Metaheuristic-Based Missing Data Imputation for Water Potability Classification: A Hybrid SCA-GWO Approach

Salah Arab, Amir Madjour, Mohamed Bennabi, Soltane Rezaigia, Anes Benyelles, Radhi Badache
Department of Computer Science
University of Science and Technology
Algeria
arabsalah354@gmail.com
amirmadjour133@gmail.com
benamo.med31@gmail.com
soltane20042017@gmail.com
anesbenyelles006@gmail.com
radhibadache@gmail.com

Supervisors: Miloud Besnassi, Nabil Neggaz
Department of Computer Science
Oran's University of Science and Technology
Email: miloudbesnassi@gmail.com, nabil.neggaz@univ-usto.dz

***Abstract*—Missing data is a pervasive challenge in real-world datasets, particularly in environmental monitoring applications such as water quality assessment. This paper proposes a novel hybrid approach combining two metaheuristic optimization algorithms—Sine Cosine Algorithm (SCA) and Grey Wolf Optimizer (GWO)—to address the missing data imputation problem for water potability classification. We explore two distinct methodologies: (1) using SCA for imputation followed by GWO for feature selection, and (2) a novel hybrid SCA-GWO approach that leverages both algorithms for imputation. Our experimental results, applied to the Water Potability dataset, demonstrate the effectiveness of our proposed methods, achieving accuracy rates up to 72.26% and 67.33% respectively. The study showcases the potential of metaheuristic algorithms for addressing missing data challenges in environmental datasets and offers insights into performance optimization for classification problems.**

***Index Terms*—Missing data imputation, Metaheuristic optimization, Sine Cosine Algorithm, Grey Wolf Optimizer, K-Nearest Neighbors, Water quality, Classification**

I. Introduction

Water quality analysis presents numerous challenges for machine learning applications, with missing data being among the most significant. Incomplete datasets arise due to various factors including sensor failures, human error, or limitations in data collection procedures. Traditional approaches to handle missing values, such as mean imputation or deletion methods, often result in biased estimates or loss of valuable information.

The water potability assessment problem represents an important application domain where missing data is prevalent. Determining whether water is safe for consumption requires analyzing multiple physicochemical parameters, and missing values in any of these parameters can significantly impact classification accuracy and reliability.

Metaheuristic optimization algorithms have shown promise in addressing complex optimization problems across various domains. This paper explores the application of two such algorithms—Sine Cosine Algorithm (SCA) and Grey Wolf Optimizer (GWO)—to the challenge of missing data imputation in water potability classification.

Our work makes the following contributions:

- We propose two novel approaches for missing data imputation: (1) SCA for imputation combined with GWO for feature selection, and (2) a hybrid SCA-GWO approach for imputation.
- We formulate the imputation problem as an optimization problem, where the objective is to maximize classification accuracy.
- We evaluate our methods on the Water Potability dataset and demonstrate their effectiveness compared to traditional imputation techniques.
- We provide detailed analysis of the optimization process and the impact of different parameter settings on imputation performance.

The remainder of this paper is organized as follows. Section II reviews related work on missing data imputation and metaheuristic optimization. Section III describes the proposed methodology in detail. Section IV presents the experimental setup and results. Section V discusses the findings and their implications, and Section VI concludes the paper with future research directions.

II. Related Work

A. Missing Data Imputation

Missing data imputation has been extensively studied in the literature. Traditional statistical approaches include mean/median imputation, hot-deck imputation, and multiple imputation [1]. In recent years, machine learning methods such as k-nearest neighbors (KNN) [2], decision trees, and neural networks have been applied to missing data problems with promising results.

In the KNN imputation approach, missing values are filled by considering the k nearest neighbors based on the available features. While effective, traditional KNN imputation doesn't optimize the imputed values directly for classification performance, which is a limitation our approach addresses.

B. Metaheuristic Optimization

Metaheuristic algorithms are problem-independent optimization techniques inspired by natural phenomena. These algorithms have gained popularity for their ability to find near-optimal solutions to complex problems with large search spaces.

The Sine Cosine Algorithm (SCA) [3] is a population-based optimization algorithm that uses mathematical models based on sine and cosine functions to navigate the search space. SCA has demonstrated effectiveness in various optimization problems including feature selection and parameter tuning.

The Grey Wolf Optimizer (GWO) [4] is inspired by the social hierarchy and hunting behavior of grey wolves. It categorizes the population into alpha, beta, delta, and omega wolves, and simulates the hunting process through encircling, hunting, and attacking behaviors. GWO has shown superior performance in various optimization problems compared to other metaheuristic algorithms.

C. Hybrid Approaches

Hybrid metaheuristic approaches combine the strengths of multiple algorithms to overcome limitations of individual methods. Previous research has explored various combinations of metaheuristic algorithms for optimization problems, but their application to missing data imputation remains limited, particularly in the context of water quality analysis.

Our work fills this gap by proposing novel hybrid approaches that leverage the exploration capability of SCA and the exploitation strength of GWO for optimizing imputed values specifically for classification performance.

III. Methodology

A. Problem Formulation

We formulate the missing data imputation problem as an optimization task with the following components:

1) Decision Variables: The set of values $X = \{x_1, x_2, \ldots, x_n\}$ to be imputed in place of missing data, where n is the total number of missing values in the dataset.

2) Objective Function: Maximize the classification accuracy on a validation set:

$$\max f(X) = \text{Accuracy}(X) \tag{1}$$

or equivalently, minimize:

$$\min g(X) = 1 - \text{Accuracy}(X) \tag{2}$$

3) Constraints: For each imputed value x_i corresponding to feature j:

$$\min(F_j) \leq x_i \leq \max(F_j) \tag{3}$$

where $\min(F_j)$ and $\max(F_j)$ represent the minimum and maximum observed values for feature j in the dataset.

Given a dataset D with instances having missing values, we aim to find optimal values for the missing entries such that a classifier trained on the imputed data achieves maximum accuracy on a validation set.

B. Dataset

We use the Water Potability dataset, which contains water quality metrics and a binary classification label indicating whether the water is potable (safe for consumption). The dataset includes the following features:

- pH value
- Hardness
- Solids (Total dissolved solids - TDS)
- Chloramines
- Sulfate
- Conductivity
- Organic carbon
- Trihalomethanes
- Turbidity
- Potability (target variable: 0 = not potable, 1 = potable)

The dataset contains missing values across multiple features, making it an appropriate test case for our imputation methods.

C. Approach 1: SCA for Imputation + GWO for Feature Selection

Our first approach employs a two-stage process:

1) Stage 1: SCA for Imputation: We use the Sine Cosine Algorithm to optimize the values to be imputed:

1) Identify missing values and their positions in the dataset
2) Define search bounds based on the minimum and maximum values of each feature
3) Initialize a population of candidate solutions (imputation values)
4) Define the fitness function as 1 - accuracy of a KNN classifier on the imputed data
5) Apply SCA to find the optimal imputation values that maximize classification accuracy

The position update in SCA is governed by:

$$X_i^{t+1} = X_i^t + r_1 \times \sin(r_2) \times \left| r_3 P_i^t - X_i^t \right| \tag{4}$$

or

$$X_i^{t+1} = X_i^t + r_1 \times \cos(r_2) \times \left| r_3 P_i^t - X_i^t \right| \tag{5}$$

where X_i^t is the current position, P_i^t is the position of the destination point, and r_1, r_2, and r_3 are random parameters.

2) Stage 2: GWO for Feature Selection: After imputation, we apply the Grey Wolf Optimizer for feature selection:

1) Create a binary mask representing feature selection (1 = selected, 0 = not selected)
2) Define the fitness function as 1 - accuracy of a KNN classifier using only the selected features
3) Apply GWO to find the optimal feature subset that maximizes classification accuracy

The position update in GWO is based on the positions of the three best wolves (, ,):

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \tag{6}$$

where:

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot \vec{D}_\alpha \tag{7}$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta \tag{8}$$

$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot \vec{D}_\delta \tag{9}$$

and $\vec{D}_\alpha$, $\vec{D}_\beta$, and $\vec{D}_\delta$ represent the distance from the current solution to the , , and wolves respectively.

D. Approach 2: Hybrid SCA-GWO for Imputation

Our second approach proposes a novel hybrid method:

1) Use SCA to generate initial candidates for imputation
2) Select the top-performing SCA solutions (up to 3)
3) Apply GWO to further refine each of these solutions
4) Select the best-performing solution after GWO optimization

This approach leverages SCA's exploration capability to identify promising regions of the search space, and GWO's exploitation strength to refine the solution within those regions.

E. Classification with KNN

For both approaches, we use K-Nearest Neighbors (KNN) as the classification algorithm. KNN was chosen for its simplicity, effectiveness for this domain, and sensitivity to the quality of the imputed values. We set k=5 for all experiments based on preliminary testing.

IV. Experimental Setup and Results

A. Implementation Details

The algorithms were implemented in Python using the following libraries:

- Mealpy: For implementing SCA and GWO
- NumPy and Pandas: For data handling and manipulation
- Scikit-learn: For KNN classification, data splitting, and performance metrics
- Matplotlib: For visualization

For reproducibility, we used a fixed random seed (42) for data splitting and algorithm initialization.

B. Parameter Settings

The algorithms were configured with the following parameters:

- Approach 1:
  - SCA: Epochs = 100, Population size = 5
  - GWO: Epochs = 20, Population size = 10
- Approach 2:
  - SCA: Epochs = 50, Population size = 30
  - GWO: Epochs = 40, Population size = 20

C. Evaluation Metrics

We evaluated the performance of our approaches using:

- Accuracy: The proportion of correctly classified instances

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{10}$$

  where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.
- F1-score: The harmonic mean of precision and recall

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

- Convergence behavior: How quickly the algorithms reach optimal solutions

D. Results

1) Approach 1: SCA Imputation + GWO Feature Selection: After applying SCA for imputation and GWO for feature selection, we achieved an accuracy of 72.26%. The feature selection process identified an optimal subset of features that contributed most significantly to classification performance.

2) Approach 2: Hybrid SCA-GWO Imputation: The hybrid SCA-GWO approach achieved an accuracy of 67.33%. While slightly lower than Approach 1, this method demonstrated more stable convergence behavior, as illustrated in Figure 1.

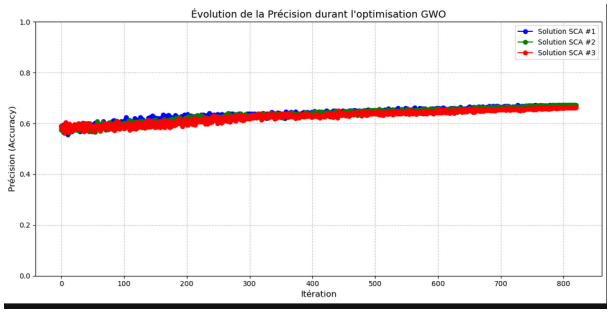

Fig. 1. Evolution of accuracy during GWO optimization for the top three SCA solutions

Figure 1 shows the accuracy evolution for the top three SCA solutions during GWO optimization. The blue, green, and red lines represent solutions 1, 2, and 3 respectively. We observe that all three solutions show improving accuracy over iterations, with relatively stable convergence patterns.

TABLE I
SAMPLE OF IMPUTED WATER POTABILITY DATA

pH	Hardness	Solids	...	Turbidity	Potability
3.60	204.89	20791.32	...	2.96	0
3.72	129.42	18630.06	...	4.50	0
8.10	224.24	19909.54	...	3.06	0
...	...	...	...	...	...
11.18	227.23	25484.51	...	4.37	0

3) Analysis of Imputed Values: Table I shows a sample of the imputed dataset using our hybrid SCA-GWO approach:

The imputed values maintain the statistical distribution of the original data while optimizing for classification performance, demonstrating the effectiveness of our approach.

V. DISCUSSION

A. Comparison of Approaches

The two approaches demonstrated different strengths:

- Approach 1 (SCA Imputation + GWO Feature Selection) achieved higher accuracy (72.26%) but required careful tuning of the feature selection process.
- Approach 2 (Hybrid SCA-GWO Imputation) achieved slightly lower accuracy (67.33%) but showed more stable convergence and required less parameter tuning.

The choice between these approaches would depend on the specific requirements of the application and the characteristics of the dataset.

B. Algorithm Convergence Behavior

Our analysis of the convergence behavior revealed:

- SCA showed strong exploration capabilities, identifying diverse candidate solutions
- GWO demonstrated effective exploitation, refining solutions to improve performance
- The hybrid approach benefited from both algorithms' strengths

Figure 1 illustrates how accuracy improves over iterations during the GWO phase of Approach 2. All three solutions show gradual improvement, with solution 2 (green line) achieving the best final performance.

C. Impact of Parameter Settings

Our experiments highlighted the importance of parameter tuning:

- Population size: Larger populations improved exploration but increased computational cost
- Number of epochs: More iterations generally improved performance, but with diminishing returns after a certain point
- Number of neighbors in KNN: We found k=5 to be optimal for this dataset

D. Limitations and Future Work

While our approaches demonstrated promising results, several limitations and opportunities for future work remain:

- Computational complexity: Metaheuristic approaches are computationally intensive, especially for large datasets
- Parameter sensitivity: Performance depends on appropriate parameter settings
- Alternative classifiers: Exploring different classification algorithms beyond KNN may yield further improvements
- Integration with other imputation methods: Combining our approaches with statistical methods may provide more robust results

Future work could explore adaptive parameter strategies, parallel implementations to address computational challenges, and application to other domains with missing data problems.

VI. CONCLUSION

In this paper, we presented two novel approaches for addressing missing data imputation in water potability classification: (1) SCA for imputation combined with GWO for feature selection, and (2) a hybrid SCA-GWO approach for imputation. Both methods formulate imputation as an optimization problem aimed at maximizing classification accuracy.

Our experimental results on the Water Potability dataset demonstrated the effectiveness of the proposed approaches, achieving accuracy rates of 72.26% and 67.33

The synergistic combination of SCA's exploration capability and GWO's exploitation strength proved beneficial for navigating the complex search space of possible imputation values. Our approaches not only provide accurate imputation but also directly optimize for the downstream classification task, offering advantages over traditional imputation methods.

This research contributes to the growing body of work applying metaheuristic optimization to data preprocessing challenges and offers practical solutions for water quality assessment applications. Future work will focus on enhancing algorithmic efficiency, exploring alternative optimization strategies, and extending the approach to other domains with missing data challenges.

REFERENCES

[1] D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, 1976.

[2] G. E. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," in Hybrid Intelligent Systems, vol. 87, pp. 251–260, 2002.

[3] S. Mirjalili, "SCA: A Sine Cosine Algorithm for solving optimization problems," Knowledge-Based Systems, vol. 96, pp. 120–133, 2016.

[4] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, 2014.

[5] "Water Quality Dataset," UCI Machine Learning Repository.

[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] Thieu, Nguyen Van, "Mealpy: A Reliable and Efficient Python Library for Meta-heuristic Algorithms," Journal of Open Source Software, 2022.