Best Features Selection for Biomedical Data Classification Using Seven Spot Ladybird Optimization Algorithm

Noria Bidi, Department of Science and Technology, University Mustapha Stamboli, Mascara, Algeria
Zakaria Elberrichi, Department of Computer Science, University Djillali Liabes, Sidi Bel Abbes, Algeria

ABSTRACT

This article presents a new adaptive algorithm called FS-SLOA (Feature Selection-Seven Spot Ladybird Optimization Algorithm) which is a meta-heuristic feature selection method based on the foraging behavior of a seven spot ladybird. The new efficient technique has been applied to find the best subset features, which achieves the highest accuracy in classification using three classifiers: the Naive Bayes (NB), the Nearest Neighbors (KNN) and the Support Vector Machine (SVM). The authors’ proposed approach has been experimented on four well-known benchmark datasets (Wisconsin Breast cancer, Pima Diabetes, Mammographic Mass, and Dermatology datasets) taken from the UCI machine learning repository. Experimental results prove that the classification accuracy of FS-SLOA is the best performing for different datasets.

KEYWORDS

Biomedical Dataset, Classification, Feature Selection, Meta-Heuristic, Naive Bayes, Nearest Neighbors, Seven Spot Ladybird Optimization Algorithm, Support Vectors Machines

1. INTRODUCTION

A diverse set of data mining algorithms have been used to extract useful information from biomedical data. The most important technique in biomedical data mining is the classification; it plays a significant role in the diagnosis of disease. In a classification problem, the objective is to learn the decision surface that accurately maps an input feature space to an output space of class labels (Babu & Suresh, 2013) with excluding of redundant and irrelevant features which can improve the performance of classification algorithm. Many feature selection methods exist in literature because there are several statements of the feature selection problem, some methods are specialized to particular cases and other methods can be used only for small numbers of features or only for large numbers of features. Among these algorithms; someones used meta-heuristics strategies to find the best feature subset.

Meta-heuristics have been successfully proposed to solve difficult optimization problems and have proved their effectiveness in wide areas of applications. These methods are often bio-inspired and can be classified into trajectory-based methods and population-based methods according to the operating principle during the search of the solution.

SLOA is a novel population based meta-heuristic approach developed by Peng Wang, Zhouquan Zhu and Shuai Huang in 2013 (Wang, Zhu & Huang, 2013). It is inspired from the seven-spot ladybird’s foraging behavior. All nature inspired meta-heuristic algorithms have something in common...
in the sense that they are intensification and diversification. The diversification strategy of SLOA is presented by exploring more effectively the space of solutions, while the intensification strategy is implicitly controlled by converging rapidly to an optimal solution.

With the aim of building an accurate classification model, an optimization algorithm (FS-SLOA) is proposed. This wrapper model approach uses the classification accuracy to evaluate the performance of a selected feature subset. It uses three classifiers: Naive Bayes (NB), Nearest Neighbors (KNN) and Support Vector Machines (SVM) classifiers for fitness evaluation. To demonstrate the efficiency of the proposed method, it is tested on four well-known benchmark datasets (Wisconsin Breast cancer, Pima Diabetes, Mammographic Mass, and Dermatology datasets) which are taken from the UCI machine learning repository. The results obtained are compared with previous methods reported in the literature and found to be encouraging.

The paper is organized as follows: Brief review of the related work in using machine learning techniques is discussed in section 2. Section 3 presents feature selection methods. The FS-SLOA method is presented in section 4 and section 5. The selected classifiers are presented in section 6. In section 7, our experimental results for the implementation of FS-SLOA on different datasets are reported. Finally, we conclude this work.

2. RELATED WORK

In order to obtain better classification accuracy for the diagnosis of diseases with the Wisconsin Breast Cancer, Pima Diabetes, Mammographic Mass, and Dermatology datasets, many machine learning algorithms have been successfully applied. In our research work, we are interested in the use of meta-heuristic for feature selection to improve the performance of classification algorithm. In this field, many researchers have developed various approaches.

Firstly, we cite several studies on Wisconsin breast cancer dataset, in (Sweilam, Tharwat & Abdel Moniem, 2010) particle swarm optimization, Quantum-behave Particle Swarm, least square support vector machine (LSSVM) and active set strategy are introduced. The effectiveness of the four different methods for training support vector machine will be evaluated and compared, the SVM-PSO shows the highest accuracy of 93, 06% on a Wisconsin breast cancer dataset. In (Prasad, Biswas & Jain, 2010) and based on SVM classifier, heuristics methods, including genetic algorithm (GA), ant colony optimization (ACO) and particle swarm optimization (PSO), were used to select the best feature subset for SVM training rather than using the exhaustive search. Those methods used the training accuracy as a criterion to evaluate different feature combinations. In their research, the results reduced the computation time for the training significantly and showed that the GA-SVM was better than the others. (Kaur & Kalra, 2016) proposed a model where feature selection is used to pre-process the data by selecting the most informative features. These selected best features obtaining by F-Score, Genetic Algorithm, K-means, ReliefF, and SVM-RFE are used by SVM classifier for classifying WDBC dataset. In this research, detection of cancer with optimal cost and better performance is done with high accuracy. It has been analyzed that the performance of SVM is better enhanced by using SVM-RFE (Support Vector Machine- Recursive Feature Elimination) which is an embedded approach that recursively removes unimportant features SVM-RFE. In (Afzan, 2012) GAwNN model was developed by combining genetic algorithm and Back-propagation neural network. In the aim to reduce the diagnosis time as well as increasing the accuracy in classifying WDBC, two different cleaning processes were carried out. The first eliminated records with missing values and the second trained with normal statistical cleaning process to identify any noisy or missing values. The author has found that medical data are best kept in its original value as it gives high accuracy percentage as compared to altered data. In (El-Hasnony, El-Bakry and al) an effective hybrid system combines K-means clustering algorithm, fuzzy rough feature selection (FRFS), and discernibility k nearest neighbor (D-KNN) classifier for breast cancer classification is presented. The proposed system achieves accuracy up to 98.9%.
Metaheuristic Algorithms for Supply Chain Management Problems
www.igi-global.com/chapter/metaheuristic-algorithms-supply-chain-management/69883?camid=4v1a

An Efficient Adaptive Strategy for Melody Search Algorithm
www.igi-global.com/article/an-efficient-adaptive-strategy-for-melody-search-algorithm/129009?camid=4v1a