State-of-the-Art Study on Search Algorithms

“ our title over here”

Put names and formatting

# Chapter 1

## Introduction To the Project

In the realm of real-world problem-solving, a common challenge emerges: the management of vast datasets with numerous attributes or features. Negotiating this data labyrinth can be an intricate task. Within these datasets lie attributes, some of which may be redundant, or irrelevant, all of which can significantly impede the performance of machine learning models.

Dimensionality too poses issues, potentially leading to overfitting and prolonged learning times. The crux of the matter is to reduce the dimensionality of these datasets while retaining their predictive accuracy. This is where the concept of feature reduction comes into play.

Feature reduction, branches into two key facets: feature construction and feature selection. Feature construction entails the creation of novel features derived from the original dataset, while feature selection involves the process of picking the most pertinent attributes from the dataset. We mainly focus on feature selection.

When dealing with a dataset containing 'n' features, 2^n subsets become potentially relevant. As 'n' grows larger, assessing the performance of the model for each subset becomes impractical. To mitigate this, various methodologies have been introduced. Exhaustive search, greedy search, random search, and more have been deployed to tackle feature selection challenges. However, many of these methods suffer from issues such as premature convergence, unnecessary complexity, and high computational costs. Consequently, the spotlight has shifted towards metaheuristic algorithms as a means of addressing these challenges.

These algorithms offer efficiency and effectiveness, capable of finding optimal feature subsets while preserving accuracy. Pruning features via these algorithms mitigates overfitting, reduces vulnerability to missing data, and facilitates better model interpretation and generalization.

For instance, in image analysis, feature selection helps discern the salient visual components, such as pixels and colours.

Mathematically, the feature selection problem can be framed as :

Given: dataset 'S' containing 'd' features.

Goal: to select the most relevant features.,

Consider dataset S = {f1, f2, …, fd}, the objective is to extract a subset D = {f1, f2, ..., fn},

where n < d, and f1, f2, f3, …, fn represent the attributes of the dataset.

Feature extraction for images encompasses the derivation of higher-level features from raw data, such as edges, corners, or image segments. Some lower-level primary features include size, colour values, intensity.

Yet, the inherent challenge in image datasets lies in their vastness, as they often comprise thousands of features due to the multitude of pixels. The quality of an image is directly correlated with the number of pixels, which, in turn, amplifies computational complexity.

Feature selection for images can be achieved through wrapper methods, filters, and embeddings. Wrapper methods evaluate feature subsets by training models, making them suitable when labelled data is available. Filters, on the other hand, select features based on statistical criteria or domain knowledge, making them computationally efficient.

Embeddings transform high-dimensional image data into lower dimensions, often using techniques like PCA (Principal Component Analysis) or other embeddings. The choice depends on the task, dataset size, and computational resources, with a combination of methods often yielding the best results. Pre-trained networks are commonly used for extracting image embeddings, offering a powerful feature selection approach.

Given the size of the feature space, feature selection can be framed as an optimization problem, thereby making it an ideal candidate for the application of metaheuristic-based optimizers. These algorithms have demonstrated their prowess in solving complex optimization challenges.

A growing area of research pertains to the use of image classification for disease detection, particularly through X-ray and CT-scan images. Over the past few years, metaheuristic algorithms have gained traction for optimizing deep neural network architectures. Their simplicity, flexibility, and problem-agnostic nature render them ideal candidates for tackling intricate optimization problems.

Metaheuristic algorithms constitute a class of optimization methods that excel at seeking optimal or near-optimal solutions for a variety of optimization problems. These algorithms are notable for their lack of reliance on derivatives and their stochastic nature. They initiate their optimization process by generating random solutions, obviating the need for derivative calculations typical of gradient-based techniques. What sets metaheuristic algorithms apart is their simplicity and flexibility, making them open to customization for specific problems. A defining characteristic of these algorithms is their ability to avoid premature convergence, a pitfall often associated with optimization. Their stochastic nature treats the problem as a black box, enabling them to explore the search space thoroughly and evade local optima effectively. Metaheuristic algorithms strike a delicate balance between exploration and exploitation, investigating promising search spaces before delving into local searches in the regions identified during exploration.

To delve into the metaheuristic algorithms applied in feature selection, we delve into the realm of binary vector representations. In this paradigm, each solution vector is represented as a sequence of binary values, such as (10101100...). Here, a '1' indicates the selection of a particular feature, while '0' signifies its exclusion from the subset. This binary representation forms the foundation for metaheuristic algorithms striving to identify optimal feature subsets.

# Chapter 2

## Literature Survey

|  |  |  |  |
| --- | --- | --- | --- |
| SNo | TITLE | METHOD/DESCRIPTION | RESULTS |
| [1] | An Effective Feature Selection Scheme for Healthcare Data Classification Using Binary Particle Swarm Optimization  Conference Proceedings:  2018 9th International Conference on Information Technology in Medicine and Education (ITME)  Author: Yiyuan Chen  Publisher: IEEE  Date: October 2018 | This research paper introduces an innovative approach for feature selection called CCFS (Confidence-based and Cost-effective Feature Selection). CCFS leverages binary particle swarm optimization to enhance the efficiency of feature selection while minimizing costs.  To achieve this, CCFS introduces a novel update mechanism that explicitly considers the confidence of each feature. This confidence calculation takes into account the correlation between each feature and the categories of interest, as well as the historical frequency of selection for each feature.  Furthermore, the design of the fitness function in CCFS is comprehensive, taking into account multiple criteria .It considers not only classification accuracy but also the reduction ratio of features and the associated feature costs.  To validate the effectiveness of CCFS, experiments were conducted using the UCI cancer classification dataset <http://archive.ics.uci.edu/ml/datasets/Lung+Cancer> , specifically focusing on lung cancer classification. The results of these experiments demonstrate the efficacy of the proposed method in improving feature selection for this critical task. | With 30 individuals in the population and a termination after 100 iterations, CCFS achieved an impressive 84.375% classification accuracy at the 87th iteration. This outperformed ACC\_BPSO by 3.125% and reached this peak accuracy four iterations earlier, all while maintaining a similar total feature cost.  Incorporating feature confidence values during position updates not only improved classification accuracy but also accelerated convergence to the global optimum, avoiding local maxima.  By considering both feature count and cost in the fitness function, CCFS enhanced feature selection efficiency, resulting in more accurate and cost-effective feature subsets. |
| [2] | [Computers in Biology and Medicine](https://www.sciencedirect.com/journal/computers-in-biology-and-medicine) [Volume 144](https://www.sciencedirect.com/journal/computers-in-biology-and-medicine/vol/144/suppl/C), May 2022, 105344 Metaheuristics based COVID-19 detection using medical images: A review Authors:Mamoona Riaz, Maryam Bashir, Irfan Younas | In this paper 23 publicly available separate datasets are used to analyse how metaheuristics can be applied at different steps in the classification of medical images for COVID-19.  It uses chest images with various preprocessing methods such as Normalization techniques including MinMax scaling, image resizing, and Gray scaling, while augmentation methods involve rotation, translation, SMOTE.  Mamoona et Al. compare accuracy of classification when features are selected using versions of nature inspired metaheuristic search algorithms such as Whale Optimization Algorithm (WOA),Cuckoo Search algorithm (CSA), Salp swarm algorithm (SSA),Marine predators algorithm (MPA). | The paper concludes that many of these algorithms increase classification accuracy for these datasets.  WOA -- 99.22%.  CSA -- 96.72%  SSA -- 95.91%  FO-MPA -- 99.80%  The limitations encompass challenges related to dataset size, class imbalance, noisy data, the time complexity of metaheuristics.  The authors interestingly also noted the need for interdisciplinary collaboration to make significant progress. |
| [3] | Feature Selection using Artificial Bee Colony Algorithm for Medical Image Classification  Authors: Vartika Agrawal , Satish Chandra  **Published in:**[2015 Eighth International Conference on Contemporary Computing (IC3)](https://ieeexplore.ieee.org/xpl/conhome/7337021/proceeding).  **Date of Conference:**20-22 August 2015 | This study employs the Artificial Bee Colony (ABC) metaheuristic algorithm to conduct feature selection on Computed Tomography (CT Scan) images of cervical cancer. Data set (CT-Scan images of cervical cancer) are 271 in number with 108 instances as negative and 163 as positive. Images were collected from various diagnostic centres of Delhi and Noida. The primary objective is to discern whether the input data indicates the presence of cancer. The process begins with image segmentation, achieved through the implementation of the Active Contour Segmentation (ACM) algorithm. A semi-automated system is devised to extract the Region of Interest (ROI) from the segmented images. Subsequently, textural features are extracted from the ROI.  The algorithm draws inspiration from the intelligent foraging behaviour of honey bees and is structured around three essential components: food sources, employed bees, and unemployed bees. This model of forage selection harnesses the collective intelligence of bees to converge on solutions.  One distinctive feature of ABC is its ability to address both local and global search aspects. Employed and onlooker bees handle local search, while scout bees manage global search, which balances exploration and exploitation tasks. | For classification, a hybrid approach combines ABC with both k-Nearest Neighbours (k-NN) and Support Vector Machine (SVM) algorithms.  Notably, the results indicate that the fusion of ABC with SVM (using a Gaussian kernel) outperforms combinations involving ABC with an SVM that employs a linear kernel and ABC with a k-NN classifier.  In the case of k-NN, it achieved 97% accuracy with a biased dataset and 100% accuracy with an unbiased dataset. SVM with a linear kernel achieved 93% accuracy with a biased dataset and 99% accuracy with a radial basis kernel. Consequently, SVM with a radial basis kernel emerged as the optimal choice for classification, although dataset bias significantly impacted results. |
| [4] | AHA-AO: Artificial Hummingbird Algorithm with Aquila Optimization for Efficient Feature Selection in Medical Image Classification Author: Mohamed Abd Elaziz  Published: 27th September 2022 in Special Issue "The Applications of Machine Learning in Biomedical Science" | This paper introduces a novel system for medical image diagnosis that combines transfer learning (TL) with a new feature selection algorithm, the Artificial Hummingbird Algorithm based on Aquila Optimization (AHA-AO).  This algorithm's unique approach draws inspiration from the precision and efficiency of hummingbirds in selecting nectar from flowers. AHA-AO demonstrates a remarkable ability to sift through complex image data and select only the critical features essential for accurate diagnosis.  It works by iteratively selecting, combining, and evaluating subsets of features, it uses a fitness function to measure the contribution of each feature subset to the performance of a deep learning model for medical diagnosis. AHA-AO continuously refines these subsets through selection, crossover, and mutation processes until it identifies the most effective feature subset.  Its main benefit is the major improvement in efficiency while not compromising in accuracy. | The approach was evaluated on four datasets (ISIC-2016, PH2, Chest-XRay, Blood-Cell) and compared it to five other feature selection algorithms. The AHA-AO achieved superior results, with accuracy percentages of 87.30%, 97.50%, 86.90%, and 88.60% for the respective datasets. Importantly, AHA-AO also demonstrated faster feature selection compared to other methods, successfully improving both the performance and efficiency of deep learning models for medical image diagnosis. |
| [5] | An Efficient Moth Flame Optimization Algorithm using Chaotic Maps for Feature Selection in the Medical Applications Authors: Ruba Abu Khurma, Ibrahim Aljarah and Ahmad Sharieh  Published in: Conference: 9th International Conference on Pattern Recognition Applications and Methods | This paper introduces and compares multiple variants of the Binary Moth Flame Optimization Algorithm (BMFO) incorporating chaotic maps as search strategies within a wrapper feature selection framework. Chaotic maps are employed primarily to enhance the initialization of solutions, aiding the optimizer in escaping local minima and converging towards global optima. These approaches are novel as they are applied for the first time to feature selection (FS) problems, which are crucial for mitigating dimensionality issues impacting learning processes, including data overfitting and prolonged learning times.  Among these, the Moth Flame Optimization (MFO) algorithm stands out as an effective choice for solving various optimization problems across diverse applications.  The proposed approaches are rigorously tested on 23 medical datasets achieving competitive results when benchmarked against other state-of-the-art metaheuristic algorithms. | The paper compares with three established metaheuristic algorithms: Binary Grey Wolf Optimization (BGWO), Binary Cuckoo Search (BCS), and Binary Bat Algorithm (BBA).  Results showed that chaotic operators improved BMFO's classification accuracy.  BBA was the top-performing method in 78% of datasets, followed by BCS at 22%. BMFO variants didn't surpass these methods.  BBA and BCS were strong performers, while BMFO variants showed effectiveness across various datasets. |
| [6] | Feature Selection in Life Science Classification: Metaheuristic Swarm Search Authors: Simon Fong, Suash Deb, Xin-She Yang, Jinyan Li  Published in: IT Professional ( Volume: 16, Issue: 4, July-Aug. 2014) | The SS(swarm search)-FS framework is a versatile design compatible with various metaheuristic and classification algorithms. It serves as a wrapper-based model, using accuracy as the fitness function to evaluate feature subset candidates for a given classifier. The workflow begins with an initial random feature subset selection and refines the classification model's accuracy through stochastic searches, aiming to converge between the selected feature subset and the classifier.  The classifier acts as a black-box evaluator, assessing feature subsets based on accuracy. The optimization function supports different metaheuristics, focusing on finding the optimal feature subset for maximum classification accuracy. The default brute-force approach tests all possible subsets, but this is computationally intensive, especially with many features. To address this, we use stochastic search strategies like swarm-based metaheuristics with parallel search agents, seeking the optimal feature subset without exhaustive exploration.  However, finding the absolute global best solution may be computationally infeasible, as Xin-She Yang pointed out. In their experiments, they combined three metaheuristics (PSO, BAT, WSA) with three popular classification algorithms (PN, DT, NB) to create diverse SS-FS methodologies. | SS-FS methods consistently outperformed other methods in our evaluations, except when paired with Naive Bayes. For instance, on the Arrhythmia dataset, FS-Cfs reduced the error rate from 40% to 30%, while non-FS methods had around 10% error.  FS-SS methods achieved high accuracy but had longer computation times. In contrast, FS-Cfs was computationally efficient, taking less than 0.5 seconds in most cases. FS-PSO-PN and FS-BAT-PN achieved almost zero error rates with FS-SS, while their FS-Cfs counterparts had error rates between 10% to 30%.  FS-SS methods retained about half of the original features for dimensionality reduction, making them valuable for high-dimensional biomedical datasets. |
| [7] | Hybrid Whale Optimization Algorithm with simulated annealing for feature selection  Authors: Majdi M. Mafarja, Seyedali Mirjalili  Published: Neurocomputing Volume 260 | Recent developments in optimization and memetic algorithms have sparked significant interest in hybrid metaheuristics. This paper explores two hybridization models that leverage the Whale Optimization Algorithm (WOA) to create distinct feature selection techniques.  In the first model, the Simulated Annealing (SA) algorithm is integrated into the WOA algorithm, while in the second model, SA is utilized to refine the best solution discovered after each iteration of the WOA algorithm. The primary objective of incorporating SA is to augment the exploitation phase by actively exploring the most promising regions identified by the WOA algorithm.  To assess the performance of these proposed approaches, experiments are conducted on 18 standard benchmark datasets sourced from the UCI repository. These approaches are then compared against three well-established wrapper feature selection methods from the existing literature. The experimental results conclusively demonstrate the efficacy of the proposed techniques in enhancing classification accuracy compared to other wrapper-based algorithms. These results underscore the WOA algorithm's capability in effectively navigating the feature space and selecting the most informative attributes for various classification tasks. | The proposed algorithm, implemented in Matlab, (KNN) classifier with K=5 and Euclidean distance. Cross-validation (K-fold) was conducted with 100 iterations and a population size of 10.  Hybrid algorithms outperformed the native one in classification accuracy and feature selection.  Tournament Selection (TS)-based approaches, WOASAT-1 and WOASAT-2, enhanced exploration and produced better results than other methods. WOASAT-2 achieved the best approach, offering high accuracy and minimal feature subsets. The WOA algorithm demonstrated robustness and balanced exploration and exploitation, making it a reliable choice for feature selection. |
| [8] | Metaheuristic Search Based Feature Selection Methods for Classification of Cancer  Authors: L. Meenachi a, S. Ramakrishnanb  Published in: Pattern Recognition  Volume 119, November 2021, 108079 | This research paper presents innovative methods for predicting cancer using microarray gene expression data through the application of metaheuristic search algorithms for feature selection. While global feature selection methods such as ant colony optimization and genetic algorithms excel at identifying optimal features, challenges arise in selecting nearby features. To overcome these challenges, two new feature selection algorithms are introduced: the ACTFRO Algorithm, which combines Ant Colony Optimization and Tabu Search with Fuzzy Rough Set for Optimal Feature Selection, and the GATFRO Algorithm, which combines Genetic Algorithm and Tabu Search with Fuzzy Rough Set for Optimal Feature Selection. These algorithms are evaluated using a fuzzy rough nearest neighbour classifier with ten-fold cross-validation on four cancer-related medical datasets and one non-medical dataset.  Classification accuracy, computation time, sensitivity, specificity, F-measure, are among the metrics of evaluation. | Results are obtained by combining these algorithms with the FRNN classifier.  The hybrid approach of global, local feature selection yields good results across different training/testing ratios. Tenfold cross-validation confirms their robustness, with ACTFRO achieving an average accuracy of 88.54%, and GATFRO reaching 90.53%.  These algorithms consistently outperform existing methods in terms of accuracy, sensitivity, specificity, f-measure, shorter computation times, making them promising for practical applications in cancer diagnosis and other fields. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Chapter 3

## Conclusion and Plan