

Übungen zur Vorlesung  
**Wissensentdeckung in Datenbanken**  
 Sommersemester 2006

Blatt 5

**Aufgabe 5.1**

Sie haben in der Vorlesung den Algorithmus Apriori zur Bestimmung von Assoziationsregeln kennengelernt (Folien Teil 13, Nr. 31 bis 37). Im folgenden sollen Assoziationsregeln über das Fernsehverhalten von Testpersonen gefunden werden. In der unten angegebenen Tabelle sehen Sie, welche Zuschauer welche Sendungen verfolgt haben.

Sendung	$z_1$	$z_2$	$z_3$	$z_4$	$z_5$	$z_6$	$z_7$	$z_8$	$z_9$	$z_{10}$	$z_{11}$	$z_{12}$
Wer will Millionen?	x	x					x		x	x		
Quizzy			x		x	x		x			x	x
Quiz Express		x			x		x			x		
Birkenallee	x							x			x	x
Gut Josepha			x		x	x		x			x	x
Gutes Bier, Schlechtes Bier	x	x	x				x		x	x	x	
Bettina Schläfer				x	x					x		
Nora am Abend		x		x	x		x	x		x		
Olli-G.-Show		x		x	x			x		x		

- (a) Bestimmen Sie alle Mengen von Fernsehsendungen mit einem support größer gleich 0.4, sowie mit support größer gleich 0.25. Geben Sie dabei für jeden Schritt die Kandidatenmenge sowie die Menge der large item sets (d.h. derjenigen Mengen, die tatsächlich den Mindestsupport erfüllen) an.
- (b) Geben Sie alle Regeln mit minimalem support 0.25/0.4 und minimaler confidence 0.8 an. Berechnen Sie für die Regeln mit minimalem support 0.4 außerdem jeweils die Konfidenz.

## Aufgabe 5.2

Zeigen Sie für folgende Aussagen, ob sie gelten oder nicht. Geben Sie dabei immer einen Beweis oder ein Gegenbeispiel an. Dabei bezeichnet  $conf(r)$  jeweils die Konfidenz der Regel  $r$ ,  $s(r)$  bezeichnet deren Support.

- (a)  $(conf(A \rightarrow B) = \alpha) \wedge (conf(B \rightarrow C) = \beta) \Rightarrow conf(A \rightarrow C) = \alpha\beta$
- (b)  $conf(A \rightarrow B) = conf(B \rightarrow A) \Rightarrow (h(A) = h(B))$ , wobei  $h(A) > 0$  und  $h(B) > 0$  die Anzahl der Transaktionen bezeichnen, in denen  $A$  bzw.  $B$  vorkommen.
- (c)  $s(X \rightarrow Y) \geq s(X \rightarrow \emptyset)s(Y \rightarrow \emptyset)$

## Aufgabe 5.3

Zur Bearbeitung dieser Aufgabe verwenden Sie bitte die in der Vorlesung vorgestellte Lernumgebung YALE, die in den Statistik-Pools vorinstalliert ist, die Sie aber auch unter der URL <http://yale.cs.uni-dortmund.de/> beziehen können. Unter dieser Adresse finden Sie auch eine Installationsanleitung und Dokumentation.

- (a) Laden Sie sich den Datensatz `mushrooms` unter der URL  
<http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/kdd2006/kdd.html>  
herunter.  
Erstellen Sie ein Experiment, das lediglich aus einer `ArffExampleSource` besteht und lesen die Daten ein. Welche Ausprägung des Attributes `ring-type` tritt im Datensatz am häufigsten auf? Wie heißen die beiden Klassen des Datensatzes?
- (b) Fügen Sie nun den Regellerner `ConjunctiveRule` als weiteren Operator in das Experiment ein, und lassen Sie eine Entscheidungsregel mit Standardeinstellungen lernen. Hierfür ist es nötig, ins Feld `label_attribute` der `ArffExampleSource` den Wert "class" einzutragen. Welche Regel wurde gelernt?
- (c) Zur Evaluierung der Güte des Lernverfahrens auf dem Datensatz erstellen Sie nun bitte ein Experiment mit einer `SimpleValidation`. Lernen Sie auf einer Trainingsmenge von 70% und evaluieren Sie die Maße "Accuracy" und "Precision" auf den verbleibenden 30% der Daten. Hierfür ist es notwendig, das Model mit einem `ModelApplier` anzuwenden und mit einem `PerformanceEvaluator` zu evaluieren.

Speichern Sie für alle Teilaufgaben jeweils die YALE-Experimente. Geben Sie die XML-Dateien bitte mit ab.

Hinweis: Falls Ihnen bei der Erstellung eines Experiments Parameter fehlen, so befinden Sie sich vermutlich im Modus "Beginner" (B). Klicken Sie auf den Buchstaben oben rechts, um in den Modus "Expert" (E) zu wechseln.