Pre-publication draft of a paper which appeared in LNAI State-of-the-Art Survey series, 'Data Mining: Theory, Methodology, Techniques, and Applications', 1-13, Berlin: Springer.

Generality is Predictive of Prediction Accuracy

Geoffrey I. Webb[1] and Damien Brain[2]

[1] Faculty of Information Technology
Monash University, Clayton, Vic, 3800, Australia
webb@infotech.monash.edu.au
[2] UTelco Systems
Level 50/120 Collins St Melbourne , Vic, 3001, Australia
damien.brain@utelcosystems.com.au

Abstract. During knowledge acquisition it frequently occurs that multiple alternative potential rules all appear equally credible. This paper addresses the dearth of formal analysis about how to select between such alternatives. It presents two hypotheses about the expected impact of selecting between classification rules of differing levels of generality in the absence of other evidence about their likely relative performance on unseen data. We argue that the accuracy on unseen data of the more general rule will tend to be closer to that of a default rule for the class than will that of the more specific rule. We also argue that in comparison to the more general rule, the accuracy of the more specific rule on unseen cases will tend to be closer to the accuracy obtained on training data. Experimental evidence is provided in support of these hypotheses. These hypotheses can be useful for selecting between rules in order to achieve specific knowledge acquisition objectives.

1 Introduction

In many knowledge acquisition contexts there will be many classification rules that perform equally well on the training data. For example, as illustrated by the version space [1], there will often be alternative rules of differing degrees of generality all of which agree with the training data. However, even when we move away from a situation in which we are expecting to find rules that are strictly consistent with the training data, in other words, when we allow rules to misclassify some training cases, there will often be many rules all of which cover exactly the same training cases. If we are selecting rules to use for some decision making task, we must select between such rules with identical performance on the training data. To do so requires a learning bias [2], a means of selecting between competing hypotheses that utilizes criteria beyond those strictly encapsulated in the training data.

All learning algorithms confront this problem. This is starkly illustrated by the large numbers of rules with very high values for any given interestingness measure that are typically discovered during association rule discovery. Many systems that learn rule sets for the purpose of prediction mask this problem by making arbitrary choices between rules with equivalent performance on the

training data. This masking of the problem is so successful that many researchers appear oblivious to the problem. Our previous work has clearly identified that it is frequently the case that there exist many variants of the rules typically derived in machine learning, all of which cover exactly the same training data. Indeed, one of our previous systems, The Knowledge Factory [3, 4] provides support for identification and selection between such rule variants.

This paper examines the implications of selecting between such rules on the basis of their relative generality. We contend that learning biases based on relative generality can usefully manipulate the expected performance of classifiers learned from data. The insight that we provide into this issue may assist knowledge engineers make more appropriate selections between alternative rules when those alternatives derive equal support from the available training data.

We present specific hypotheses relating to reasonable expectations about classification error for classification rules. We discuss classification rules of the form $Z \rightarrow y$, which should be interpreted as all cases that satisfy conditions Z belong to class y. We are interested in learning rules from data. We allow that evidence about the likely classification performance of a rule might come from many sources, including prior knowledge, but, in the machine learning tradition, are particularly concerned with *empirical* evidence—evidence obtained from the performance of the rule on sample (training) data. We consider the learning context in which a rule $Z \rightarrow y$ is learned from a *training set* $D'=(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_n, y'_n)$ and is to be applied to a set of previously unseen data called a *test set* $D=(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$. For this enterprise to be successful, D' and D should be drawn from the same or from related distributions. For the purposes of the current paper we assume that D' and D are drawn independently at random from the same distribution and acknowledge that violations of this assumption may affect the effects that we predict.

We utilize the following notation.

$Z(I)$ represents the set of instances in instance set I covered by condition Z.

$E(Z \rightarrow y, I)$ represents the number of instances in instance set I that $Z \rightarrow y$ misclassifies (the absolute error).

$\varepsilon(Z \rightarrow y, I)$ represents the proportion of instance set I that $Z \rightarrow y$ misclassifies (the error) $= \frac{E(Z \rightarrow y, I)}{|I|}$.

$W \gg Z$ denotes that the condition W is a proper generalization of condition Z. $W \gg Z$ if and only if the set of descriptions for which W is true is a proper superset of the set of descriptions for which Z is true.

$NODE(W \rightarrow y, Z \rightarrow y)$ denotes that there is no other distinguishing evidence between $W \rightarrow y$ and $Z \rightarrow y$. This means that there is no available evidence, other than the relative generality of W and Z, indicating the likely direction (negative, zero, or positive) of $\varepsilon(W \rightarrow y, D) - \varepsilon(Z \rightarrow y, D)$. In particular, we require that the empirical evidence be identical. In the current research the learning systems have access only to empirical evidence and we assume that $W(D')=Z(D') \rightarrow NODE(W \rightarrow y, Z \rightarrow y)$. Note that $W(D')=Z(D')$ does not preclude W and Z from covering different test cases at classification time and hence having different test set error. We utilize the notion of

other distinguishing evidence to allow for the real-world knowledge acquisition context in which evidence other than that contained in the data may be brought to bear upon the rule selection problem.

We present two hypotheses relating to classification rules $W \rightarrow y$ and $Z \rightarrow y$ learned from real-world data such that $W \gg Z$ and $NODE(W \rightarrow y, Z \rightarrow y)$.

1. $Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(true \rightarrow y, D)| < |\varepsilon(Z \rightarrow y, D) - \varepsilon(true \rightarrow y, D)|) > Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(true \rightarrow y, D)| > |\varepsilon(Z \rightarrow y, D) - \varepsilon(true \rightarrow y, D)|)$. That is, the error of the more general rule, $W \rightarrow y$, on unseen data will tend to be closer to the proportion of cases in the domain that do not belong to class y than will the error of the more specific rule, $Z \rightarrow y$.
2. $Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(W \rightarrow y, D')| > |\varepsilon(Z \rightarrow y, D) - \varepsilon(Z \rightarrow y, D')|) > Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(W \rightarrow y, D')| < |\varepsilon(Z \rightarrow y, D) - \varepsilon(Z \rightarrow y, D')|)$. That is, the error of the more specific rule, $Z \rightarrow y$, on unseen data will tend to be closer to the proportion of negative training cases covered by the two rules[3] than will the error of the more general rule, $W \rightarrow y$.

Another way of stating these two hypotheses is that of two rules with identical empirical and other support,

1. the more general can be expected to exhibit classification error closer to that of a *default rule*, $true \rightarrow y$, or, in other words, of assuming all cases belong to the class, and
2. the more specific can be expected to exhibit classification error closer to that observed on the training data.

It is important to clarify at the outset that we are not claiming that the more general rule will invariably have closer generalization error to the default rule and the more specific rule will invariably have closer generalization error to the observed error on the training data. Rather, we are claiming that relative generality provides a source of evidence that, in the absence of alternative evidence, provides reasonable grounds for believing that each of these effects is more likely than the contrary.

Observation. With simple assumptions, hypotheses (1) and (2) can be shown to be trivially true given that D' and D are idd samples from a single finite distribution $\mathcal{D}$.

Proof.

1. For any rule $X \rightarrow y$ and test set D, $\varepsilon(X \rightarrow y, D) = \varepsilon(X \rightarrow y, X(D))$, as $X \rightarrow y$ only covers instances $X(D)$ of D.
2. $\varepsilon(Z \rightarrow y, D) = \frac{E(Z \rightarrow y, Z(D \cap D')) + E(Z \rightarrow y, Z(D - D'))}{|Z(D)|}$
3. $\varepsilon(W \rightarrow y, D) = \frac{E(W \rightarrow y, W(D \cap D')) + E(W \rightarrow y, W(D - D'))}{|W(D)|}$
4. $Z(D) \subseteq W(D)$ because Z is a specialization of W.

[3] Recall that both rules have identical empirical support and hence cover the same training cases.

5. $Z(D \cap D') = W(D \cap D')$ because $Z(D') = W(D')$.
6. $Z(D - D') \subseteq W(D - D')$ because $Z(D) \subseteq W(D)$.
7. from 2-6, $E(Z \rightarrow y, Z(D \cap D'))$ is a larger proportion of the error of $Z \rightarrow y$ than is $E(W \rightarrow y, W(D \cap D'))$ of $W \rightarrow y$ and hence performance on D' is a larger component of the performance of $Z \rightarrow y$ and performance on $D - D'$ is a larger component of the performance of $W \rightarrow y$.

□

However, in most domains of interest the dimensionality of the instance space will be very high. In consequence, for realistic training and test sets the proportion of the training set that appears in the test set, $\frac{|D \cap D'|}{|D|}$, will be small. Hence this effect will be negligible, as performance on the training set will be a negligible portion of total performance. What we are more interested in is off-training-set error. We contend that the force of these hypotheses will be stronger than accounted for by the difference made by the overlap between training and test sets, and hence that they do apply to off-training-set error. We note, however, that it is trivial to construct no-free-lunch proofs, such as those of Wolpert [5] and Schaffer [6], that this is not, in general, true. Rather, we contend that the hypotheses will in general be true for 'real-world' learning tasks. We justify this contention by recourse to the similarity assumption [7], that in the absence of other information, the greater the similarity between two objects in other respects, the greater the probability of their both belonging to the same class. We believe that most machine learning algorithms depend upon this assumption, and that this assumption is reasonable for real-world knowledge acquisition tasks. Test set cases covered by a more general but not a more specific rule are likely to be less similar to training cases covered by both rules than are test set cases covered by the more specific rule. Hence satisfying the left-hand-side of the more specific rule provides stronger evidence of likely class membership.

A final point that should be noted is that these hypotheses apply to individual classification rules — structures that associate an identified region of an instance space with a single class. However, as will be discussed in more detail below, we believe that the principle is nonetheless highly relevant to 'complete classifiers,' such as decision trees, that assign different regions of the instance space to different classes. This is because each individual region within a 'complete classifier' (such as a decision tree leaf) satisfies our definition of a classification rule, and hence the hypotheses can cast light on the likely consequences of relabeling subregions of the instance space within such a classifier (for example, generalizing one leaf of a decision tree at the expense of another, as proposed elsewhere [8]).

2 Evaluation

To evaluate these hypotheses we sought to generate rules of varying generality but identical empirical evidence (no other evidence source being considered in the research), and to test the hypotheses' predictions with respect to these rules.

Table 1. Algorithm for generating a random rule

1. Randomly select an example x from the training set.
2. Randomly select an attribute a for which the value of a for x (a_x) is not *unknown*.
3. If a is categorical, form the rule $IF\, a = a_x\, THEN\, c$, where c is the most frequent class in the cases covered by $a = a_x$.
4. Otherwise (if a is ordinal), form the rule $IF\, a \# a_x\, THEN\, c$, where $\#$ is a random selection between $\leq$ and $\geq$ and c is the most frequent class in the cases covered by $a \# a_x$.

We wished to provide some evaluation both of whether the predicted effects are general (with respect to rules with the relevant properties selected at random) as well as whether they apply to the type of rule generated in standard machine learning applications. We used rules generated by C4.5rules (release 8) [9], as an exemplar of a machine learning system for classification rule generation.

One difficulty with employing rules formed by C4.5rules is that the system uses a complex resolution system to determine which of several rules should be employed to classify a case covered by more than one rule. As this is taken into account during the induction process, taking a rule at random and considering it in isolation may not be representative of its application in practice. We determined that the first listed rule was least affected by this process, and hence employed it. However, this caused a difficulty in that the first listed rule usually covers few training cases and hence estimates of its likely test error can be expected to have low accuracy, reducing the likely strength of the effect predicted by Hypothesis 2.

For this reason we also employed the C4.5rules rule with the highest cover on the training set. We recognized that this would be unrepresentative of the rule's actual deployment, as in practice cases that it covered would frequently be classified by the ruleset as belonging to other classes. Nonetheless, we believed that it provided an interesting exemplar of a form of rule employed in data mining.

To explore the wider scope of the hypotheses we also generated random rules using the algorithm in Table 1.

From the *initial rule*, formed by one of these three processes, we developed a *most specific rule*. The most specific rule was created by collecting all training cases covered by the initial rule and then forming the most specific rule that covered those cases. For a categorical attribute a this rule included a clause $a \in X$, where X is the set of values for the attribute of cases in the random selection. For ordinal attributes, the rule included a clause of the form $x \leq a \leq z$, where x is the lowest value and z the highest value for the attribute in the random sample.

Next we found the set of all *most general rules*—those rules R formed by deleting clauses from the most specific rule S such that $cover(R) = cover(S)$ and there is no rule T that can be formed by deleting a clause from R such that

Table 2. Generality relationships between rules

More Specific	More General
most specific rule	combined rule
most specific rule	random most general rule
most specific rule	initial rule
combined rule	random most general rule

$cover(T) = cover(R)$. The search for the set of most general rules was performed using the OPUS complete search algorithm [10].

Then we formed the:

random most general rule: a single rule selected at random from the most general rules.

combined rule: a rule for which the condition was the conjunction of all conditions for rules in the set of most general rules.

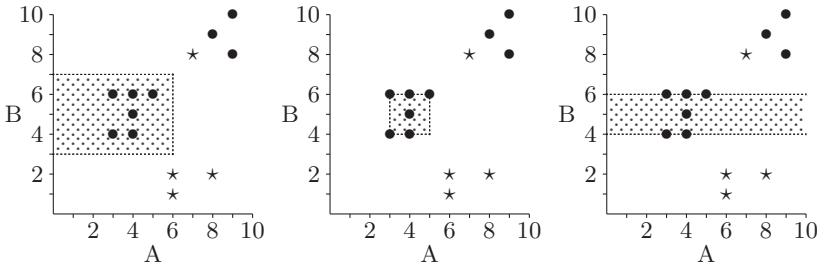
default rule: a rule with the antecedent *true*.

For all rules, the class was set to the class with the greatest number of instances covered by the initial rule. All rules other than the default rule covered exactly the same training cases. Hence all rules other than the default rule had identical empirical support.

We present an example to illustrate these concepts. We utilize a two dimensional instance space, defined by two attributes, A and B, and populated by training examples belonging to two classes denoted by the shapes $\bullet$ and $\star$. This is illustrated in Fig. 1. Fig. 1(a) presents the hypothetical initial rule, derived from some external source. Fig. 1(b) shows the most specific rule, the rule that most tightly bounds the cases covered by the initial rule. Note that while we have presented the initial rule as covering only cases of a single class, when developing the rules at differing levels of generality we do not consider class information. Fig. 1(c) and (d) shows the two most general rules that can be formed by deleting different combinations of boundaries from the most specific rule. Fig. 1(d) shows the combined rule, formed from the conjunction of all most general rules. The generality relationships between these rules are presented in Table 2.

Note that it could not be guaranteed that any pair of these rules were strictly more general or more specific than each other as it was possible for the most specific and random most general rules to be identical (in which case the set of most general rules would contain only a single rule and the initial and combined rules would also both be identical to the most specific and random most general rules. It was also possible for the initial rule to equal the most specific rule even when there were multiple most general rules. Also, it was possible for no generality relationship to hold between an initial and the combined or the random most general rule developed therefrom.

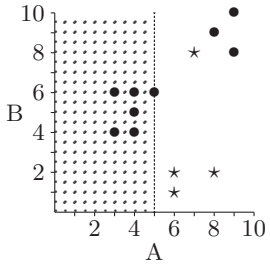
We wished to evaluate whether the predicted effects held between the rules of differing levels of generality so formed. It was not appropriate to use the normal



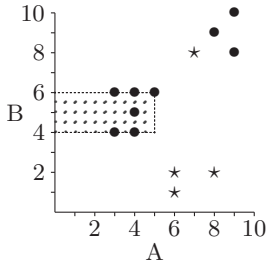
a) Initial rule: $IF\, A \leq 6 \wedge 3 \leq B \leq 7$ $THEN \bullet$

b) Most specific rule: $IF\, 3 \leq A \leq 5 \wedge 4 \leq B \leq 6$ $THEN \bullet$

c) Most General Rule 1: $IF\, 4 \leq B \leq 6$ $THEN \bullet$



d) Most General Rule 2: $IF\, A \leq 5$ $THEN \bullet$



e) Combined Rule: $IF\, A \leq 5 \wedge 4 \leq B \leq 6$ $THEN \bullet$

Fig. 1. Types of rule generated

machine learning experimental method of averaging over multiple runs for each of several data sets, as our prediction is not about relationships between average outcomes, but rather relationships between specific outcomes. Further, it would not be appropriate to perform multiple runs on each of several data sets and then compare the relative frequencies with which the predicted effects held and did not hold, as this would violate the assumption of independence between observations relied on by most statistical tools for assessing such outcomes. Rather, we applied the process once only to each of the following 50 data sets from the UCI repository [11]:

> abalone, anneal, audiology, imports-85, balance-scale, breast-cancer, breast-cancer-wisconsin, bupa, chess, cleveland, crx, dermatology, dis, echocardiogram, german, glass, heart, hepatitis, horse-colic, house-votes-84, hungarian, allhypo, ionosphere, iris, kr-vs-kp, labor-negotiations, lenses, long-beach-va, lung-cancer, lymphography, new-thyroid, optdigits, page-blocks, pendigits, pima-indians-diabetes, post-operative, promoters, primary-tumor, sat, segmentation, shuttle, sick, sonar, soybean-large, splice, switzerland, tic-tac-toe, vehicle, waveform, wine.

These were all appropriate data sets from the repository to which we had ready access and to which we were able to apply the combination of software tools employed in the research. Note that there is no averaging of results. Statistical analysis of the outcomes over the large number of data sets is used to compensate for random effects in individual results due to the use of a single run.

3 Results

Results are presented in Tables 3 to 5. Each table row represents one of the combinations of a more specific and more general rule. The right-most columns present win/draw/loss summaries of the number of times the relevant difference between values is respectively positive, equal, or negative. The first of these columns relates to Hypothesis 1. The second relates to Hypothesis 2. Each win/draw/loss record is followed by the outcome of a one-tailed sign test representing the probability of obtaining those results by chance. Where rules $\mathbf{x}$ and $\mathbf{y}$ are identical for a data set, or where one of the rules made no decisions on the unseen data, no result has been recorded. Hence not all win/draw/loss records sum to 50.

As can be seen from Table 3, with respect to the conditions formed by creating an initial rule from the C4.5rules rule with the greatest cover, all win/draw/loss comparisons but one significantly (at the 0.05 level) support the hypotheses. The one exception is marginally significant ($p = 0.055$).

Where the initial rule is the first rule from a C4.5rules rule list (Table 4), all win/draw/loss records favor the hypotheses, but some results are not significant at the 0.05 level. It is plausible to attribute this outcome to greater unpredictability in the estimates obtained from the performance of the rules on

Table 3. Results for initial rule is C4.5rules rule with most coverage

		$\lvert\alpha - x\rvert > \lvert\alpha - y\rvert$		$\lvert\beta - x\rvert < \lvert\beta - y\rvert$	
x	**y**	w:d:l	p	w:d:l	p
Most Specific	Combined	27:15: 5	< 0.001	21:15:11	0.055
Most Specific	Random MG	29:14: 4	< 0.001	23:14:10	0.017
Most Specific	Initial	33:10: 4	< 0.001	28:10: 9	0.001
Combined	Random MG	8: 9: 0	0.004	8: 9: 0	0.004

Note: x represents the accuracy of rule **x** on the test data. y represents the accuracy of rule **y** on the test data. β represents the accuracy of rules **x** and **y** on the training data (both rules cover the same training cases and hence have identical accuracy on the training data). α represents the accuracy of the default rule on the test data.

Table 4. Results for initial rule is C4.5rules first rule

		$\lvert\alpha - x\rvert > \lvert\alpha - y\rvert$		$\lvert\beta - x\rvert < \lvert\beta - y\rvert$	
x	**y**	w:d:l	p	w:d:l	p
Most Specific	Combined	16:13: 9	0.115	17:13: 8	0.054
Most Specific	Random MG	19:10: 9	0.044	20:10: 8	0.018
Most Specific	Initial	20: 9: 9	0.031	21: 9: 8	0.012
Combined	Random MG	5: 5: 1	0.109	5: 5: 1	0.109

See Table 3 for abbreviations.

the training data when the rules cover fewer training cases, and due to the lower numbers of differences in rules formed in this condition.

Where the initial rule is a random rule (Table 5), all of the results favor the hypotheses, except for one comparison between the combined and random most general rules for which a difference in prediction accuracy was only obtained on one of the fifty data sets. Where more than one difference in prediction accuracy was obtained, the results are significant at the 0.05 level with respect to Hypothesis 1, but not Hypothesis 2.

These results appear to lend substantial support to Hypothesis 1. For all but one comparison (for which only one domain resulted in a variation in performance between treatments) the win/draw/loss record favors this hypothesis. Of these eleven positive results, nine are statistically significant at the 0.05 level. There appears to be good evidence that of two rules with equal empirical and other support, the more general can be expected to obtain prediction accuracy on unseen data that is closer to the frequency with which the class is represented in the data.

The evidence with respect to Hypothesis 2 is slightly less strong, however. All conditions result in the predicted effect occurring more often than the reverse. However, only five of these results are statistically significant at the 0.05 level.

Table 5. Results for initial rule is random rule

		$\lvert\alpha - x\rvert > \lvert\alpha - y\rvert$		$\lvert\beta - x\rvert < \lvert\beta - y\rvert$	
x	**y**	w:d:l	p	w:d:l	p
Most Specific	Combined	26: 5:12	0.017	21: 5:17	0.314
Most Specific	Random MG	26: 5:12	0.017	21: 5:17	0.314
Most Specific	Initial	26: 5:12	0.017	21: 5:17	0.314
Combined	Random MG	0: 2: 1	1.000	1: 2: 0	1.000

See Table 3 for abbreviations.

The results are consistent with an effect that is weak where the accuracy of the rules on the training data differs substantially from the accuracy of the rules on unseen data. An alternative interpretation is that they are manifestations of an effect that only applies under specific constraints that are yet to be identified.

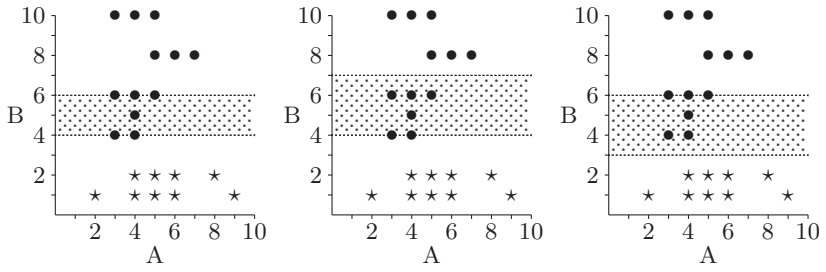
4 Discussion

We believe that our findings have important implications for knowledge acquisition. We have demonstrated that in the absence of other suitable biases to select between alternative hypotheses, biases based on generality can manipulate expected classification performance. Where a rule is able to achieve high accuracy on the training data, our results suggest that very specific versions of the rule will tend to deliver higher accuracy on unseen cases than will more general alternatives with identical empirical support. However, there is another trade-off that will also be inherent in selecting between two such alternatives. The more specific rule will make fewer predictions on unseen cases.

Clearly this trade-off between expected accuracy and cover will be difficult to manage in many applications and we do not provide general advice as to how this should be handled. However, we contend that practitioners are better off aware of this trade-off than making decisions in ignorance of their consequences.

Pazzani, Murphy, Ali, and Schulenburg [12] have argued with empirical support that where a classifier has an option of not making predictions (such as when used for identification of market trading opportunities), selection of more specific rules can be expected to create a system that makes fewer decisions of higher expected quality. Our hypotheses provide an explanation of this result. When the accuracy of the rules on the training data is high, specializing the rules can be expected to raise their accuracy on unseen data towards that obtained on the training data.

Where a classifier must always make decisions and maximization of prediction accuracy is desired, our results suggest that rules for the class that occurs most frequently should be generalized at the expense of rules for alternative classes. This is because as each rule is generalized it will trend towards the accuracy of a



a) Initial rule: $IF\, 4 \leq B \leq 6$ $THEN \bullet$

b) First generalization: $IF\, 4 \leq B \leq 7$ $THEN \bullet$

c) Second generalization: $IF\, 3 \leq B \leq 6$ $THEN \bullet$

Fig. 2. Alternative generalizations to a rule

default rule for that class, which will be highest for rules of the most frequently occurring class.

Another point that should be considered, however, is alternative sources of information that might be brought to bear upon such decisions. We have emphasized that our hypotheses relate only to contexts in which there is no other evidence available to distinguish between the expected accuracy of two rules other than their relative generality. In many cases we believe it may be possible to derive such evidence from training data. For example, we are likely to have differing expectations about the likely accuracy of the two alternative generalizations depicted in Fig. 2. This figure depicts a two dimensional instance space, defined by two attributes, A and B, and populated by training examples belonging to two classes denoted by the shapes $\bullet$ and $\star$. Three alternative rules are presented together with the region of the instance space that each covers. In this example it appears reasonable to expect better accuracy from the rule depicted in Fig. 2b than that depicted in Fig. 2c as the former generalizes toward a region of the instance space dominated by the same class as the rule whereas the latter generalizes toward a region of the instance space dominated by a different class.

While our experiments have been performed in a machine learning context, the results are applicable in wider knowledge acquisition contexts. For example, interactive knowledge acquisition environments [3, 13] present users with alternative rules all of which perform equally well on example data. Where the user is unable to bring external knowledge to bear to make an informed judgement about the relative merits of those rules, the system is able to offer no further advice. Our experiments suggest that relative generality is a factor that an interactive knowledge acquisition system might profitably utilize.

Our experiments also demonstrate that the effect that we discuss is one that applies frequently in real-world knowledge acquisition tasks. The alterna-

tive rules used in our experiments were all rules of varying levels of generality that covered exactly the same training instances. In other words, it was not possible to distinguish between these rules using traditional measures of rule quality based on performance on a training set, such as information measures. The only exception was the data sets for which the rules at differing levels of generality were all identical. In all such cases the results were excluded from the win/draw/loss record reported in Tables 3 to 5. Hence the sum of the values in each win/draw/loss record places a lower bound on the number of data sets for which there were variants of the initial rule all of which covered the same training instances. Thus, for at least 47 out of 50 data sets, there are variants of the C4.5rules rule with the greatest cover that cover exactly the same training cases. For at least 38 out of 50 data sets, there are variants of the first rule generated by C4.5rules that cover exactly the same training cases. This effect is not a hypothetical abstraction, it is a frequent occurrence of immediate practical import.

In such circumstances, when it is necessary to select between alternative rules with equal performance on the training data, one approach has been to select the least complex rule [14]. However, some recent authors have argued that complexity is not an effective rule quality metric [8, 15]. We argue here that generality provides an alternative criterion on which to select between such rules, one that allows for reasoning about the trade-offs inherent in the choice of one rule over the other, rather than providing a blanket prescription.

5 On the difficulty of measuring degree of generalization

It might be tempting to believe that our hypotheses could be extended by introducing a measure of magnitude of generalization together with predictions about the magnitude of the effects on prediction accuracy that may be expected from generalizations of different magnitude.

However, we believe that it is not feasible to develop meaningful measures of magnitude of generalization suitable for such a purpose. Consider, for example, the possibility of generalizing a rule with conditions $age < 40$ and $income < 50000$ by deleting either condition. Which is the greater generalization? It might be thought that the greater generalization is the one that covers the greater number of cases. However, if one rule covers more cases than another then there will be differing evidence in support of each. Our hypotheses do not relate to this situation. We are interested only in how to select between alternative rules when the only source of evidence about their relative prediction performance is their relative generality.

If it is not possible to develop measures of magnitude of generalization then it appears to follow that it will never be possible to extend our hypotheses to provide more specific predictions about the magnitude of the effects that may be expected from a given generalization or specialization to a rule.

6 Conclusion

We have presented two hypotheses relating to expectations regarding the accuracy of two alternative classification rules with identical supporting evidence other than their relative generality. The first hypothesis is that the accuracy on unseen data of the more general rule will be more likely to be closer to the accuracy on unseen data of a default rule for the class than will the accuracy on unseen data of the more specific rule. The second hypothesis is that the accuracy on previously unseen data of the more specific rule will be more likely to be closer to the accuracy of the rules on the training data than will the accuracy of the more general rule on unseen data.

We have provided experimental support for those hypotheses, both with respect to classification rules formed by C4.5rules and random classification rules. However, the results with respect to the second hypothesis were not statistically significant in the case of random rules. These results are consistent with the two hypotheses, albeit with the effect of the second being weak when there is low accuracy for the error estimate for a rule derived from performance on the training data. They are also consistent with the second hypothesis only applying to a limited class of rule types. Further research into this issue is warranted.

These results may provide a first step towards the development of useful learning biases based on rule generality that do not rely upon prior domain knowledge, and may be sensitive to alternative knowledge acquisition objectives, such as trading-off accuracy for cover. Our experiments demonstrated the frequent existence of rule variants between which traditional rule quality metrics, such as an information measures, could not distinguish. This shows that the effect that we discuss is not an abstract curiosity but rather is an issue of immediate practical concern.

Acknowledgements

We are grateful to the UCI repository donors and librarians for providing the data sets used in this research. The breast-cancer, lymphography and primary-tumor data sets were donated by M. Zwitter and M. Soklic of the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

References

1. Mitchell, T.M.: Version spaces: A candidate elimination approach to rule learning. In: Proceedings of the Fifth International Joint Conference on Artificial Intelligence. (1977) 305–310
2. Mitchell, T.M.: The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, Department of Computer Science, New Brunswick, NJ (1980)

3. Webb, G.I.: Integrating machine learning with knowledge acquisition through direct interaction with domain experts. Knowledge-Based Systems **9** (1996) 253–266
4. Webb, G.I., Wells, J., Zheng, Z.: An experimental evaluation of integrating machine learning with knowledge acquisition. Machine Learning **35** (1999) 5–24
5. Wolpert, D.H.: On the connection between in-sample testing and generalization error. Complex Systems **6** (1992) 47–94
6. Schaffer, C.: A conservation law for generalization performance. In: Proceedings of the 1994 International Conference on Machine Learning, Morgan Kaufmann (1994)
7. Rendell, L., Seshu, R.: Learning hard concepts through constructive induction: Framework and rationale. Computational Intelligence **6** (1990) 247–270
8. Webb, G.I.: Further experimental evidence against the utility of Occam's razor. Journal of Artificial Intelligence Research **4** (1996) 397–417
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993)
10. Webb, G.I.: OPUS: An efficient admissible algorithm for unordered search. Journal of Artificial Intelligence Research **3** (1995) 431–465
11. Blake, C., Merz, C.J.: UCI repository of machine learning databases. [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine, CA. (2004)
12. Pazzani, M.J., Murphy, P., Ali, K., Schulenburg, D.: Trading off coverage for accuracy in forecasts: Applications to clinical data analysis. In: Proceedings of the AAAI Symposium on Artificial Intelligence in Medicine. (1994) 106–110
13. Compton, P., Edwards, G., Srinivasan, A., Malor, R., Preston, P., Kang, B., Lazarus, L.: Ripple down rules: Turning knowledge acquisition into knowledge maintenance. Artificial Intelligence in Medicine **4** (1992) 47–59
14. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Occam's Razor. Information Processing Letters **24** (1987) 377–380
15. Domingos, P.: The role of Occam's razor in knowledge discovery. Data Mining and Knowledge Discovery **3** (1999) 409–425