Prof. Dr. Katharina Morik, Prof. Dr. Uwe Ligges
Gero Szepannek, Felix Jungermann

Dortmund, 17. April
Abgabe: bis 24.04., 23.59 Uhr an
jungerma@ls8.cs.uni-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2007

Blatt 2

Aufgabe 2.1

In der Vorlesung haben Sie den `apriori`-Algorithmus zur Bestimmung von Assoziationsregeln kennengelernt. Im Folgenden sollen Assoziationsregeln für die Sympathien von Fussballfans gefunden werden. In der unten angegebenen Tabelle (Beachten Sie bitte die im Gegensatz zur Vorlesung geänderte Anordnung von Zeilen und Spalten!) sehen Sie, mit welchen Vereinen ein Fan sympathisiert. Dies ist natürlich nur eine Übungsaufgabe, die in keinster Weise der Realität entspricht, da ein Fan eigentlich nur einem Verein zugetan ist.

Verein	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Schalke	X	X	X	X	X	X	X	X	X		X	X
Gladbach		X		X	X	X			X	X	X	
Bochum	X					X	X	X				
Hamburg		X			X	X	X					
Berlin				X				X				
Hannover	X				X		X		X			
Nuernberg	X		X	X	X	X	X	X	X		X	X
Stuttgart			X	X				X	X			X
Bayern										X		

(a) Bestimmen Sie alle Mengen von Vereinen mit einem Support größer gleich 0.5, sowie mit einem 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 (und danach 0.5) und minimaler Konfidenz 0.9 an. Berechnen Sie für die Regeln mit minimalem Support 0.5 außerdem jeweils die Konfidenz. Welche zwei Fangruppen sind am stärksten befreundet?

Aufgabe 2.2
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 2.3
Zur Bearbeitung dieser Aufgabe verwenden Sie bitte die bereits bekannte Lernumgebung YALE.

(a) Laden Sie sich den Datensatz `mushrooms` unter der URL
`http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS07/MATERIAL/mushrooms.xrff`
herunter. Erstellen Sie ein Experiment, das lediglich aus einem `XrffExampleSource` besteht, und lesen Sie die somit die Daten ein. Welche Ausprägung des Attributes `ring-type` tritt im Datensatz am häufigsten auf? Wie oft tritt dieses Ausprägung im Datensatz 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 (d.h. Sie müssen keine Einstellungen ändern) lernen. 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% (`split_ratio` = 0.7) 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` (hier müssen Sie accuracy und precision auswählen) zu evaluieren. Beachten Sie, dass Sie die benötigten Operatoren in den Operator `SimpleValidation` "einhängen" müssen. `SimpleValidation` erlaubt nur zwei Unteroperatoren. Einen für die erste Datenmenge, den anderen für die restliche Datenmenge. Benutzen Sie den Operator `OperatorChain` um mehr als einen Operator für die jeweilige Datenmenge zu benutzen.

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