CLIP4 Inductive Machine Learning Algorithm

Lukasz A. Kurgan
Advisor: Dr. Krzysztof J. Cios

Intelligent Systems Laboratory
Department of Bioengineering

Presentation Outline

• Goal

• Introduction

• CLIP4 Algorithm

• Results and Summary
Goal

• Development of a new inductive machine learning algorithm CLIP4 (Cover Learning using Integer Programming)

• This algorithm is a descendent of CLIP3 algorithm (Cios, Wedding and Liu in 1997)
Introduction

CLIP4 Algorithm:

• First, it divides data into noise-free subsets by using integer-programming mode, and a pruning technique to partition the data.

• Then, it generates rules from these subsets using a set covering technique.

CLIP4 Inductive Machine Learning Algorithm, Lukasz Kurgan
Introduction

CLIP4 algorithm can be used for:

- **classification problems**
  (generation of rules describing classification of the data)
- **data analysis problems**
  (finding significant attributes in the data)

Example applications:

- generation of diagnostic rules for heart perfusion classification, cancer type classification
- generation of compact and highly accurate user satisfaction instruments

CLIP4 Algorithm

- Defining research goals
- Initial implementation of the algorithm
- Literature search
- Iterative process of goals realization and re-definition based on literature and own ideas
- Validation and adjustments of the algorithm based on its performance
- User-friendly implementation of the algorithm
- Publication and popularization of the algorithm
CLIP4 Algorithm

Improvements and new features

1. User-friendly interface
2. Multi-class data handling
3. Nominal data encoding
4. Implementation

Outcome

- Faster data processing that allows it to analyze large data sets
- Generation of more accurate rules
- Ability to work with missing-value data, multi-class problems, to discretize the data
- Easy to use software
CLIP4 Algorithm

CLIP3 vs. CLIP4

• MONKS data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MONKS 1</th>
<th>MONKS 2</th>
<th>MONKS 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nr rules</td>
<td>Accuracy</td>
<td>Nr rules</td>
</tr>
<tr>
<td>CLIP3</td>
<td>4</td>
<td>100</td>
<td>18</td>
</tr>
<tr>
<td>CLIP4</td>
<td>4</td>
<td>100</td>
<td>9</td>
</tr>
<tr>
<td>CLIP3 (threshold 1)</td>
<td>4</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>CLIP3 (threshold 2)</td>
<td>4</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>IB3 with windowing</td>
<td>28</td>
<td>98.6</td>
<td>110</td>
</tr>
<tr>
<td>CN2</td>
<td>10</td>
<td>100</td>
<td>58</td>
</tr>
<tr>
<td>C4.5 decision tree</td>
<td>----</td>
<td>75.7</td>
<td>----</td>
</tr>
</tbody>
</table>

• Breast Cancer data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Breast Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nr rules</td>
</tr>
<tr>
<td>CLIP3</td>
<td>3</td>
</tr>
<tr>
<td>CLIP4</td>
<td>1</td>
</tr>
<tr>
<td>CLIP3 (threshold 0)</td>
<td>----</td>
</tr>
<tr>
<td>CLIP3 (threshold 1)</td>
<td>----</td>
</tr>
<tr>
<td>CLIP3 (threshold 2)</td>
<td>----</td>
</tr>
<tr>
<td>C4.5</td>
<td>----</td>
</tr>
</tbody>
</table>

Summary

The new algorithm was developed and implemented that is
• more accurate
• more data compatible
• very user-friendly

Feature goals
• To do more data benchmarking on large datasets
• To compare its results with other algorithms
• To publish the findings in a machine learning journal
• To popularize CLIP4 by offering it as a freeware to the research community
Major references


