Hybrid Genetic Algorithms in Data Mining Applications

Sancho Salcedo-Sanz
Universidad de Alcalá, Spain

Gustavo Camps-Valls
Universitat de València, Spain

Carlos Bousoño-Calzón
Universidad Carlos III de Madrid, Spain

INTRODUCTION

G**enetic algorithms** (**GAs**) are a class of problem solving techniques which have been successfully applied to a wide variety of hard problems (Goldberg, 1989). In spite of conventional GAs are interesting approaches to several problems, in which they are able to obtain very good solutions, there exist cases in which the application of a conventional GA has shown poor results. Poor performance of GAs completely depends on the problem. In general, problems severely *constrained* or problems with difficult *objective functions* are hard to be optimized using GAs. Regarding the difficulty of a problem for a GA there is a well established theory. Traditionally, this has been studied for binary encoded problems using the so called Walsh Transform (Liepins & Vose, 1991), and its associated *spectrum* (Hordijk & Stadler, 1998), which provides an idea of the distribution of the important schemas (building blocks) in the search space.

Several methods to enhance the performance of GAs in difficult applications have been developed. Firstly, the encoding of a problem determines the search space where the GA must work. Therefore, given a problem, the selection of the best *encoding* is an important pre-processing step. *Operators* which reduce the search space are then interesting in some applications. Secondly, variable length or transformed encodings are schemes, which can be successfully applied to some difficult problems. The hybridization of a GA with *local search* algorithms can also improve the performance of the GA in concrete applications. There are two types of hybridization:

- If the GA is hybridized with a local search heuristic in order to tackle the problem constraints, it is usually known as a *hybrid genetic algorithm*.
- If the GA is hybridized with a local search heuristic in order to improve its performance, then it is known as a *memetic algorithm*.

In this chapter we revise several hybrid methods involving GAs that have been applied to data mining problems. First, we provide a brief background with several important definitions on genetic algorithms, hybrid algorithms and operators for improving its performance. In the Main Trust section, we present a survey of several hybrid algorithms, which use GAs as search heuristic, and their main applications in data mining. Finally, we finish the chapter giving some conclusions and future trends.

BACKGROUND

Genetic Algorithms

Genetic algorithms are robust problems' solving techniques based on natural evolution processes. They are population-based algorithms, which encode a set of possible solutions to the problem, and evolve it through the application of the so called *genetic operators* (Goldberg, 1989). The standard genetic operators in a GA are:

- *Selection*, where the individuals of a new population are selected from the old one. In the standard implementation of the Selection operator, each individual has a probability of surviving for the

next generation proportional to its associated fitness (objective function) value. This procedure of selection is usually called roulette wheel selection mechanism.

- *Crossover*, where new individuals are searched starting from couples of individuals in the population. Once the couples are randomly selected, the individuals have the possibility of swapping parts of themselves with its couple, the probability of this happens is usually called *crossover probability*, *P_c*.
- *Mutation*, where new individuals are searched by randomly changing bits of current individuals with a low probability *P_m* (*probability of mutation*).

Operators for Enhancing Genetic Algorithms' Performance

Enhancing genetic algorithms with different new operators or local search algorithms to improve their performance in difficult problems is not a new topic. From the beginning of the use of these algorithms, researchers noted that there were applications too difficult to be successfully solved by the conventional GA. Thus, the use of new operators and the hybridization with other heuristics were introduced as mechanisms to improve GAs performance in many difficult problems.

The use of new operators is almost always forced by the problem encoding, usually in problems with constraints. The idea is to use the local knowledge about the problem by means of introducing these operators. Usually they are special *Crossover* and *Mutation* operators which results in feasible individuals. As an example, a *partially mapped crossover* operator is introduced when the encoding of the problem are permutations. Another example of special operators is the restricted search operators, which reduces the search space size, as we will describe later.

Another possibility to enhance the performance of GAs in difficult applications is their hybridization with local search heuristics. There are a wide variety of local search heuristics to be used in hybrid and memetic algorithms, neural networks, simulated annealing, tabu search or just hill-climbing, are some of the local heuristics used herein (Krasnogor & Smith, 2005). When the GA individual is modified after the application of the local search heuristic, the hybrid or memetic algorithm is called *Lamarckian* algorithm[1]. If the local heuristic does not modify the individual, but only its fitness value, it is called a *Baldwin* effect algorithm.

MAIN THRUST

Restricted Search in Genetic Algorithms

Encoding is an important point in genetic algorithms. Many problems in *data mining* require special encodings to be solved by means of GAs. However, the main drawback of not using a traditional binary encoding is that special operators must be used to carry on the crossover and mutation operations. Using binary encodings also have some drawback, mainly the large size of the search space in some problems. In order to reduce the search space size when using GAs with binary encoding, a restricted search operator can be used. Restricted search have been applied in different problems in data mining, such as feature selection (Salcedo-Sanz et al, 2002), (Salcedo-Sanz et al. 2004) or web mining (Salcedo-Sanz & Su, 2006). It consists of an extra operator to be added to the traditional selection, crossover and mutation. This new operator fixes the number of 1s that a given individual can have to a maximum of *p*, reducing the number of 1s if there are more than *p*, and adding 1s at random positions if there are less than *p*. This operation reduces the search space from 2^m to

$$\binom{m}{p},$$

being *m* the length of the binary strings that the GA encodes. Figure 1 shows the outline of the restricted search operator.

Hybrid and Memetic Algorithms

Hybrid algorithms are usually constituted by a genetic algorithm with a local search heuristic to repair infeasible individuals or to calculate the fitness function of the individual. They are specially used in constrained problems, where obtaining feasible individuals randomly is not a trivial task. No impact of the local search in the improvement fitness function value of