Novel Methods for Graph Mining in Databases of Small Molecules

Andreas Maunz, andreas@maunz.de
Retreat Spitzingsee, 05.-06.04.2011
Tradeoff Overview

Data Mining Exercise: Find patterns / motifs in large graph databases as descriptors for classification or regression.

Integrated Algorithms
BBRC [1], LAST-PM [2]
Hypothesis Spaces

Frequent Subgraph Mining is anti-monotonic w.r.t. refinement. Subgraphs form a partial order.
Hypothesis Spaces

- Schematic depiction of the partial order.
- Not restricted to sub-graphs. (applies to general sets with the subset relation)

Novel Methods for Graph Mining in Databases of Small Molecules
OPEN Patterns

- State of the art compression method.
- Yields one subgraph for each level of frequency.
- Subsumes the set of all frequent subgraphs in terms of occurrences (lossless compression).
Typical Substructure Frequencies for Databases of Small Molecules

Real Trees; 85%

Cycle-closing Graphs; 10%

Paths; 5%

How to break up this large piece?
Backbone Refinement
Class Mining (BBRC)
The search space of Trees

The Backbone

Branching
Can derive **upper bound** for $\chi^2$ values of specializations of any subgraph [14].
BBRC Hypothesis Space

- Forms **classes** by prohibiting backbone changes during refinement.
- Search is **structurally partitioned**.
- Uses the **convex \( \chi^2 \) measure** to find most significant member of each class (allows for efficient pruning).
Coverage compared to all frequent and significant Trees.
Nearest neighbor predictions
all: all predictions
AD: top 80% confidence predictions
wt.: predictions weighted by confidence

<table>
<thead>
<tr>
<th></th>
<th>Sign. Tr.</th>
<th>Open Tr.</th>
<th>BBRC-R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>74.6</td>
<td>75.5</td>
<td>74.6</td>
</tr>
<tr>
<td>SM</td>
<td>80.7</td>
<td>80.6</td>
<td>79.4</td>
</tr>
<tr>
<td>wt.</td>
<td>86.8</td>
<td>84.5</td>
<td>85.4</td>
</tr>
<tr>
<td>all</td>
<td>64.4</td>
<td>64.5</td>
<td>67.2</td>
</tr>
<tr>
<td>RC</td>
<td>70.0</td>
<td>68.7</td>
<td>70.4</td>
</tr>
<tr>
<td>wt.</td>
<td>81.8</td>
<td>80.0</td>
<td>82.2</td>
</tr>
<tr>
<td>all</td>
<td>73.3</td>
<td>71.5</td>
<td>71.7</td>
</tr>
<tr>
<td>MoC</td>
<td>75.7</td>
<td>74.4</td>
<td>76.5</td>
</tr>
<tr>
<td>wt.</td>
<td>83.7</td>
<td>80.8</td>
<td>82.0</td>
</tr>
<tr>
<td>all</td>
<td>71.9</td>
<td>70.2</td>
<td>70.3</td>
</tr>
<tr>
<td>MuC</td>
<td>75.6</td>
<td>73.5</td>
<td>74.1</td>
</tr>
<tr>
<td>wt.</td>
<td>83.5</td>
<td>81.3</td>
<td>84.9</td>
</tr>
</tbody>
</table>

Novel Methods for Graph Mining in Databases of Small Molecules
Latent Structure Pattern Mining (LAST-PM)
LAST-PM Hypothesis Space

- **Observation**: Subgraphs significant for the same target class form *patches* in the search space.

- Subgraphs within a patch are often very similar, patches may have branches.

- **Idea**: Merge patch members to a single pattern.
Chemical Example:

**Conclusion**: The electrostatic properties decrease the ability of both drugs to cross membranes in the body, such as the blood-brain barrier.

Indeed, the two molecules are inactive for the blood-brain barrier endpoint, see Hu et al., *J Chem Inf*, 2005 [3].

Two negatively charged sites is a clear indication for a polar molecule. The difference in atoms (arrow) can be neglected.
LAST-PM executes a three-step pipeline on ground Patterns:

Align ground Patterns

Stack to weighted edge graph.

Compress by SVD to reveal latent structure.

The latent structure pattern contains characteristics of the ensemble of ground Patterns, such as weight, core size, and "heavy" regions.
Conflicting Patterns:
Distinct nodes/edges with the same position.

Resolved by logical OR  ⇔  Ambiguous positions.
## LAST-PM Experiments

<table>
<thead>
<tr>
<th>ID</th>
<th>Biological Endpoint</th>
<th>Study Reference</th>
</tr>
</thead>
</table>
LAST-PM Experiments

Crossvalidation: repeated 10-fold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LAST-PM %Test</th>
<th>ALL %Test</th>
<th>BBRC %Test</th>
<th>MOSS %Test</th>
<th>SLS %Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>bloodbarr</td>
<td>72.20</td>
<td>70.49</td>
<td>68.50</td>
<td>67.49</td>
<td>70.4</td>
</tr>
<tr>
<td>nctrer</td>
<td>80.22</td>
<td>79.13</td>
<td>80.22</td>
<td>77.17</td>
<td>78.4</td>
</tr>
<tr>
<td>yoshida</td>
<td>69.81</td>
<td>65.19</td>
<td>65.96</td>
<td>66.46</td>
<td>63.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LAST-PM # Patterns (t)</th>
<th>ALL # Patterns (t)</th>
<th>Ratio #Patterns/t</th>
</tr>
</thead>
<tbody>
<tr>
<td>bloodbarr</td>
<td>249 (1.23s)</td>
<td>1613 (0.36s)</td>
<td>0.51</td>
</tr>
<tr>
<td>nctrer</td>
<td>193 (12.49s)</td>
<td>22942 (0.13s)</td>
<td>0.81</td>
</tr>
<tr>
<td>yoshida</td>
<td>124 (0.28s)</td>
<td>462 (0.09s)</td>
<td>0.84</td>
</tr>
</tbody>
</table>

<1: positive tradeoff

Novel Methods for Graph Mining in Databases of Small Molecules
LAST-PM Experiments

Performance on yoshida and bloodbarr test sets (Hu et al.).

- bloodbarr (Hou and Xu): LAST-PM improved phys-chem descriptor model in 10-fold CV.
Backbone Refinement Class Representatives

• Structurally heterogeneous descriptors, compression by structural invariant (backbone constraint)
• Good dataset coverage, robust against increasing minimum frequencies
• Applicable to large-scale graph databases through a novel statistical pruning technique
• Discriminative potential similar to complete set of trees, but significantly better than open trees.
Latent Structure Pattern Mining

- Describing latent patterns in the data.
- More expressive than most other subgraph descriptors, including the complete set of ground Patterns from which they were derived.
- Very good performance on external test sets that have been difficult for structural Patterns.
- Favorable tradeoff between Pattern reduction and runtime.
Acknowledgements

Supported by the EU seventh framework programme under contract no Health-F5-2008-200787 (OpenTox).

Visit the project websites (source code, API, how to reproduce results,…):

http://bbrc.maunz.de | http://last-pm.maunz.de

Thank you for your attention. I will be happy to answer your questions.
References


