An Evolutionary Approach to Feature Set Selection

David W. Opitz
Department of Computer Science
University of Montana
Missoula, MT 59812
opitz@cs.umt.edu

Abstract

This paper investigates an ensemble feature selection algorithm that is based on genetic algorithms. The task of ensemble feature selection is harder than traditional feature selection in that one not only needs to find features germane to the learning task and learning algorithm, but one also needs to find a set of feature subsets that will promote disagreement among the ensemble's classifiers. Our algorithm shows improved performance over the popular and powerful ensemble approaches of AdaBoost and Bagging.

INTRODUCTION

Feature selection algorithms attempt to find and remove the features which are unhelpful or destructive to learning (Kohavi & John 1997). Previous work on feature selection has focused on finding the appropriate subset of relevant features to be used in constructing *one* inference model; however, recent "ensemble" work has shown that combining the output of a *set* of models that are generated from separately trained inductive learning algorithms can greatly improve generalization accuracy (Opitz 1999). Research has shown that an effective ensemble should consist of a set of models that are not only highly correct, but ones that make their errors on different parts of the input space as well (Opitz & Shavlik 1996). Varying the feature subsets used by each member of the ensemble (which we refer to as *ensemble feature selection*) should help promote this necessary diversity. Thus, while traditional feature-selection algorithms have the goal of finding the best feature subset that is germane to both the learning task and the selected inductive-learning algorithm, the task of ensemble feature selection has the additional goal of finding a *set* of features subsets that will promote disagreement among the component members of the ensemble.

The search space of sets of feature subsets is enormous and quickly becomes impractical to do hill-climbing searches. In this paper, we present a genetic algorithm approach for searching this space. Genetic algorithms are a logical choice since they have been shown to be effective global optimization techniques. Our approach works by first creating an initial population of classifiers where each classifier is generated by randomly selecting a different subset of features. We then continually produce new candidate classifiers by using the genetic operators of crossover and mutation on the feature subsets. Our algorithm defines the overall fitness of an individual to be a combination of accuracy and diversity. The most fit individuals make up the population which in turn comprise the ensemble. Using neural networks as our classifier, results on 21 datasets show that our simple and straight-forward algorithm for creating the initial population produces comparable ensembles on average to the popular and powerful ensemble approaches of Bagging and Boosting. Results also show that further running the algorithm with the genetic operators improves performance. Refer to Opitz 1999 for more details.

Acknowledgements

This work was partially supported by National Science Foundation grant IRI-9734419 and a University of Montana MONTS grant.

References

Kohavi, F., and John, G. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97(1):273–324.

Opitz, D., and Shavlik, J. 1996. Actively searching for an effective neural-network ensemble. *Connection Science* 8(3/4):337–353.

Opitz, D. 1999. Feature selection for ensembles. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.