Visualisierung von Clustern in multivariaten Daten unter Einsatz von R

Georg Ohmayer, Michael Sieger

Fakultät für Gartenbau und Lebensmitteltechnologie
Hochschule Weihenstephan-Triesdorf (HSWT)
Am Staudengarten 10
85354 Freising
georg.ohmayer@hswt.de

Abstract: Es wird gezeigt, wie die Ergebnisse verschiedener Cluster-Methoden für multivariate Daten mit Darstellungstechniken visualisiert und damit besser interpretiert werden können. Zu diesem Zweck wurde von den Autoren das Package *VisuClust* für das freie Statistik-System R entwickelt.

1 Einleitung und Hintergrund

Unter dem Oberbegriff Clusteranalyse sind zahlreiche Verfahren zur Suche nach eventuell vorhandenen Clustern (Gruppen von ähnlichen Beobachtungen) in multivariaten Daten bekannt. Zur Visualisierung der Ergebnisse werden aber meist nur die bekannten Dendrogramme verwendet, welche allerdings die komplexen Ähnlichkeitsverhältnisse der Beobachtungen nur eindimensional, d.h. sehr häufig unzureichend, abbilden können.

Um die Beschreibung der Methoden zu veranschaulichen, wird ein einfacher Datensatz, bestehend aus den 4 wichtigsten Inhaltsstoffen der Milchen von 20 Lebewesen, verwendet (siehe Tab. 1). Dieser Datensatz ist Teil der schon in [OS1985] verwendeten Daten.

Tab. 1: Beispiels-Datensatz „Milch-Inhaltsstoffe“

	Fett	Eiweiß	Laktose	Asche
Mensch	3,8	1,0	7,0	0,2
Orang-Utan	3,5	1,5	6,0	0,2
Schimpanse	3,7	1,2	7,0	0,2
Pavian	5,0	1,6	7,3	0,3
Esel	1,4	2,0	7,4	0,5
Pferd	1,9	2,5	6,2	0,5
Kamel	5,4	3,9	5,1	0,7
Rind	3,7	3,4	4,8	0,7
Ziege	4,5	2,9	4,1	0,8
Schaf	7,4	5,5	4,8	1,0
Hund	12,9	7,9	3,1	1,2
Wolf	9,6	9,2	3,4	1,2
Schakal	10,5	10,0	3,0	1,2
Braunbär	22,6	7,9	2,1	1,4
Eisbär	33,1	10,9	0,3	1,4
Biber	11,7	8,1	2,6	1,1
Ratte	10,3	8,4	2,6	1,3
Maus	13,1	9,0	3,0	1,3
Buckelwal	33,0	12,5	1,1	1,6
Delphin	33,0	6,8	1,1	0,7

2 Kurzbeschreibung der statistischen Verfahren

Die meisten Clusterverfahren setzen die Berechnung von Distanzwerten d_{ij} für gegebene multivariate Beobachtungen i, j (jeweils ≤ n) voraus. Je nach Art der Daten stehen dafür verschiedene Distanzmaße zur Verfügung (nach Euklid, Manhattan, Mahalanobis etc.). Mit dem Paket *VisuClust* können Ergebnisse von Clustermethoden visualisiert werden, sofern sie entweder vom Typ „disjunkt" (jede Beobachtung gehört zu genau einem von m Clustern) oder vom Typ „fuzzy" (unscharf: für jede Beobachtung i liegt ein Vektor u_i vor, wobei die Komponente u_{ic} mit c = 1, … m die Zugehörigkeit der i.ten Beobachtung zum Cluster c mit der Maßgabe $\Sigma\, u_{ic} = 1$ angibt).

Disjunkte Gruppierungen liefern einerseits die bekannten hierarchischen Verfahren, sofern die gewünschte Clusteranzahl m vorgegeben wird, oder beispielsweise auch die sog. Austauschverfahren. Verwendbare R-Routinen hierzu sind *hclust, kmeans* u.a. aus dem R-Package *cluster*.

Basis der zu beschreibenden Visualisierung ist die Methode „Nichtlineare Abbildung (NLM = NonLinear Mapping) nach SAMMON" [Sa1969], mit deren Hilfe alle Beobachtungen so in einer Ebene dargestellt werden, dass die Abstandsverhältnisse im multivariaten Raum bestmöglich approximiert werden. Im NLM-Diagramm können dann zunächst die Ergebnisse der Clusterung durch farbliche Kennzeichnung der Cluster gekennzeichnet werden. Zusätzlich werden dann durch die Eintragung von Verbindungen zwischen allen ähnlichen Beobachtungspaaren einerseits die Homogenität der Cluster und andererseits auch Ähnlichkeiten zwischen den Clustern visualisiert. Wenn dabei für verschiedene Schranken t_k im NLM-Diagramm alle Punktepaare (i, j), für die ($d_{ij} \le t_1$ bzw. $t_1 < d_{ij} \le t_2$ usw.) gilt, mit unterschiedlichen Linierungen verbunden werden, entsteht ein sog. Vernetzungsdiagramm (Linkage Map).

In Abb. 1 wird beispielhaft ein solches Vernetzungsdiagramm gezeigt. Dabei wird deutlich, dass i.w. 4 Cluster vorhanden sind (I: Mensch, Primaten, Esel Pferd; II: Rind, Ziege, Schaf, Kamel; III: Biber bis Wolf; IV: Eisbär, Buckelwal), während der Delphin ziemlich isoliert mit geringen Ähnlichkeiten zu den Clustern III, IV und der Braunbär zwischen den Clustern II bis IV liegt.

Eine Schätzung der Dichtefunktion für die Distanzen d_{ij} – beispielsweise mit der Methode „Kerndichteschätzung" über die R-Routine *density* – kann bei der Suche nach den optimalen Grenzen t_k (= Täler in der Dichtefunktion) helfen, um die Cluster im Vernetzungsdiagramm zu detektieren (siehe Beispiel in Abb. 2). Die *VisuClust*-Routine *LinkageMap* ermöglicht dem Nutzer über Schieber für die Parameter t_k eine dynamische Generierung solcher Vernetzungsdiagramme. Dabei ist zu empfehlen, die über die Dichteschätzung gewonnenen Werte t_k noch zu variieren, um die Cluster in den Daten und ihre Beziehungen zueinander möglichst gut sichtbar zu machen.

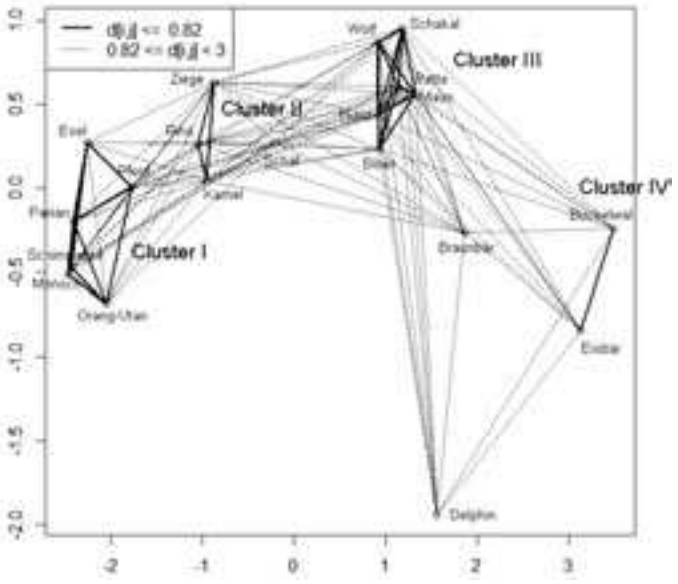

Abb.1: Vernetzungsdiagramm (Linkage Map) für die Milch-Daten mit den Schranken $t_1 = 0.82$ und $t_2 = 3$

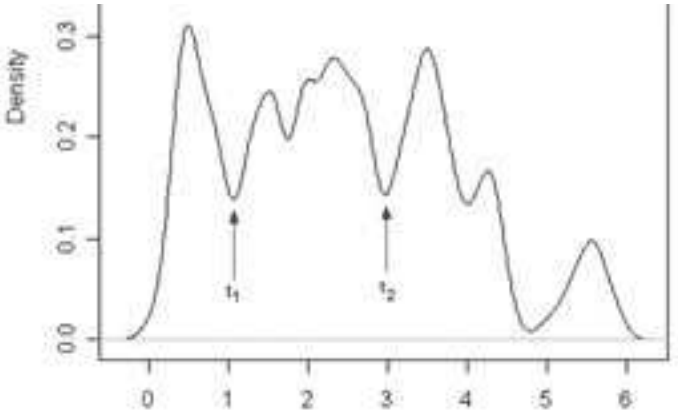

Abb. 2: Kerndichteschätzung für die euklidischen Distanzen der Milch-Daten

Die Routine *FuzzyPlot* im Package *VisuClust* unterstützt die Visualisierung einer unscharfen Gruppierung. Dabei werden im NLM-Plot bei Auswahl eines bestimmten Clusters alle Beobachtungen entsprechend ihrer Zugehörigkeit zu diesem Cluster im Helligkeitswert des Punktes sowie der Schriftgröße der Bezeichnung verändert und damit die Kern- bzw. Randpunkte eines Clusters sichtbar gemacht. Im Beispiel der Abb. 3 werden die folgenden Zugehörigkeitswerte der Lebewesen zum Cluster III über den Helligkeitswert und die Schriftgröße gezeigt: $u_{i3} > 0.65$ für die Lebewesen i des Clusters III (Biber bis Wolf), $u_{Braunbär,3} = 0.3$, $u_{Schaf,3} = 0.2$ und $u_{i3} < 0.1$ für alle anderen Lebewesen i.

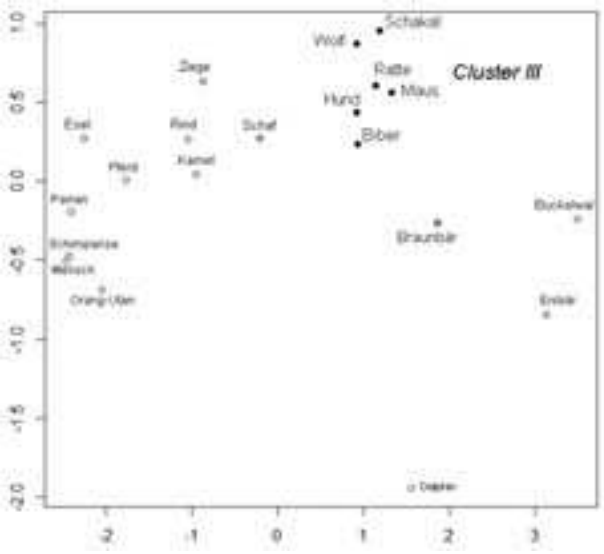

Abb. 3: FuzzyPlot für den Cluster III der Milch-Daten

3 Das R-Package *VisuClust* und potenzielle Einsatzbereiche

Wie viele andere R-Packages kann *VisuClust* mit folgenden R-Befehlen von einem der CRAN-Server[1] installiert und die Routinen verfügbar gemacht werden:

install.packages("VisuClust") und *library(VisuClust)*

Über *example(VisuClust)* erhält der Nutzer das hier beschriebene Demobeispiel „Milch-Daten". Die Befehle *help(LinkageMap)* bzw. *help(FuzzyPlot)* beschreiben und erläutern jeweils die Reihenfolge und die Bedeutung der notwendigen bzw. optionalen Parameter der beiden *VisuClust*-Routinen.

Mögliche Einsatzgebiete von *LinkageMap* und *FuzzyPlot* im Agrarbereich reichen beispielsweise von züchterischen Fragestellungen, bei denen die genetischen Verwandtschaftsbeziehungen von Pflanzen-Gattungen/–Arten oder Tier-Rassen dargestellt werden sollen, bis hin zu sozio-ökonomischen Studien, in denen spezifische Betriebstypen oder charakteristische Führungsstile ermittelt werden sollen.

Literaturverzeichnis

[He11] Hellbrück, R.: Angewandte Statistik mit R – Eine Einführung für Ökonomen und Sozialwissenschaftler, 2. Auflage. Gabler Verlag Wiesbaden, 2011, S. 215 ff.

[OS85] Ohmayer, G.; Seiler, H.: Numerische Gruppierung und graphische Darstellung von Daten – Ein Methodenvergleich. EDV in Medizin und Biologie 16 (2), 1985, S. 65-73.

[Sa69] Sammon, J.W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers, C-18, 1969, S. 401-409.

[1] CRAN steht für "Comprehensive R Archive Network" und bezeichnet ein Netzwerk von Servern mit identischem Angebot von Download-Möglichkeiten (Programm, Packages, Informationen) zum System R.