# Assigned Project - Feature Selection for MDP

**Zhe Yu**
200109973, zyu9@ncsu.edu

## 1 Discretization

- Features with value type less than 10, keep as it is;
- Otherwise: 0 if <= median, 1 if > median.

## 2 Methods

Three different methods are tested in this project: Greedy Forward Feature Selection, Genetic Algorithm, and Decision Tree. Greedy Forward Feature Selection and Genetic Algorithm run on discretized data and use the feedback of ECR value to guide the feature selection. On the other hand, Decision Tree run on continuous data and need not the ECR feedback. Decision Tree learns for discretization as well as feature selection.

### 2.1 Greedy Forward Feature Selection

1. Initiate an empty feature set $S = []$, and a candidate feature $C = [f_1, f_2, \cdots, f_n]$.
2. For every feature $f \in C$, test ECR value of $S + f$.
3. Find the feature with largest ECR value $f^*$, $S = S + f^*$, $C = C - f^*$.
4. Go back to 2 until $S$ has 8 features.

### 2.2 Genetic Algorithm

1. **Initialization:** 100 random feature sets $F = [f_1, f_2, \cdots, f_8]$.
2. **Evaluation:** Evaluate the ECR values for the 100 candidate feature sets.
3. **Selection:** Select 10 feature sets with highest ECR values.
4. **Crossover:** Use the selected feature sets to make 100 baby feature sets. Baby feature sets are subsets of the joint of the selected feature sets.
5. **Mutation:** $5\%$ of chance one baby feature set will be randomly mutated.
6. Go back to 2 until stop rule is satisfied (max generation reached or best ECR does not change for several generations).

### 2.3 Decision Tree

1. No discretization.
2. Set Reward value to 0 or 1 depending on whether the reward at the end of the problem is negative or positive.
3. Train a Decision Tree Learner (CART) to predict the new reward.
4. Use the Decision Tree model to decide which feature to use and how to discretize the features.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

# 3 Results

## 3.1 Greedy Forward Feature Selection

- **Feature Selected:** {probDiff, Level, SolvedPSInLevel, cumul_NonPSelements, cumul_TotalWETime, TotalTime, cumul_AppCount, cumul_avgstepTimeWE}
- **ECR:** 79.3

## 3.2 Genetic Algorithm

- **Feature Selected:** {cumul_AppRatio, cumul_deletedApp, CurrPro_avgProbTimeWE, OptionalCount, difficultProblemCountSolved, SolvedPSInLevel, probIndexPSinLevel , DirectProofActionCount}
- **ECR:** 216.9

## 3.3 Decision Tree

- **Feature Selected:** {cumul_NonPSelements, cumul_WrongSyntaxApp, ruleScoreMP, cumul_WrongSemanticsApp cumul_englishSymbolicSwitchCount, CurrPro_avgProbTimeDeviationPS, F1Score}
- **ECR:** 25.0

## 3.4 Summary

- **Runtime:** Genetic Algorithm $>$ Greedy Forward Feature Selection $>>$ Decision Tree
- **ECR:** Genetic Algorithm $>$ Greedy Forward Feature Selection $>$ Decision Tree
- **Best method:** Genetic Algorithm
- **Best ECR:** 216.9

# 4 Discussion

**Why is Genetic Algorithm better than Greedy Forward Feature Selection?** Because Greedy Forward Feature Selection has a much much higher probability stuck in local optimum.

**Why is Decision Tree not working?** The ECR score for Decision Tree selected features is even not higher than some randomly selected feature set. This indicates that the correlation between feature and reward might not be helpful for building MDP model. This effect requires more experiments to validate.

# 5 Source Code

This project is open source on Github at https://github.com/azhe825/Machine-assisted-Testing/tree/master/assigned.