From *Proceedings of the AAAI Fall Symposium on Relevance* (1994). New Orleans, LA: AAAI Press.

Selection of Relevant Features in Machine Learning

Pat Langley (Langley@flamingo.stanford.edu)
Institute for the Study of Learning and Expertise
2451 High Street, Palo Alto, CA 94301

Abstract

In this paper, we review the problem of selecting relevant features for use in machine learning. We describe this problem in terms of heuristic search through a space of feature sets, and we identify four dimensions along which approaches to the problem can vary. We consider recent work on feature selection in terms of this framework, then close with some challenges for future work in the area.

1. The Problem of Irrelevant Features

The selection of relevant features, and the elimination of irrelevant ones, is a central problem in machine learning. Before an induction algorithm can move beyond the training data to make predictions about novel test cases, it must decide which attributes to use in these predictions and which to ignore. Intuitively, one would like the learner to use only those attributes that are 'relevant' to the target concept.

There have been a few attempts to define 'relevance' in the context of machine learning, as John, Kohavi, and Pfleger (1994) have noted in their review of this topic. Because we will review a variety of approaches, we do not take a position on this issue here. We will focus instead on the task of selecting relevant features (however defined) for use in learning and prediction.

Many induction methods attempt to deal directly with the problem of attribute selection, especially ones that operate on logical representations. For instance, techniques for inducing logical conjunctions do little more than add or remove features from the concept description. Addition and deletion of single attributes also constitute the basic operations of more sophisticated methods for inducing decision lists and decision trees. Some nonlogical induction methods, like those for neural networks and Bayesian classifiers, instead use weights to assign *degrees* of relevance to attributes. And some learning schemes, such as the simple nearest neighbor method, ignore the issue of relevance entirely.

We would like induction algorithms that scale well to domains with many irrelevant features. More specifically, we would like the sample complexity (the number of training cases needed to reach a given level of accuracy) to grow slowly with the number of irrelevant attributes. Theoretical results for algorithms that search restricted hypothesis spaces are encouraging. For instance, the worst-case number of errors made by Littlestone's (1987) Winnow method grows only logarithmically with the number of irrelevant features. Pazzani and Sarrett's (1992) average-case analysis for Wholist, a simple conjunctive algorithm, and Langley and Iba's (1993) treatment of the naive Bayesian classifier, suggest that their sample complexities grow at most linearly with the number of irrelevant features.

However, the theoretical results are less optimistic for induction methods that search a larger space of concept descriptions. For example, Langley and Iba's (1993) average-case analysis of simple nearest neighbor indicates that its sample complexity grows exponentially with the number of irrelevant attributes, even for conjunctive target concepts. Experimental studies of nearest neighbor are consistent with this conclusion, and other experiments suggest that similar results hold even for induction algorithms that explicitly select features. For example, the sample complexity for decision-tree methods appears to grow linearly with the number of irrelevants for conjunctive concepts, but exponentially for parity concepts, since the evaluation metric cannot distinguish relevant from irrelevant features in the latter situation (Langley & Sage, in press).

Results of this sort have encouraged machine learning researchers to explore more sophisticated methods for selecting relevant features. In the sections that follow, we present a general framework for this task, and then consider some recent examples of work on this important problem.

2. Feature Selection as Heuristic Search

One can view the task of feature selection as a search problem, with each state in the search space specifying a subset of the possible features. As Figure 1 depicts, one can impose a partial ordering on this space, with each child having exactly one more feature than its parents. The structure of this space suggests that any feature selection method must take a stance on four basic issues that determine the nature of the heuristic search process.

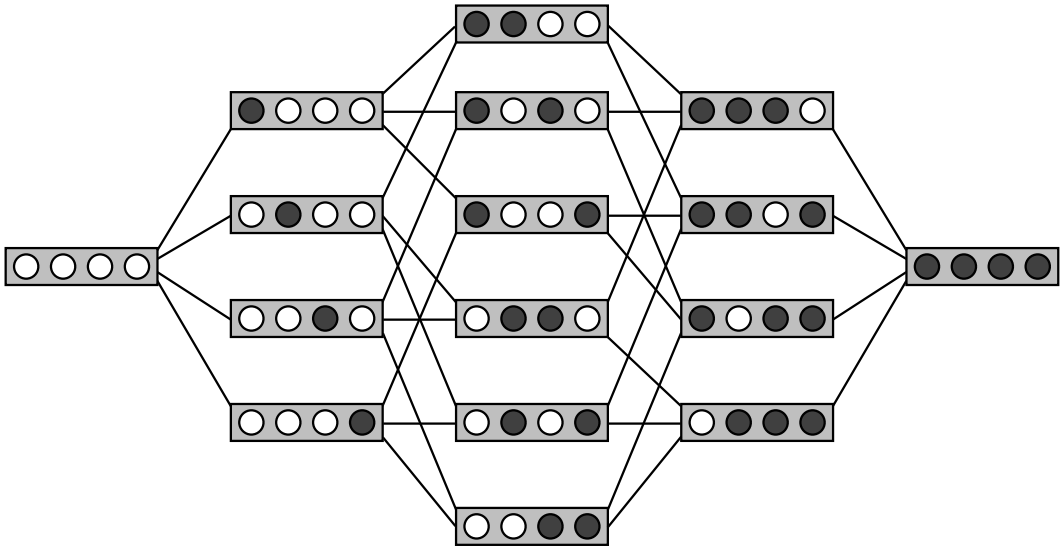

Figure 1. Each state in the space of feature subsets specifies the attributes to use during induction. Note that the states in the space (in this case involving four features) are partially ordered, with each of a state's children (to the right) including one more attribute (dark circles) than its parents.

First, one must determine the starting point in the space, which in turn determines the direction of search. For instance, one might start with no features and successively add attributes, or one might start with all attributes and successively remove them. The former approach is sometimes called *forward selection*, whereas the latter is known as *backward elimination*. One might also select an initial state somewhere in the middle and move outward from this point.

A second decision involves the organization of the search. Clearly, an exhaustive search of the space is impractical, as there exist 2^a possible subsets of a attributes. A more realistic approach relies on a greedy method to traverse the space. At each point in the search, one considers local changes to the current set of attributes, selects one, and then iterates, never reconsidering the choice. A related approach, known as *stepwise* selection or elimination, considers both adding and removing features at each decision point, which lets one retract an earlier decision without keeping explicit track of the search path. Within these options, one can consider all states generated by the operators and then select the best, or one can simply choose the first state that improves accuracy over the current set. One can also replace the greedy scheme with more sophisticated methods, such as best-first search, which are more expensive but still tractable in some domains.

A third issue concerns the strategy used to evaluate alternative subsets of attributes. One broad class of strategies considers attributes independently of the induction algorithm that will use them, relying on general characteristics of the training set to select some features and exclude others. John, Kohavi, and Pfleger (1994) call these *filter* methods, because they filter out irrelevant attributes before the induction process occurs. They contrast this approach with *wrapper* methods, which generate a set of candidate features, run the induction algorithm on the training data, and use the accuracy of the resulting description to evaluate the feature set. Within this approach, one must still pick some estimate for accuracy, but this choice seems less central than settling on a filter or wrapper scheme.

Finally, one must decide on some criterion for halting search through the space of feature subsets. Within the wrapper framework, one might stop adding or removing attributes when none of the alternatives improves the estimate of classification accuracy, one might continue to revise the feature set as long as accuracy does not degrade, or one might continue generating candidate sets until reaching the other end of the search space and then select the best. Within the filter framework, one criterion for halting notes when each combination of values for the selected attributes maps onto a single class value. Another alternative simply orders the features according to some relevancy score, then uses a system parameter to determine the break point.

Note that the above methods for feature selection can be combined with *any* induction algorithm to increase its learning rate in domains with irrelevant attributes. The effect on behavior may differ for different induction techniques and for different target concepts, in some cases producing little benefit and in others giving major improvement. But the basic idea of searching the space of feature sets is conceptually and practically distinct from the specific induction method that benefits from the feature-selection process.

3. Recent Work on Feature Selection

The problem of feature selection has long been an active research topic within statistics and pattern recognition (e.g., Devijver & Kittler, 1982), but most work in this area has dealt with linear regression. In the past few years, feature selection has received considerable attention from machine learning researchers interested in improving the performance of their algorithms.

The earliest approaches to feature selection within machine learning emphasized filtering methods. For example, Almuallim and Dietterich's (1991) Focus algorithm starts with an empty feature set and carries out breadth-first search until it finds a minimal combination of features that predicts pure classes. The system then passes the reduced feature set to ID3, which constructs a decision tree to summarize the training data. Schlimmer (1993) described a related approach that carries out a systematic search (to avoid revisiting states) through the space of feature sets, again starting with the empty set and adding features until it finds a combination consistent with the training data.

Kira and Rendell (1992) used a quite different scheme for filtering attributes. Their Relief algorithm assigns a weight to each feature that reflects its ability to distinguish among the classes, then selects those features with weights that exceed a user-specified threshold. The system then uses ID3 to induce a decision tree from the training data using only the selected features. Relief does not quite fit into our framework, as it imposes a linear ordering on the features rather than searching the partially ordered space of feature sets. Kononenko (1994) reports two extensions to the method that handle non-Boolean attributes, and Doak (1992) has explored similar approaches to the problem.

Although Focus and Relief follow feature selection with decision-tree construction, one can also combine the former with other induction methods. For instance, Cardie (1993) used a filtering approach to identify a subset of features for use in nearest neighbor retrieval, whereas Kubat, Flotzinger, and Pfurtscheller (1993) filtered features for use with a naive Bayesian classifier. Both used C4.5 to construct a decision tree from the data, but only to determine the features to be passed to their primary induction methods.

Most recent research on feature selection differs from these early methods by relying on wrapper strategies rather than filtering schemes. The general argument for wrapper approaches is that the induction method that will use the feature subset should provide a better estimate of accuracy than a separate measure that may have an entirely different inductive bias. John, Kohavi, and Pfleger (1994) were the first to present the wrapper idea as a general framework for feature selection. Their own work has emphasized its combination with decision-tree methods, but they also encourage its use with other induction algorithms.

The generic wrapper technique must still use some measure to select among alternative features. One natural scheme involves running the induction algorithm over the entire training data using a given set of features, then measuring the accuracy of the learned structure on the training data. However, John et al. argue convincingly that a cross-validation method, which they use in their implementation, provides a better measure of expected accuracy on novel test cases.

John et al. also review existing definitions of relevance in the context of machine learning and propose a new definition that overcomes some problems with earlier ones. In addition, they describe feature selection in terms of heuristic search and review a variety of methods that, although designed for filter schemes, also work within the wrapper approach. Finally, they carry out systematic experiments on a variety of search methods within the wrapper model, varying the starting point and the available operators.

The major disadvantage of wrapper methods over filter methods is the former's computational cost, which results from calling the induction algorithm for each feature set considered. This cost has led some researchers to invent ingenious techniques for speeding the evaluation process. In particular, Caruana and Freitag (1994) devised a scheme for caching decision trees that substantially reduces the number of trees considered during feature selection, which in turn lets their algorithm search larger spaces in reasonable time. Moore and Lee (1994) describe an alternative scheme that instead speeds feature selection by reducing the percentage of training cases used during evaluation.

Like John et al., Caruana and Freitag review a number of greedy methods that search the space of feature sets and report on comparative experiments that vary the starting set and the operators. However, their concern with efficiency also led them to examine the tradeoff between accuracy and computational cost. Moreover, their motivation for exploring feature-selection methods was more strict than dealing with irrelevant attributes. Their aim was to find sets of attributes that are *useful* for induction and prediction.

Certainly not all work within the wrapper framework has focused on decision-tree induction. Langley and Sage's (1994a) Oblivion algorithm combines the wrapper idea with the simple nearest neighbor method. Their system starts with all features and iteratively removes the one that leads to the greatest improvement in accuracy, continuing until the estimated accuracy actually declines. Aha and Bankert (1994) take a similar approach to augmenting nearest neighbor, but their system starts with a randomly selected subset of features and includes an option for beam search rather than greedy decisions. Skalak's (1994) work on nearest neighbor also starts with a random feature set, but replaces greedy search with random hill climbing that continues for a specified number of cycles.

Table 1. Characterization of recent work on feature selection in terms of heuristic search through the space of feature sets.

Authors (System)	Starting Point	Search Control	Evaluation Scheme	Halting Criterion
Aha and Bankert (Beam)	Random	Comparison	Comparison	No Better
Almuallim/Dietterich (Focus)	None	Breadth First	Filter	Consistency
Cardie	None	Greedy	Filter	Consistency
Caruana and Freitag (CAP)	Comparison	Greedy	Wrapper	All Used
Doak	Random	Ordering	Filter	Threshold
John, Kohavi, and Pfleger	Comparison	Greedy	Comparison	No Better
Kira and Rendell (Relief)	—	Ordering	Filter	Threshold
Kubat et al.	None	Greedy	Filter	Consistency
Langley/Sage (Oblivion)	All	Greedy	Wrapper	Worse
Langley/Sage (Selective Bayes)	None	Greedy	Wrapper	Worse
Moore and Lee (Race)	Comparison	Greedy	Wrapper	No Better
Schlimmer	None	Systematic	—	Consistency
Skalak	Random	Mutation	Wrapper	Enough Times
Townsend-Weber and Kibler	All	Comparison	Wrapper	No Better

Most research on wrapper methods has focused on classification, but both Moore and Lee (1994) and Townsend-Weber and Kibler (1994) have combined this idea with k nearest neighbor for numeric prediction. Also, most work has emphasized the advantages of feature selection for induction methods that are sensitive to irrelevant features, but Langley and Sage (1994b) have shown that the naive Bayesian classifier, which is sensitive to *redundant* attributes, can benefit from the same basic approach. This suggests that techniques for feature selection can improve the behavior of induction algorithms in a variety of situations, not only in the presence of irrelevant attributes.

4. Challenges for Future Research

Despite the recent activity, and the associated progress, in methods for selecting relevant features, there remain many directions in which machine learning can improve its study of this important problem. One of the most urgent involves the introduction of more challenging data sets. Almost none of the domains studied to date have involved more than 40 features. One exception is Aha and Bankert's study of cloud classification, which used 204 attributes, but typical experiments have dealt with many fewer features.

Moreover, Langley and Sage's results with the nearest neighbor method suggest that many of the UCI data sets have few if any irrelevant attributes. In hindsight, this seems natural for diagnostic domains, in which experts tend to ask about relevant features and ignore other ones. However, we believe that many real-world domains do not have this character, and that we must find data sets with a substantial fraction of irrelevant attributes if we want to test our ideas on feature selection adequately.

Experiments with artificial data also have important roles to play in the study of feature-selection methods. Such data sets can let one systematically vary factors of interest, such as the number of relevant and irrelevant attributes, while holding other factors constant. In this way, one can directly measure the sample complexity of algorithms as a function of these factors, showing their ability to scale to domains with many irrelevant features. However, we distinguish between the use of artificial data for such systematic experiments and reliance on isolated artificial data sets (such as the Monks problems), which seem much less useful.

More challenging domains, with more features and a higher proportion of irrelevant ones, will require more sophisticated methods for feature selection. Although further increases in efficiency would increase the number of states examined, such constant-factor improvements cannot eliminate problems caused by exponential growth in the number of feature sets. However, viewing these problems in terms of heuristic search suggests some places to look for solutions. In general, we must:

- invent more intelligent techniques for selecting an initial set of features from which to start the search;
- formulate search-control methods that take advantage of structure in the space of feature sets;
- devise improved frameworks (better even than the wrapper method) for evaluating the usefulness of alternative feature sets;
- design better halting criteria that will improve efficiency without sacrificing useful feature sets.

Naturally, the details of these extensions remain to be discovered, but each holds significant potential for increasing the ability of selection methods to handle realistic domains with many irrelevant features.

Future research in the area should also compare feature selection to attribute-weighting schemes. In the limit, attribute weighting should outperform selection in domains that involve different degrees of relevance, but the introduction of weights also increases the number of hypotheses considered during induction, which can slow learning. Thus, each approach has some advantages, leaving an open question that is best answered by experiment, but preferably by *informed* experiments designed to test specific hypotheses about these two approaches to relevance. Resolving such basic issues promises to keep the field of machine learning occupied for many years to come.

Acknowledgements

This research was supported in part by Grant Number N00014-94-1-0505 from the Office of Naval Research. Many of the researchers active in the area of feature selection contributed, directly or indirectly, to the ideas presented in this paper.

References

Aha, D. W., & Bankert, R. L. (1994). Feature selection for case-based classification of cloud types. *Working Notes of the AAAI94 Workshop on Case-Based Reasoning* (pp. 106–112). Seattle, WA: AAAI Press.

Almuallim, H., & Dietterich, T. G. (1991). Learning with many irrelevant features. *Proceedings of the Ninth National Conference on Artificial Intelligence* (pp. 547–552). San Jose, CA: AAAI Press.

Cardie, C. (1993). Using decision trees to improve case-based learning. *Proceedings of the Tenth International Conference on Machine Learning* (pp. 25–32). Amherst, MA: Morgan Kaufmann.

Caruana, R. A., & Freitag, D. (1994). Greedy attribute selection. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 28–36). New Brunswick, NJ: Morgan Kaufmann.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach.* New York: Prentice-Hall.

Doak, J. (1992). *An evaluation of feature-selection methods and their application to computer security* (Technical Report CSE-92-18). Davis: University of California, Department of Computer Science.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 121–129). New Brunswick, NJ: Morgan Kaufmann.

Kira, K., & Rendell, L. (1992). A practical approach to feature selection. *Proceedings of the Ninth International Conference on Machine Learning* (pp. 249–256). Aberdeen, Scotland: Morgan Kaufmann.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of Relief. *Proceedings of the 1994 European Conference on Machine Learning.*

Kubat, M., Flotzinger, D., & Pfurtscheller, G. (1993). Discovering patterns in EEG signals: Comparative study of a few methods. *Proceedings of the 1993 European Conference on Machine Learning* (pp. 367–371). Vienna: Springer-Verlag.

Langley, P., & Iba, W. (1993. Average-case analysis of a nearest neighbor algorithm. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 889–894). Chambery, France.

Langley, P., & Sage, S. (1994a). Oblivious decision trees and abstract cases. *Working Notes of the AAAI-94 Workshop on Case-Based Reasoning* (pp. 113–117). Seattle, WA: AAAI Press.

Langley, P., & Sage, S. (1994b). Induction of selective Bayesian classifiers. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 399–406). Seattle, WA: Morgan Kaufmann.

Langley, P., & Sage, S. (in press). Scaling to domains with many irrelevant features. In R. Greiner (Ed.), *Computational learning theory and natural learning systems* (Vol. 4). Cambridge, MA: MIT Press.

Littlestone, N. (1987). Learning quickly when irrelevant attributes abound: A new linear threshold algorithm. *Machine Learning, 2*, 285–318.

Moore, A. W., & Lee, M. S. (1994). Efficient algorithms for minimizing cross validation error. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 190–198). New Brunswick, NJ: Morgan Kaufmann.

Pazzani, M. J., & Sarrett, W. (1992). A framework for the average case analysis of conjunctive learning algorithms. *Machine Learning, 9*, 349–372.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Schlimmer, J. C. (1987). Efficiently inducing determinations: A complete and efficient search algorithm that uses optimal pruning. *Proceedings of the Tenth International Conference on Machine Learning* (pp. 284–290). Amherst, MA: Morgan Kaufmann.

Skalak, D. B. (1994). Prototype and feature selection by sampling and random mutation hill-climbing algorithms. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 293–301). New Brunswick, NJ: Morgan Kaufmann.

Townsend-Weber, T., & Kibler, D. (1994). Instance-based prediction of continuous values. *Working Notes of the AAAI94 Workshop on Case-Based Reasoning* (pp. 30–35). Seattle, WA: AAAI Press.