Prof. Dr. Claus Weihs, Prof. Dr. Katharina Morik
Julia Schiffner, Marco Stolpe

Dortmund, 14. Juli
Abgabe: bis zum 21.07. um 10.00 Uhr an
marco.stolpe@tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2009

Blatt 13

Aufgabe 13.1 (2 Punkte)
Der Cluster-Algorithmus k-Means partitioniert eine Menge unklassifizierter Beispiele so, dass sich Objekte innerhalb von Clustern ähnlicher sein sollen als solche aus unterschiedlichen Clustern.

(a) Warum kann man allein anhand dieses Kriteriums den Parameter k nicht mit Hilfe einer herkömmlichen Parameter-Optimierung bestimmen?

(b) Der k-Means-Algorithmus lässt sich gut auf Beispiele anwenden, die naturgemäß Gruppierungen bilden. Was passiert jedoch, wenn die Beispiele im Extremfall gleichverteilt sind? Lassen Sie dazu in RapidMiner einige 1000 gleichverteilte Beispiele in der Ebene erzeugen und clustern Sie diese mit dem Operator `KMeans` (z. B. für $k = 10$). Sehen Sie sich das resultierende Cluster-Modell an und erstellen Sie für die Data Table einen Scatter-Plot, der die Cluster in der Ebene anhand von Farben verdeutlicht. Interpretieren Sie das Ergebnis!

Aufgabe 13.2 (3 Punkte)
In der Vorlesung wurden mit Hilfe des Apriori-Algorithmus die häufigen Mengen einer binären Transaktionsdatenbank gefunden. Auf Basis dieser häufigen Mengen sind dann Assoziationsregeln generiert worden.

Gegeben sei die nachfolgende fiktive Aufstellung von Filmen, die von Zuschauern $z_1, \ldots, z_{10}$ besucht worden sind.

Titel	Jahr	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
Sissi	1955	1	0	1	0	0	1	0	1	0	0
Star Wars	1977	1	1	0	1	0	0	0	1	1	1
E.T. der Außerirdische	1982	1	1	0	1	0	1	1	1	1	1
Indiana Jones	1989	0	1	0	1	1	0	0	1	1	1
Otto - der Außerfriesische	1989	1	0	1	1	1	1	1	1	0	0
Waynes's World	1992	0	0	0	1	1	1	1	1	1	0
Bang Boom Bang	1999	0	0	0	1	1	1	1	1	0	1
Bridget Jones	2001	1	0	1	0	0	1	0	1	1	0
Simpsons (Film)	2007	0	1	0	1	0	0	1	1	1	1

(a) Bestimmen Sie mit Hilfe des Apriori-Algorithmus die häufigen Mengen mit minimalem Support von 0.4 und 0.6. Geben Sie dabei für jeden Schritt die Kandidatenmenge sowie die Menge der *large itemsets* (d. h. diejenigen Mengen, die den minimalen Support erfüllen) an.

(b) Geben Sie alle Regeln mit minimalem Support von 0.4 und einer Konfidenz von mindestens 0.8 an.

Aufgabe 13.3 (5 Punkte)
Betrachten Sie erneut die Datenbank der Kinogänger aus Aufgabe 2. Es sei ein minimaler Support von $\frac{2}{5}$ gegeben, der für alle nachfolgenden Aufgaben gelten soll. Für ein besseres Verständnis von FP-Growth könnte Ihnen der folgende Artikel von Nutzen sein: Han et. al. (2001) - *Mining frequent patterns without candidate generation - A frequent-pattern tree approach*. Data Mining and Knowledge Discovery (8), 53–87.

(a) Geben Sie die Transaktionstabelle mit nach Häufigkeit sortierten Items (innerhalb der Transaktionen) an!

(b) Bestimmen Sie die Header-Tabelle sowie den *FP-Tree* aus der angegebenen Transaktionstabelle.

(c) Bestimmen Sie alle *conditional pattern bases* zum *FP-Tree*.

(d) Bestimmen Sie nun zu den *conditional pattern bases* die *conditional FP-Trees*.

(e) Bestimmen Sie anhand der *conditional FP-Trees* rekursiv die *frequent patterns*. Zeigen Sie die Erfassung der *frequent patterns* jeweils an der Entwicklung der *conditional pattern bases* sowie den *conditional FP-Trees*.