Comparison of clustering optimization for classification with PSO algorithms

Klemen Berkovič
University of Maribor, Faculty of Electrical Engineering and Computer Science
Koroška cesta 46
Maribor, Slovenia
klemen.berkovic@student.um.si

Uroš Mlakar
University of Maribor, Faculty of Electrical Engineering and Computer Science
Koroška cesta 46
Maribor, Slovenia
uros.mlakar@um.si

Borko Bošković
University of Maribor, Faculty of Electrical Engineering and Computer Science
Koroška cesta 46
Maribor, Slovenia
borko.boskovic@um.si

Iztok Fister
University of Maribor, Faculty of Electrical Engineering and Computer Science
Koroška cesta 46
Maribor, Slovenia
iztok.fister@um.si

Janez Brest
University of Maribor, Faculty of Electrical Engineering and Computer Science
Koroška cesta 46
Maribor, Slovenia
janez.brest@um.si

ABSTRACT

In this paper, we compare Particle Swarm Optimization (PSO) algorithms for classification based on clustering. Clustering is presented within the proposed PSO algorithms as an optimization problem, and captured in a so called fitness function. Clusters are presented with centers that should represent the classes hidden in data. With the help of PSO algorithms and proper fitness function, we optimize centers of the clusters. Because clustering belongs to a field of unsupervised learning methods, we redefine the clustering problem to supervised learning with a new fitness function that helps PSO algorithms in finding clusters centers that can classify instances of data to the right class. Two experiments are performed in our study. In the former, various fitness functions for classification based on clustering are tested. The latter measures the performance of PSO algorithms using the best fitness function from the first experiment. For testing the PSO algorithms, we use three different datasets. Friedman non-parametric statistical test, and Nemenyi and Wilcoxon post-hoc tests are conducted to properly compare the PSO algorithms.

Categories and Subject Descriptors

I.5.3 [**Pattern recognition**]: Clustering; G.1.6 [**Numerical analysis**]: Optimization

Keywords

clustering optimization, classification, particle swarm optimization, statistical comparison

1. INTRODUCTION

In many fields, including physics, bioinformatics, engineering and economics, we can encounter problems, where from a plethora of solutions we are interested in finding the most suitable. These so called optimization problems (OPs) can be sub-categorized into: discrete, continuous, and/or mixed-variable problems [18]. The difference between three mentioned groups of OPs are the type of variables in their solution space. Discrete problems work with discrete variables that are elements of $\mathbb{N}^+$ set, continuous problems have variables with elements of $\mathbb{R}$ set, while the mixed-variable problems can capture variables of both aforementioned sets.

Since we are usually limited by various computational resources (e.g., computational power and space, problem's constraints), the complexity of OPs are increased and therefore we are forced to be satisfied with a "good enough" solutions for the real-world application. Definition of a "good enough" depends exclusively on the type of a problem. Majority of the OPs have a huge, yet finite solution spaces, which can be solved approximately using meta-heuristics. In our work we are focused on meta-heuristics algorithms that are inspired by nature and are a part of Swarm Intelligence (SI) algorithms [4]. From the field of SI algorithms, we use be using PSO algorithms.

In this work, we are focused on minimization of single-objective OPs that can be defined mathematically, as follows:

$$f\left(\mathbf{x}^*\right) \leq \min_{\mathbf{x}\in\Omega} f\left(\mathbf{x}\right), \tag{1}$$

where $f : \mathbb{R}^D \mapsto \mathbb{R}$ represents the fitness function of the problem to be solved, $\mathbf{x}$ is D dimensional vector consisting of problem variables and $\mathbf{x}^*$ is a global minimum or the best solution of the problem. The fitness function f is realized as a mathematical formula or a set of empirical data that refer to the problem of interest. In order to make the problem simpler, we demand that the solutions should be in some

DOI: https://doi.org/10.26493/978-961-7055-82-5.35-42