a simple method for subset selection which assists learning schemes in the execution of their task by relieving them of the burden of determining relevancy without impairing their ability to learn useful concepts.

References

[1] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–90, 1993.

[2] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129, New Jersey, July 1994.

[3] D. Michie. Methodologies from machine learning in data analysis and software. *The Computer Journal*, 34(6):559–565, 1991.

dataset used	fraction of features in subset	accuracy using subset	accuracy using all
CH	12/36	82.5%	99.1%
LY	15/18	72.8%	73.5%
GL	7/9	70.7%	65.6%
G2	7/9	73.7%	73.7%

Table 2: C4.5 accuracy using selected subset and using all features.

```
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- H>1.8, D>1.

'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- L>46, M<=1.
'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- L>42, J>49.
'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- K>3.8, H<=1.5.
'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- I>155, I<=230, J>30.
```

When the variables from the rule sets are matched with their feature labels it can be seen that the two sets are identical. In this instance our subset selection method has allowed FOIL to produce the same rule sets using only 3/5 of the features.

The results outlined above are an encouraging indication that the premise of our subset selection technique is a good one. Holte [1] has pointed out, however, that it appears difficult to avoid producing a particular result from the commonly used datasets, thus these results may not be as promising as such a cursory analysis indicates.

To improve our evaluation, we ran a series of 25 cross-validation experiments on four other datasets, CH (chess-end-games), GL (glass) and G2 (GL modified) from University of California at Irvine, and LY (lymphography) from the University Medical Center, Institute of Oncology in Ljubljana. The results were compared against running C4.5 on all attributes. Table 2 summarises the results.

The figures shown in Table 2 show that our subset selection method improves the results from C4.5 for GL, does not significantly affect C4.5's performance for LY and G2, and impairs its performance for CH. These conflicting outcomes need not disparage the technique, for we have seen enough positive results—where the technique aids the learning scheme in finding a good classification. The shortcomings might be overcome with some minor adjustment, such as including the results from a medial range query on the dataset into the relevancy measure, or weighting the relevancy in proportion to the number of records returned on a query.

Conclusions

Feature subset selection is an issue of growing concern as the products of machine learning research are directed on to the databases of the real world. We have described

attributes used	accuracy before pruning	accuracy after pruning
top 1	85.0%	83.8%
top 2	92.5%	92.4%
top 3	91.2%	90.0%
top 13	97.5%	93.8%
all	97.5%	93.8%

Table 1: Results from hepatitis dataset using C4.5.

```
 10 predicts         42 field values. ()
  1 predicts          0 field values. ()
  3 predicts          0 field values. ()
  8 predicts          0 field values. ()
  9 predicts          0 field values. ()
 11 predicts          0 field values. ()
 13 predicts          0 field values. ()
 14 predicts          0 field values. ()
```

The decision tree produced by running C4.5 on only the 13 attributes shown to be relevant (i.e. non-zero prediction) was identical to the decision tree produced by running C4.5 on all features. The accuracy was 97.5% before pruning and 93.8% after pruning. This seems to suggest that the selection method is successful as it can reduce by 2/3 the number of features needed for C4.5 to produce the same decision tree. Table 1 summarises other the results obtained from all tests on HE.

When FOIL was run on all features of MU it produced the following rule sets:

```
 'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- R<=46, S>1, N>1.
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- O>230.
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- R<=31, R>23, A>32.
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- N>1.8, E>1.

'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- R>46, S<=1.
'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- R>42, P>49.
'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- Q>3.8, N<=1.5.
'LIVE'(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S) :- O>155, O<=230, P>30.
```

When FOIL was run on all and only those selected by our relevancy criteria it produced the following rule sets:

```
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- L<=46, M>1, H>1.
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- I>230.
'DIE'(A,B,C,D,E,F,G,H,I,J,K,L,M) :- L<=31, L>23, A>32.
```

This table shows that the fourth and fifth features are never high when the third feature is low. We now know that the value of the third field always allows more accurate prediction of the fourth and fifth features than through pure chance, thus the third field is relevant. By collating the results of these queries on all fields we can assert relevancy for every feature that finds no representative values in a subrange of another feature.

Subset selection

Once relevant features have been identified, they can be ordered according to their ability to predict the value of other features. The number of feature values predicted by a particular feature offers a reasonable measure of its utility. This is obtained by summing the product of the number of records found to be correlated to the feature in question and the number of features for which such a correlation exists—that is, those features with a zero in their corresponding row. From the tables above, for example, the third feature is correlated to two fields in 75 records (excluding correlation to itself) and two fields in 79 records—its "degree of relevancy" is thus $2\times75+2\times79 = 308$. Clearly the number of records found that satisfy the initial query has a large impact on the relevancy metric, but it is not clear if this is undesirable.

Results

The subset selection method outlined above has been applied to two standard datasets, HE (hepatitis) and MU (mushroom) from the collection distributed by the University of California at Irvine. In this section, we outline the results obtained from applying two machine learning schemes, C4.5 and FOIL, to the suggested subsets and compare them with those obtained from applying the same schemes to the complete datasets.

When run on HE, our subset selection technique produced the following ranking:

```
15 predicts        96 field values. ()
16 predicts        89 field values. ()
 2 predicts        83 field values. ()
17 predicts        82 field values. ()
19 predicts        82 field values. ()
 5 predicts        59 field values. ()
12 predicts        55 field values. ()
 6 predicts        52 field values. ()
 7 predicts        49 field values. ()
20 predicts        47 field values. ()
18 predicts        44 field values. ()
 4 predicts        42 field values. ()
```

all other feature values lie when the first feature's value lies within a particular subset of its own quartiles. For example, we could ask "within what quartiles do the values of all other features lie whenever feature a is high (i.e. in its last two quartiles)?" The dataset is scanned and whenever feature a's value is high we record the quartile that all other feature values reside in. The following results are obtained when the Iris dataset is examined for the conditions under which the third feature is high.

```
=============================================================
  File:IR.raw  Field: 3  Quality: hig  Found: 75/150
-------------------------------------------------------------
   @field     @vlo     @low     @mid     @hi      @vhi
       1        1       10       39       68       41
       2       28       50       53       42       10
       3        0        0       41       75       42
       4        0        0       41       73       46
       5        0        0       75       75       50
=============================================================
```

The table shows that when the third feature's value lay in its third and fourth quartile, 1 record had a very low (first quartile) value for its first feature, 10 records had a low (first two quartiles) valued first feature, 39 had a medial (second and third quartile) valued first feature, 68 had a high (third and fourth quartile) first feature and 41 had a very high (fourth quartile) first feature (note that overlap in subrange boundaries allows the total of a single row to exceed the number of records found). While this table shows a tendency for the first feature to be high when its third feature is also high, the existence of values in all of the first feature's quartiles roughly indicates that the first feature's value may be anything. We cannot conclude any relevance for the third feature based on this result. However, the table shows that the fourth and fifth features are never low when the third feature is high. This constraint on the range of values for the fourth and fifth feature when the third feature's value is known to be low indicates a possible relevancy for the third feature. What is the situation when the third field is low?

```
=============================================================
  File:IR.raw  Field: 2  Quality: low  Found: 79/150
-------------------------------------------------------------
   @field     @vlo     @low     @mid     @hi      @vhi
       0       40       71       46       12        3
       1       21       36       30       53       32
       2       44       79       42        0        0
       3       41       75       45        0        0
       4       50       79       79        0        0
=============================================================
```

neither region is as broad as the full range of values observed for that feature.

The need for a constraint on the range of values merits some discussion. Consider a training set where the value of attribute b is 3 whenever the value of attribute a is 7. We may in this instance say that a predicts the value of b. Using the terminology of information theory, the entropy of the data is lower than a training set without this correlation (all other things being equal) because the observation that a is 7 allows us to infer that the value of b is 3. The entropy is also lower if it is the case that the value of b is always, say, in the upper half of its observed range whenever a is 7, though the entropy would be not be as low as in the first example because the precision of our estimate for b is lower. It doesn't matter what subrange for b we use, though it is important we realise as it approaches the full range of observed values for b the decrease in entropy approaches zero (with respect to these features)—that is, our ability to predict b given a approaches pure chance.

Attribute a is deemed relevant if our ability to predict b given a is better than pure chance, but *only* if it is also better than our ability to predict b not given a. For example, if it is the case that whenever attribute a is, say, either 7 or 8 then attribute b is always 3, then the number of instances for which we can precisely predict b is (all other things unchanged) greater than when we could make this prediction only when a was 7. In fact, as the subrange for a approaches its full range of observed values, with the value of b still fixed, then our ability to predict b approaches perfect accuracy. However, if b is fixed throughout the full range of values for a then a has also proven itself irrelevant.

We conclude that a feature may be considered relevant if a particular subrange of its values can be used to predict the value of another feature more accurately than by pure chance, provided the subrange is narrower than the full range of observed values. Precisely what subrange is used is for the most part unimportant, but we can use general notions of what we expect a correlation to be as a guide. For example, if one feature is always low whenever another feature is high then the second feature allows us some ability to predict the first and may thus be considered relevant. Finding such a correlation requires that we simply decide on what constitutes a low or high value.

Rough dependency

We regard a value to be low if it falls within the first two quartiles of its observed range, and high if it falls within the second two quartiles. We also regard any values within the second and third quartiles to be medial. To aid in more precise prediction we also define values within the first quartile as very low and those in the fourth quartile as very high. The first step to determining relevancy is the construction of histograms from which quartile boundaries may be ascertained.

To determine the relevancy of a single feature we tabulate the quartiles in which

cesses are applied recursively and in parallel throughout the learning process. As noted, this can lead to a factorial explosion in the number of hypotheses that must be tested.

Relevancy

Gennari et al. (1989) state that "Features are relevant if their values vary systematically with category membership." John et al (1994) formalise this definition as

Definition 1 *X_i is relevant iff there exists some x_i and y for which $p(X_i = x_i) > 0$ such that $p(Y = y|X_i = x_i) \neq p(Y = y)$,*

where x_i is an instance of feature X_i, and y is one of the set Y of possible category labels.

John et al (1994) demonstrate that, among other things, this definition is not an adequate measure of relevance in situations where features demonstrate an XOR relationship. A proposed refinement of the definition is given as

Definition 2 *X_i is relevant iff there exists some x_i and y for which $p(X_i = x_i, S_i = s_i) > 0$ such that $p(Y = y|X_i = x_i, S_i = s_i) \neq p(Y = y|S_i = s_i)$,*

where S_i is a set of features and s_i is the corresponding set of features values for a training instance. John et al argue that situations may exist where even this definition may give unexpected results, maintaining that a further distinction between *weak* and *strong* relevance is required, but this is not necessary for our purposes.

The problem these definitions give for learning schemes is that possible interdependency between a set of features entails that a completely correct assessment of relevancy cannot be carried out without exhaustive permutation in subset selection—a situation that exacerbates as the number of features increases. Heuristic subset selection is therefore necessary if exhaustive hypothesis testing is to be avoided. Fortunately practical machine learning on real datasets has shown that mutual interdependency between more than two features is not common, and that where it does occur the features involved usually demonstrate some pairwise interdependency.

Definition 2 claims that a feature is relevant if its value contributes to the prediction of the value for another (n.b. as training sets usually include the class label as one the features, explicit reference in the definition to prediction of the label is not necessary). For subset selection one need only detect that a correlation exists, leaving precise formulation of the relationship to the learning scheme itself. This means that a crude and simple method of detecting dependencies is just as valid as any more refined method if it gets the job done.

We use the following broad concept of dependency to define relevance:

> When the value of a feature falls within a particular region of its observed range, if the value of another feature always falls within a particular region of its observed range, then the first can be regarded as relevant, *provided*

Introduction

A large percentage of computer systems are presently dedicated to the collection of an incomprehensibly vast amount of information. The daily world-wide accrual of data has long since exceeded the ability for human analysis, and machine processing has been largely limited to collation, summation, sorting and a variety of basic statistical analyses.

Artificial intelligence research in the form of expert systems has attempted to relieve much of the burden of analysis from the shoulders of human beings by developing automated reasoning systems that can respond quickly to the flood of incoming data. Computer users have countered with the collection of an even broader range of information for which expert systems have either not yet been built or whose domains have proven as yet too difficult to formalise.

In the past decade another area of computer science has emerged aimed at the automatic construction of rule-based decision systems from sample training sets. Though still in its infancy, machine learning has shown a great deal of promise as a means of garnering expert knowledge from raw material [3]. Even so, it is becoming increasingly clear to researchers in the field that these early successes have largely been obtained from small datasets from which correct inferences are pretty much difficult to avoid. In fact, Holte [1] found that a classification scheme based on a single feature would frequently perform comparably to many of the more elaborate "traditional" schemes.

The problem is that real datasets have thousands, even millions, of records, each consisting of perhaps hundreds of fields, thus even heuristically guided hill-climbing algorithms can face an astronomically large number of hypotheses that must be evaluated and compared in their search for a satisfactory set of classification rules. In practice, machine learning researchers have had to pre-select a subset of the data over which the learning is to proceed in order to obtain results within reasonable time constraints.

Pre-selection allows for possible oversight of key data fields during the learning process. However, this does not imply that a method for avoiding pre-selection must be sought. Rather it may indicate that large-scale learning requires at least two stages: one that identifies the information that is to be considered salient, and another that constructs its classification rules from that information.

A majority of learning schemes use some sort of filtering approach for identifying relevant attributes and suitable boundary values for constructing classification rules. For example, sequential backward elimination observes the effects of removing individual features during the classification process, while forward methods add features to an initially empty set. A summary of these and various hybrid approaches, including some using search techniques and genetic algorithms, can be found in John, Kohavi & Pfleger [2].

What makes existing methods intractable is that the filtering and learning pro-

Subset Selection Using Rough Numeric Dependency

Tony C. Smith

Department of Computer Science, University of Waikato, Hamilton, New Zealand
Email: tcs@waikato.ac.NZ; phone: +64 (7) 838–4453; fax: +64 (7) 838–4155

Geoff Holmes

Department of Computer Science, University of Waikato, Hamilton, New Zealand
Email: geoff@waikato.ac.NZ; phone: +64 (7) 838–4021; fax: +64 (7) 838–4155