Feature Selection using Multi-objective CHC Genetic Algorithm

UNIVERSITY OF GHANA

UNIVERSITY OF GHANA

Group 4

Department of Computer Science – CSCD 612 – Intelligent Systems

Members' IDs: 11010649; 11008901; 11005257; 11004749

Overview

Introduction: This paper proposes a Multi-objective (MO) Cooperative Coevolutionary Algorithm (CHC) for Feature Selection (FS) which includes the design of a CHC in a multi-objective framework with the help of non-dominating sorting genetic algorithm II (NSGA II) with trade-offs between accuracy and reduction rate. The MOCHC-FS algorithm has the potential to improve ML algorithms and can be used in classification and clustering.

Related Work: Various studies on feature selection, including filter, wrapper and hybrid approaches, as well as multi-objective optimization algorithms such as NSGA-II, MOEA/D, and SPEA2. The limitations to existing feature selection methods, such as the inability to handle large datasets, the lack of diversity in the selected features and the difficulty in finding the optimal trade-offs between accuracy and reduction rate.

Proposed Algorithm: A hybrid implementation of a MOCHC by using CHC and NSGA-II to maximize classification accuracy and minimizing the number of selected features.

Results and Analysis: For *moving window* analysis, we recommend the sampling of *Bellwether* projects to be considered for the construction of software effort estimation models.

Relevance of Algorithm

The relevance of proposed algorithm lies in its ability to handle large datasets with many features and provide a diverse set of non-dominated solutions with trade-offs between accuracy and reduction rate, which can help improve the performance and interpretability of machine learning a models.

Problem Statement

The problem statement of the paper is to address the challenges of feature selection in ML, including the difficulty of finding the optimal trade-offs between accuracy and reduction rate, the lack of diversity in the selected features and the inability to handle large datasets.

Research Contributions

1. The proposal of a novel multi-objective coevolutionary algorithm for feature selection that simultaneously optimizes two conflicting objecitves.

2. The development of a unique coding scheme that ensures diversity and prevents premature convergence in the MOCHC algorithm

3. The demonstration of the relevance of the proposed algorithm in handling large datasets with many features and providing a diverse set of non-dominated solutions.

Terminologies

This section provides techniques and strategies used within GA presented in the paper:

1. Multi-objective GA: Aims to find a set of Pareto-optimal solutions that represent the trade-offs between conflicting objectives
2. CHC GA: Uses cross-generational elitist selection, heterogeneous recombination and cataclysmic mutation to generate new candidates.
3. NSGA II: Assignes each solution to a front based on its dominance relationship with other solution
4. Feature Selection: Used to reduce the dimensionality of the data and improve the accuracy and efficiency of ML algorithm
5. KNN: A classification algorithm that assigns a class label to a new data point based on the class labels of its k nearest neighbors in the training data.
6. Pareto-optimal solutions: A set of candidate feature subsets that represent the trade-offs between conflicting objectives in feature selection.

Experimental Approach

•The proposed algorithm is validated on twenty datasets available on the UCI dataset repository.

•Preprocessing of datasets by removing missing values, normalizing the features and converting categorical features to numerical ones

•Applying the proposed algorithm on each dataset and compared its performance with other feature selection models.

•Using ML models to determine accuracies.

Limitation w/ Exercise

- Datasets not found at provided location.
- More functionalities were given than what was provided.
- Computer used was not powerful enough to process all computations
- It is believed that, more than one ML classifier was used in processing the data.

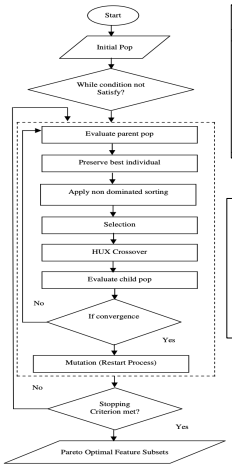

Fig. 1. Flowchart of proposed algorithm

No	Datasets	#Features	#Instances	#Classes	No.	Datasets	#Features	#Instances	#Classes
1	Spambase	58	4601	2	11	German	20	1000	8
2	Waveform	41	5000	3	12	Zoo	18	101	7
3	Ionosphere	35	351	2	13	Mushroom	23	8123	2
4	WDBC	32	568	2	14	Chess	6	175358	4
5	Vehicle	26	12684	2	15	Splice	93	6071	3
6	Wine	12	1599	6	16	Vote	15	10000000	2
7	Breast Cancer	31	569	2	17	Connect-4	43	50667	3
8	WBC	31	569	2	18	Flare	12	12455	2
9	Glass	10	214	6	19	Tic-Tac-Toe	10	958	2
10	Iris	5	150	3	20	Lymph	13	1000	4

Fig. 2. Team's Dataset Description

Sr. No.	Datasets	#Features	#Instances	#Classes	Sr. No.	Datasets	#Features	#Instances	#Classes
1.	Spambase	57	4701	2	11.	German	24	1000	2
2.	Waveform	40	5000	3	12.	Zoo	17	101	7
3.	Ionosphere	34	351	2	13.	Mushroom	22	5644	2
4.	WDBC	30	569	2	14.	Chess	36	3196	2
5.	Vehicle	24	1000	2	15.	Splice	60	3190	3
6.	Wine	13	178	3	16.	Vote	16	232	2
7.	Breast Cancer	10	683	2	17.	Connect-4	42	67557	3
8.	WBC	9	699	2	18.	Flare	11	1066	6
9.	Glass	9	214	6	19.	Tic-Tac-Toe	9	958	2
10.	Iris	4	150	3	20.	Lymph	18	148	4

Fig. 3. Article's Dataset Description

$$Fitness = \{f1 = Acuracy; f2 = \left(1 - \frac{|RS|}{|DS|}\right) * 100\}$$

Fig. 4. Fitness Functions used

Analysis – Result Replication (Team Results)

Datasets	MOCHC-FS Accuracy %	% Reduction	Total #Features	Selected #Features
Spambase	89.14	56.14	58	25
Waveform	76.4	52.5	41	19
Ionosphere	84.51	52.94	35	16
WDBC	73.57	53.33	32	14
Vehicle	62.51	48	26	13
Wine	55.63	27.27	12	8
Breast Cancer	74.89	43.33	31	17
WBC	69.15	46.67	31	16
Glass	69.77	44.44	10	5
Iris	83.33	50	5	2
German	27.00	50	20	4
Zoo	100	17.65	18	14
Mushroom	66.22	50.00	23	11
Chess	67.30	60.00	6	2
Splice	75.67	53.26	93	43
Vote	87.39	50.00	15	2
Connect-4	66.60	38.10	43	26
Flare	90.53	64.71	12	6
Tic- Tac -Toe	67.71	44.44	10	5
Lymph	66.33	41.64	13	6

Comparing the results from the algorithm followed from the paper and the actual results from the paper, it seems the algorithm does indeed address the objective of multi-objective in feature selection. Unfortunately, the team's replication of the code the not yield higher results compared to that of the actual paper.

Analysis – Article Results

Datasets	KNN Accuracy	MOCHC-FS Accuracy	% Feature Reduction	Total #Features	Selected #Features
Spambase	81.89	99.26	98.43	57	2
Waveform	89.90	98.78	97.56	40	2
Ionosphere	90.00	97.56	96.96	34	2
WDBC	97.76	99.79	88.88	30	4
Vehicle	65.17	78.52	94.73	24	2
Wine	100	100	92.30	13	2
Breast Cancer	97.46	98.67	72.34	10	3
WBC	96.32	97.31	88.88	9	2
Glass	95.34	96.19	66.66	9	3
Iris	98.33	98.66	50.00	4	2
German	61.00	100	95.00	24	2
Zoo	100	100	93.75	17	2
Mushroom	95.78	98.18	90.90	22	2
Chess	99.52	98.56	83.33	36	6
Splice	72.17	75.90	76.66	60	14
Vote	98.92	98.20	89.50	16	2
Connect-4	97.99	84.70	86.79	42	5
Flare	95.99	78.36	90.90	11	2
Tic-Tac-Toe	88.68	82.75	88.88	9	3
Lymph	86.66	98.43	88.86	18	3

Algorithm Parameters

- ❖ Chromosome Length = number of features within a dataset
- ❖ Generation Count = 80
- ❖ Population Size = 40
- ❖ Difference threshold = ¼ of length of chromosome
- ❖ Mutation rate = 0.35
- ❖ Number of Elite kids = 10%
- ❖ Stalling generations = 10
- ❖ Criterion: if size of the new population is larger than the original size of population, chromosones chosen based on crowding distance

Conclusion

The results show that the MOCHC algorithm outperforms the other methods in terms of classification accuracy and reduction rate, and provides a diverse set of non-dominated solutions with trade-offs between accuracy and reduction rate. The proposed algorithm is relevant in handling large datasets with many features and providing a better understanding of the underlying data patterns. The paper contributes to the field of feature selection by introducing a new algorithm that can improve the performance and interpretability of machine learning models.