Characteristic Relational Patterns

Arne Koopman & Arno Siebes
Algorithmic Data Analysis, Universiteit Utrecht
Characterising the Database

- Relational database models
  - Local models: frequent pattern mining
  - Global models: probabilistic relational model

- Characterising the database
  - Combine patterns to form a global model

- Experiments
Relational Databases

KDD cup relational databases
- Genes
- Hepatitis
- Financial

Financial database

ORDER
- orderID
- loanID
- accountID
- Bank-To
- Amount
- K-Symbol

682 records

ACCOUNT
- loanID
- accountID
- Frequency
- Date

827 records

CARD
- cardID
- loanID
- dispositionID
- Type
- IssueDate

36 records

DISPOSITION
- loanID
- accountID
- dispositionID
- clientID
- Type

827 records
Local Models

Frequent pattern mining: too many patterns!
Global Model

<table>
<thead>
<tr>
<th>ACCOUNT</th>
<th></th>
<th>LOAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>accountID</td>
<td>Date</td>
<td>loanID</td>
</tr>
<tr>
<td>10</td>
<td>2009</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>2009</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>102</td>
</tr>
</tbody>
</table>

Probabilistic Relational Model

- Probabilistic Relational Models: local co-occurrence information lost
Combined Model

Relational Code Table: compact and lossless description of the complete database
RDB-Krimp selects patterns that describe the database well

Candidates are frequent relational patterns

Describing patterns are placed in a code table
- Shortness code length proportional to usage
Compression - 1

- Code tables encode the database
  - Increasingly better encoded sizes for lower minimum supports
Code tables encode the database
- Increasingly better encoded sizes for lower minimum supports
- Candidate set grow exponentially
- Code tables stay compact
Pattern Languages

**local table**

- **ACCOUNT**
  - Frequency = 3

**WARMR**

- **ACCOUNT** (Frequency = 3)
- **LOAN** (Duration = 12)

**ALL:** **FARMER w/o target**

- **ACCOUNT**
  - Frequency = 3
  - Duration = 12
- **LOAN**
  - Duration = 24
- **ORDER**
  - Amount = 1000
  - Duration = 6
- **LOAN**
  - Duration = 12

[Faculty of Science
Information and Computing Sciences]
Pattern Complexity

More complex patterns lead to better descriptions.

Thus, they encode MDL-relevant structure.
Conclusions

- Code tables describe the database while preserving local information

- Code tables stay compact
  - Stay compact for low minimum support values
  - Reductions up to 4 orders of magnitude

- Richer patterns lead to better models
  - Smaller encodings
  - Better descriptions without target tables
Questions?
Database Encoding

Reordering allows for a lossless encoding
We obtain better encodings without a target table
In all cases reductions are obtained
Additional rich patterns lead better encodings