Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2007
Blatt 2

Aufgabe 2.1
In the lecture you heard about the apriori-algorithm to create association-rules. Following
you should find association rules for the preferences of soccer-fans. In the table below (Note
the changed order of columns and rows in contrast to the lecture) you see which club a fan
prefers. This of course is just an excercise because in reality a fan just supports one club.

<table>
<thead>
<tr>
<th>club</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>F11</th>
<th>F12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schalke</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Gladbach</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bochum</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamburg</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlin</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hannover</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuernberg</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stuttgart</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Bayern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

(a) Create the sets of clubs which support is greater or equal 0.5, do the same for support
greater or equal 0.25. Give the canditates and large item sets (which really achieve the
minimum support) for every step.

(b) Give all rules with minimum support 0.25 (and then 0.5) and minimal confidence 0.9.
Calculate the confidence for every rule with minimum support of 0.5. Which fangroups
like each other the most?
Aufgabe 2.2
Show for the following conclusions if they hold. Give a proof or a counter-example. \( conf(r) \) stands for the confidence of rule \( r \), \( s(r) \) stands for the support.

(a) \( (conf(A \rightarrow B) = \alpha) \land (conf(B \rightarrow C) = \beta) \Rightarrow conf(A \rightarrow C) = \alpha\beta \)

(b) \( conf(A \rightarrow B) = conf(B \rightarrow A) \Rightarrow (h(A) = h(B)) \), where \( h(A) > 0 \) and \( h(B) > 0 \) mean the count of transactions in which \( A \) and \( B \) respectively occur.

(c) \( s(X \rightarrow Y) \geq s(X \rightarrow \emptyset)s(Y \rightarrow \emptyset) \)

Aufgabe 2.3
Use YALE to solve this exercise.

(a) Download the dataset mushrooms from http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS07/MATERIAL/mushrooms.xrff. Make an experiment, that just consists of an XrffExampleSource, and load the data. Which occurrence of the attribute ring-type is the most frequent one? How frequent is that occurrence? How are the classes of that dataset named?

(b) Add the rulelearner ConjunctiveRule as next operator and learn one association rule (use the standard settings). Which rule was learned?

(c) To evaluate the performance of the learner, you should use SimpleValidation. Learn on a training-set of 70\% (\( split\_ratio = 0.7 \)) and evaluate “Accuracy” and “Precision” on the rest of 30\% of the data. It is necessary to apply the learned model by the ModelApplier and evaluate the performance with the PerformanceEvaluator (you have to mark accuracy and precision). You have to mind that the needed operators have to be “hung” under the SimpleValidation operator. SimpleValidation allows only two sub-operators: one for the first and one for the second dataset. Use the OperatorChain-operator to use more than one operator.

Save all Yale-experiments and deliver them with your exercise.