Chapter 3
A Comparative Study Among Recursive Metaheuristics for Gene Selection

Nassima Dif
https://orcid.org/0000-0002-8683-3163
EEDIS Laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria

Zakaria Elberrichi
https://orcid.org/0000-0002-3391-6280
EEDIS Laboraory, Djillali Liabes University, Sidi Bel Abbes, Algeria

ABSTRACT

This chapter compares 4 variants of metaheuristics (RFA, EMVO, RPSO, and RBAT). The purpose is to test the impact of refinement on different types of metaheuristics (FA, MVO, PSO, and BAT). The refinement helps to enhance exploitation and to speed up the search process in multidimensional spaces. Moreover, it presents a powerful tool to solve different issues such as slow convergence. The different methods have been used for gene selection on 11 microarrays datasets to solve their various issues related to the presence of irrelevant genes. The obtained results reveal the positive impact of refinement on FA, MVO, and PSO, where all performances have been improved. On the other hand, this process harmed the BAT algorithm. The comparative study between the 4 variants highlights the efficiency of EMVO and FA in terms of precision and dimensionality reduction, respectively. Overall, this study suggests drawing attention to the choice of embedded metaheuristics in the refinement procedure, where powerful methods in exploration are recommended. Moreover, metaheuristics that risk form fast convergence are not advised.

DOI: 10.4018/979-8-3693-3026-5.ch003

1. INTRODUCTION

The DNA microarrays technology helps researchers to measure the expression level of thousands of genes(Harrington et al., 2000). Cancer identification is among the most important applications in the microarrays field (Almugren, & Alshamlan, 2019). The extracted biomarkers assist in diagnosis, prognosis, and treatment (Baliarsingh et al., 2019). In such applications, machine learning (ML) methods are exploited to analyze the generated biomedical datasets and to extract meaningful knowledge. However, these datasets suffer from the curse of dimensionality and the presence of redundant and irrelevant genes, which can result false diagnoses because of the presence of indiscriminate features. This issue becomes crucial especially for some machine learning algorithms that don't perform feature selection during training. Moreover, these datasets can result overfitting during training because of the large difference between the number of genes and samples. Thus to handle these volumes effectively, preprocessing techniques such as gene selection have been largely exploited.

The feature selection process set out to select M relevant subset of features from the initial set N (M <= N). The purpose of this method is to reduce the computational complexity of the ML algorithm and to enhance its precision.

Feature (gene) selection techniques are categorized into four strategies: filters(Hancer et al., 2018), wrappers (Jiang et al., 2019), embedded (Zhu et al., 2007) and hybrid methods (Alomari et al., 2018). Filters are based on statistic methods to evaluate the selected set. Whereas, wrappers depend on the performance of the machine learning algorithm, where a training step is required for each subset, which makes them computationally expensive compared to filters but more accurate (Inza et al., 2004). To take advantage of these two methods, hybrid approaches between filters and wrappers are proposed. In general, a filter strategy is performed first to reduce the high-dimensional space, and then the wrapper method is applied to the obtained result to select effectively the relevant subset. Last, embedded methods are characterized by their embedded future selection process within the training process.

For feature selection, first, a subset generation process is performed to generate the candidate subsets. Then, theses subsets are evaluated according to the cited strategies above. In exact methods, the generation step generates 2^N subset for N features in the initial set, and then these subsets are evaluated to select finally the best one among all possibilities. This procedure is computationally expensive, especially for high-dimensional datasets. As a solution, stochastic methods have been largely exploited such as probabilistic (Roffo et al., 2017), heuristic (Min, & Xu, 2016) and metaheuristic (Gu et al., 2018) strategies.

Metaheuristics are stochastic methods that have received considerable attention in several optimization problems: feature selection (Gu et al., 2018), classification (Abd-Alsabour, & Ramakrishnan, 2016), parameters and hyper-parameters optimization (Rojas-Delgado et al., 2019). The main purpose of these methods is to find good solutions in a reasonable run time complexity compared to exact methods. We should note that metaheuristics do not guarantee to find the best global solution as exact methods but it helps to find good solutions close to the best global one. Traditionally, genetic algorithm (GA) (Holland et al., 1992), particle swarm optimization (PSO) (Eberhart, & Kennedy, 1995) and ant colony optimization (ACO) (Dorigo et al., 1996) metaheuristics have been used in various classic studies to solve different NP-hard problems. Lately, there has been a large amount of publication on recent metaheuristics such as: cuckoo search (CS) (Yang, & Deb, 2009), firefly algorithm (FA) (Yang, 2009), bat algorithm (BAT) (Yang, 2010) and seven-spot ladybird optimization (SLO) (Wang et al., 2013).

These methods can be categorized into two groups: evolutionary and swarm intelligence algorithms (Fister et al., 2013). The first category points out on the exploration of the search space due to the evo-