Chapter IX

A Building Block Approach to Genetic Programming for Rule Discovery

A.P. Engelbrecht and L. Schoeman
University of Pretoria, South Africa

Sonja Rouwhorst
Vrije Universiteit Amsterdam, The Netherlands

Genetic programming has recently been used successfully to extract knowledge in the form of IF-THEN rules. For these genetic programming approaches to knowledge extraction from data, individuals represent decision trees. The main objective of the evolutionary process is therefore to evolve the best decision tree, or classifier, to describe the data. Rules are then extracted, after convergence, from the best individual. The current genetic programming approaches to evolve decision trees are computationally complex, since individuals are initialized to complete decision trees.

This chapter discusses a new approach to genetic programming for rule extraction, namely the building block approach. This approach starts with individuals consisting of only one building block, and adds new building blocks during the evolutionary process when the simplicity of the individuals cannot account for the complexity in the underlying data.

Experimental results are presented and compared with that of C4.5 and CN2. The chapter shows that the building block approach achieves very good accuracies compared to that of C4.5 and CN2. It is also shown that the building block approach extracts substantially less rules.
A BUILDING BLOCK APPROACH TO GENETIC PROGRAMMING FOR RULE DISCOVERY

Recently developed knowledge extraction tools have their origins in artificial intelligence. These new tools combine and refine approaches such as artificial neural networks, genetic algorithms, genetic programming, fuzzy logic, clustering and statistics. While several tools have been developed, this chapter concentrates on a specific evolutionary computing approach, namely genetic programming (GP).

Evolutionary computing approaches to knowledge discovery have shown to be successful in knowledge extraction applications. They are, however, computationally expensive in their nature by starting evolution on large, complex structured individuals. This is especially true in the case of genetic programming where complex decision trees are evolved. This chapter presents a building-block approach to genetic programming, where conditions (or conjuncts) and sub-trees are only added to the tree when needed. The building-block approach to genetic programming (BGP) starts evolution with a population of the simplest individuals. That is, each individual consists of only one condition (the root of the tree), and the associated binary outcomes – thus representing two simple rules. These simple individuals evolve in the same way as for standard GP. When the simplicity of the BGP individuals fails to account for the complexity of the data, a new building block (condition) is added to individuals in the current population, thereby increasing their representation complexity. This building-block approach differs from standard GP mainly in the sense that standard GP starts with an initial population of individuals with various complexities.

The remainder of the chapter is organized as follows: the next section offers background on current knowledge extraction tools, and gives a motivation for the building-block approach. A short overview of standard GP is also given. The section that follows discusses the building-block approach in detail, with experimental results in the last section.

BACKGROUND

This section gives a short overview of well-known knowledge extraction tools, motivates the building-block approach and presents a summary of standard GP.

Knowledge Extraction Tools

The first knowledge extraction tools came from the machine learning community, grouped in two main categories based on the way that a classifier is constructed: decision tree approaches and rule induction approaches. The most popular decision tree algorithm was developed by Quinlan (1992), namely ID3. Subsequent improvement of ID3 resulted in C4.5 (Quinlan, 1993), which was later further extended, with the improved version called C5 (Quinlan, 1998). These decision tree approaches construct a decision tree, from which if-then rules are extracted (using e.g.,