Feature Selection Using Optimization Metaheuristics in Machine Learning

Loshan Rasan
Safae Elkhaoui
Sonimith Hang
Laurent Letrouit
Noé Flandre

October 22, 2024

1 Introduction

Feature selection is a critical step in machine learning, especially when dealing with high-dimensional data such as genomics datasets. It aims to select a subset of relevant features that contribute the most to the predictive modeling task, improving model performance and reducing computational cost.

2 Approach Description

In this project, we implemented feature selection methods using optimization metaheuristics in both unsupervised and supervised settings. The metaheuristics compared are:

- **Genetic Algorithm (GA)**
- **Particle Swarm Optimization (PSO)**
- **Simulated Annealing (SA)**

2.1 Unsupervised Setting

In the unsupervised setting, the criterion to minimize was the mutual information score between selected features. The goal was to reduce redundancy and select features that provide unique information.

2.2 Supervised Setting

In the supervised setting, we used the mean Area Under the ROC Curve (AUC) of Logistic Regression classifier as the fitness function.

2.3 Implementation Details

The metaheuristics were implemented as follows:

- **Genetic Algorithm**: Initialized a population of binary chromosomes representing feature subsets, evolved over generations using selection, crossover, and mutation.
- **Particle Swarm Optimization**: Initialized particles representing feature subsets, updated velocities and positions based on personal and global best positions.

- **Simulated Annealing**: Started with an initial solution, explored neighboring solutions by flipping feature selection bits, accepted new solutions based on a probability that decreases over time.

Each method was run separately in both settings, and the selected features were evaluated using cross-validation to compute the AUC scores for the following classifiers : Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors.

3 How to Run the Script

To run the feature selection script using optimization metaheuristics, use the following command format:

```
python script/feature_selection.py CSV/dataset.csv --output CSV/results.csv
```

4 Results

The results of the experiments are summarized in Table 1. The table presents the mean AUC, standard deviation of AUC, and execution time for each combination of metaheuristic and classifier in both settings.

Table 1: Performance Comparison of Metaheuristics in Feature Selection

Method	**Classifier**	**AUC Mean**	**AUC Std**	**Execution Time (s)**
GA Unsupervised	Logistic Regression	0.6588	0.0399	5.62
GA Unsupervised	Random Forest	0.6681	0.0189	5.62
GA Unsupervised	Support Vector Machine	0.6398	0.0330	5.62
GA Unsupervised	K-Nearest Neighbors	0.6519	0.0217	5.62
GA Supervised	Logistic Regression	0.7512	0.0143	4.09
GA Supervised	Random Forest	0.7334	0.0151	4.09
GA Supervised	Support Vector Machine	0.7345	0.0221	4.09
GA Supervised	K-Nearest Neighbors	0.6280	0.0433	4.09
PSO Unsupervised	Logistic Regression	0.6848	0.0467	3.04
PSO Unsupervised	Random Forest	0.6517	0.0290	3.04
PSO Unsupervised	Support Vector Machine	0.6655	0.0664	3.04
PSO Unsupervised	K-Nearest Neighbors	0.6237	0.0434	3.04
PSO Supervised	Logistic Regression	0.7459	0.0278	4.12
PSO Supervised	Random Forest	0.6625	0.0348	4.12
PSO Supervised	Support Vector Machine	0.7286	0.0169	4.12
PSO Supervised	K-Nearest Neighbors	0.5936	0.0572	4.12
SA Unsupervised	Logistic Regression	0.6701	0.0519	9.11
SA Unsupervised	Random Forest	0.7050	0.0357	9.11
SA Unsupervised	Support Vector Machine	0.6561	0.0526	9.11
SA Unsupervised	K-Nearest Neighbors	0.6632	0.0281	9.11
SA Supervised	Logistic Regression	0.7191	0.0245	10.79
SA Supervised	Random Forest	0.6773	0.0244	10.79
SA Supervised	Support Vector Machine	0.7085	0.0291	10.79
SA Supervised	K-Nearest Neighbors	0.6327	0.0888	10.79

5 Analysis and Discussion

From the results, we observe the following:

- **Execution Time**: PSO was the fastest metaheuristic, especially in the unsupervised setting, with an average execution time of approximately 3 seconds. SA had the longest execution times due to its iterative nature and high number of iterations.
- **Classification Performance**: In the supervised setting, the GA achieved the highest AUC mean with Logistic Regression (0.7512) and relatively low standard deviation, indicating consistent performance. PSO also performed well with an AUC mean of 0.7459.
- **Speed-Performance Tradeoff**: Considering both execution time and performance, the **PSO in the supervised setting** offers the best tradeoff. It achieved high AUC scores comparable to GA but with shorter execution times.
- **Unsupervised vs. Supervised**: Supervised metaheuristics generally outperformed their unsupervised counterparts in terms of AUC, as expected since they utilize label information during feature selection.
- **Classifier Comparison**: Logistic Regression and Support Vector Machine classifiers consistently achieved higher AUC scores compared to Random Forest and K-Nearest Neighbors across most methods.

6 Conclusion

In conclusion, the Particle Swarm Optimization (PSO) metaheuristic in the supervised setting provided the best speed-performance tradeoff for feature selection on the genomics dataset. It achieved high classification performance with reduced computational time compared to other methods. This demonstrates the effectiveness of PSO for feature selection tasks in machine learning, particularly when computational resources are a constraint.