Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungsprofessur
Wilhelmshöher Allee 73
34121 Kassel

Email: hotho@cs.uni-kassel.de
Tel.: ++49 561 804-6252

Dr. Andreas Hotho
Prof. Gerd Stumme

16.06.04

Knowledge Discovery

Übungsblatt 7

Sommersemester 2004

Aufgabe 1 Cluster-Verfahren

Ein Kaufhaus, das seine Kunden in fünf Gruppen klassifiziert hat, möchte eine Werbekampagne durchführen. Da es zu aufwendig wäre, für jede der fünf Gruppen ein spezifisches Werbekonzept zu konzipieren, sollen sie in zwei Hauptgruppen eingeteilt werden. Dazu hat man für die Repräsentanten der einzelnen Gruppe {1,2,3,4,5} die folgenden Abstände *d* ermittelt:

D(*x*,*y*)	**1**	**2**	**3**	**4**	**5**
1	0	2	2	17	16
2	2	0	4	9	10
3	2	4	0	13	10
4	17	9	13	0	1
5	16	10	10	1	0

a) Entwerfen Sie ein Verfahren, welches ausgehend von einer Anfangsklassifikation K^0 durch den Austausch von Elementen die Klassifikation iterativ bezüglich eines Güteindexes optimiert (Austauschverfahren).

b) Ausgehend von der Anfangsklassifikation K^0={{1,2},{3,4,5}} soll mit Hilfe des Austauschverfahrens die bestmögliche Klassifikation *K* mit dem Güteindex

$$b(K) = \sum_{C \in K} \left(\frac{1}{|C|} \sum_{x,y \in C} d(x,y) \right)$$

erstellt werden.

c) Welche anderen Verfahren hätte man zur Lösung der Aufgabe auch verwenden können? (Führen Sie ein Verfahren durch, und vergleichen Sie die Ergebnisse. Zusatzaufgabe!!!)

d) Wie kann das Kaufhaus die Ergebnisse zur Aufstellung der Marketingstrategien verwenden?

Aufgabe 2 verteilter K-Means

Das Clusterverfahren K-Means lässt sich sehr einfach und mit sehr gutem Laufzeitverhalten implementieren. Für sehr große Datenmengen möchte man aber nicht nur einen Rechner nutzen, sondern auf Rechnercluster zurückgreifen.

a) Ist dies möglich?
b) Erläutern Sie kurz die Hauptidee.
c) Geben Sie einen verteilten K-Means Algorithmus an!

Aufgabe 3: Hierarchisches Clustern

a) Was ist der Unterschied zwischen agglomerativen und divisiven Cluster-Verfahren?

b) Gegeben Sei die folgende (binäre) Attribut-Wert Matrix:

	preparable	buyable	cookable	fryable	drinkable
chicken	1	1	1	1	0
spaghetti	1	1	1	0	0
salad	1	1	0	0	0
juice	0	1	0	0	1
wine	0	1	0	0	1

Berechnen Sie das Ergebnis eines agglomerativen hierarchischen Cluster-Verfahrens jeweils mit single-linkage und complete-linkage als Strategien zur Berechnung der Ähnlichkeit zwischen zwei Clustern. Als Ähnlichkeitsmaß soll der Dice-Koeffizient verwendet werden:

$$sim_{DICE}(t_1, t_2) = \frac{2C}{A+B}$$

wobei C die Anzahl der gemeinsamen Features, A, B jeweils die Anzahl der Features von t_1 und t_2.

c) Versuchen Sie eine Interpretation dieser Clusterergebnisse zu geben. Welche Strategie (single linkage, complete linkage) liefert ihrer Meinung nach in Bezug auf diese Interpretation die beste Lösung?

Aufgabe 4 (Clusteranalyse)

Sie haben in der Vorlesung verschiedene Clusterverfahren kennengelernt. Erläutern Sie für jeden dieser Algorithmen das Hauptprinzip:

a) FCA
b) Hierarchisches Clustern (agglomerativ/divisiv)
c) K-Means
d) EM