Distribution and Replication for Feature Selection

(ESR 2.2)

Uchechukwu Njoku

2nd DEDS Winter school

Aalborg, Denmark

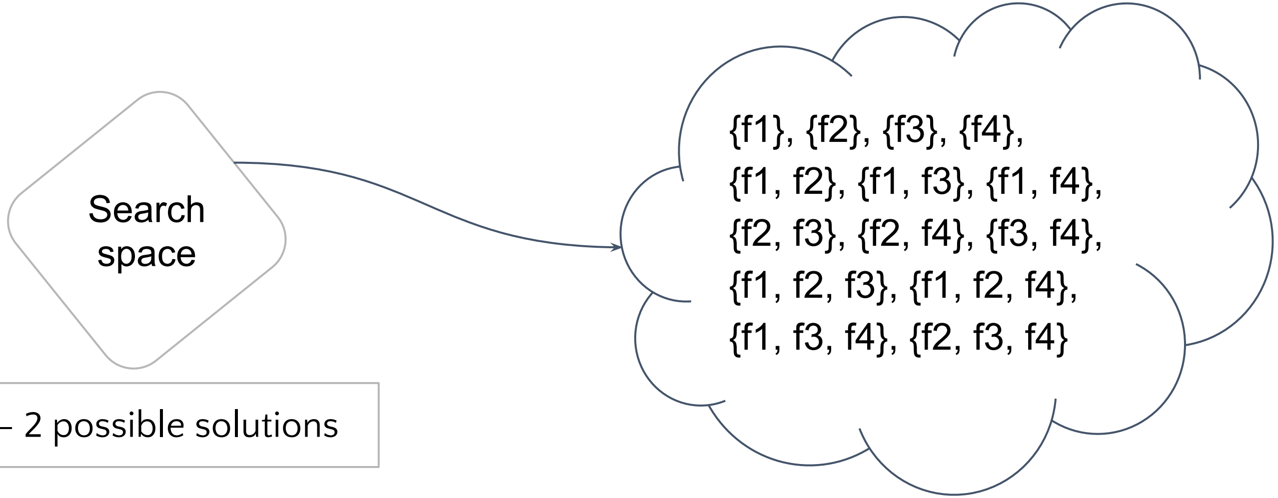
8th March, 2022

Supervisors: Alberto Abelló (UPC), Besim Bilalli (UPC), Gianluca Bontempi (ULB)

Feature selection (FS)

Feature selection is **a search problem** of **detecting the relevant features and discarding the irrelevant and redundant ones** with the goal of obtaining a subset of features that accurately describe a given problem with a **minimum degradation of performance**

Instance of a dataset with **four features**: {f1, f2, f3, f4}



$2^n - 2$ possible solutions

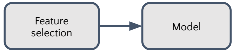
Exploring the search space

- Starting point
- Search strategy
    - Exhaustive search
    - Sequential search
    - Population-based search
- Feature subsets evaluation
- Halting criterion

Feature selection classification

Feature selection methods are popularly classified based on their relationship with the learning algorithm

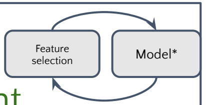
Filter methods



- Fast execution time
- Good generalization
- Robust to overfitting
- Possible redundancy
- Model independent
- Non 'optimal' selection

E.g: Gini, ReliefF, MRMR, CFS

Wrapper methods



- Model dependent
- High accuracy
- Captures dependencies
- Poor generalization
- Risk of overfitting
- Computationally intense

E.g: SFS, SBS

Embedded methods

- Model dependent
- Moderate execution time
- Captures dependencies
- Poor generalization

E.g: Tree based algorithms

Objectives

1. Study existing feature selection methods to:

   a. compare the several search algorithms of wrapper feature selection and existing tools

   b. evaluate their scalability, stability, and impact on performance (e.g., accuracy)

2. Propose a novel approach for **multi-criteria** wrapper feature selection

3. **Optimize** the distribution and parallelism of feature engineering for Big Data

   a. Analyze the scalability of feature engineering methods

4. Optimize wrapper feature selection methods by adopting frameworks for distribution, parallel computing, load partitioning, and communication methods for **scalable** feature selection

Work done

Objective 1: Study existing feature selection methods to:

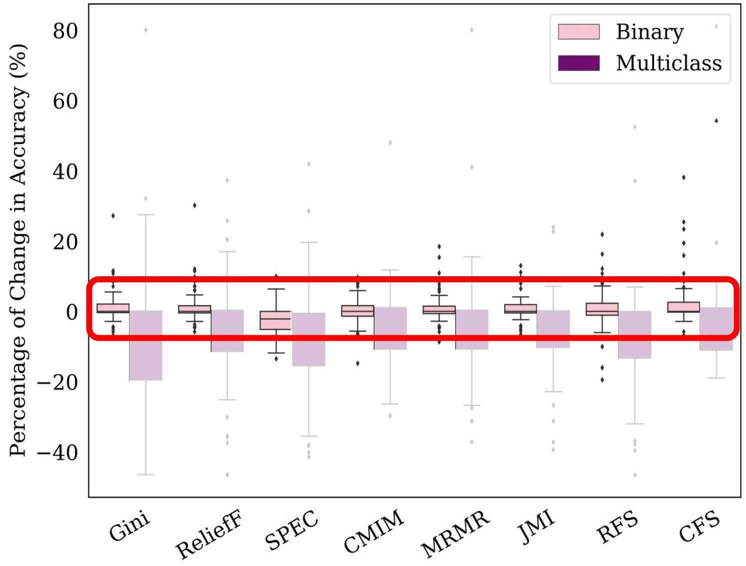
- Extensive evaluation of the predictive performance and stability of existing wrapper and filter methods in Python libraries
- Empirical comparison of multi-objective and mono-objective wrapper FS by considering two well-known metrics, accuracy and AUC
- Analysis of the scalability and memory footprint of wrapper methods

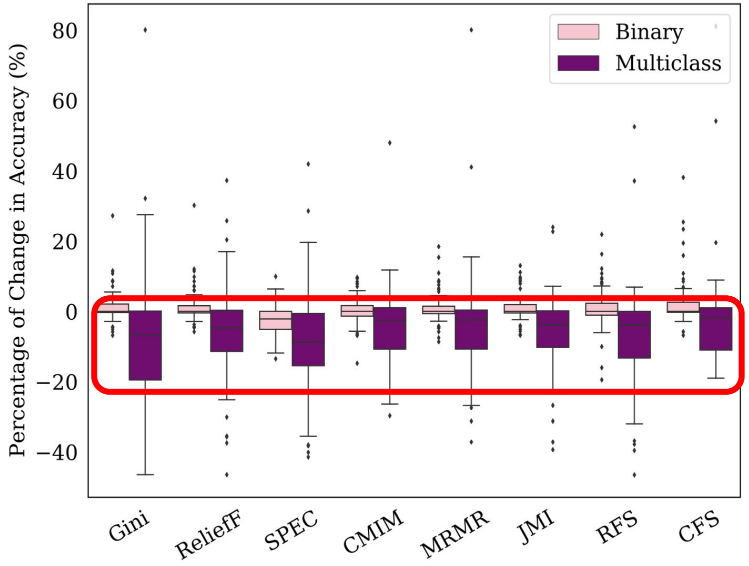
Outcome:

- Njoku, Uchechukwu Fortune, et al. "Impact of filter feature selection on classification: an empirical study." Proceedings of the 24rd International Workshop on DOLAP 2022
- Njoku, Uchechukwu Fortune, et al. "Wrapper methods for multi-objective feature selection." 26th International Conference on EDBT 2023

Results

The improvement in accuracy after feature selection is different for binary and multiclass classifications for **filter methods**





Set-up

- 32 datasets (binary and multi-class)
- Four classification algorithms: KNN, NB, DT, SVM
- Six filter methods: Gini, MRMR, Relief, JMI, CMIM, SPEC
- Four wrapper methods: SFS_{KNN}, SFS_{NB}, SFS_{DT}, SFS_{SVM}
- Metrics: Accuracy and AUC

Results

- Wrapper methods show superior results in predictive performance

Method	NB	KNN	DT	SVM	Average Rank
SFS	10	9	4	9	1
JMI	4	4	8	6	2
Gini	5	2	6	5	3
CMIM	3	2	4	5	4
MRMR	3	3	4	3	5
Relief	2	1	4	3	6
SPEC	0	2	1	2	7

Accuracy change in binary problems

Method	NB	KNN	DT	SVM	Average Rank
SFS	5	13	7	8	1
Gini	7	4	3	1	2
MRMR	5	4	2	1	3
Relief	2	5	2	3	2
JMI	3	3	2	4	5
CMIM	3	2	2	2	6
SPEC	2	2	0	1	7

AUC change in binary problems

Existing implementations

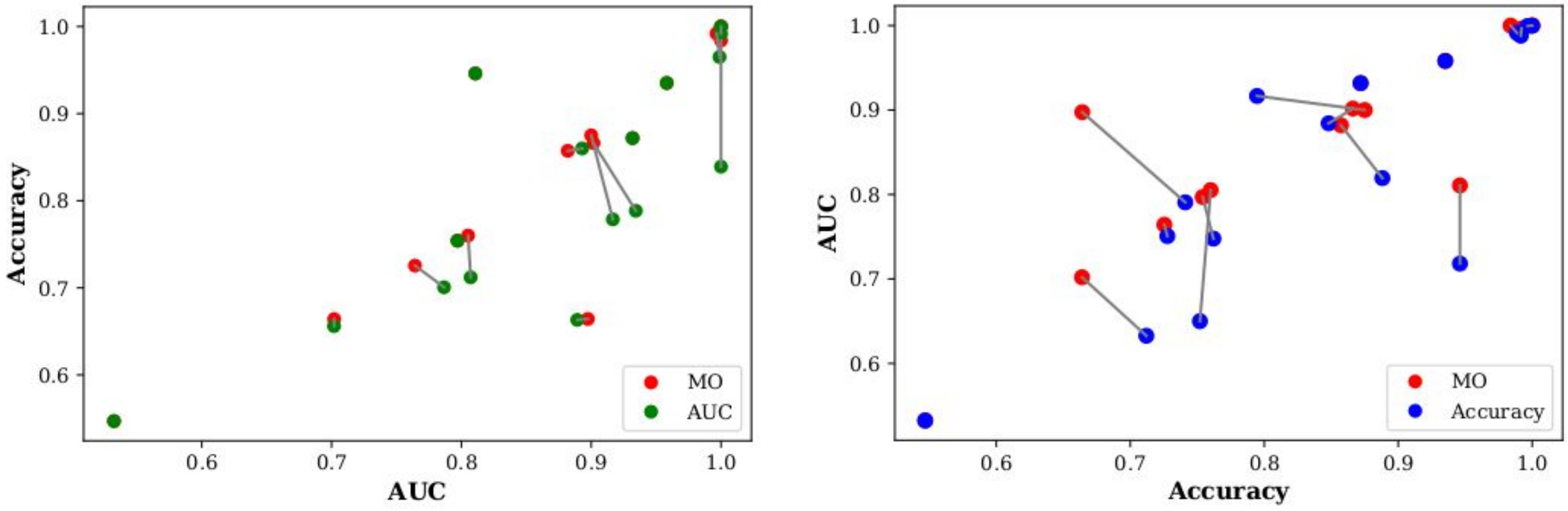
- Less focus on population-based wrapper methods

Tool	Lang.	Exhaustive	Population-based	Sequential	Parallelism
Weka	Java	X	X	✓	Partial
Scikit-Learn(SKL)	Python	X	X	✓	Partial
Rapidminer	Java	✓	✓	✓	Partial
Mlxtend(MLX)	Python	✓	X	✓	Yes
Scikit-Feature(SKF)	Python	X	X	✓	No
FeatureSelect	Matlab	X	✓	X	No

Feature selection tools and libraries

Results

- Wrapper method are unique for multi-criteria feature selection



MO-multi-objective, **ACC**-Accuracy, **AUC**- Area under the ROC Curve

Current work

Objective 2: Propose a novel approach for multi-criteria wrapper feature selection

Traditionally, multi-criteria feature selection is limited to **two objectives** – number of features and model performance

We propose the use of **more than two objectives** simultaneously for feature selection and demonstrate the advantage and trade-off through an interactive visualization board using **population-based search**

Multi-criteria feature selection

Criteria:

Internal evaluation criteria

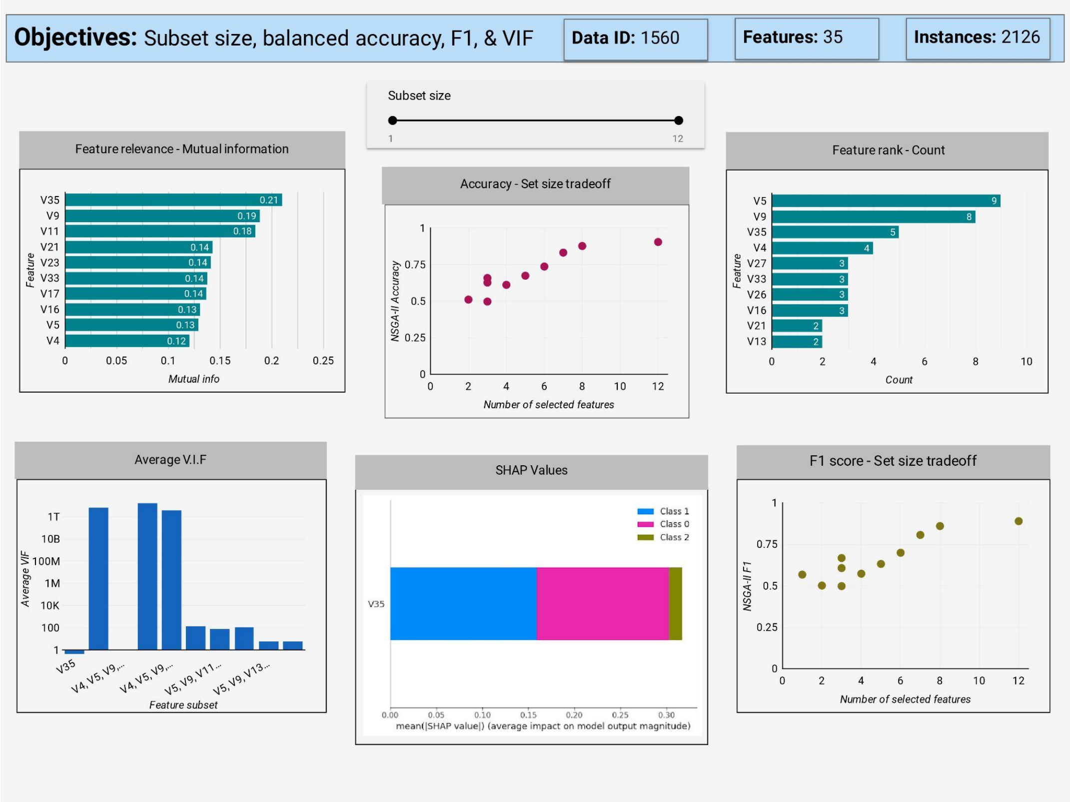
- AUC
- Precision
- Accuracy
- Redundancy
- Number of features

External evaluation criteria

- Relevance
- Shapley function

Outcome:

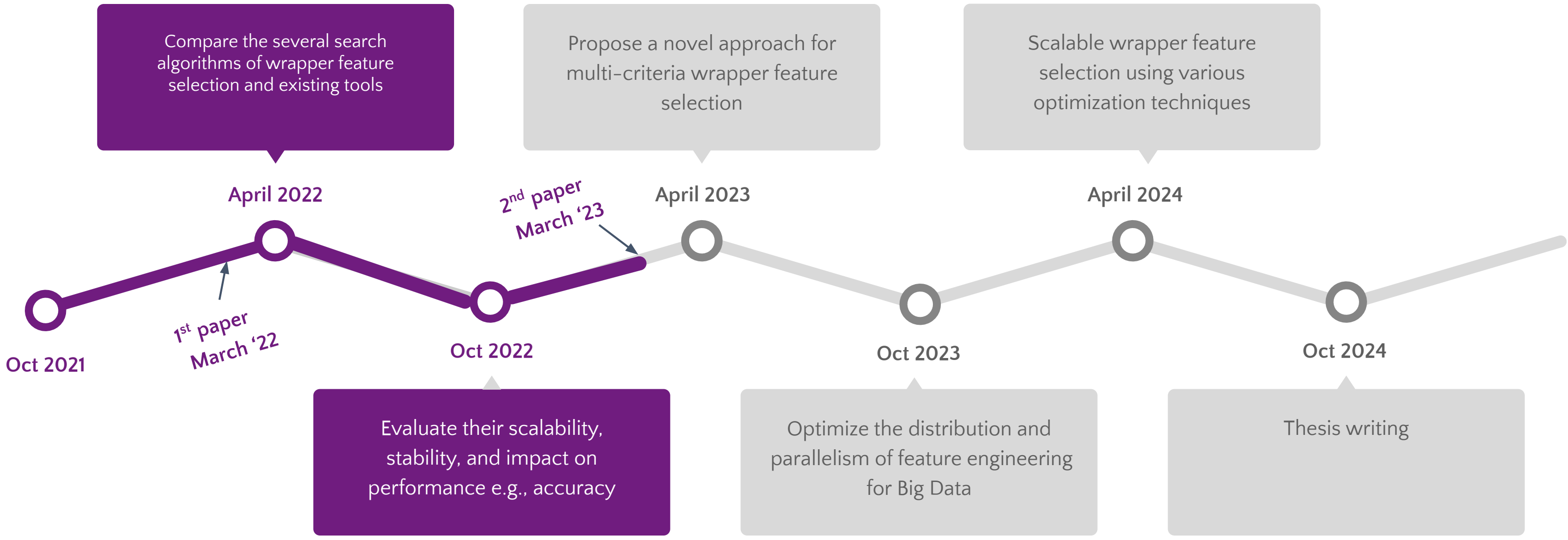
- Set of near-optimal feature subsets
- Interactive dashboard of results for explainability


Objectives: Subset size, balanced accuracy, F1, & VIF
Data ID: 1560
Features: 35
Instances: 2126
Subset size
1
12
Feature relevance - Mutual information
V35 0.21
V9 0.19
V11 0.18
V21 0.14
V23 0.14
V33 0.14
V17 0.14
V16 0.13
V5 0.13
V4 0.12
Feature
Mutual info
Accuracy - Set size tradeoff
NSGA-II Accuracy
Number of selected features
Feature rank - Count
V5 9
V9 8
V35 5
V4 4
V27 3
V33 3
V26 3
V16 3
V21 2
V13 2
Feature
Count
Average V.I.F
1T
10B
100M
1M
10K
100
1
Average VIF
V35
V4, V5, V9,...
V4, V5, V9,...
V5, V9, V11...
V5, V9, V13...
Feature subset
SHAP Values
Class 1
Class 0
Class 2
V35
mean(|SHAP value|) (average impact on model output magnitude)
F1 score - Set size tradeoff
NSGA-II F1
Number of selected features

Future work

- Feature engineering
  - External stay
  - 3rd April – 30th June at Orange
  - Dataset of 200 columns and 2.5 million rows
- Optimization through distribution and parallelism

Timeline



Compare the several search algorithms of wrapper feature selection and existing tools
April 2022
Oct 2021
1st paper March '22
Oct 2022
Evaluate their scalability, stability, and impact on performance e.g., accuracy
2nd paper March '23
April 2023
Propose a novel approach for multi-criteria wrapper feature selection
Oct 2023
Optimize the distribution and parallelism of feature engineering for Big Data
April 2024
Scalable wrapper feature selection using various optimization techniques
Oct 2024
Thesis writing

Thanks for your attention

Q & A