SLIMMER, outsmarting SLIM

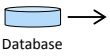
Manan Gandhi
Saarland University
Cluster of Excellence, MMCI

Jilles Vreeken
Saarland University
Cluster of Excellence, MMCI
Max-Planck Institute for Informatics

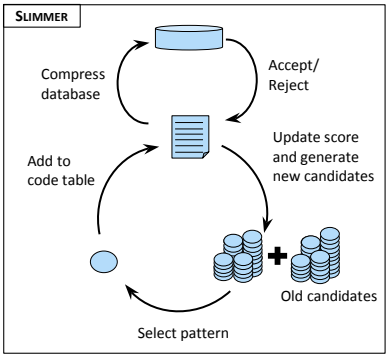
CLUSTER OF EXCELLENCE

MDL and SLIM

- **Minimum Description Length (MDL) principle:** Given a dataset D and a collection of models H, the best model H is the one that minimizes $L(H) + L(D|H)$



- **SLIM:** For a transaction database, finds the set of itemsets that together describe the data best



Database	SLIM		SLIMMER	
	Accuracy	Time (s)	Accuracy	Time (s)
Adult	**80.6**	139	80.5	7
Chess	**52.8**	20	**52.8**	3
Connect	65.2	1045	**65.6**	65
Ionosphere	89.5	32	**91.7**	1
Mushroom	**100.0**	9	**100.0**	3
Pen Digits	**95.6**	26	95.3	3
Waveform	73.5	46	**74.1**	3

As good, way faster!

SLIMMER

Stricter Candidate Estimation

- **SLIM** Considers all candidates that reduce $L(D|CT)$
- **SLIMMER** Considers only candidates that reduces $L(D, CT)$

Caching Candidate Scores

- **SLIM** Re-generates **all** candidates after every acceptance
- **SLIMMER** Generates only new candidates, **efficiently updates** scores for old candidates

Altogether, an **order of magnitude** faster than SLIM

Thresholding

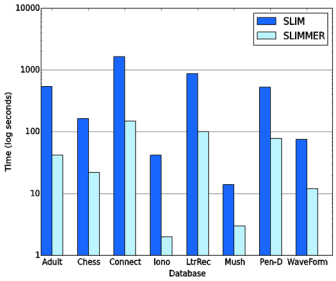
- **SLIM** Converges slowly: evaluates overly many candidates
- **SLIMMER** Avoids bad candidates by requiring minimal quality

Stop compression when gain in $L(D|CT) <$ threshold

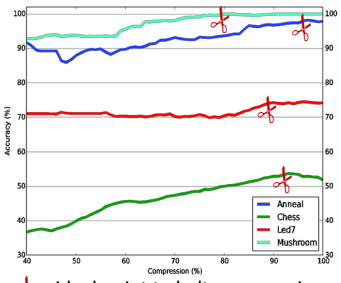
Thresholding **halves** run time—*without* harming classification

Classification
SLIMMER with 1 bit threshold is **10 up to 20 times faster** than **SLIM** *without* harming accuracy significantly

Compression time

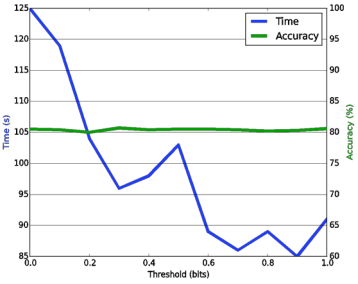


Classification

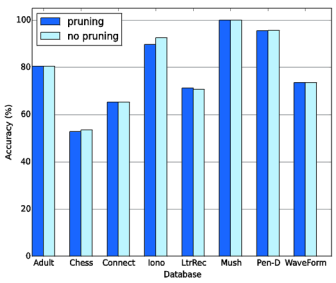


→ ideal point to halt compression

Time, Accuracy vs. Threshold



Accuracy



Classification

- Generate code tables per class, assign label of best compressor
- Accuracy **on par with state of the art classifiers**

Problems in Paradise

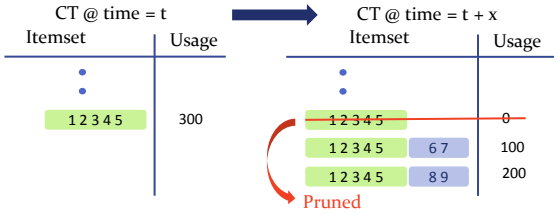
- Slow in practice: Long time to converge
- SLIM is heuristic: can overfit

Solution to both:

- Early-stop compression: avoids overfitting

Pruning

Pruning removes patterns that harm compression. However, these patterns are often helpful for classification



SLIMMER at work

- **Pruning** : No
- **Threshold** : 1 bit

Result:
Outperforms (on average) **every other** SLIM classifier

UNIVERSITÄT DES SAARLANDES
max planck institut informatik
Deutsche Forschungsgemeinschaft DFG