Frequent Relational Pattern Mining

Martin Svatoš

October 13, 2016
Motivation for frequent pattern mining

- find set of items that are used frequently together
- market basket analysis
- fraud detection, technical dependence analysis, building classifiers
- association rules
item base $B = \{i_1, \ldots, i_m\}$

any subset of $B$ is called an item set, $I \subseteq B$

transaction database $T$ is a vector of transactions over $B$ (item sets)

transaction $t$ covers item set $i$ iff $i \subseteq t$.

support of an item set $X$: $\text{sup}(X) = \# \text{ transactions of } T \text{ covering } X$

association rule (AR) is $X \implies Y$ where $X$ and $Y$ are item sets and $X \cap Y = \emptyset$

confidence of AR $X \implies Y$: $\text{conf}(X \implies Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$

AR $X \implies Y$ informally: if the transaction covers $X$, then it is most likely that it also covers $Y$
Apriori algorithm [1]

- two phase algorithm for mining frequent patterns
- returns all item sets with support at least $sup_{min}$ (user specified parameter)
- 1) candidate generation
- 2) pruning
- level-wise algorithm by cardinality of item set
- anti-monotone property: no superset of an infrequent item set can be frequent
- another used approach, different from Apriori, is pattern growth
Motivation for relational domain

- mining frequent subgraphs [3]
- use gSpan, Gaston, LEAP, MoFa, Subdue, ... (pattern growth approach)
- or FSG, FFSM, SPIN, FTOSM, ... (Apriori approach)
So gSpan, mine frequent patterns from Bongard’s problems [2]

<table>
<thead>
<tr>
<th>Left Side</th>
<th>Right Side</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Pattern 1" /></td>
<td><img src="image2.png" alt="Pattern 2" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Pattern 3" /></td>
<td><img src="image4.png" alt="Pattern 4" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Pattern 5" /></td>
<td><img src="image6.png" alt="Pattern 6" /></td>
</tr>
</tbody>
</table>

- Find out pattern that is common in the left side but is missing in the right side.
- Converting the picture to graph?
ILP method based on Apriori *generate & prune* approach
finds relational frequent patterns, precisely Datalog queries
*atom set* instead of item set
support equals to number of different bindings of the query
monotone specialization
candidates are generated by extending older ones by allowed extension (language bias)
does not follow level-wise generation of candidates
*occurrence check*
Warmer [6]

- aimed to be more efficient than Warmr
- diagonally contained query within another
- to be more level-wise
- introduces operations for candidate generation: extension, join, selection, projection
- generate a small superset of all possible candidate queries and remove each query of which a generalization is not known to be frequent
Farmer [7]

- the same goal as Warmr
- use different *occurrence check*
- different structure for storing queries
- faster than Warmr
1. find $n$ frequent patterns
2. construct a boolean matrix $M$ such that $m_{ij}$ is true iff example $i$ contains pattern $j$; otherwise false
3. learn decision tree from the matrix
Descriptive language bias

- narrows the space of hypothesis
Beyond

- another ILP methods: SPADA
- application of relational frequent patterns: networking, healthcare, sales domains
- relational frequent patterns in stream data: Star FP Stream [8], SWARM


