Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07836-8

ORIGINAL ARTICLE

An Efficient High-dimensional Feature Selection Approach Driven By Enhanced Multi-strategy Grey Wolf Optimizer for Biological Data Classification

Majdi Mafarja[1] · Thaer Thaher[2,3] · Jingwei Too[4] · Hamouda Chantar[5] · Hamza Turabieh[6] · Essam H. Houssein[7] · Marwa M. Emam[7]

Received: 6 March 2022 / Accepted: 12 September 2022

Abstract
Biological data generally contain complex and high-dimensional samples. In addition, the number of samples in biological datasets is much fewer than the number of features, so the vast number of features should be selected carefully and determine the optimal subset of features. Feature selection (FS) is a vital stage in biological data mining applications (e.g., classification) for dealing with the curse of dimensionality problems and finding highly informative features. This work proposes an effective FS approach based on a new version of Gray Wolf Optimizer (GWO) called Multi-strategy Gray Wolf Optimizer (MSGWO) for better features selection for biological data classification. The use of MSGWO in feature selection is to find the optimal subset of features between classes, solve premature convergence, and enhance the local search ability of the GWO algorithm. Multiple exploration and exploitation strategies are proposed to enhance the global search and local search abilities of the GWO algorithm through the optimization process. The support vector machine (SVM) classifier is used to evaluate the proposed GWO-based FS approaches. MSGWO was evaluated on thirteen high-dimensional biological datasets obtained from the UCI repository with a smaller number of instances. The reported results confirm that employing multiple exploration and multiple exploitation strategies is highly useful for enhancing the search tendency of the MSGWO in the FS domain. Statistical tests proved that the superiority of the proposed approach is statistically significant as compared to the basic GWO and similar wrapper-based FS techniques, including binary particle swarm optimization (BPSO), binary bat algorithm (BBA), binary gravitational search algorithm (BGSA), and binary whale optimization algorithm (BWOA). In terms of classification accuracy, MSGWO yielded better accuracy rates than the standard GWO algorithm on 84% of applied biological datasets. MSGWO also recorded better accuracy rates than its other competitors in all 13 cases. In terms of the lowest number of selected features, MSGWO yielded excellent reduction rates compared to its peers.

Keywords Feature selection · Binary grey wolf optimizer · Classification · Meta-heuristics · Biological data

1 Introduction

The continuous growth of the internet and the advanced information systems have resulted in a huge amount of data in all fields [1, 2]. For example, in the biological areas, an excessive number of biological features are normally collected from the data acquisition stage. As such, finding useful information from biological data becomes challenging and time-consuming [3]. Moreover, the performance of the learning algorithm is highly affected by the size (dimensionality) of the datasets. That is to say, having a dataset with a large number of features may negatively affect the performance. Therefore, a pre-processing step, such as the dimensionality reduction, is needed to deal with such datasets in order to reduce their dimensionality, and hence enhance the learning process [4–6].

Studies and technologies for understanding complex biological systems have enabled the generation of biological data in different dimensions such as Deoxyribonucleic acid (DNA) sequence [7], single-cell gene expression, epigenomic states activity, and proteomics [8]. When developing machine learning techniques to classify

Extended author information available on the last page of the article

Published online: 04 October 2022

biological and biomedical data, many difficulties appear. In addition to large sizes of biological and medical datasets (e.g., thousands of features), there are other deep-rooted complexities such as heterogeneous, incomplete, imbalanced, dynamic, and noisy [8, 9]. Furthermore, the number of acquired samples in biological studies is mostly limited and very few in contrast with the number of variables (features). This may lead to problems such as sparse of data and over-fitting [9, 10]. In biological data, there is complicated reliance among features such as relevance and redundancy [3] [11]. In addition, physiological and pathological changes in complex biological systems are expressed via the interaction between features such as genes and metabolites [3]. Along with the challenges mentioned above, features dependencies and relevance between features and category labels are two important aspects that must be considered when developing FS approaches for biological data classification. Feature Selection (FS) is a preprocessing technique that aims to reduce the dimensionality of a dataset by eliminating the unnecessary (irrelevant and redundant) features in order to improve the performance of the learning process. This also leads to increasing the learning accuracy, simplifying the learning results, and reducing the computational time of the learning process [12]. FS in biological data has a great role in avoiding, diagnosis, and treatment of several diseases [3].

The key to the FS approach is to reduce the number of features by eliminating redundant features from the sample data, thereby selecting the most effective features in the sample data. This can simplify model inference and improve the accuracy of the classification model. In general, methods for selecting the optimal subset of features can be divided into three categories, namely filtering methods [13], wrapper methods [14] and embedded methods [15]. The filtering method sets thresholds to initially filter the dataset and then trains the model. It does not use any learning algorithms to evaluate the relevance of features, which makes classification fast enough, but its classification accuracy is dependent on the setting of thresholds. In contrast, the wrapper approach finds the best solution by formulating an objective function and using iterative search to continuously add or subtract features. This model involves a learning model in the iterative process and tends to consume more computational time. The first two feature selection methods are not directly related to the algorithm chosen, while the embedded methods are a combination of filter and wrapper methods and are related to the specific optimization algorithm. An example of embedded FS is the ridge regression, which has penalization inbuilt. FS is considered an NP-hard problem with many possible solutions, especially for a large feature space [15]. Some embedded methods perform feature weighting based on regularization models with objective functions that minimize fitting errors and in the mean time force the feature coefficients to be small or to be exactly zero. To be specific, the feature selection that is going to be used in this study is going to be the wrapper-based one.

In the wrapper FS methods, the evaluation criterion is the performance of the learning algorithm (e.g., classification accuracy) in addition to the number of selected features. The aim of those methods is to produce a feature subset with the maximum classification accuracy and the minimum number of features. In comparison with the filter approaches, wrappers often obtain higher accuracy rates. However, the selected subset, using a specific classifier, may not be suitable for a different classifier [16]. In addition, involving a learning algorithm in the FS process increases the computational time. Thus, the selection of the FS model depends on the target from reducing a dataset. If the problem that has to be solved needs to make feature selection for a general purpose, and not for a specific learning technique, with low computational time, a filter approach can be used. If one is seeking to make feature selection for a general purpose, and not for a specific learning technique, with low computational time, a filter approach can be used. If the target is to enhance the performance of a specific learning algorithm, however, wrappers are more desirable [17].

In this work, the main target is to develop an efficient wrapper FS approach for enhancing the performance of the support vector machine (SVM) classifier. SVM classifier has been chosen based on initial classification results obtained from testing several machine learning classifiers for classifying thirteen high dimensions biological datasets with different feature sizes. Another criterion that is used to classify FS methods is the searching mechanism for the best subset of features [5]. One possible strategy is to generate all possible feature subsets exhaustively. This approach is impractical when dealing with high-dimensional datasets since the search space is extremely big [18]. Generally speaking, if a dataset contains M features, $2^M - 1$ feature subsets should be explored and evaluated to select only one subset. To combat these issues, random search strategies can be used to only search promising regions of the search space. This strategy may find the best subset according to the evaluation criterion, but in the worst case, it may behave like an exhaustive search strategy.

Heuristics and Metaheuristics (MHs) are other popular methods [19], which are stochastic and able to find the (near) optimal solutions for an optimization problem in a reasonable time [20]. The general idea behind MHs is that they generate a random (single) solution or a set of random solutions (the so-called population). They then improve that solution/population iteratively until the satisfaction of