Metaheuristics in Data Mining

Miguel García Torres
Universidad de La Laguna, Spain

Belén Melián Batista
Universidad de La Laguna, Spain

José A. Moreno Pérez
Universidad de La Laguna, Spain

José Marcos Moreno-Vega
Universidad de La Laguna, Spain

INTRODUCTION

The *Metaheuristics* are general strategies for designing heuristic procedures with high performance. The term metaheuristic, which appeared in 1986 for the first time (Glover, 1986), is compound by the terms: "meta", that means over or behind, and "heuristic". Heuristic is the qualifying used for methods of solving optimization problems that are obtained from the intuition, expertise or general knowledge (Michalewicz & Fogel, 2000).

Nowadays a lot of known strategies can be classified as metaheuristics and there are a clear increasing number of research papers and applications that use this kind of methods. Several optimization methods that already existed when the term appeared have been later interpreted as metaheuristics (Glover & Kochenberger, 2003). Genetic Algorithms, Neural Networks, Local Searches, and Simulated Annealing are some of those classical metaheuristics. Several modern metaheuristics have succeeded in solving relevant optimization problems in industry, business and engineering. The most relevant among them are Tabu Search, Variable Neighbourhood Search and GRASP. New population based evolutionary metaheuristics such as Scatter Search and Estimation Distribution Algorithms are also quite important. Besides Neural Networks and Genetic Algorithms, other nature-inspired metaheuristics such as Ant Colony Optimization and Particle Swarm Optimization are also now well known metaheuristics..

BACKGROUND

The ***Metaheuristic*** methods are general strategies for designing heuristic procedures for solving an optimization problem. An optimization problem is characterized by a search space S of feasible solutions and an objective function f. Solving the problem consists of finding an *optimal* solution s^*; i.e., a feasible solution that optimizes f in S. Given a set of transformations or moves on the solution space, the *neighbourhood* of s, denoted by $N(s)$, is the set of solutions that are reachable from s with one of these moves. A *local optimum* is a solution s that optimizes f in its neighbourhood $N(s)$. A *Local Search* is a procedure that iteratively applies an improving move to a solution (Pirlot, 1996; Yagiura & Ibaraki, 2002). The main objection to local searches is that they are trapped in a local optimum. The first metaheuristics arose looking for ways to escape from local optima in order to reach an optimal solution. There are an increasing number of books and reviews on the whole field of Metaheuristics (Reeves, 1993, Michalewicz & Fogel, 2000; Glover & Kochenberger, 2003; Blum & Roli, 2003)

Data mining (DM) is a constantly growing area. DM tools are confronted to a particular problem: the great number of characteristics that qualify data samples. They are more or less victims of the abundance of information. DM needs benefits from the powerful metaheuristics that can deal with huge amounts of data in Decision Making contexts. Several relevant tasks in DM; such as clustering, classification, feature selection and data reduction, are formulated as optimization problems. The solutions for the corresponding problem consist of the values for the parameters that specify the role designed for performing the task. In nearest-neighbour clustering and classification, the solutions consist of the possible selections of cases for applying the rule. The objective functions are the corresponding performance measures. In Feature Selection and Data Reduction, the solutions are set of variables or cases and, if the size of set of features or the amount of data is fixed, the objective is to maximize the (predictive) performance. However in general, there are, at least, two objectives: the accuracy and the simplicity. They are usually contradictory and generally referred by the performance and the amount of information used for prediction. The accuracy is to be maximized and the amount of information is to be minimized. Therefore, multi-objective metaheuristics are appropriated to get the adequate tradeoff.

MAIN FOCUS

The main focus in the metaheuristics field related to DM is in the application of the existing and new methods and in the desirable properties of the metaheuristics. Most metaheuristic strategies have already been applied to DM tasks but there are still open research lines to improve their usefulness.

Main Metaheuristics

The *Multi-start* considers the ways to get several initial solutions for the local searches in order to escape from local optima and to increase the probability of reaching the global optimum (Martí, 2003; Fleurent & Glover, 1999). *GRASP* (*Greedy Randomized Adaptive Search Procedures*) comprises two phases, an adaptive construction phase and a local search (Feo & Resende, 1995; Resende & Ribeiro, 2003). The distinguishing feature of *Tabu Search* (Glover, 1989, 1990, Glover & Laguna, 1997) is the use of adaptive memory and special associated problem-solving strategies. *Simulated Annealing* (Kirkpatrick et al., 1983; Vidal, 1993) is derived from a local search by allowing also, probabilistically controlled, not improving moves. *Variable Neighbourhood Search* is based on systematic changes of neighbourhoods in the search for a better solution (Mladenović & Hansen, 1997; Hansen and Mladenović, 2003). *Scatter Search* (Glover, 1998; Laguna & Martí, 2002) uses an evolving reference set, with moderate size, whose solutions are combined and improved to update the reference set with quality and dispersion criteria. *Estimation of Distribution Algorithms* (Lozano & Larrañaga, 2002) is a population-based search procedure in which a new population is iteratively obtained by sampling the probability distribution on the search space that estimates the distribution of the good solutions selected from the former population. *Ant Colony Optimization* (Dorigo & Blum, 2005; Dorigo & Di Caro, 1999; Dorigo & Stützle, 2004) is a distributed strategy where a set of agents (artificial ants) explore the solution space cooperating by means of the pheromone. *Particle Swarm Optimization* (Clerc, 2006, Kennedy & Eberhart, 1995; Eberhart & Kennedy, 1995; Kennedy & Eberhat, 2001) is an evolutionary method inspired by the social behaviour of individuals within swarms in nature where a swarm of particles fly in the virtual space of the possible solutions conducted by the inertia, memory and the attraction of the best particles.

Most metaheuristics, among other optimization techniques (Olafsson et al., 2006), have already been applied to DM, mainly to Clustering and Feature Selection Problems. For instance, *Genetic Algorithms* has been applied in (Freitas, 2002), *Tabu Search* in (Tahir et al., 2007; Sung & Jin, 2000), *Simulated Annealing* in (Debuse & Rayward-Smith, 1997, 1999), *Variable Neighbourhood Search* in (Hansen and Mladenović, 2001; Belacel et al., 2002; García-López et al., 2004a), *Scatter Search* in (García-López et al., 2004b, 2006; Pacheco, 2005), *Estimation of Distribution Algorithms* in (Inza et al., 2000, 2001), *Ant Colony Ooptimization* in (Han & Shi, 2007; Handl et al., 2006; Admane et al., 2004; Smaldon & Freitas, 2006) and *Particle Swarm Optimization* in (Correa et al., 2006; Wang et al., 2007). Applications of *Neural Networks* in DM are very well known and some review or books about modern metaheuristics in DM have also already appeared (De la Iglesia et al., 1996; Rayward-Smith, 2005; Abbass et al., 2002)

Desirable Characteristics

Most authors in the field have used some of desirable properties of metaheuristics to analyse the proposed methods and few of them collected a selected list of them (Melián et al., 2003). The desirable characteristics