An expressive dissimilarity measure for relational clustering using neighbourhood trees

Sebastijan Dumančić, Hendrik Blockeel

DTAI, CS Department, KU Leuven

ECML PKDD 2017, Journal track
1. Overture

2. How do we do it now?

3. An expressive dissimilarity for relational data

4. Experiments and results

5. Summary
1 – Identifying groups in data

Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
1 – Identifying groups in data

Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
1 – Which clustering is correct?

Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
Clustering is fundamental, but ill-defined problem
1 – What about relational data?

Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
Machine learning with a powerful knowledge representation language
- usually based on first-order logic

Common representation for:
- vectors
- graphs
- sequences
- ...

... with a unifying reasoning and learning engine
1 – Many faces of relational data

Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
1 Overture

2 How do we do it now?

3 An expressive dissimilarity for relational data

4 Experiments and results

5 Summary
<table>
<thead>
<tr>
<th>Hybrid similarities</th>
<th>Graph kernels</th>
<th>Relational similarities</th>
</tr>
</thead>
<tbody>
<tr>
<td>incorporate link information into attribute-based similarity</td>
<td>structural similarities of graphs</td>
<td>comparing logical constructs</td>
</tr>
<tr>
<td>measure the similarity of connected vertices</td>
<td>random walks, propagation of information</td>
<td>logical formulas in common, matching terms</td>
</tr>
</tbody>
</table>

Relational clustering over neighbourhood trees – S. Dumančić, H. Blockeel
<table>
<thead>
<tr>
<th>Hybrid similarities</th>
<th>Graph kernels</th>
<th>Relational similarities</th>
</tr>
</thead>
<tbody>
<tr>
<td>incorporate link information into attribute-based similarity</td>
<td>structural similarities of graphs</td>
<td>comparing logical constructs</td>
</tr>
<tr>
<td>measure the similarity of connected vertices</td>
<td>random walks, propagation of information</td>
<td>logical formulas in common, matching terms</td>
</tr>
</tbody>
</table>

Impose a fixed bias
3 – Outline

1. Overture
2. How do we do it now?
3. An expressive dissimilarity for relational data
4. Experiments and results
5. Summary
3 – How similar are ProfA and ProfB?
A similarity measure for relational data should:

- incorporate multiple views of similarity
- be easily adaptable
- take attributes and relationships into account
- be insensitive to neighbourhood size
- be efficient
Neighbourhood trees summarize the neighbourhood of an instance/example.
Neighbourhood trees summarize the neighbourhood of an instance/example.

**Similarity of instances** = similarity of their neighbourhood trees.
Decompose NTs into semantic parts

\[ \text{similarity} = \text{linear combination of similarities of individual semantic parts} \]

\[ (w_1, w_2, w_3, s_4, w_5) \]
Decompose NTs into semantic parts

\[ \text{similarity} = \text{linear combination of similarities of individual semantic parts} \]

\[ (w_1, w_2, w_3, s_4, w_5) \]
Decompose NT in multisets of:

- attribute
- edge labels
- vertex identities

per level and vertex type

Multiset of edge labels (level 1):
\{ (Advised,2), (Advised,2), (TaughtBy,2) \}

Compare two multisets, \( A \) and \( B \) with \( \chi^2 \) distance

\[
\chi^2(A, B) = \sum_{x \in A \cup B} \frac{(f_A(x) - f_B(x))^2}{f_A(x) + f_B(x)}
\]
Many of the existing similarities are a special case:

- hybrid similarities
- relational similarities

... or they can be defined over neighbourhood trees (graph kernels) with different biases:

- makes it easier to compare the imposed biases
Many of the existing similarities are a special case:

- hybrid similarities
- relational similarities

... or they can be defined over neighbourhood trees (graph kernels) with different biases:

- makes it easier to compare the imposed biases

Additionally: effective - linear in the number of unique elements in a multiset
1. Overture

2. How do we do it now?

3. An expressive dissimilarity for relational data

4. Experiments and results

5. Summary
Datasets:
- IMDB
- UWCSE
- Mutagenesis
- WebKB
- TerroristAttacks

Questions:
- Quality of the obtained clustering?
- Are different views really necessary?
- Can we learn the bias from data?
- Can we learn the bias from labels?

- combined with spectral and hierarchical clustering
- a wide range of existing similarity measures
- performance measure: ARI/Accuracy
Takeaway message: incorporating multiple biases consistently performs well
Takeaway message: relational data requires multiple views of similarity in order to find informative clusters.
ReCeNT with $w_i = 0.2$

vs.

AASC + ReCeNT

AASC - given multiple similarity matrices, find an optimal combination for clustering

barely any benefit

Huang, Chuang, Chen: Affinity Aggregation for Spectral Clustering
Similarity measure in combination with a kNN (parameters optimised with CV)

Takeaway message: when labels are provided, ReCeNT outperforms the competing similarities
1. Overture

2. How do we do it now?

3. An expressive dissimilarity for relational data

4. Experiments and results

5. Summary
A similarity measure for relational data that:

- is versatile (meta-similarity)
- easily adaptable
- efficient
- generalization of many existing structured/relational sims
- works well across many different tasks
A similarity measure for relational data that:

- is versatile (meta-similarity)
- easily adaptable
- efficient
- generalization of many existing structured/relational sims
- works well across many different tasks

**Code:** [https://dtai.cs.kuleuven.be/software/recent](https://dtai.cs.kuleuven.be/software/recent)

S. Dumancic, H. Blockeel: *Clustering-Based Unsupervised Relational Representation Learning with an Explicit Distributed Representation*, IJCAI ’17

S. Dumancic, H. Blockeel: *Demystifying Relational Latent Representations*, ILP ’17