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

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

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

---

**Accuracy**

<table>
<thead>
<tr>
<th>Database</th>
<th>SLIM</th>
<th>SLIMMER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>80.6</td>
<td>80.5</td>
</tr>
<tr>
<td>Chess</td>
<td>52.8</td>
<td>52.8</td>
</tr>
<tr>
<td>Connect</td>
<td>65.2</td>
<td>65.6</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>89.5</td>
<td>91.7</td>
</tr>
<tr>
<td>Mushroom</td>
<td>100.0</td>
<td>99.0</td>
</tr>
<tr>
<td>Pen Digits</td>
<td>95.6</td>
<td>95.3</td>
</tr>
<tr>
<td>Waveform</td>
<td>73.5</td>
<td>74.1</td>
</tr>
</tbody>
</table>

**Time, Accuracy vs. Threshold**

---

**Compression time**

---

**Classification**

---

**Pruned**

---

**Usage**

---