ABC-based Feature Selection and KNN Classification

Zeynep Altiparmak

January 2024

1 Abstract

This study explores the application of an Artificial Bee Colony (ABC)-based feature selection approach coupled with k-Nearest Neighbors (KNN) classification on the Breast Cancer Diagnostic Dataset (BCD). The primary objective is to enhance the accuracy and interpretability of breast cancer diagnosis by identifying the most relevant features through the ABC algorithm and subsequently employing KNN for effective classification. The ABC algorithm iteratively refines feature subsets using employed and onlooker bees, while a "scout bee" introduces randomness for exploration. The selected features are then utilized in the KNN classification model. The dataset, comprising 30 attributes and a classification label, is divided into training and testing sets for model training and evaluation. By integrating feature selection and classification techniques, this study aims to contribute to the improvement of breast cancer diagnostic models.

Keywords: Artificial Bee Colony Algorithm, Classification, Optimization, Breast Cancer Diagnosis

2 Materials and Methods

2.1 Classification

Classification is a fundamental aspect of machine learning models, particularly in the realm of medical diagnostics such as breast cancer detection. In the context of the Breast Cancer Diagnostic Dataset (BCD), classification involves the process of training a model to categorize instances into distinct classes, indicating the presence or absence of breast cancer.

2.2 Feature selection

Feature selection is a critical aspect of machine learning models, particularly in the context of the Breast Cancer Diagnostic Dataset (BCD). With 30 attributes

available, the goal is to identify a subset of features that significantly contribute to the model's predictive performance regarding breast cancer diagnosis. In this study, an Artificial Bee Colony (ABC) algorithm is employed, utilizing employed bees for iterative feature selection, onlooker bees for performance evaluation, and a "scout bee" for introducing randomness. The ABC algorithm dynamically refines the feature subset over iterations, optimizing the model's interpretability, reducing dimensionality, and enhancing predictive accuracy for more effective breast cancer diagnosis.

2.3 Dataset

The Breast Cancer Diagnostic Dataset (BCD) serves as the foundational dataset for this study, focusing on the diagnosis of breast cancer. Comprising 30 attributes representing various features of breast cells and a classification label, each attribute plays a potentially significant role in breast cancer diagnosis. The dataset includes two separate CSV files: "BCDTrain.csv," utilized for model training, and "BCDTest.csv," employed for evaluating the trained model's performance. The classification label assumes values of 1 or 0, signifying the presence or absence of breast cancer, respectively.

2.4 ABC (Artificial Bee Colony) Algorithm

Artificial Bee Colony (ABC) is one of the most recently defined algorithms by Derviş Karaboğa in 2005 [1], motivated by the intelligent behavior of honey bees. Since 2005, Karaboğa and his research group have studied on ABC algorithm and its applications to real world-problems. The ABC Algorithm is a swarm based meta-heuristic algorithm. It based on the foraging behavior of honey bee colonies. The artificial bee colony contains three groups.

2.4.1 Employed bees

An employed bee produces a modification on the position (solution) in her memory depending on the nectar amount (fitness value) of the new source (new solution). Provided that the nectar amount of the new one is higher than that of the previous one, the bee memorizes the new position and forgets the old one. After all employed bees complete the search process they share the nectar information of the food sources and their position information with the onlooker bees on the dance area.

2.4.2 Onlooker bees

An onlooker bee evaluates the nectar information taken from all employed bees and chooses a food source with a probability related to its nectar amount. As in the case of the employed bee, it produces a modification on the position in its memory and checks the nectar amount of the candidate source. Providing that its nectar is higher than that of the previous one, the bee memorizes the new position and forgets the old one.

2.4.3 Scout bees

The food source of which the nectar is abandoned by the bees is replaced with a new food source by the scouts. In ABC, providing that a position can not be improved further through a predetermined number of cycles, which is called "limit" then that food source is assumed to be abandoned.

2.5 KNN (K-Nearest Neighbors) Algorithm

For this study, the k-Nearest Neighbors (KNN) classification algorithm is employed. KNN is chosen for its simplicity and effectiveness in handling classification tasks. The algorithm classifies instances based on the majority class among their k-nearest neighbors, making it suitable for medical diagnoses where patterns may exist in proximity.

3 Technologies and Libraries

3.1 Python

The project was developed using the Python programming language. Python's flexibility and extensive library support make it an ideal choice for data analysis and machine learning applications.

3.2 Pandas Library

The Pandas library [2] was used for data manipulation and analysis. This library is ideal for performing fast and efficient operations on data frames.

3.3 Scikit-Learn Library

In this study, the Scikit-Learn library [3], an open-source and user-friendly machine learning toolkit, has been utilized. Specifically, modules such as KNeighborsClassifier for creating a k-NN classification model and accuracy score for assessing model performance were employed

4 Results

Before the feature selection, initial classification using the k-Nearest Neighbors (KNN) algorithm yielded a baseline accuracy score of 0.90. Following the application of the Artificial Bee Colony (ABC) algorithm for feature selection, the accuracy scores exhibited improvement, ranging between 0.92 and 0.95. Notably, the highest achieved accuracy was 0.95. These results suggest that the integration of the ABC-based feature selection process enhanced the discriminatory power of the model, resulting in more accurate predictions for breast cancer diagnosis. The incremental accuracy gains underscore the effectiveness

of the feature selection strategy in refining the dataset for optimal utilization by the KNN classification algorithm.

References

[1] Dervis Karaboga and Bahriye Basturk. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm. *Journal of global optimization*, 39:459–471, 2007.

[2] https://pandas.pydata.org/docs/.

[3] https://scikit-learn.org/stable/getting$_s$*tarted.html*.