Evaluation of a heuristic search algorithm based on sampling and clustering

Maria Harita Alvaro Wong Dolores Rexachs and Emilio Luque

Computer Architecture and Operating System Department,
Universitat Autonoma de Barcelona, Barcelona, Spain.
maria.haritar@gmail.com, alvaro.wong@uab.es, dolores.rexachs@uab.es,
emilio.luque@uab.es

Abstract. Systems have evolved in such a way that today's parallel systems are capable of offering high capacity and better performance. The design of approaches seeking for the best set of parameters in the context of a high-performance execution is fundamental. Although complex, heuristic methods are strategies that deal with high-dimensional optimization problems. We are proposing to enhance the evaluation method of a baseline heuristic that uses sampling and clustering techniques to optimize a complex, large and dynamic system. To carry out our proposal we selected the benchmark test functions and perform a density-based analysis along with k-means to cluster into feasible regions, discarding the non-relevant areas. With this, we aim to avoid getting trapped in local minima. Ultimately, the recursive execution of our methodology will guarantee to obtain the best value, thus, getting closer to method validation without forgetting the future lines, e.g. its distributed parallel implementation. Preliminary results turned out to be satisfactory, having obtained a solution quality above 99%.

Keywords: Optimization, Heuristic methods, Clustering, Benchmark.

1 Introduction

As optimization problems become more complicated and extensive, parameterization becomes complex, resulting in a laborious, complicated task that requires a significant amount of time and resources, besides the fact that the number of possible solutions can become prohibitive in an exhaustive search. It is the reason why optimization algorithms play an important role in this transformation that usually attempt to characterize the type of search strategy through an improvement on simple local search algorithms [2]. In cases where the search space is large, metaheuristic ideas [1], which are sometimes classified global search algorithms, can often find good solutions with less computational effort. Some other approaches to achieve the optimization objectives are based on the extraction of data from probability distributions aiming for a reduction of the search space. Probabilistic distribution methods, such as Montecarlo offer flexible forms of approximation, with some advantages regarding cost. There are other approaches

that use similarity or metaheuristic algorithms to solve high-dimensional optimization problems which are validated using large-scale functions [4]. However, they are prone to fall into local optimum values. In order to solve global optimization problems, making use of global exploration there are, e.g. the naturally inspired approaches such as Genetic algorithms, Particle swarm, Grey Wolf or Ant Colony optimization algorithms [6].

Regarding the clustering methods, these are the techniques that group a series of objects, it is a pattern recognition technique which has a broad range of applications.Cluster analysis algorithms are a key element of exploratory data analysis used in e.g. data mining. Between the most widely used clustering algorithms is k-means. Here, a cluster is defined as a set of data characterized by a small distance to the cluster centers. Among the combinatorial optimization methods, for example [3], the efficiency gains regarding the application of sampling and grouping techniques are explored to solve a problem of a complex nature, because it is a dynamic and strongly human-dependent system.

In this paper, we propose an evaluation methodology of a heuristic method, based on sampling and clustering techniques comprising Montecarlo sampling, useful to obtain samples in multi-dimensional spaces, a density-based spatial analysis, and the k-means algorithm, crucial to determine and classify the data into feasible regions. For our evaluation proposal, we have selected the benchmark test functions [5] which allow testing algorithms and are useful when measuring relevant features, and can also be especially useful for understanding the algorithms applied to large-scale and continuous optimization problems. First we generate an initial sample of the benchmarks through the Montecarlo methods. Then, we apply an efficient clustering to locate feasible regions where the optimal solution might exist and can be found. The elements of the domain, which are known as candidate solutions, define such feasible regions. In our approach, we perform a density analysis to find patterns that will handle efficient grouping based on euclidian distances. Also, the recursive application of our proposal guarantees an almost optimal solution by reducing the search area, and enhancing the selection and grouping of feasible regions.

This proposal organizes as follows: the next Section presents our approach with an overview of the methodology, explaining definitions such as feasible regions, a justification of our proposal, the results obtained and the last Section discusses the open lines.

2 Proposal methodology and parameterization analysis

The design of optimization algorithms requires to make several decisions ranging from implementation details to the setting of parameter values for testing intermediate designs. Proper parameter setting can be crucial because a bad parameter setting can make an algorithm perform poorly. For our evaluation proposal, we are using the optimization benchmark functions along with an efficient grouping that performs spatial clustering by density in order to find patterns and/or variations in the landscape, in addition to k-means.

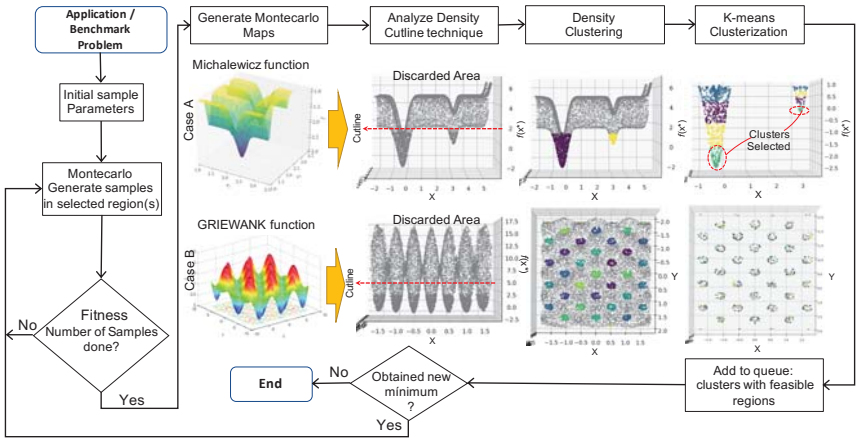

Fig. 1. Flowchart of our evaluation methodology.

As seen in Fig. 1, we generate an initial sample using the Montecarlo methods making successive iterations of the algorithm until it converges. This parameter will be linked to the precision within the range of the search space. If the sample is satisfactory, we generate the Montecarlo map, a landscape visualization that will help to analyze variations in the parameters, such as roughness. Since we are dealing with high-dimensional data with large variations in density because of the fixed global parameters (such as distance radius), we propose to make a cutline along the $f(x*)$ axis, which represents the value associated to each solution that is evaluated by computing the value of the objective function. It enables an efficient grouping by detecting arbitrary shapes, and automatically discovering the number of clusters through a density threshold allowing to find one cluster surrounded (although not connected) by another different cluster. It needs to have a notion of noise and be robust in detecting outliers. In the same Fig. 1 we can see that the data groups based on connected density objects, form different shapes and variations are detected. In this way, we will obtain the feasible regions, for which we will adapt the classical k-means algorithm locating the centroids along $f(x*)$. This type of grouping will be very useful to manage the dense areas, locating the clusters in which the best values are.

When the algorithm is able to find more than one feasible region, a queue forms to analyze clusters from each region, or the ones containing the best values, until the end of the queue resulting in selecting a single cluster. In this way, we ensure that we will not be trapped in a local minimum. If the single selected cluster contains the optimal (or near-optimal) value the simulation ends, otherwise, we return to the previous step of obtaining a new sample, analyzing density and grouping. It will execute recursively until finding the optimal value or until it is not possible to find a lower value than that obtained in the last iteration. From the preliminary results we obtained when testing the Michalewicz function, the problem size is in the range of $x* = (1.0000, 3.50000)$ and the sample size was of 69,700 which is less than 0.5%. The total search space

therefore reaches 6.2500E+08, and its optimal value $f(x*) = -1.8013$ located in $x* = (2.2000, 1.57000)$. Through our proposal, the best result we obtained was $f(x**) = -1.8012$ located at $x** = (2.2024, 1.5709)$. The quality of such solution was of 99.9944%, which is very promising, so we believe that further testing is necessary, as well as parallel systems exploration.

3 Discussion and Open Lines

The effectiveness of heuristic methods in dealing with challenging optimization problems is a widely studied field. Understanding the limitations of existing approaches and identifying areas for improvement contributes to evaluate a system, validate the method and allow its comparison to real-world problems. For our proposal, we selected the benchmark test functions to enhance a methodology that evaluates a heuristic based on sampling and clustering.

We are proposing a calibration of the parameters involved in the density-based analysis, as well as adapting the k-means clustering to select the feasible regions, creating a model that is based on the parameters of the best solution concerning the optimal value. The preliminary results that we obtained, gave a solution quality of 99.9944%, which encourages to expand the method.

Regarding the limitations about metaheuristic methods, one issue we may find has to do with the high-dimensionality of real problems, making it difficult to characterize. Still, it is a very useful tool when it is possible to achieve high precision in the evaluation phase. Nevertheless, there are some open lines that need to be explored, such as the increase in the dimensionality, which will increase the ranges and consequently the search space. To conclude, we believe that the extrapolation of combinatorial optimization techniques along with heuristics in distributed parallel systems is a step forward in the process that allows decision-making in real-time.

References

1. Bianchi, L., Dorigo, M., Gambardella, L.M., Gutjahr, W.: A survey on metaheuristics for stochastic combinatorial optimization. Natural Computing **8** (06 2009) 239–287
2. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. Siam Review **60**(2) (2018) 223–311
3. Cabrera, E., Taboada, M., Iglesias, M.L., Epelde, F., Luque, E.: Optimization of healthcare emergency departments by agent-based simulation. Procedia Computer Science **4** (2011) 1880 – 1889 Proceedings of the International Conference on Computational Science, ICCS 2011.
4. Eftimov, T., Korošec, P.: A novel statistical approach for comparing meta-heuristic stochastic optimization algorithms according to the distribution of solutions in the search space. Information Sciences **489** (2019) 255 – 273
5. Hussain, K., Salleh, M., Cheng, S., Naseem, R.: Common benchmark functions for metaheuristic evaluation: A review. International Journal on Informatics Visualization **1** (11 2017) 218–223
6. Vikhar, P.A.: Evolutionary algorithms: A critical review and its future prospects. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). (2016) 261–265