Feature Selection Methods: Genetic Algorithms vs. Greedy-like Search

Haleh Vafaie and Ibrahim F. Imam
George Mason University, Fairfax, VA, 22030

Abstract. This paper presents a comparison between two feature selection methods, the Importance Score (IS) which is based on a greedy-like search and a genetic algorithm-based (GA) method, in order to better understand their strengths and limitations and their area of application. The results of our experiments show a very strong relation between the nature of the data and the behavior of both systems. The Importance Score method is more efficient when dealing with little noise and small number of interacting features, while the genetic algorithms can provide a more robust solution at the expense of increased computational effort.
Keywords. feature selection, machine learning, genetic algorithms, search.

1. INTRODUCTION

Feature selection is a problem that has to be addressed in many areas, especially in artificial intelligence. The main issues in developing feature selection techniques are choosing a small feature set in order to reduce the cost and running time of a given system, as well as achieving an acceptably high recognition rate. This has led to the development of a variety of techniques for selecting an optimal subset of features from a larger set of possible features. These feature selection techniques fall into two main categories. In the first approach problem specific strategies are developed based on the domain knowledge in order to reduce the number of features used to a manageable size (Dom 89). The second approach is used when the domain knowledge is unavailable or expensive to exploit. In this case, generic heuristics, essentially greedy algorithms, are applied to select a subset "d" of the available "m" features (Kittler 78).

The experiments reported in this paper compare two techniques which belong to the second category, in order to better understand their strengths and limitations and their area of application. The selected feature selection algorithms (Important Score method and GA-based technique) contain the basic components as shown in Figure 1 (Vafaie 93). The search procedures used by the Importance Score (IS) technique and the genetic algorithm-based (GA) method require no domain knowledge to assist the search process. Both IS and GA systems perform a search for achieving the highest predictive accuracy of rules produced by the AQ learning system as evaluated by their criterion function. The IS method performs a greedy-like search to obtain the minimum set of features that maximizes the recognition of AQ learned rules, while the genetic algorithm method

explores, in an efficient way, the space of all possible subsets to obtain the set of features that maximizes the predictive accuracy of the learned rules.