Lehrstuhl für Statistik und Ökonometrie
Übung zur Datenanalyse

Aufgabenserie 7: Clusteranalyse

Aufgabe 16

Die Betreiber eines sozialen Netzwerks vermuten systematische Nutzungsunterschiede zwischen den registrierten Nutzern. Im Rahmen der Untersuchung dieser Systematik soll eine Clusteranalyse durchgeführt werden.
Zunächst werden hierfür die folgenden Variablen betrachtet:

Bezeichnung	Beschreibung
`member`	Dauer der Mitgliedschaft (in Tagen)
`use`	Durchschnittliche tägliche Nutzungsdauer (in Min.)
`con.ord`	Anzahl Kontakte mit *1*: "unter 20", *2*: "20 bis 200", *3*: "über 200"
`age`	Alter (in Jahren)
`sex`	Geschlecht mit *0*: "weiblich", *1*: "männlich"
`area`	Wohnsitz mit *0*: "Gemeinde auf dem Land", *1*: "Stadt"

Für zwei Nutzer liegen Ihnen die folgenden Beobachtungen vor:

	`member`	`use`	`con.ord`	`age`	`sex`	`area`
Nutzer 1	100	10	2	29	1	1
Nutzer 2	550	6	3	24	1	0

1. Bestimmen Sie auf Basis aller quantitativen Merkmale
   (a) die Distanz zwischen beiden Nutzern mittels der Minkowski-2-Metrik,
   (b) die normierte Distanz zwischen beiden Nutzern mittels der Minkowski-1-Metrik.
2. (a) Nennen Sie zwei Möglichkeiten, die Ähnlichkeit zweier Nutzer basierend auf den Variablen `con.ord`, `sex` und `area` zu bestimmen, und berechnen Sie jeweils den Ähnlichkeitswert.
   (b) Welchen Einfluss hat es auf die beiden Ähnlichkeitswerte, wenn die Variable `sex` derart umkodiert wird, dass *1* "weiblich" und *0* "männlich" bedeutet?

Um mehr Informationen zu nutzen, wird anstelle von `con.ord` die Variable `con` verwendet, welche die exakte Anzahl der Kontakte eines Nutzers angibt. Somit stehen `member`, `use`, `con` und `age` als metrische Variablen für eine Clusteranalyse zur Verfügung.

3. Nachfolgende Matrix $\boldsymbol{D}$ beinhaltet die normierten euklidischen Distanzen von fünf Nutzern basierend auf den metrischen Variablen:

$$\boldsymbol{D} = \begin{pmatrix} 0.00 & 3.67 & 3.07 & 3.43 & 2.98 \\ 3.67 & 0.00 & 2.93 & 2.88 & 3.76 \\ 3.07 & 2.93 & 0.00 & 2.40 & 3.85 \\ 3.43 & 2.88 & 2.40 & 0.00 & 2.21 \\ 2.98 & 3.76 & 3.85 & 2.21 & 0.00 \end{pmatrix}.$$

Führen Sie eine Clusteranalyse unter Verwendung des Average-Linkage-Verfahrens durch, und erstellen Sie das zugehörige Dendrogramm.

4. In `R` liegen Ihnen im Datensatz `users` die Beobachtungen für 20 Nutzer vor. Führen Sie eine Segmentierung auf Basis der Ihnen bekannten agglomerativen Clusterverfahren und des K-Means-Verfahrens durch, und stellen Sie die Ergebnisse grafisch gegenüber.

5. In welche Gruppen lassen sich hierarchische Clusterverfahren unterteilen? Erläutern Sie den Unterschied zwischen diesen Gruppen.

6. Durch Unterschiede bei der Bestimmung der Clusterabstände haben verschiedene agglomerative Verfahren unterschiedliche Fusionierungseigenschaften. Was versteht man hierbei unter dilatierenden, kontrahierenden und konservativen Verfahren?