

## **Metaheuristics in Data Mining**

**Miguel García Torres**

Departamento de Estadística, I.O. y Computación  
Instituto Universitario de Desarrollo Regional  
Universidad de La Laguna  
38271 La Laguna  
Spain  
Voice: +34 922318186  
Fax: +34 922318170  
e-mail: mgarcia@ull.es

**Belén Melián Batista**

Departamento de Estadística, I.O. y Computación  
Instituto Universitario de Desarrollo Regional  
Universidad de La Laguna  
38271 La Laguna  
Spain  
Voice: +34 922318637  
Fax: +34 922318170  
e-mail: mbmelian@ull.es

**José A. Moreno Pérez (\*)**

Departamento de Estadística, I.O. y Computación  
Instituto Universitario de Desarrollo Regional  
Universidad de La Laguna  
38271 La Laguna  
Spain  
Voice: +34 922318186  
Fax: +34 922318170  
e-mail: jamoreno@ull.es

**José Marcos Moreno-Vega**

Departamento de Estadística, I.O. y Computación  
Instituto Universitario de Desarrollo Regional  
Universidad de La Laguna  
38271 La Laguna  
Spain  
Voice: +34 922318175  
Fax: +34 922318170  
e-mail: jmmoreno@ull.es

(\* Corresponding author)

# 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 & Eberhart, 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 Optimization* 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 of the metaheuristics, from the theoretical and practical points of view, provide ways for the improvement in the field. A list of them follows.

1. ***Simplicity***: the metaheuristics should be based on a simple and clear principle, easy to understand.
2. ***Precise***: the steps or phases of the metaheuristic must be stated in precise terms, without room for the ambiguity.
3. ***Coherence***: the steps of the algorithms for particular problems should follow naturally from the principles of the metaheuristic;
4. ***Efficiency***: the procedures for particular problems should provide good solutions (optimal or near-optimal) in moderate computational time;
5. ***Efficacy***: the algorithms should solve optimally most problems of benchmarks, when available;
6. ***Effectiveness***: the procedures for particular problems should provide optimal or near-optimal solutions for most realistic instances.

7. **Robustness**: the metaheuristics should have good performance for a variety of instances, i.e., not just be fine-tuned to some data and less good elsewhere;
8. **Generality**: the metaheuristics should lead to good heuristics for a large variety of problems.
9. **Adaptable**: the metaheuristics should include elements to adapt to several contexts or field of applications or different kind of models
10. **User-friendliness**: the metaheuristics should be easy to use; without parameters or such they are easily understood and tuned.
11. **Innovation**: the principles of the metaheuristics, and/or their use, should lead to new types of applications.
12. **Interactivity**: the metaheuristics should allow the user to incorporate his knowledge in order to improve the performance of the procedure.
13. **Multiplicity**: the methods should be able to present several near optimal solutions among which the user can choose.
14. **Autonomous**: the metaheuristics should allow implementations without parameters or such that they are automatically tuned.
15. **Applicability**: the metaheuristics should be widely applicable to several fields.

## **FUTURE TRENDS**

Several of the future trends in metaheuristics will have a big impact in DM because they incorporate the methodologies of intelligent systems to solve the difficulties for efficiently dealing with high amount of data.

### **Hybrid Metaheuristics**

Metaheuristics can get the benefits from the hybridization methodologies (Almeida et al., 2006; Talbi, 2002). This new emerging field includes combinations of components from different metaheuristics, low-level and high-level hybridization, portfolio techniques, expert systems, co-operative search and co-evolution techniques.

### **Cooperative Metaheuristics**

The cooperative metaheuristics consider several search agents that implement metaheuristics and cooperate by interchanging information on the search. The cooperative scheme can be centralized or decentralized depending on the existence of a control of the communications and the way the agents use the information. They are usually obtained from the population search strategies where each individual gets search capabilities and communicates with other individuals in the population (García-Pedrajas et al., 2001; Huang, 2006; Grundel et al., 2004; Melián-Batista et al., 2006).

### **The Learning Metaheuristics**

The performance of the metaheuristics improves by using Machine Learning tools that incorporate the knowledge obtained by the algorithm while it runs. These tools should allow a metaheuristic to tune their parameters to get the best possible performance on a set of instances (Cadenas et al., 2006; Guo, 2003).

### **Nature-inspired Metaheuristics**

The number and success of the metaheuristics derived from the studies of some natural phenomena are increasing in the last years. In addition to the classical *Neural Networks* and *Genetic Algorithms* other metaheuristic strategies are being consolidated in the field such as *Ant Colony Optimization* and *Particle Swarm Optimization*.

However recent proposals like *Artificial Immune Systems* (Timmis, 2006), *Membrane Systems* (Paun, 2002) or *Swarm Intelligence* (Engelbrecht, 2005) have not been enough studied (Forbes, 2005).

### **Multiple Objective Metaheuristics**

Since most of the real problems in DM are multiobjective problems (where several contradictory objective functions involved), the adaptation or capabilities of metaheuristics for these problems are very relevant in real applications (Baños et al., 2006).

### **Parallel Metaheuristics**

With the proliferation of parallel computers and faster community networks, parallel metaheuristics (Alba, 2005) is already being now an effective alternative to speed up metaheuristics searches in DM and allow efficiently dealing with high amount of data.

## **CONCLUSIONS**

Modern metaheuristics have succeeded in solving optimization problems in relevant fields. Several relevant tasks in DM are formulated as optimization problems. Therefore, metaheuristics should provide promising tools for improving DM tasks. Some examples of this already have appeared in the specialized literature. In order to extend this success, the knowledge of the good characteristics of the successful metaheuristic is important. The relevance of the future trends of the field for DM applications depends of this knowledge.

## REFERENCES

- Abbass, H.A., Newton, C.S. & Sarker, R. (Eds.) (2002) *Data Mining: A Heuristic Approach*. Idea Group Publishing
- Admane, L., Benatchba, K., Koudil, M., Drias, M., Gharout, S. & Hamani, N. (2004) Using ant colonies to solve data-mining problems. *IEEE International Conference on Systems, Man and Cybernetics*, 4 (10-13), 3151–3157
- Alba, E. (ed.) (2005) *Parallel Metaheuristics. A New Class of Algorithms*. Wiley.
- Almeida, F., Blesa Aguilera, M.J., Blum, C., Moreno Vega, J.M., Pérez Pérez, M., Roli, A., Sampels, M. (Eds.) (2006) *Hybrid Metaheuristics*. LNCS 4030, pp. 82-93, Springer
- Baños, R., Gil, C., Paechter, B., Ortega, J. (2007) A Hybrid Meta-heuristic for Multi-objective Optimization: MOSATS. *Journal of Mathematical Modelling and Algorithms*, 6/2, 213-230.
- Belacel, N., Hansen, P. & Mladenovic, N. (2002) Fuzzy J-means: a new heuristic for fuzzy clustering. *Pattern Recognition*, 35(10), 2193-2200
- Blum, C. & Roli, A. (2003) Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, 35(3), 268-308
- Cadenas, J.M., Canós, M.J., Garrido, M.C., Liern, V., Muñoz, E., Serrano, E. (2007) Using Data Mining in a Soft-Computing based Cooperative Multi-agent system of Metaheuristics. *Journal of Applied soft Computing* (in press).
- Clerc, M. (2006). *Particle Swarm Optimization*. ISTE.
- Correa, E.S., Freitas, S.A. & Johnson, C.G. (2006) A new Discrete Particle Swarm Algorithm Applied to Attribute Selection in a Bioinformatic Data Set. *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, 35–42

- Debuse, J.C.W. & Rayward-Smith, V.J. (1997) Feature subset selection within a simulated annealing data mining algorithm. *J. Intelligent Information Systems*, 9(1), 57-81
- Dorigo, M. & Blum, C. (2005) Ant colony optimization theory: A survey. *Theoretical Computer Science*, 344(2-3), 243-278
- Dorigo, M. & Di Caro, G. (1999) The Ant Colony Optimization Meta-Heuristic. In Corne, D., Dorigo, M. F. & Glover, F. (Eds), *New Ideas in Optimization*, McGraw-Hill, 11-32
- Dorigo, M. & Stützle, T. (2004) *Ant Colony Optimization*. MIT Press
- Eberhart, R.C. & Kennedy, J. (1995) A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, Nagoya, Japan, 39–43.
- Engelbrecht, P. (2005) *Fundamentals of Computational Swarm Intelligence*. Wiley
- Feo, T.A & Resende, M.G.C. (1995) Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6, 109-133
- Fleurent, C. & Glover, F. (1999) Improved constructive multistart strategies for the quadratic assignment problem using adaptive memory. *INFORMS Journal on Computing*, 11, 198-204
- Forbes, N. (2005) *Imitation of life: How Biology is Inspiring Computing*. MIT Press
- Freitas, A. (2002) *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer
- García-López, F., García-Torres, M., Melián, B., Moreno-Pérez, J.A. & Moreno-Vega, J.M. (2004a) Solving feature subset selection problem by a hybrid metaheuristic. In *Proceedings of First International Workshop in Hybrid Metaheuristics at ECAI2004*, 59-69

- García-López, F., García-Torres, M., Moreno-Pérez, J.A. & Moreno-Vega, J.M. (2004b) Scatter Search for the feature selection problem. *Lecture Notes in Artificial Intelligence*, 3040, 517-525
- García-López, F., García-Torres, M., Melián, B., Moreno, J.A. & Moreno-Vega, J.M. (2006) Solving feature subset selection problem by a parallel scatter search. *European Journal of Operational Research*, 169, 477-489
- García-Pedrajas, N., Sanz-Tapia, E., Ortiz-Boyer, D. & Cervás-Martínez, C. (Eds.) (2001) *Introducing Multi-objective Optimization in Cooperative Coevolution of Neural Networks*. LNCS, 2084.
- Glover, F. & Kochenberger, G. (Eds.) (2003) *Handbook on MetaHeuristics*. Kluwer
- Glover, F. & Laguna, M. (1997) *Tabu Search*. Kluwer.
- Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 5:533-549.
- Glover, F. (1989) Tabu search. part I. *ORSA Journal on Computing*, 1:190-206.
- Glover, F. (1990) Tabu search. part II. *ORSA Journal on Computing*, 2:4-32.
- Glover, F. (1998) A template for scatter search and path relinking. In J.-K. Hao and E. Lutton, editors, *Artificial Evolution*, volume 1363, 13-54. Springer-Verlag.
- Grundel, D., Murphey, R. & Pardalos, P.M. (Eds.) (2004) *Theory and Algorithms for Cooperative Systems*. Series on Computers and Operations Research, 4. World Scientific.
- Guo, H. (2003) A Bayesian Approach to Automatic Algorithm Selection. IJCAI03 Workshop on AI and Automatic Computing.
- Han, Y.F. & Shi, P.F. (2007) An improved ant colony algorithm for fuzzy clustering in image segmentation. *Neurocomputing*, 70, 665-671

- Handl, J., Knowles, J. & Dorigo, M (2006) Ant-based clustering and topographic mapping. *Artificial Life*, 12, 35-61
- Hansen, P. & Mladenović, N. (2001) J-means: A new local search heuristic for minimum sum-of-squares clustering. *Pattern Recognition*, 34(2), 405-413
- Hansen, P. & Mladenovic, N. (2003) Variable Neighborhood Search. In F. Glover and G. Kochenberger (Eds.), *Handbook of Metaheuristics*, Kluwer, 145--184.
- Huang, X. (2006) From Cooperative Team Playing to an Optimization Method. *NICSO-2006 Granada*.
- Iglesia, B. de la, Debuse, J.C.W. & Rayward-Smith, V.J. (1996). Discovering Knowledge in Commercial Databases Using Modern Heuristic Techniques. *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining KDD96*, 44-49
- Inza, I., Larrañaga, P. & Sierra, B. (2000) Feature subset selection by bayesian network based optimization. *Artificial Intelligence*, 123, 157-184
- Inza, I., Larrañaga, P. & Sierra, B. (2001) Feature subset selection by bayesian networks: a comparison with genetic and sequential algorithms. *International Journal of Approximate Reasoning*, 27(2), 143-164
- Kennedy, J. & Eberhart, R. (1995) Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks*, IV, 1942–1948.
- Kennedy, J. & Eberhart, R. (2001) *Swarm Intelligence*. Morgan Kaufmann.
- Kirpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983) Optimization by Simulated Annealing. *Science*, 220, 671-679
- Laguna, M. & Martí, R. (2002) Scatter Search Methodology and Implementations in C. Kluwer.
- Lozano, J.A. & Larrañaga, P. (2002). Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation. Kluwer.

- Martí, R. (2003) Multistart Methods. In *Handbook on MetaHeuristics*, Glover, F. & Kochenberger, G. (Eds.) Kluwer, 355-368
- Mladenovic, N. & Hansen, P. (1997) Variable neighborhood search. *Computers and Operations Research*, 24, 1097-1100.
- Melián-Batista, B., Moreno-Pérez, J.A. & Moreno Vega, J.M. (2006) Nature-inspired Decentralized Cooperative Metaheuristic Strategies for Logistic Problems. *NiSIS-2006*, Tenerife.
- Michalewicz, Z. & Fogel, D.B. (2000). *How to Solve It: Modern Heuristics*. Springer.
- Olafsson, S., Lia, X. & Wua, S. (2006) Operations research and data mining. *European Journal of Operational Research*, (doi:10.1016/j.ejor.2006.09.023), to appear
- Paun, G. (2002) *Membrane Computing. An Introduction*. Springer
- Pacheco, J. (2005) A Scatter Search Approach for the Minimum Sum-of-Squares Clustering Problem. *Computers and Operations Research*, 32, 1325-1335
- Pirlot, M. (1996) General local search methods. *European Journal of Operational Research*, 92(3):493-511.
- Rayward-Smith, V.J. (2005) Metaheuristics for clustering in KDD. *Evolutionary Computation*, 3, 2380-2387
- Reeves, C.R. (ed.) (1993) *Modern Heuristic Techniques for Combinatorial Problems*. Blackwell.
- Resende, M.G.C. & Ribeiro, C.C. (2003) Greedy randomized adaptive search procedures. In Glover, F. & Kochenberger, G. (Eds.) *Handbook of Metaheuristics*, Kluwer, 219-249
- Smaldon, J. & Freitas, A. (2006) A new version of the ant-miner algorithm discovering unordered rule sets. *Proceedings of GECCO '06*, 43-50

- Sung, C.S. & Jin, H.W. (2000) A tabu-search-based heuristic for clustering. *Pattern Recognition*, 33, 849-858
- Talbi, E-G. (2002) A Taxonomy of Hybrid Metaheuristics, *Journal of Heuristics* 8(5), 541-564
- Tahir, M.A., Bouridane, A. & Kurugollu, F. (2007) Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. *Pattern Recognition Letters*, 28, 438-446
- Timmis, J. (2006). Artificial Immune Systems - Today and Tomorrow. To appear in *Special Issue on Natural and Artificial Immune Systems. Natural Computation*.
- Yagiura M. and Ibaraki T. (2002) Local search. In P.M. Pardalos and M.G.C. Resende, (Eds.), *Handbook of Applied Optimization*, 104-123. Oxford University Press.
- Vidal, R.V.V. (1993) *Applied simulated annealing*. Springer, 1993.
- Wang, X.Y., Yang, J., Teng, X.L., Xia, W.J. & Jensen, R. (2007) Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, 28, 459-471

## **KEY TERMS AND THEIR DEFINITIONS:**

**Optimization problem:** Given a set  $S$  of alternative solutions and an objective function  $f$ , the corresponding optimization problem consists in finding the solution  $s$  that optimizes  $f$ .

**Heuristic:** A procedure to provide a good solution of an optimization problem that is obtained from the intuition, expertise or general knowledge.

**Metaheuristics:** The metaheuristics are general strategies for designing heuristic procedures with high performance.

**Local Search:** A heuristic method consisting of iteratively applying an improving move to a solution until a stopping criterion is met.

**Neighbourhood:** Given a set of moves or transformations in the solution space of an optimization problem, the neighbourhood of a solution is the set of solutions that can be obtained from it by one of these moves or transformations

**Local optimum:** A solution of an optimization problem that is better than any other solution of its neighbourhood

**Memory:** The capability of some search methods for using the information on the solutions examined in the search process and their objective values.

**Intensification:** The capability of the search methods for improving the solutions of an optimization problem.

**Diversification:** The capability of the search methods to explore different zones of the search space.

**Population Method:** A solution method, based on a set of solutions (the population), that searches for the optimum in the solution space of an optimization problem.