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

24.09.04

Knowledge Discovery

Lösungsblatt 7

Sommersemester 2004

Aufgabe 3: Hierarchisches Clustern

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

Agglomerative Verfahren funktionieren bottom-up, d.h. sie fügen Elemente oder Cluster zu neuen Clustern zusammen. Am Anfang ist also jedes Element in einem Cluster. Divisive Verfahren operieren hingegen top-down d.h. sie spalten bestehende Cluster in Teilcluster. Dementsprechend sind am Anfang alle Element in einem Cluster.

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 ist und A, B jeweils die Anzahl der Attribute von t_1 und t_2 sind.

Ersteinmal müssen die Ähnlichkeiten zwischen den einzelnen Objekten berechnet werden (dabei kann ausgenutzt werden, dass der DICE-Koeffizient symmetrisch ist).

$$sim_{DICE}(chicken, spaghetti) = \frac{2*3}{7} \approx 0.86$$

$$sim_{DICE}(chicken, salad) = \frac{2*2}{6} \approx 0.66$$

$$sim_{DICE}(chicken, juice) = \frac{2*1}{6} \approx 0.33$$

$$sim_{DICE}(chicken, wine) = \frac{2*1}{6} \approx 0.33$$

$$sim_{DICE}(spaghetti, salad) = \frac{2*2}{5} = 0.8$$

$$sim_{DICE}(spaghetti, juice) = \frac{2*1}{5} = 0.4$$

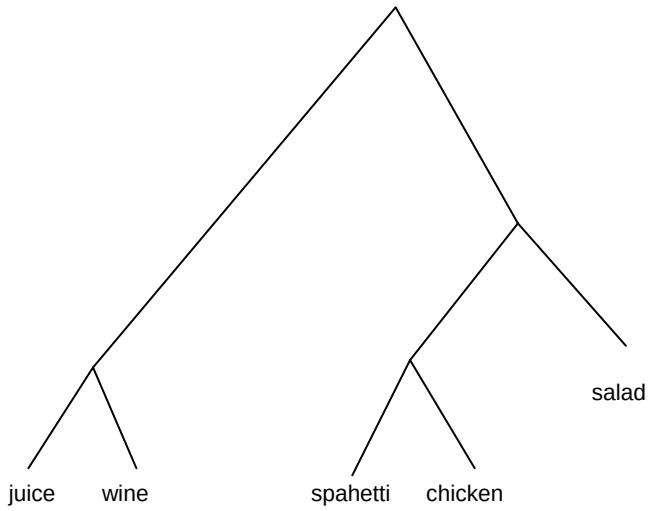
$$sim_{DICE}(spaghetti, wine) = \frac{2*1}{5} = 0.4$$

$$sim_{DICE}(salad, juice) = \frac{2*1}{4} = 0.5$$

$$sim_{DICE}(salad, wine) = \frac{2*1}{4} = 0.5$$

$$sim_{DICE}(juice, wine) = \frac{2*2}{4} = 1$$

Nun, unabhängig davon, ob man nun single-linkage oder complete-linkage, erhält man folgenden Cluster-Baum:



c) Versuchen Sie eine Interpretation dieser Clusterergebnisse zu geben. Welche Strategie (single linkage, complete linkage) liefert Ihrer Meinung für das konkrete Szenario die beste Lösung?

Das Clusterergebniss könnte als eine Hierarchie zwischen den gegebenen Termen interpretiert werden, d.h. das Cluster, was spaghetti und chicken enthält beinhaltet kochbare Lebensmittel. Das übergeordnete Cluster, was spagehtti, chicken und salad enthält, beinhaltet damit Lebensmittel im Allgemeinen und ist damit ein generelleres Cluster. So können also Objekte entsrpechend ihrer Eigenschaften durch die Cluster-Analyse hierarchisch angeordnet werden. Im konkreten Beispiel liefern single-linkage und complete-linkage das gleiche Ergebnis; es ist jedoch emprisch bewiesen, dass complete-linkage homogenere (d.h. in sich ähnlichere Cluster) bildet.

Aufgabe 4 (Clusteranalyse)

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

a) FCA ist ein konzeptuelles Cluster-Verfahren, wobei konzeptuell hier bedeutet, dass es nicht nur Objekte clustert, sondern auch erklärt, welche Eigenschaften die geclusterten Objekte gemeinsam haben. FCA basiert auf einem mengen-theoretischen Ansatz, d.h. es erzeugt eine Ordnung auf der Basis der Inklusionsbeziehungen zwischen den Attributmengen der Objekte.

b) Hierarchische Cluster-Verfahren operieren entweder top-down (divisiv) oder bottom-up (agglomerativ). Beim top-down Vorgehen startet man mit einem grossen Cluster und teilt diesen immer weiter auf bis jedes Element in einem Cluster landet. Beim bottom-up Vorgehen startet man hingegen mit einem Cluster pro Element und fügt so lange Cluster zusammen bis wieder ein allgemeiner Cluster entsteht.

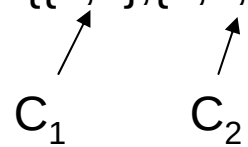
c) K-Means ist ein Zentroid-basiertes Verfahren. Beim K-Means wird die Anzahl der Cluster von Anfang an vorgegeben und die Zentroide (Mittelpunkte) jedes Clusters werden initialisiert (z.B. zufällig). Dann werden die Elemente diesen Zentroiden zugewiesen. Dann werden die Clustermittelpunkte neu berechnet und das Verfahren so lange iteriert bis die Mittelpunkte sich nicht mehr ändern oder die Lösung gut genug ist oder wenn sonstige Abbruchkriterien erfüllt sind.

d) EM (Expectation Maximization) ist ein probabilistisches Clusterverfahren. Im Gegensatz zu den vorhergehenden Verfahren wird ein Element nicht deterministisch einem Cluster zugewiesen, sondern für jeden Cluster die Wahrscheinlichkeit angegeben, dass ein Objekt dazugehört. Dabei soll eine gegebene Wahrscheinlichkeits- verteilung approximiert werden. Im so genannten Expectation-Schritt werden die Wahrscheinlichkeiten berechnet, im Maximization-Schritt wird dann die Wahrscheinlichkeitsverteilung selbst neu berechnet.

Übung 7 - Aufgabe 1 (Lösungsvorschlag)

a) z.B. gegenseitiges Vertauschen eines Elements

- K^0={{1,2},{3,4,5}}



- Vertausche Element aus C_1 mit Element aus C_2 so, dass Güteindex sich verbessert. (siehe b))

Übung 7 - Aufgabe 1

(Lösungsvorschlag)

b) Klassifizierung der Objektmenge {1,2},{3,4,5} mit D=

0	2	2	17	16
2	0	4	9	10
2	4	0	13	10
17	9	13	0	1
16	10	10	1	0

$$b(K^0) = \frac{1}{2}2 + \frac{1}{3}(13+10+1) = 1+8 = 9$$

Element	Permutation	b()	
{1 }	{2},{1,3,4,5}	17,25	>9
{2}	{1},{2,3,4,5}	12,25	>9
{3}	**{1,2,3},{4,5}**	**3,17**	**<9**
{4}	{1,2,4},{3,5}	14,33	>9
{5}	{1,2,5},{3,4}	15,83	>9

K^1={{1,2,3},{4,5}}

Keine weitere Verbesserung möglich!

Übung 7 - Aufgabe 1 (Lösungsvorschlag)

c) z.B. Verwendung des hierarchischen **single-linkage Verfahrens** mit

Verschiedenheitsfunktion $D(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y)$

Klassifizierung der Objektmenge {1,2,3,4,5} mit D=

0	2	2	17	16
2	0	4	9	10
2	4	0	13	10
17	9	13	0	1
16	10	10	1	0

Iterationsschritt 1:

$$K^0 = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$$

$$v\{C_{t_1}^0, C_{t_2}^0\} = d_{t_1 t_2} \Rightarrow \min_{C_{t_1}^0, C_{t_2}^0 \in K^0} = v(C_{t_1}^0, C_{t_2}^0) = d_{45} = 1$$

$$\Rightarrow K^1 = \{\{1\},\{2\},\{3\},\{4,5\}\} = \{K_1^1, ..., K_4^1\}$$

Übung 7 - Aufgabe 1 (Lösungsvorschlag)

Iterationsschritt 2: $K^1 = \{\{1\},\{2\},\{3\},\{4,5\}\}$

$$(v(C^1_{t_1}, C^1_{t_2}))_{\substack{t_1=1\ldots4 \\ t_2=1\ldots4}} = \begin{pmatrix} 0 & & & \\ 2 & 0 & & \\ 2 & 4 & 0 & \\ 16 & 9 & 10 & 0 \end{pmatrix}$$

$$\min_{C^1_{t_1}, C^1_{t_2} \in K^1} = v(C^1_{t_1}, C^1_{t_2}) = v(C^1_1, C^1_2) = v(C^1_1, C^1_3) = 2$$

$$\Rightarrow \textit{willkürlich}[1,2]: K^2 = \{\{1,2\},\{3\},\{4,5\}\} = \{K^2_1, \ldots, K^2_3\}$$

Iterationsschritt 3: $K^2 = \{\{1,2\},\{3\},\{4,5\}\}$

$$(v(C^2_{t_1}, C^2_{t_2}))_{\substack{t_1=1,2,3 \\ t_2=1,2,3}} = \begin{pmatrix} 0 & & \\ 2 & 0 & \\ 9 & 10 & 0 \end{pmatrix}$$

$$\min_{C^2_{t_1}, C^2_{t_2} \in K^2} = v(C^2_{t_1}, C^2_{t_2}) = v(C^2_1, C^2_2) = 2$$

$\Rightarrow K^3 = \{\{1,2,3\},\{4,5\}\}$ (gleiches Ergebnis!)

d) Verwendung der Ergebnisse:

Entwerfe zwei Werbekonzepte, eines für Kunden der Gruppen 1+2+3 und eines für die Kunden der Gruppen 4+5.
Eine gruppenspezifische Marketingstrategie lässt sich mit den vorliegenden Daten nicht aufstellen, da über die Kaufeigenschaften und Charakteristiken der Gruppen keine Informationen vorliegen.

Idee: Verteile die Objekte (die in sehr großer Zahl vorliegen) gleichmäßig auf k Rechner.
Jeder der Rechner berechnet für die ihm zugeteilten Objekte Clusterzentren und ordnet die Objekte den tatsächlichen Clusterzentren zu, die sich durch Mittelwert-bildung aus den einzeln errechneten Zentren ergeben.

Bildung von Rechnercluster ist möglich aufgrund der Assoziativität der Clusterzentren:

Cluster C = $\{\vec{x}_1;...;\vec{x}_n\}$

Clusterzentrum $$\bar{x}_c^i = \frac{1}{n}\sum_{j=1}^{n} x_j^i = \frac{1}{n}\sum_{j=1}^{n_1} x_j^i + \frac{1}{n}\sum_{j=n_1+1}^{n_2} x_j^i + ... + \frac{1}{n}\sum_{j=n_{(k-1)}+1}^{n} x_j^i$$

Rechner 1 Rechner 2 ... Rechner k