Improving Generalization of Evolutionary Feature Construction with Minimal Complexity Knee Points in Regression

Hengzhe Zhang

Supervisor: Mengjie Zhang, Bing Xue, Qi Chen, Wolfgang Banzhaf (MSU)

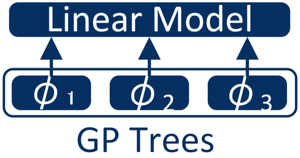
Victoria University of Wellington

02/01/2023

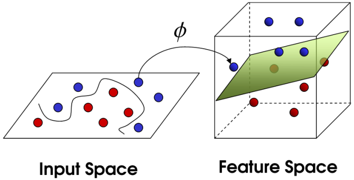
Table of Contents

Background

- **Objective:** Construct a set of new features, $\{\phi_1, \ldots, \phi_m\}$, to *enhance learning* on the dataset $\{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\}$ compared to learning on the original features $\{x^1, \ldots, x^p\}$.
- **Approaches:**
  - Kernel Methods: Black-box, non-parametric.
  - Neural Networks: Black-box, gradient-based.
  - Genetic Programming: Interpretable, gradient-free.



(a) Feature Construction on Linear Regression



(b) New Feature Space

Balancing Accuracy and Complexity

- **High accuracy** is crucial but may lead to complex models prone to overfitting.
- **Tree size** in genetic programming (GP) serves as a measure of model complexity.
- The key is to find an optimal balance: Maximize accuracy while minimizing tree size.

Occam's Razor in Modeling

Simpler models (smaller trees) are more likely to generalize well to unseen data, reducing the risk of overfitting.

- Utilize **multi-objective optimization** to address the dual objectives of high accuracy and low complexity.
- Solutions on the **Pareto front** represent the best trade-offs between competing objectives.

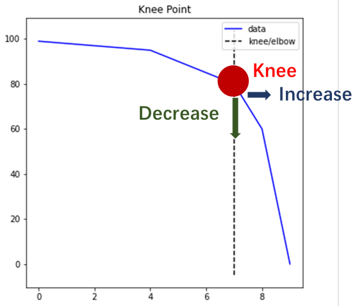
How to Select the Final Model?

Choose a model from the Pareto front that offers a balanced compromise, such as **knee point**, to ensure robust generalization.

Knee Point

Definition

A knee point in the context of multi-objective optimization is a point on the Pareto front where a small improvement in one objective would lead to a significant deterioration in another objective.



Potential Knee Points

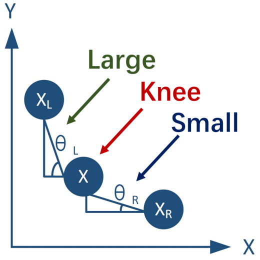
Point	Improvement in Obj. 1	Deterioration in Obj. 2
A	3%	10%
B	4%	20%
C	1%	3%

- Each point represents a potential knee point.
- The "significance" of improvements and deteriorations is subjective, leading to ambiguity in knee point selection.

Key Insight

The identification of knee points is influenced by subjective interpretations of small improvement and significant deterioration.

- Angle-based Method [1]
- Utility Function [2]
- Distance To Extreme Line [3]



Bend Angle Calculation

[1] Deb and Gupta 2011
[2] Rachmawati and Srinivasan 2009
[3] Schütze, Laumanns, and Coello 2008

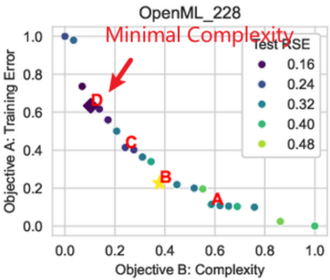
Method

Core Hypothesis

A significant increase in model complexity for small improvements in training accuracy may signal overfitting.

Minimal Complexity Knee Points

Opt for models at knee points with the least complexity to mitigate overfitting risks.

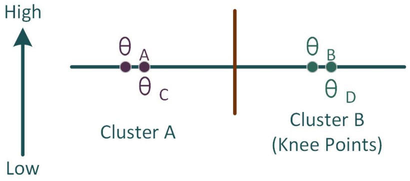


Using Clustering to Determine Knee Points

- Calculate the bend angle of each model on the Pareto front.
- Apply a clustering algorithm to group models.
- Identify clusters with largest bend angles.

Model	Bend Angle
D	θ_D
B	θ_B
C	θ_C
A	θ_A

Non-dominated Solutions



Key Advantage

Clustering automates the identification of thresholds for knee points, eliminating the need for manual threshold setting.

Task: Multi-tree GP for feature construction on a linear regression model.
Objectives: Minimize cross-validation loss and tree size.
Process:

- Initialize population with GP trees.
- Evaluate individuals using cross-validation loss and tree size.
- Select parents and generate offspring with GP operators.
- Use NSGA-II for environmental selection.

Final Model: Selected from Pareto front based on knee point strategy.

Settings

Datasets Overview

- 58 real-world datasets from the Penn Machine Learning Benchmark (PMLB).
- Excluded synthetic datasets to focus on real-world datasets.

Baseline Algorithms

Model Selection Methods:

- Angle Knee Selection (AKS)
- Four Angle Knee Selection (FAKS)
- Bended Angle Knee Selection (BAKS)
- Utility Function Knee Selection (UFKS)
- Distance To Extreme Line Knee Selection (DELKS)
- Best Training Accuracy
- Best Harmonic Mean Rank
- Standard GP (without Model Size as Objective)

Machine Learning Methods:

- SVR, KNN, Ridge, and DT

Results

- MCKP strategy **enhances generalization** on 32 datasets compared to standard GP (STD-GP).
- MCKP outperforms traditional knee point selection strategies, such as AKS, indicating better overfitting control.

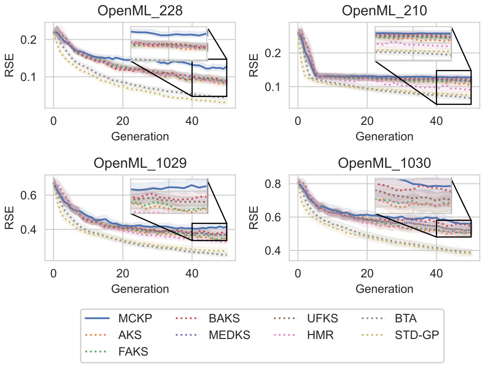
Statistical comparison of test RSE.

	AKS	FAKS	BAKS	MEDKS
MCKP	20(+)/31(∼)/7(-)	15(+)/36(∼)/7(-)	10(+)/37(∼)/11(-)	10(+)/40(∼)/8(-)
AKS	—	0(+)/58(∼)/0(-)	0(+)/56(∼)/2(-)	0(+)/54(∼)/4(-)
FAKS	—	—	0(+)/55(∼)/3(-)	0(+)/56(∼)/2(-)
BAKS	—	—	—	0(+)/58(∼)/0(-)
MEDKS	—	—	—	—
UFKS	—	—	—	—
HMR	—	—	—	—
BTA	—	—	—	—

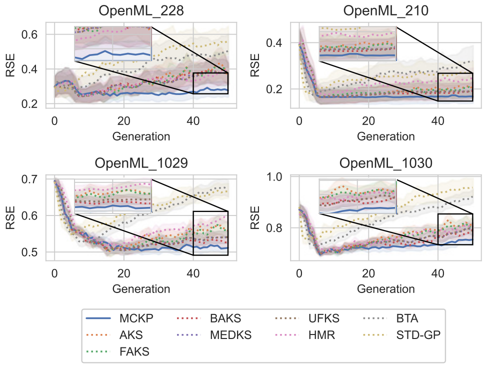
	UFKS	HMR	BTA	STD-GP
MCKP	11(+)/37(∼)/10(-)	16(+)/30(∼)/12(-)	31(+)/15(∼)/12(-)	32(+)/15(∼)/11(-)
AKS	0(+)/55(∼)/3(-)	2(+)/51(∼)/5(-)	23(+)/28(∼)/7(-)	21(+)/29(∼)/8(-)
FAKS	0(+)/56(∼)/2(-)	2(+)/52(∼)/4(-)	22(+)/28(∼)/8(-)	24(+)/26(∼)/8(-)
BAKS	0(+)/58(∼)/0(-)	3(+)/52(∼)/3(-)	24(+)/29(∼)/5(-)	24(+)/27(∼)/7(-)
MEDKS	0(+)/58(∼)/0(-)	3(+)/52(∼)/3(-)	24(+)/28(∼)/6(-)	26(+)/26(∼)/6(-)
UFKS	—	3(+)/52(∼)/3(-)	23(+)/29(∼)/6(-)	25(+)/27(∼)/6(-)
HMR	—	—	24(+)/31(∼)/3(-)	27(+)/26(∼)/5(-)
BTA	—	—	—	1(+)/53(∼)/4(-)

Results

- MCKP exhibits good generalization performance on unseen data, unlike some other knee point selection strategies that may overfit.

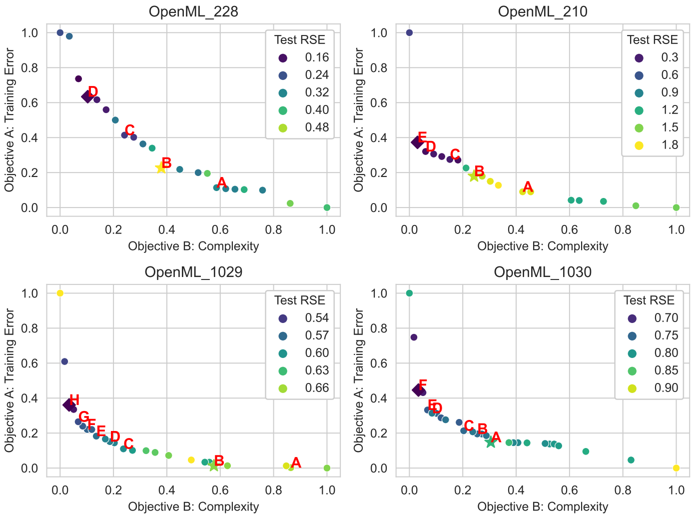


(a) Training RSE

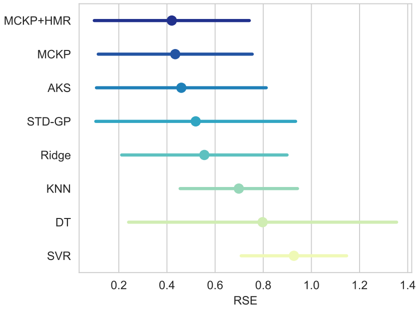


(b) Test RSE

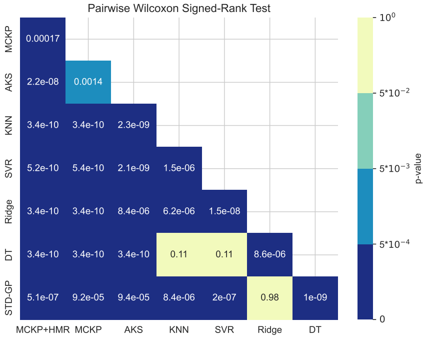
- *Knee points with minimal complexity* yield optimal results in many cases.



- GP with MCKP outperforms SVR, KNN, Ridge, and DT in sample-limited scenarios.
- Ensemble Learning: Combining MCKP with HMR yields even better results.



(a) Median RSE of different learning methods.



(b) Pairwise statistical comparison of different learning methods.

Key Takeaways

- **Ensemble Model:**
  - If an ensemble model is acceptable, combine MCKP with any existing knee point selection method for enhanced performance.
- **Single Model:**
  - If an ensemble model is not suitable, use cross-validation on the training set to decide between MCKP and existing knee point selection methods.

Thanks for listening!

Email: Hengzhe.zhang@ecs.vuw.ac.nz

GitHub Project: https://github.com/hengzhe-zhang/KneePoint-Python