

# Data Mining with Scatter Search\*

I.J. García del Amo, M. García Torres\*\*, B. Melián Batista, J.A. Moreno Pérez, J.M. Moreno Vega and Raquel Rivero Martín

Dep. de Estadística, I.O. y Computacin, Universidad de La Laguna,  
38271 La Laguna, Spain

**Abstract.** Most Data Mining tasks are performed by the application of Machine Learning techniques. Metaheuristic approaches are becoming very useful for designing efficient tools in Machine Learning. Metaheuristics are general strategies to design efficient heuristic procedures. Scatter Search is a recent metaheuristic that has been successfully applied to solve standard problems in three central paradigms of Machine Learning: Clustering, Classification and Feature Selection. We describe the main components of the Scatter Search metaheuristic and the characteristics of the specific designs to be applied to solve standard problems in these tasks.

## 1 Introduction

Processing Intelligent Information requires efficient tools to extract the useful information stored in databases. Data Mining and Knowledge Discovery are powerful techniques for the extraction of information from large databases. Heuristic approaches are already quite relevant in Data Mining [1]. Most of the data mining tasks are solved by the application of Machine Learning techniques. Three central paradigms for the application of Machine Learning [9] in Data Mining are Clustering, Instance-Based Classification and Feature Selection. The Scatter Search metaheuristic has been tested for the kind of problems that appear in these tasks ([3], [2]). We describe the main components of this metaheuristic and their specific designs to solve standard problems in these contexts.

Given a set of instances characterized by several features, the clustering or grouping problem consists in grouping similar instances in the same cluster and dissimilar instances in different clusters. If in addition to the description of the objects, their classes in a training set are given, the classification problem consists of obtaining the optimal classification rule to assign the class to the new examples based on their description. Finally, the feature selection problem consisting in selecting a subset of features in order to best perform the classification task.

Scatter Search [8] is a population-based metaheuristic that constructs solutions by combining others in an evolving set of solutions, named reference set (*RefSet*). The procedure combines solutions of the reference set and runs a local

---

\* This research has been partially supported by the Ministerio de Ciencia y Tecnología through the project TIC2002-04242-C03-01; 70% of which are FEDER funds

\*\* This author has been partially supported by the Cajacanarias grant

search procedure to reach a local optimum that would be used to update the reference set depending on the results of the improvements. The two main differences between Scatter Search and other classical population-based procedures in Data Mining [7] like Genetic Algorithms [6] are the size of the evolving set of solutions and the way the method combines the existing solutions to provide new ones. The evolving set *RefSet* in scatter search has a relatively small or moderate size (typical sizes are 10 or 15, see [8]). Scatter Search combines good solutions to construct others exploiting the knowledge of the problem at hand in an intelligent way. Genetic Algorithms are also evolutionary algorithms in which a population of solutions evolves by using the mutation and crossover operators, which have a significant reliance on randomization to create new solutions.

## 2 Scatter Search Metaheuristic

The principles of the Scatter Search metaheuristic were first introduced in the 1970s as an extension of formulations for combining decision rules and problem constraints. This initial proposal generates solutions taking account of characteristics in several parts of the solution space [4].

In a Scatter Search algorithm [8], a moderate-sized set of solutions, the reference set *RefSet*, evolves due to mechanisms of intelligent combination of solutions. Unlike other strategies of combination of existing rules like genetic algorithms, the search for a local optimum is a guided task. In order to perform this strategy the set of reference solutions, (*RefSet*), is selected from a population of solutions. The *RefSet* is generated and then iteratively updated attempting to intensify and diversify the search. After intelligently combining the solutions in the reference set, a local search procedure is applied to improve the resulting solution, and the *RefSet* is updated to incorporate both good and disperse solutions. These steps are repeated until a stopping condition is met. The method provides not only a single heuristic solutions, like other metaheuristics, but a reduced set of disperse high quality solutions.

The Scatter Search metaheuristic includes five main methods or component processes:

1. Diversification Generation Method This method is used to generate a wide set of diverse solutions.
2. Improvement Method This component process improves the solutions to reach better ones; usually local optima.
3. Reference Set Update Method This is the method that builds and updates the reference set, which consists of a reduced set of good and disperse solutions.
4. Subset Generation Method This is the method applied to select the subsets of solutions from the reference set to be combined.
5. Solution Combination Method This process combines the solutions in the selected subsets to produce new solutions

A comprehensive description of the fundamentals of Scatter Search can be found in [5].

A simple implementation of the basic Scatter Search algorithm based in these methods is shown in Figure 1.

---

```
procedure Scatter Search
begin
    Diversification Generation Method;
    Improvement Method;
    repeat
        Reference Set Update Method;
        Subset Generation Method;
        Solution Combination Method;
        Improvement Method;
    until (StoppingCriterion);
end.
```

---

**Fig. 1.** A Scatter Search Metaheuristic Pseudocode

The algorithm starts generating a population of solutions by running the *Diversification Generation Method*. This procedure creates a large set of disperse solutions that are improved by the *Improvement method*. A representative set of  $RefSetSize$  good solutions are chosen to be included in the reference set (*RefSet*). These solutions are not limited to those with the best objective function values; the reference set must also include diverse solutions. The reference set is initially generated by selecting the  $RefSetSize_1$  best solutions according to the objective function values that are chosen to be in *RefSet*. Then  $RefSetSize_2$  times, the most disperse solution with respect to *RefSet* is found and added to *RefSet* (the final size of *RefSet* is  $RefSetSize = RefSetSize_1 + RefSetSize_2$ ). Several subsets of solutions from the *RefSet* are then systematically selected by the *Subset Generation Method*. The *Solution Combination Method* combines the solutions in each subset taking account their good features without reliance on randomization. Then, the *Improvement Method* is applied to the result of the combination to get an improved solution. Finally, the *Reference Set Update Method* uses the obtained solution to update the reference set following both intensification and diversification criteria.

### 3 Application of Scatter Search in Data Mining

Metaheuristic searches are becoming very important for Machine Learning applications in Data Mining, Medical Record, Software Engineering, Autonomous

Driving, Speech Recognition and Self Customizing programs. Three main paradigms in Machine Learning applications in Data Mining are: clustering, instance-based learning and feature selection. We describe the specific design of the main components of the Scatter Search to solve standard problems in these three tasks.

Clustering is the main paradigm of unsupervised learning. The objective of clustering is to find groups of instances constituted by similar instances. Given a set of instances described by a series of features, the problem is to find a partition of the whole set in subsets or classes in such a way that instances in the same class are very similar and instances in different classes are very dissimilar. The distance based approach considers a distance between the instance descriptions to evaluate the similarity and dissimilarity among them. A wide set of distance functions appropriated for different kinds of instance descriptions have been proposed and analyzed in the literature (see [10]). Then, a very usual way to define the partition consists of finding some representative instances for the classes (in the simplest case only one instance is chosen for each class). Then each instance is assigned to the class of the nearest of these representative instances. Scatter Search has been successfully applied to the  $p$ -median location problem that is very similar to the Clustering Problem [3]. The  $p$ -median problem consists in choosing the  $p$  points that minimize the sum of distances to the remainder instances.

In instance-based supervised learning, in addition to the features that describe the instances, an additional variable that represents the class of the instance that is to be predicted from its description is considered. From a training set of instances with known classes, we want to get a classification rule to obtain the unknown classes of a set of test instances or examples. The distance-based classification approach also selects a set of representative instances from the training set and classifies the test instances taking into account the class of the nearest selected instance. A wide set of possible distance functions among descriptions can also be applied for this tasks. This problem is also similar to the  $p$ -median problem since it also consists of selecting a number of instances with a different optimization function. They belong to the wide set of the named  $p$ -selection problems, for which most of the heuristic procedures are based on swaps in the solutions. The scatter search approach for a  $p$ -selection problem based on interchange moves can be easily adapted to other problem in this class.

However, the use of the whole set of features is not useful for being considered in this or other classification paradigms. The feature selection problem tries to get the best subset of features to perform the classification task. The appropriated selection of features has not only the advantage of taking the relevant information in the description of the instances, but also avoiding redundant information and making the classification algorithms and rules more efficient to obtain and to use. Scatter Search has been tested for this problem in [2] using a distance between solutions (now sets of features) to evaluate the diversity among a set of solutions.

The distance function between solutions plays a central role in the Scatter Search to modulate the diversification and intensification. Given a distance

between the items that constitute the solutions (instances for clustering and classification and variables for feature selection), the distance between two solutions is the sum of the distances between the items in one solution and the other solution. Similarly, the most diverse solution with respect to a set of solutions is defined in a similar way.

The most important parameter in the population creation method is the size of the population. The usual sizes are a quadratic or linear function of the number of classes for clustering and classification and the number of features to be chosen for feature selection problem. Usual procedures consist of randomly generating solutions from which a good population is obtained by quality and diversity criteria. The *Reference Set Update Method* generates and updates the reference set by following both quality and diversity criteria. Solutions for the reference set are first chosen by quality; e.g. the  $RefSetSize/2$  best solutions. Then new solutions are iteratively included in the reference set by following a diversity criterium until the whole reference set is obtained. A usual procedure is described as follows. Let  $C$  be the set of items that belong to any solution already in the reference set. The diversity of each possible new solution  $S$  is given by a distance between  $S$  and  $C$  using a corresponding distance measure between items. Then the most diverse solution is chosen  $RefSetSize/2$  times until the reference set with  $RefSetSize = RefSetSize_1 + RefSetSize_2$  solutions is obtained.

The usual *Subset Generation Method* in the applications of Scatter Search consists of considering all the subsets of a fixed size (usually two) of solutions in the current reference set of solutions. The solutions in the subsets are then combined to construct other solutions avoiding repetitions if the subset have been previously used in a combination. The *Solution Combination Method* combines good characteristics of the selected solutions to get new current solutions.

The possible combination methods for these problems are random/greedy strategies. They start with a partial solution consisting of the items common to the solutions to be combined. Then, at each iteration, one of the remaining items in some of the combined solutions is added. The criteria applied to select the items are between the pure random and greedy criteria and consist of selecting at random one of the most improving item.

The *Improvement Method* applied to the solutions of the population and those generated by the combination method are typical local searches. They are mostly based on the basic exchange method that consists in replacing an item in the solution by an item out of the solution. The solutions obtained by improving the combined solutions are used to update *RefSet* by the Reference Set Update Method.

The *Reference Set Update Method* also applies intensity and diversity criteria to update the reference set using the improved solutions. Using the strategy called *Static Update*, the improved solutions obtained after combination and improvement are recorded in a pool of solutions, *ImpSolSet*. The method selects the  $RefSetSize$  best solutions from  $RefSet \cup ImpSolSet$ . If a *Dynamic Update* strategy is used, the combination method would be applied to new solution faster

than in the static strategy. That is, instead of waiting until all the combinations have been performed to update the reference set, if a new solution is to be added to the reference set because it is better than the worst, this set is updated before the next subset of solutions combination is carried out.

## 4 Conclusions

The Scatter Search metaheuristic has been proved to be useful for the main standard tasks in Machine Learning for Data Mining: Clustering, Classification and Feature Selection. Future research will be oriented to use scatter search also for the instance pruning problem.

## References

1. H.A. Abbass, C.S. Newton, R. Sarker. *Data Mining: A heuristic Approach*. Idea Group (2002).
2. F. García López, M. García Torres, B. Melián Batista, J.A. Moreno Pérez and J.M. Moreno Vega. Solving Feature Subset Selection Problem by a Parallel Scatter Search, *European Journal of Operational Research*, 2005, to appear.
3. F. García López, B. Melián Batista, J.A. Moreno Pérez and J.M. Moreno Vega. Parallelization of the Scatter Search for the  $p$ -median problem, *Parallel Computing*, 29 (2003) 575-589.
4. Glover, F., Heuristics for Integer Programming using Surrogate Constraints, *Decision Sciences* 8, (1977) 156-166
5. Glover, F., Laguna, M., Martí, R. Fundamentals of Scatter Search and Path Relinking *Control and Cybernetics*, 39, (2000) 653-684
6. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley, (1989)
7. Ghosh, A. and Jain, L.C. (Eds.) *Evolutionary Computation in Data Mining Studies in Fuzziness and Soft Computing*, 163. Springer (2005)
8. Laguna, M. and R. Martí, *Scatter Search: Methodology and Implementations in C*, Kluwer Academic Press, (2003).
9. Mitchell, T. *Machine Learning*, Series in Computer Science, McGraw-Hill, (1997).
10. D. R. Wilson, T. R. Matinez, Improved heterogeneous distance functions, *Journal of Artificial Intelligence Research* 6 (1997) 1-34.