Universität Siegen

mbe

Learning (Partial) Boolean Configuration Spaces: Insights and Challenges

Mathis Weiß

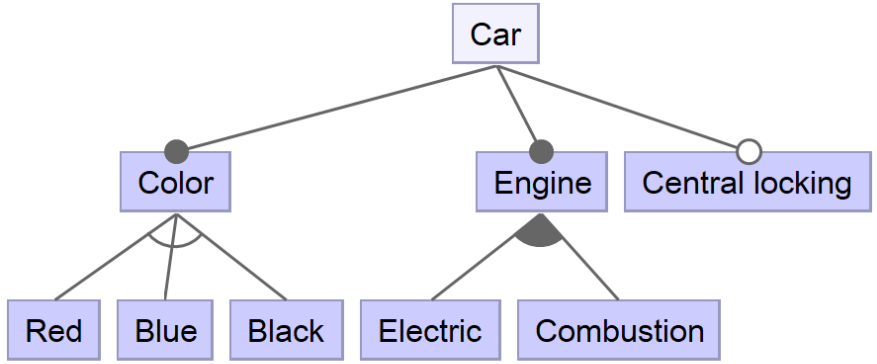
FOSD 2024 - Eindhoven
09. – 12. April 2024

Throwback FOSD 2023 – Comparing Efficiency and Effectiveness of Feature Model Synthesis and Feature Model Learning

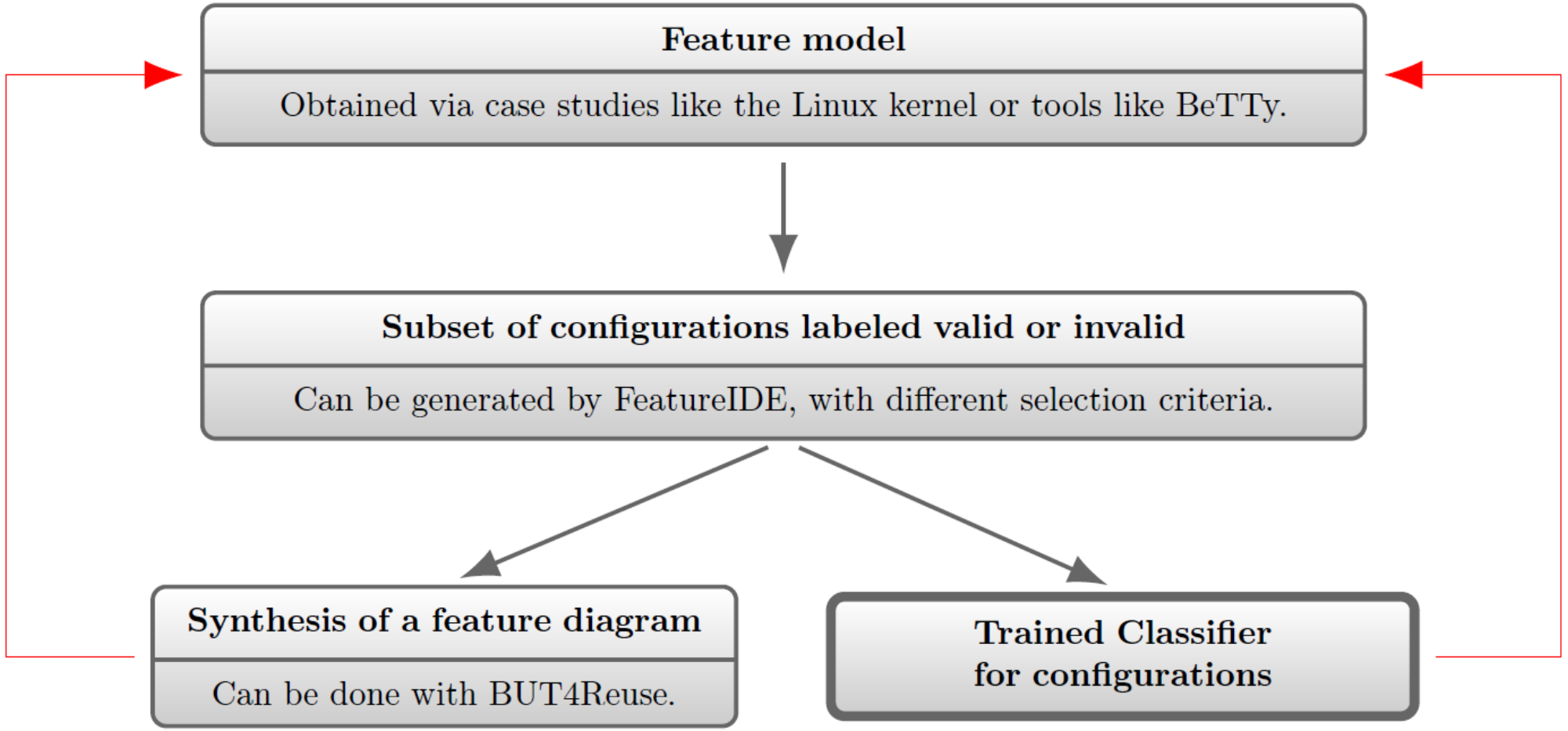
Unstructured Product Family

Reconstruction

Feature Model



Throwback FOSD 2023 – Comparing Efficiency and Effectiveness of Feature Model Synthesis and Feature Model Learning



Challenges and Insights

Challenge – Small Amount of Training Data

- Machine learning requires large amount of data

Possible solution:

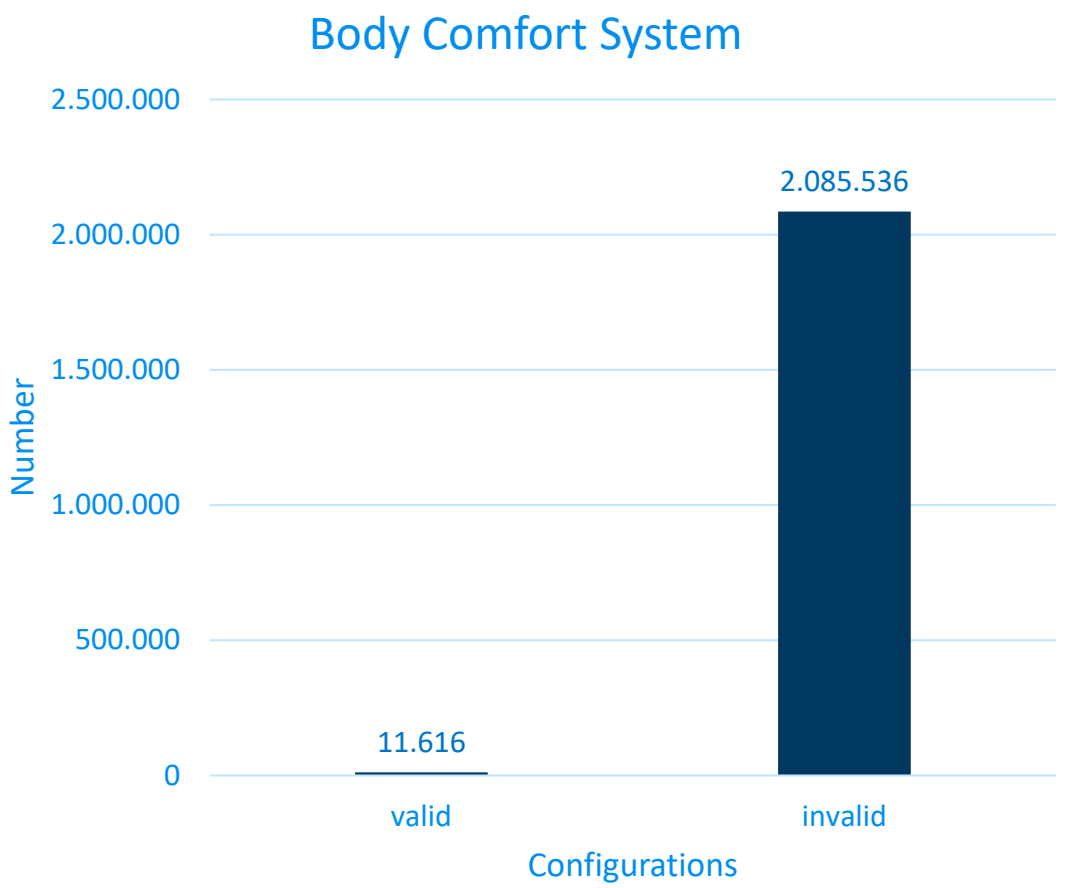
- Self-Supervised Learning
  - Re-Use learned labels as training data
  - Train in multiple iterations

- **Did not work in our use case.**

Challenge – Imbalance between Classes

- Over-Sampling of valid Configurations:
  - By duplication – no
  - Synthetic oversampling – no
- Under-Sampling of invalid Configurations

- Weighting of each instance per class
  - Use the right optimization metric:
    - At least: “Balanced Accuracy“
    - Better: Combination of precision and recall



SAT-Solver-based Analysis vs. Machine Learning-based Analysis

Types of analysis	SAT	Machine Learning (black box)
Validity of complete Configurations	Yes	Yes
Validity of partial Configurations	Yes	No
Sampling	Yes	No
Counting	Yes	No
Intersection, union, difference etc. of multiple models	Yes	No

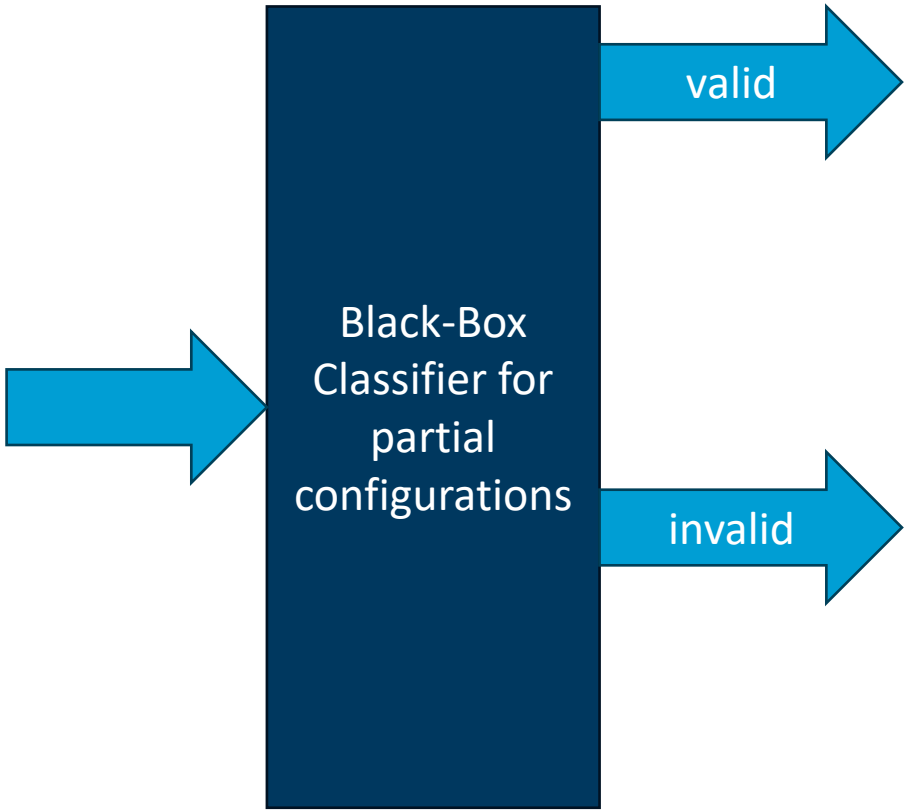
A (black-box) classifier for partial configurations would enable the analysis of partial configuration and e.g., sampling

Partial Configurations

New Black-Box Classifier for valid Configurations

Partial Configuration

A	B	C
1	0	?



If at least one valid completion exists:

A	B	C	valid
1	0	0	no
1	0	1	yes

If all completions are invalid e.g.:

A	B	C	valid
1	0	0	no
1	0	1	no

Using partial Configurations as Training Data is not usefull

Problems:

- The classifier is likely not consistent:
  - The classifier has no understanding of logic

Training Data			
A	B	C	valid
1	0	?	no

Model Prediction			
A	B	C	valid
1	0	0	yes
1	0	1	yes

- If refined knowledge about partial invalid configurations exists – ML is unnecessary.
- Use the knowledge to build a model (e.g., containing the constraint "$\neg A \vee B$")

Why Bother?

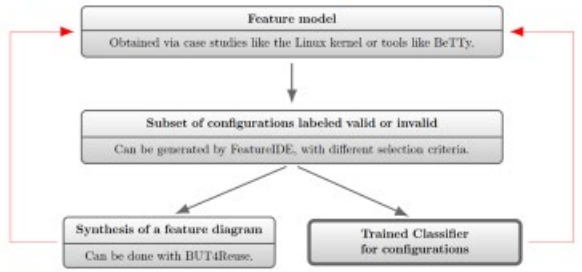
- Applicable to all kinds of configuration spaces
  - Tristate, int, hex, string -> Linux Kernel
  - Cardinalities
  - Features over continuous value domains

- Applicable for large configuration spaces

Learning (Partial) Boolean Configuration Spaces: Insights and Challenges

Introduction

Throwback FOSD 2023 – Comparing Efficiency and Effectiveness of Feature Model Synthesis and Feature Model Learning



Feature Model Learning

Challenge – Imbalance between Classes

- Over-Sampling of valid Configurations:
  - By duplication – no
  - Synthetic oversampling – no
- Under-Sampling of invalid Configurations

- Weighting of each instance per class
  - Use the right optimization metric:
    - At least: "Balanced Accuracy"
    - Better: Combination of precision and recall

Partial Configurations

SAT-Solver-based Analysis vs. Machine Learning-based Analysis

Types of analysis	SAT	Machine Learning (black box)
Validity of complete Configurations	Yes	Yes
Validity of partial Configurations	Yes	No
Sampling	Yes	No
Counting	Yes	No
Intersection, union, difference etc. of multiple models	Yes	No

A (black-box) classifier for partial configurations would enable the analysis of partial configuration and e.g., sampling

Partial Configurations

New Black-Box Classifier for valid Configurations