Pruning Irrelevant Features from Oblivious Decision Trees

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

Abstract

In this paper, we examine an approach to feature selection designed to handle domains that involve both irrelevant and interacting features. We review the reasons this situation poses challenges to both nearest neighbor and decision-tree methods, then describe a new algorithm - Oblivion - that carries out greedy pruning of oblivious decision trees. We summarize the results of experiments with artificial domains, which show that Oblivion's sample complexity grows slowly with the number of irrelevant features, and with natural domains, which suggest that few existing data sets contain many irrelevant features. In closing, we consider other work on feature selection and outline directions for future research.

1. Nature of the Problem

One of the central problems in machine induction involves discriminating between features that are relevant to the target concept and ones that are irrelevant. Presumably, many real-world learning tasks contain large numbers of irrelevant terms, and for such tasks, one would prefer to use algorithms that scale well along this dimension. More specifically, one would like the number of training instances needed to reach a given level of accuracy (the sample complexity) to grow slowly with increasing numbers of irrelevant features.

We define relevance in the context of such an induction task. Given a set of classified training instances for some target concept, the goal is to improve classification accuracy on a set of novel test instances. One way to improve accuracy involves identifying the features relevant to the target concept. Following John, Kohavi, and Pfleger (1994), we say that a feature is *relevant* if it belongs to some subset of the known features that is minimally sufficient to correctly classify instances. John et al. break their definition down further into notions of strong and weak relevance, but in this paper we will not find it necessary to distinguish the two senses.

Some previous experimental studies have examined the effect of irrelevant features on learning. For example, Aha (1990) reports experiments with a simple Boolean target concept which suggest that the sample complexity for the simple nearest neighbor method is *exponential* in the number of irrelevant features. Techniques for inducing decision trees, such as Quinlan's (1993) C4.5, do much better on conjunctive and similar target concepts because they attempt to select relevant features and eliminate irrelevant ones. However, such methods typically carry out a greedy search through the space of decision trees. This approach works well in domains where there is little interaction among the relevant attributes, as in conjunctive concepts, but the presence of attribute interactions, such as occurs in parity concepts, can cause significant problems for this scheme. Experimental studies by Almuallim and Dietterich (1991) and by Kira and Rendell (1992) show that, for some target concepts, methods for decision-tree induction also deal poorly with irrelevant features.

In response to this problem, Almuallim and Dietterich (1990) developed Focus, an algorithm which directly searches for minimal combinations of attributes that perfectly discriminate among the classes. This method begins by looking at each feature in isolation, then turns to pairs of features, triples, and so forth, halting as soon as it finds a combination that generates pure partitions of the training set (i.e., in which no instances have different classes). Their scheme then passes on the reduced set of features to ID3, which constructs a decision tree from the simplified training data. Comparative studies with ID3 and with Pagallo and Haussler's (1990) Fringe showed that, for a given number of training cases on randomly selected Boolean target concepts, Focus was almost unaffected by the introduction of irrelevant attributes, whereas the accuracy of ID3 and Fringe degraded significantly. Schlimmer (1993) has described a similar method that also starts with individual attributes and searches the space of attribute combinations, continuing until it finds a partition of the training set that has pure classes.

Both of these algorithms address the problem of attribute interaction in the presence of irrelevants by directly examining combinations of features. At least for noise-free data, this approach has the advantage of guaranteeing identification of minimal relevant feature sets, in contrast to the greedy approach used by C4.5 and its relatives. However, the price is greatly increased computational cost. Almuallim and Dietterich showed that FOCUS' time complexity is quasi-polynomial in the number of attributes, which they acknowledged is impractical for target concepts that involve many features. Schlimmer introduced techniques for pruning the search tree without losing completeness, but even with this savings, he had to limit the length of feature combinations considered (and thus the complexity of learnable target concepts) to keep search within bounds. Thus, there remains a need for more practical algorithms that can handle domains with both complex feature interactions and irrelevant attributes.

2. Pruning of Oblivious Decision Trees

Our research goal was to develop an algorithm that handled both irrelevant features and attribute interactions without resorting to expensive, enumerative search. Our response draws upon the realization that both Almuallim and Dietterich's and Schlimmer's approaches construct *oblivious* decision trees, in which all nodes at the same level test the same attribute. For example, a three-level oblivious tree might test attribute X at the top node, attribute Y in all nodes at the second level, and attribute Z in all nodes at the lowest level. This framework does not limit one's representational coverage; for every possible decision tree there exists an equivalent oblivious tree, though the former may have fewer nodes than the latter.

Although the above algorithms use forward selection (i.e., top-down search) to construct oblivious decision trees, this is not the only possible approach. Almuallim and Dietterich's FOCUS and Schlimmer's method require combinatorial search to handle attribute interactions precisely because they operate in this direction. However, experience with C4.5 and its relatives suggests that much of their power lies not in their forward selection scheme but in their use of *pruning* to eliminate unnecessary attributes. This suggests an alternative approach in which one starts with a full oblivious decision tree that includes all attributes, then uses pruning or backward elimination to remove features that do not aid classification accuracy. This scheme's advantage lies in the fact that accuracy decreases substantially when one removes a *single* relevant attribute, even if it interacts with other features, but accuracy remains unaffected when one prunes an irrelevant or redundant feature. This means that one can use greedy search rather than enumerative methods.

We have developed an algorithm, called OBLIVION, that instantiates this idea. The method begins with a full oblivious tree that incorporates all potentially relevant attributes and estimates this tree's accuracy on the entire training set, using a conservative technique like n-way cross validation. OBLIVION then removes each attribute in turn, estimates the accuracy of the resulting tree in each case, and selects the most accurate. If this best tree makes no more errors than the initial one, OBLIVION replaces the initial tree with the best one and continues the process. On each step, the algorithm tentatively prunes each of the remaining features, selects the best, and generates a new tree with one fewer attribute. This continues until the accuracy of the best pruned tree is less than the accuracy of the current one. Unlike FOCUS and Schlimmer's method, OBLIVION's time complexity is polynomial in the number of features, growing with the square of this factor.

There remain a few problematic details, such as determining the order of the retained attributes. However, one need not assign an order at all, since every order should produce equivalent behavior. Instead, one can view an oblivious decision tree as a set of disjoint rules, each using the same attributes in their condition sides. Because pruning can produce impure partitions of the training set, each rule specifies a distribution of class values. When an instance matches a rule's conditions, it simply predicts the most likely class. But sparse training data raises another issue – making predictions when a test case fails to perfectly match any rule. In this situation, we assume that one finds the best matching rules, sums the class probability distributions for each one, and predicts the most likely class.

In fact, this scheme is equivalent to using the simple nearest neighbor algorithm, but with some attributes ignored during the distance calculations. Given a test instance, this technique retrieves all those training cases that are nearest to it in the reduced space. If many features have been pruned, it becomes likely that a perfect match will occur so that the distance will be zero. Pruning also makes it probable that many training cases, though different in the original space, will appear identical in the reduced space. Given a tie, we assume that nearest neighbor takes the majority vote, which produces the same effect as predicting the most frequent class associated with an abstract rule. If no perfect matches exist, the method takes the majority vote of the nearest stored cases (which can correspond to multiple rules), giving the same result as the probabilistic scheme above. This insight into the relation between oblivious decision trees and nearest-neighbor algorithms was an unexpected benefit of our work.

3. Experimental Results with OBLIVION

We have carried out two types of experiments to evaluate OBLIVION's learning ability in comparison with nearest neighbor and decision-tree methods. In the first, we presented the three algorithms with artificial data, which let us explicitly vary the number of irrelevant Boolean features and observe the resulting degra-

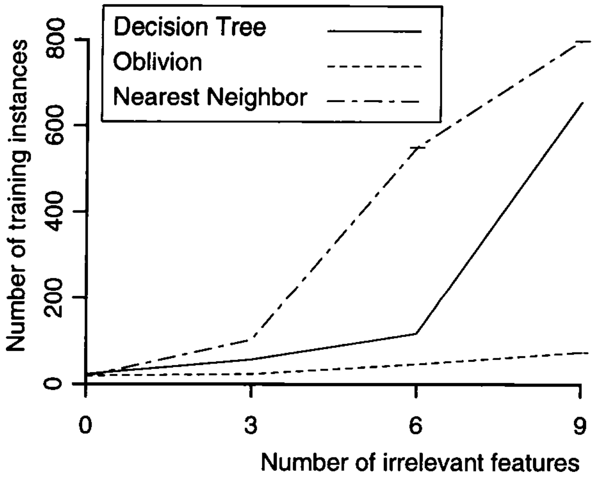

Figure 1. The number of training instances required for decision-tree induction, nearest neighbor, and OBLIVION to reach 95% accuracy on a separate test set, as a function of the number of irrelevant features, for the target concept ((A ∧ B ∧ C) ∨ (¬A ∧ ¬B ∧ ¬C)).

dation in behavior. The second type of experiment examined the ability of the algorithms in natural domains. These results demonstrated the comparative ability of OBLIVION in tackling real-world problems, and provided some clues as to the presence of irrelevant features in these domains.

We designed the studies with artificial data to test the methods' ability to scale to domains with many irrelevant features. Briefly, we found that the empirical sample complexity of nearest neighbor was exponential in the number of irrelevant features, confirming Aha's (1990) results. Moreover, this relation appears to hold across a variety of target concepts. In contrast, C4.5 scales linearly on some concepts (e.g., conjunctions and other concepts that can be stated as linear trees), but its sample complexity appears exponential on others (such as parity concepts). We present results for a parity-like concept in Figure 1. The number of training cases that OBLIVION requires to reach a given accuracy appears to be linear in the number of irrelevants, independent of the target concept, supporting our prediction that the algorithm should scale well to irrelevant features even in the presence of attribute interaction.

The experiments with natural domains provided a comparison of the three algorithms on a variety of data sets from the UCI repository and elsewhere. Although we could neither vary nor measure the number of irrelevants in these domains, we could make educated guesses about the prevalence of irrelevant features by comparing the patterns of results to those found with artificial data. In four domains – voting records, mushroom, DNA promoters, and breast cancer – we found no difference among the learning curves for the various methods. Inspection of the features selected by C4.5 and OBLIVION indicated that both algorithms used a small number of features in prediction. Since nearest neighbor's learning rate was no different in these domains, we hypothesized that the remaining attributes were probably redundant rather than irrelevant.

Different patterns emerged in two other natural domains. On chess endgames, we found that C4.5 learned more rapidly than OBLIVION, and that nearest neighbor fared the worst. On one of Cardie's (1993) natural language tasks, which involved prediction of a word's semantic class, we found that the learning curves for C4.5 and OBLIVION were indistinguishable, but that both were better than that for nearest neighbor. The poor performance of nearest neighbor in these domains suggests a reasonable number of truly irrelevant attributes. The superiority of C4.5 in the chess domain suggests that the target concept lacks feature interactions, and inspection of the induced trees confirmed that they were nearly linear in structure.

In summary, our experiments revealed a clear difference in the effect of irrelevant attributes and feature interaction on the behavior of nearest neighbor, the C4.5 algorithm, and OBLIVION. The rate of learning for the nearest neighbor method decreases drastically with the number of irrelevant dimensions, regardless of the target concept. The effect of irrelevant attributes on decision-tree induction depends on the nature of the target concept, giving a sample complexity that is linear for some and exponential for others. In contrast, the sample complexity for OBLIVION appears to be linear in the number of irrelevant terms, independent of the target concept. However, these encouraging results had little impact on six natural domains, where C4.5's learning curves were always as good or better than those for OBLIVION.

4. Concluding Remarks

Clearly, the experimental results we have presented are preliminary and must be treated with caution. In future work, we hope to replicate our findings in the presence of noise and on a broader range of target concepts, including ones that incorporate more relevant features, alternative Boolean combinations, and numeric attributes. In addition we would like to explore alternative schemes for searching the space of oblivious decision trees that may further improve OBLIVION's ability to scale to domains with many irrelevant features. We also hope to forge a stronger link between our studies of natural and artificial domains.

Another direction for future work involves comparing our approach with other methods for feature selection, including Almuallim and Dietterich's (1990) FOCUS algorithm, Kira and Rendell's (1992) more efficient RELIEF technique, and Aha's (1990) scheme for determining attribute weights for use in case retrieval. In addition, we should examine the relation of our work to similar feature-selection methods that a number of researchers have developed in parallel. These include methods for selecting attributes to use in decision-tree induction, described by John et al. (1994) and by Caruanå and Freitag (1994), as well as analogous techniques for use with nearest-neighbor methods, developed by Moore and Lee (1994), Skalak (1994), Townsend-Weber and Kibler (1994, and Aha and Bankert (1994). Like OBLIVION, all of these methods embed the induction algorithm within the feature-selection process, using estimated accuracy to direct a greedy search through the space of feature sets.

Despite the work that remains to be done, we believe that our initial studies have revealed interesting insights into the relative abilities of three different induction methods in handling two important sources of difficulty in learning. We anticipate that future experiments with OBLIVION and its relatives will produce deeper understanding of the characteristics of both algorithms for feature selection and the domains in which they operate, and we hope that other researchers will join us in their study.

Acknowledgements

Thanks to David Aha, George John, Karl Pfleger, Russ Greiner, Ronny Kohavi, Bharat Rao, and Jeff Schlimmer for discussions that improved the work reported in this paper. We also thank Ray Mooney, who provided modified C4.5 code for producing learning curves, and Claire Cardie, who provided her natural language data. Some of the material in this paper appears in the *Working Notes of the AAAI94 Workshop on Case-Based Reasoning*. This work was supported in part by ONR Grant No. N00014-94-1-0505.

References

Aha, D. (1990). *A study of instance-based algorithms for supervised learning tasks: Mathematical, empirical, and psychological evaluations*. Doctoral dissertation, Department of Information & Computer Science, University of California, Irvine.

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.

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.

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.

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.

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.

Pagallo, G., & Haussler, D. (1990). Boolean feature discovery in empirical learning. *Machine Learning*, *5*, 71–100.

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.