation technique was employed. In the second experiment the disjunct addition generalisation technique was employed.

When applying the classification rules to cases in the evaluation sets, if multiple rules applied to a case the rule with the highest value according to the preference function was employed. If no rule applied to the case the case was considered unclassified.

Table 1 presents the mean predictive accuracies for experiments 1 and 2. Note that both sets of results are presented with those for the more general rules on the right. Note also that the results of a two-tailed matched pairs test of significance is presented for each pair of results to determine the statistical significance, if any, of differences in performance. Values of 0.05 or less are considered significant.

Table 1: Mean predictive accuracy.

	Conjunct Deletion			Disjunct Addition		
Data	S	G	*p*	S	D	*p*
breast cancer	72.9	72.1	0.00	71.7	72.0	0.48
echocardiogram	73.9	71.2	0.00	76.0	76.2	0.81
glass	82.5	81.2	0.00	80.6	78.9	0.00
hepatitis	88.1	84.7	0.00	88.0	86.4	0.00
house votes	95.5	95.2	0.00	95.4	95.2	0.13
hypothyroid	99.2	99.1	0.00	99.2	97.8	0.00
iris	94.7	94.7	0.85	94.3	94.3	0.99
lymphography	84.8	81.8	0.00	84.1	81.8	0.00
primary tumor	47.3	39.9	0.00	46.8	46.3	0.13
soybean large	91.5	83.8	0.00	91.9	90.8	0.00

These results support the theorem of decreasing inductive power. Irrespective of the method used to increase generality, in the majority of cases an increase in generality leads to a statistically significant decrease in mean predictive accuracy. An increase in generality in no case leads to a significant increase in mean predictive accuracy.

Table 2 presents the mean cover for each treatment. In all cases an increase in generality leads to a statistically significant increase in cover.

Study Two

While Study 1 provides powerful support for the theorem of decreasing inductive power, it does so in a context that is rarely employed in machine learning research, one in which not all cases from the evaluation set are classified by the classifier. In most machine learning research, cases that are not covered by an inferred rule are assigned

Table 2: Mean cover.

	Conjunct Deletion			Disjunct Addition		
Data	S	G	*p*	S	D	*p*
breast cancer	85.3	89.8	0.00	84.7	92.4	0.00
echocardiogram	67.9	85.5	0.00	62.2	71.1	0.00
glass	68.6	88.2	0.00	68.0	76.7	0.00
hepatitis	76.3	89.1	0.00	75.7	81.6	0.00
house votes	97.2	98.2	0.00	97.4	99.3	0.00
hypothyroid	98.6	99.4	0.00	98.6	99.0	0.00
iris	88.9	98.4	0.00	86.7	88.3	0.00
lymphography	81.1	89.5	0.00	81.6	88.4	0.00
primary tumor	55.5	83.9	0.00	55.5	63.1	0.00
soybean large	74.2	92.5	0.00	73.5	77.3	0.00

to the most common class in the training set, using the so called default rule.

The theorem of decreasing inductive power does not make strong predictions about the performance of generalisation in this context. It predicts that more general rules will make more predictions of lower quality. Whether this will increase or decrease predictive accuracy in the presence of a default rule depends upon whether the accuracy of the additional predictions is greater than that of the default rule on those cases. This leads to a suggestion that the effect of generalisation upon predictive accuracy will be identical irrespective of the means by which that generalisation is created. However, this suggestion is a weak one in that the degree and form of generalisation wrought by each treatment may differ substantially. This provides a strong contrast to Occam’s Razor which favours generality produced through conjunct deletion but which does not favour generality produced through disjunct addition. Whereas the theorem of decreasing inductive power makes a weak prediction that generalisation will produce a single effect, Occam’s Razor states that the less complex rule will outperform the more complex irrespective of their relative generality.

To evaluate these predictions, the experimental treatments of Study 1 were replicated with the sole alteration that the default rule was employed during classification. Table 3 presents the mean predictive accuracies that were obtained.

Table 3: Mean predictive accuracy with default rule.

	Conjunct Deletion			Disjunct Addition		
Data	S	G	*p*	S	D	*p*
breast cancer	68.9	70.0	0.00	69.2	70.3	0.00
echocardiogram	69.9	70.5	0.46	69.9	70.2	0.64
glass	66.7	75.0	0.00	65.5	68.0	0.00
hepatitis	82.8	82.5	0.46	82.8	83.2	0.07
house votes	94.4	94.8	0.00	94.1	94.8	0.00
hypothyroid	98.6	98.9	0.00	98.6	97.4	0.00
iris	85.0	93.6	0.00	85.6	86.6	0.00
lymphography	77.6	78.8	0.01	77.5	77.6	0.89
primary tumor	34.1	35.0	0.01	34.2	35.2	0.00
soybean large	69.6	79.1	0.00	68.8	70.9	0.00

In most cases, an increase in generality leads to a significant increase in mean predictive accuracy. In the case of generality through conjunct deletion, the exceptions are for the echocardiogram and hepatitis data for which there is no significant change in predictive accuracy. Increase in generality through disjunct addition mirrors increase in generality through conjunct deletion with two exceptions. For the hypothyroid data it leads to a significant decrease in predictive accuracy. For the lymphography data the change, while in the same direction, is not statistically significant. Thus, the weak prediction arising from the theorem of decreasing inductive power, that the effect of generalisation in the context of a default rule should be insensitive to the means by which generalisation is produced, is borne out in all but two cases and contradicted by significant changes in opposite direction in only one case.

For fourteen out of fifteen significant differences in predictive accuracy, the more general classifier outperforms the more specific. This suggests that maximisation of generality is an effective secondary learning bias for this type of learning.

Where complexity is measured in terms of the sum of conjuncts in a rule set, in all cases the experimental manipulation leads to a significant change in rule set complexity. Conjunct deletion always decreases mean complexity while disjunct addition always increases mean complexity. A detailed table of complexities cannot be presented due to space constraints, but is available in a technical report form. It is interesting to note that for the hypothyroid data disjunct addition is almost trebling the complexity of the rule set. This suggests that it is not possible to create alternative rules of equivalent value as measured by the preference function to those in the initial rule set. Instead, large numbers of rules each covering fewer cases appear to have been generated. This may account for the manner in which an increase in generality through disjunct addition leads to a decrease in predictive accuracy while an increase in generality through conjunct deletion does not.

In contrast to the manner in which the prediction based on the theorem of decreasing inductive power was borne out, the strong prediction of Occam's Razor, that less complex rules should be preferred, is contraindicated by seven out of the ten cases of generalisation through disjunct addition and only supported by one of these ten cases, the hypothyroid data.

Conclusion

This paper has presented the theorem of decreasing inductive power: that, in the absence of other evidence to the contrary, if two classifiers *a* and *b* cover identical cases from the training set and *a* is a generalisation of *b*, the expected misclassification rate of *a* is higher than that of *b*. Countering this effect is the increase in cover that results from an increase in generality. Study 1 supported the prediction that the predictive accuracy of classification rules would decrease as generality increased irrespective of the means by which generality was created.

The theorem of decreasing inductive power suggests that generality should be a major criterion for selection between classifiers with identical empirical support. Whether greater or lesser generality should be preferred must depend upon the desired outcome. Where the maximisation of predictive accuracy is desired in the presence of a default rule, the desirability of generalisation should depend upon the relative predictive accuracy of the default rule and of the more general rules in the regions covered by the more general rules and not the more specific rules. Unfortunately, this information cannot be obtained from a training set as the regions of the instance space in question are not represented therein.

However, the theorem does suggest that generality should have the same effect on predictive accuracy whether the generality is introduced by increasing or decreasing the complexity of the classifiers. This result was observed in the majority of the experimental conditions. By contrast, Occam's Razor prefers the least complex classifier irrespective of its generality. This preference was found to be counterproductive in the majority of cases where increased complexity resulted in increased generality. However, it should be noted that for the types of classifier usually employed in computational learning, there is usually a direct correlation (either positive or negative) between generality and complexity. Thus, if it is desirable to manipulate generality (as the studies presented above suggest) it will also be desirable to manipulate complexity. This, perhaps, goes some way toward explaining the persistence of Occam's Razor despite an apparent lack of theoretical or empirical support.

The results of these studies provide strong support for the theorem of decreasing inductive power. This theorem provides a theoretical framework for investigating secondary learning biases. Where maximisation of predictive accuracy at the possible expense of cover is of primary importance, the theorem of decreasing inductive power indicates that a bias toward more specific classifiers is desirable. The results of Study 2 suggest that if the objective is to maximise predictive accuracy in the presence of a default rule, that maximisation of generality should be employed as a secondary learning bias. It is hoped that the theorem will provide a framework in which learning biases suited to other learning objectives may also be derived.

Acknowledgments

I am grateful to Doug Newlands, Charlie Clelland and Mark Kuzmycz for stimulating comments that have helped to refine the ideas and presentation herein. The lymphography, primary tumor and breast cancer data was compiled by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. This research has been supported by the Australian Research Council.

References

1. T. M. Mitchell, *The need for biases in learning generalizations* (Rutgers University, Department of Computer Science, Technical Report CBM-TR-117, 1980).
2. J. R. Quinlan, in *Machine Learning: An Artificial Intelligence Approach*, ed. R. S. Michalski, J. G. Carbonell, T. M. Mitchell (Springer-Verlag, Berlin, 1984) p. 463.
3. J. R. Quinlan, *Machine Learning* **1** (1986) 81.
4. C. S. Wallace and D. M. Boulton, *Computer Journal* **11** (1968) 185.
5. P. Clark and R. Boswell, in *Proceedings of the Fifth European Working Session on Learning*, (1991) p. 151.
6. R. S. Michalski, in *Machine Learning: An Artificial Intelligence Approach*, ed. R. S. Michalski, J. G. Carbonell, T. M. Mitchell (Springer-Verlag, Berlin , 1984) p. 83.
7. W. V. O. Quine, *Synthese* **15** (1963) 103.
8. J. Pearl, *International Journal of General Systems* **4** (1978) 255.
9. A. Blumer, A. Ehrenfeucht, D. Haussler and M. K. Warmuth, *Information Processing Letters* **24** (1987) 377.
10. T. G. Dietterich, Re: Occam's Razor (*Electronic Newsgroup, comp.ai,* 23 May 1994)
11. S. M. Weiss, R. S. Galen and P. Tadepalli, *Artificial Intelligence* **45** (1990) 47.
12. L. Rendell, *Machine Learning* **1** (1986) 177.
13. G. I. Webb, *Learning Disjunctive Class Descriptions by Least Generalisation* (Deakin University School of Computing and Mathematics, Tech.Rep. C92/9, 1992).
14. G. I. Webb, *OPUS: A systematic search algorithm and its application to categorical attribute-value data-driven machine learning* (Deakin University School of Computing and Mathematics, Technical Report C93/35, 1993).
15. P. Murphy and D. Aha, UCI repository of machine learning data bases.