Chemometrie in der analytischen Chemie[1]

4. Clusteranalyse

David Hornung, Eckhard Reh*
* Korrespondenz: Prof. Dr. E. Reh, Technische Hochschule Bingen, Email: e.reh@th-bingen.de

Einleitung

Die Clusteranalyse erfreut sich heutzutage vielerlei Anwendung, so beispielsweise bei der Verarbeitung von großen Datenmengen im Internet (big data). Natürlich findet eine solche Ermittlung von Zusammenhängen auch in diversen chemischen Disziplinen ihren Einsatz, wie z.B. in der forensischen Chemie bei der Einordnung gefälschter Medikamente (legales Generikum, kein bzw. zu geringer Wirkstoff, falscher Wirkstoff, falsche Galenik, ... [1]), der Lebensmittelchemie (Unterscheidung von Rum-Sorten [2]) bzw. der Klinischen Chemie (Gruppierung von Herzpatienten [3]).
Ziel der Clusteranalyse ist, für eine Sammlung von Objekten (z.B. Medikamenten-Chargen, Getränke-Proben, Patienten-Blutproben) anhand ihrer gemessenen Merkmale (z.B. Raman-Banden, Aromastoff-Konzentration, Blutparameter) Gruppierungen (Cluster) für Proben mit ähnlichen Eigenschaften zu finden.
Anders als bei der Klassifizierung (bei der neue Proben bereits definierten Clustern zugeordnet werden) können bei der Clusteranalyse keine Vorgaben einbezogen werden (wie viele Cluster, welche Merkmale etc.). Dies erfordert daher oftmals ein interaktives Vorgehen, bei dem auf verschiedene Methoden zurückgegriffen wird.
Dazu gehören die hierarchische Clusteranalyse, die k-Means- und MASLOC-Methode, darüber hinaus wird in der Literatur auch die Hauptkomponenten-Analyse herangezogen.

In der Literatur ist zumeist die hierarchische Clusteranalyse eingesetzt. Dabei geht man zunächst von den zwei Objekten aus, die die größte Ähnlichkeit miteinander bzw. den geringsten Abstand zueinander haben, diese bilden den ersten Cluster. In der Folge werden sukzessiv je nach Abstand weitere Objekte diesem Cluster zugeordnet oder ein neuer Cluster gebildet. Zur Visualisierung dient häufig ein Dendrogramm (z.B. Abbildung 2) in dem die Abfolge der Zusammenfassungen dargestellt ist. Daraus können erste Informationen zu den Clustern abgelesen werden.

Die k-Means-Methode baut auf willkürlich vorgegebenen Clustern auf, deren Anzahl beispielsweise aus einer vorangegangen hierarchischen Analyse resultiert. Dann werden Objekte zunächst zufällig den Clustern zugeordnet. Anschließend wird anhand der Abstände zu den verschiedenen Clustern ermittelt, ob das Objekt zu dem anfangs zugeordneten Cluster gehört oder besser zu einem anderen, zu dem der Abstand geringer ist. In der Folge werden die Objekte wiederholt neu zugeordnet, bis keine Verbesserung mehr erreicht wird.

Bei der MASLOC-Methode müssen zuvor die Clusterzahl fest vorgegeben und für jeden Cluster ein repräsentatives Referenzobjekt (centrotype) definiert sein. Daher wird diese Methode zur Clusteranalyse ohne Vorgaben zumeist nicht verwendet.

Die Hauptkomponenten-Analyse ist a priori nicht für die Clusteranalyse vorgesehen, aus der grafischen Darstellung der Hauptkomponenten (z.B. Biplot, Abbildung 1) können aber oft erste Informationen über Clusterzahl und Objektgruppierung getroffen werden.
Da zur Bedeutung der gewählten Merkmale (Signale, Konzentrationen) für die avisierte Cluster-Bildung keine Aussage gemacht werden kann, ist es sinnvoll, generell die aus den Merkmalen abgeleiteten Hauptkomponenten für die Berechnungen zu verwenden (es wird vorausgesetzt, das die 1. Hauptkomponente die größte Variabilität der Daten und damit die größte Bedeutung für die Clusteranalyse beschreibt).

Das detaillierte Vorgehen der angeführten Methoden soll an dieser Stelle nicht ausgeführt werden, hierzu wird auf einschlägige Monographien verwiesen [4, 5].

1 Hornung, D., Reh, E., GIT, 1, xyff, 2019

Resultat der Clusteranalyse ist eine minimale Clusterzahl mit einer optimalen Zuordnung der Objekte entsprechend ihren Merkmalen (z.B. Konzentrationen / Hauptkomponenten). Eine Maßzahl für die Güte der Probenzuordnung ist z.B. der Davies-Bouldin-Index. Bezogen auf 2 Cluster beschreibt dieser den Quotienten des mittleren Abstands aller Objekte in Cluster 1 (Cl_1) bzw. Cluster 2 (Cl_2) relativ zur Distanz der beiden Cluster:

$$dbi = 0{,}5 \frac{mittl.\ Abstand\ Cl_1 + mittl.\ Abstand\ Cl_2}{Distanz\ Cl_1\ zu\ Cl_2} \quad (1)$$

Der dbi-Wert sollte möglichst klein sein, dies zeigt an, das die Cluster räumlich eng begrenzt und gut differenziert sind. Die Definition findet sich im Anhang (Gleichung 2), für mehrere Cluster werden die dbi-Werte für alle Paare berechnet und der Mittelwert angegeben.

Nachfolgend soll die Thematik anhand von zwei Beispielen erläutert und eine generelles Vorgehen vorgeschlagen werden. Für die Realisierung wird die Software-Implementierung *Cluster* eingesetzt.

Beispiel 1: Botanische Herkunft verschiedener Honigproben

In der Publikation von Fernández-Torres [6] wurde mittels ICP-AES die Elementzusammensetzung (P, B, Zn, Mn, ..., Na, K) diverser Honigproben bestimmt (Tabelle 1, Daten). Anhand dieser Konzentrationen soll die Herkunft festgestellt und Honigproben mit ähnlichem oder gleichem Ursprung (Eukalyptus, Rosmarin, ...) in einem entsprechendem Cluster zusammengefasst werden.

Da bei der Clusteranalyse die Relevanz der Merkmale nicht bewertet werden kann, ist es empfehlenswert, zunächst eine Hauptkomponentenanalyse durchzuführen. Zur Festlegung der Hauptkomponenten-Zahl dient die Kreuzvalidierung (Bootstrapping [4, 7]). Im resultierenden Biplot (Abbildung 1), kann man vorab schon näherungsweise erkennen, wie viele Cluster es geben wird (hier: 3 - 4) und welche Objekte vermutlich hierzu gehören.

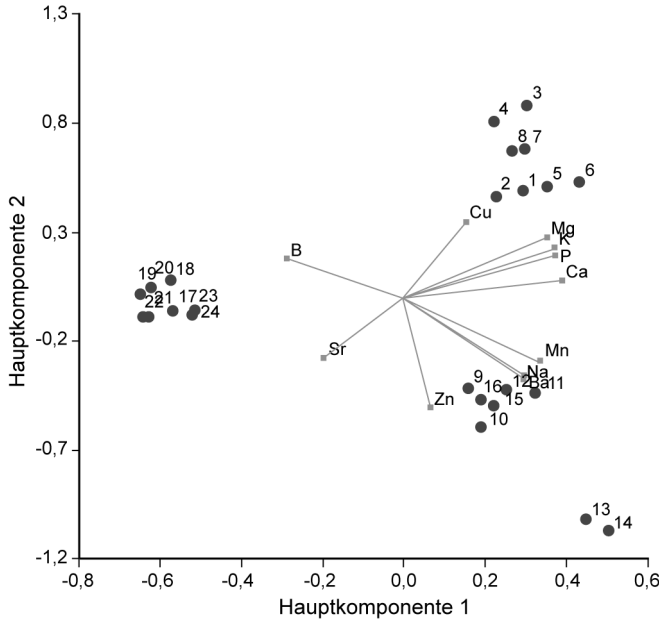

Abbildung 1: Biplot zum Datensatz der Honigproben (Objekt-Nummern der Proben, Merkmale: Konzentration Cu, Mg, K, P, Ca, Mn, Na, Ba, Zn, Sr, B)

Der nächste Schritt ist die Durchführung der hierarchischen Clusteranalyse, daraus resultiert das Dendrogramm.

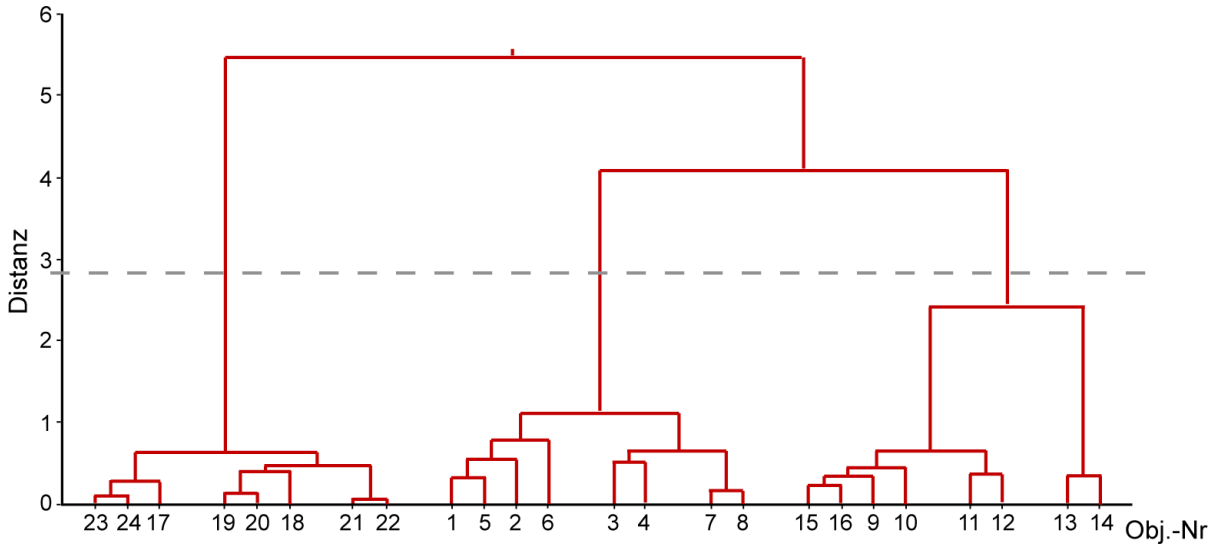

Abbildung 2: Dendrogramm zum Datensatz der Honigproben

Es wird deutlich, dass zuerst die Proben 23 und 24 zusammengefasst wurden, dann kam die Probe 17 hinzu, danach die Proben 19 bis 22, die zusammen vermutlich einen ersten Cluster bilden.
Ein Schnittpunkt einer vertikalen Distanz-Linie mit der Grenzlinie (grau gestrichelt) deutet die Objekte eines Clusters an. Problem der hierarchischen Clusteranalyse ist die Anordnung der Grenzlinie, hier sind viele Varianten in der Literatur beschrieben (*Cluster* verwendet die Mojena-Variante mit diversen Faktoren, vgl. Anhang Gleichung 3.) Empfehlenswert ist der Ausdruck des Dendrogramms und das manuelle Zeichnen einer horizontalen Grenze.
Aus dem Dendrogramm der Honigproben kann man auf vermutlich 3 Cluster schließen.

Mit diesen Vorab-Informationen wird die k-Means-Methode benutzt, um eine abschließende Clusteranalyse durchzuführen. Auf Grund der zufälligen Zuordnung der Objekte im 1. Schritt, kann die k-Means-Methode mit anderen Startparametern wiederholt werden (bei einer optimalen Implementierung ist der Algorithmus parallelisierbar, so dass die Berechnung auf einem Mehrkern-Rechner im Hintergrund erfolgt). Die Software *Cluster* erlaubt diverse Einstellungen für die k-Means-Berechnung wie in der Dialogbox in Abbildung 3 dargestellt (erwartete Clusterzahl: 3 ± 1, jeweils 1000 Wiederholungen).

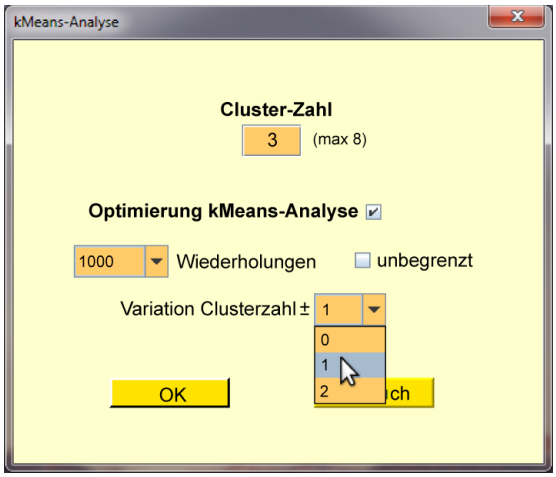

Abbildung 3: Dialog k-Means-Berechnung

Um die Güte der jeweiligen k-Means-Berechnung zu bewerten, wird der Davies-Bouldin-Index herangezogen, die aktuell beste Cluster-Anordnung wird festgehalten.
Das Resultat der k-Means-Clusteranalyse sind 3 Cluster mit definiert zugeordneten Proben (mittlerer dbi-Wert = 0,22).

Da die Berechnung von den gewählten Parametern abhängt (hier: Standardisierung der Daten, euklidischer Objektabstand, median-linkage Cluster-Distanz) sollten die Resultate visuell kritisch bewertet werden, z.B. in Form eines Distanzplot, wenn möglich mit einem interaktiven 3D-Grafen. Der resultierende 3D-Distanzplot in Abbildung 4 visualisiert das Ergebnis der durchgeführten Clusteranalyse.

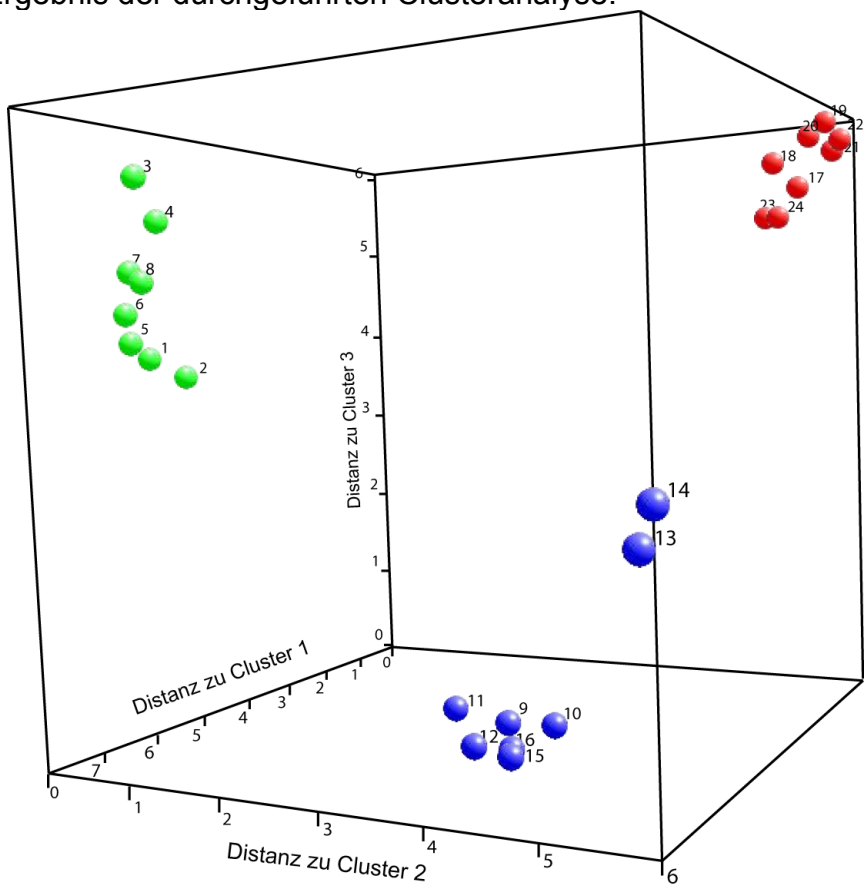

Abbildung 4: 3D-Distanzplot des Honig-Datensatz

Es wird deutlich, wie die Objekte den Clustern zugeordnet sind, so hat z.B. Objekt 3, das Cluster 2 zugeordnet wurde, eine minimale Distanz zu Cluster 2, eine große Distanz zu Cluster 1 und 3.
Es kann festgehalten werden, dass in diesem Fall die Clusteranalyse sichere und nachvollziehbare Resultate ergibt.

Beispiel 2: Produktionskontrolle von Proben des japanischen Staudenknöterichs
In diesem Beispiel wird ein Datensatz mit mehreren Proben von *Polygonum cuspidatum* verschiedener geographischer Herkunft untersucht. *Polygonum cuspidatum* wird häufig in der Pflanzenheilkunde verwendet und enthält diverse Inhaltsstoffe (Gallensäure gaa, Epicatecin epi, Resveratrol res, ...) in verschiedenen Konzentrationen. Die Quantifizierung erfolgte mittels HPLC-Trennung, die Identifizierung mittels GC-MS [8].
Die Clusteranalyse soll feststellen, ob alle gesammelten Proben diverser Herkunft für die Produktion zusammengeführt werden können oder ob einzelne Chargen sich signifikant unterscheiden (d.h. zu anderen Clustern gehören).

Das Vorgehen ist wie in Beispiel 1, zunächst erfolgt eine Hauptkomponenten-Analyse (gleiche Parameter wie oben), Abbildung 5 zeigt den resultierenden Biplot.

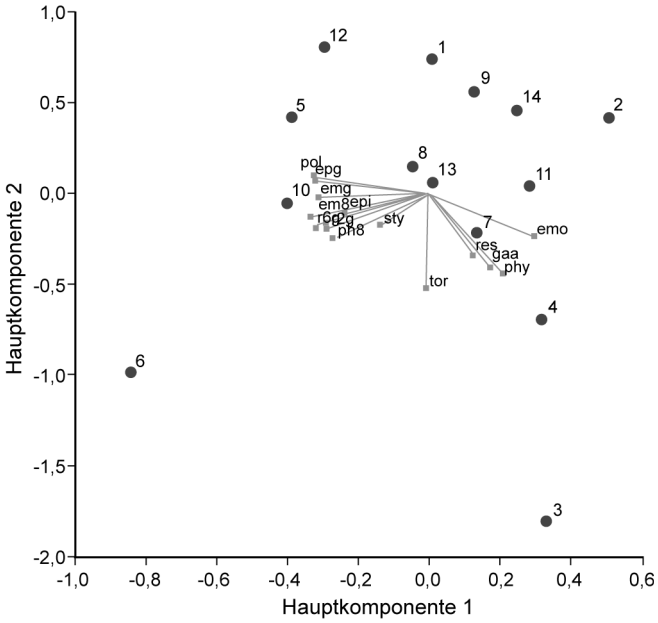

Abbildung 5: Biplot zu den Proben von *Polygonum* cuspidatum

In diesem Fall liegen die Proben weit gestreut, dennoch kann man vermuten, dass es ca. 2 bis 3 Cluster geben wird, die Proben 3, 6 sind evtl. deutlich unterschiedlich, verglichen mit den übrigen Proben. Es wird auch deutlich, dass die Inhaltsstoffe nicht optimal gewählt wurden, z.B. die Ladungen (graue Linien im Biplot), von Gallensäure gaa, Physcion phy, Resveratrol res sind sehr ähnlich, diese Substanzen beschreiben daher vergleichbare Eigenschaften.

Mit optimaler Hauptkomponenten-Zahl (1 Hauptkomponente nach Kreuzvalidierung, Bootstrapping-Variante) läßt die hierarchische Clusteranalyse anschließend auf zwei oder drei Cluster schließen (Abbildung 6, Grenzlinie a bzw. b).

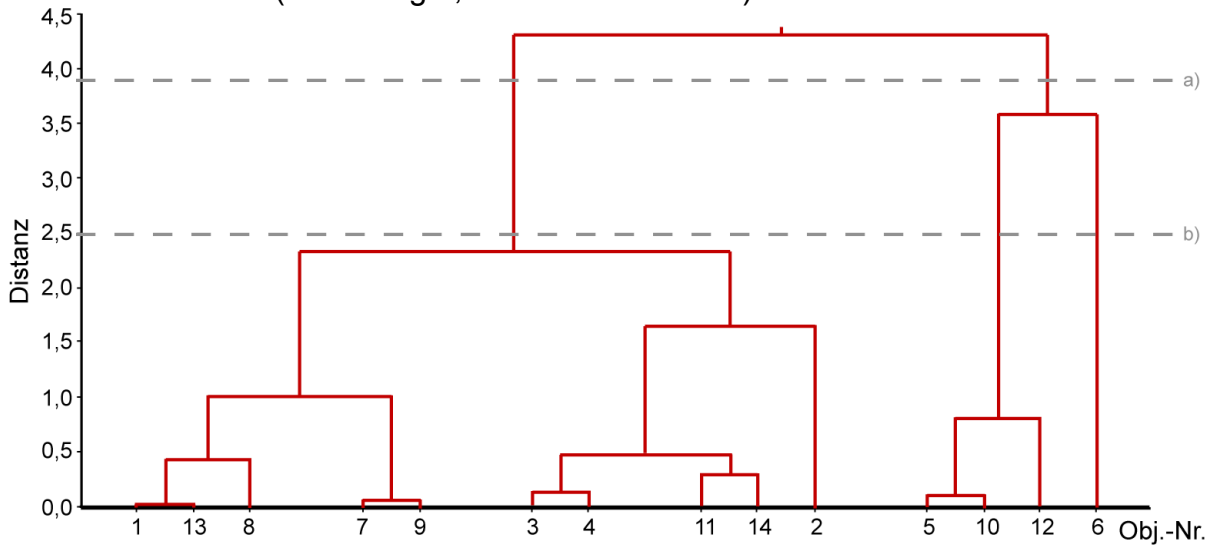

Abbildung 6: Dendrogramm nach hierarchischer Clusteranalyse von *Polygonum* cuspidatum.

Die finale Cluster-Zuordnung liefert anschließend die k-Means-Methode mit 3 Clustern, bei denen die Proben 6 bzw. 5 / 10 / 12 von den übrigen differenziert wurden.
Der 3D-Distanzplot verdeutlicht dies.

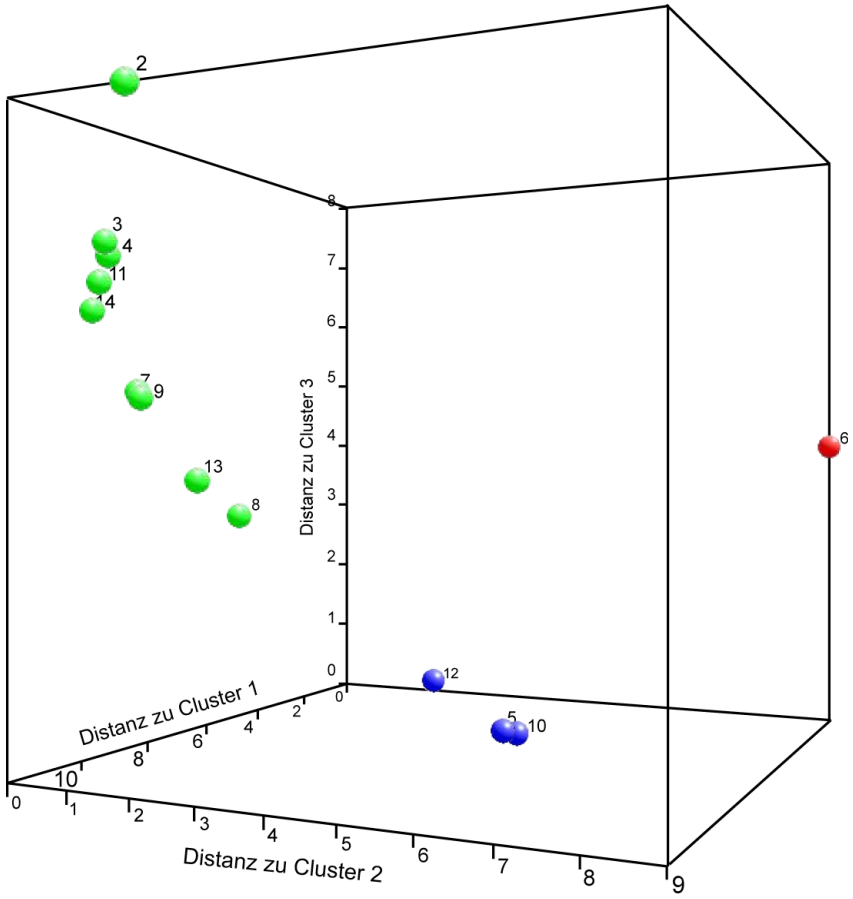

Abbildung 7: Distanzplot zu Polygonum cuspidatum Datensatz

Zusammenfassung

Die hier genannten Beispiele zeigen, inwieweit die Clusteranalyse heutzutage in der Analytischen Chemie verwendet werden kann. Günstig ist, wenn diverse chemometrische Optionen (diverse Skalierungen, Abstands-, Distanzmaße) und verschiedene Methoden zur Verfügung stehen.

Mit Einsatz des Programms *Cluster* hat sich folgendes Procedere bewährt:
- Voruntersuchung mittels Hauptkomponentenanalyse und hierarchischer Clusteranalyse,
- finale Clusteranalyse mit der k-Means-Methode,
- visuelle Bewertung der Resultate.

Insbesondere dem letzten Schritt obliegt eine besondere Bedeutung, da die Resultate oft von den gewählten Parametern (besonders der optimalen Hauptkomponenten-Zahl) abhängen. Sinnvoll ist zudem, wenn die Resultate mit weiteren Probe-Informationen in Zusammenhang gesetzt werden können.

Literatur

[1] Been, F. et al, Profiling of counterfeit medicines by vibrational spectroscopy, Forensic Sci. Int., 211, 83-100, 2011
[2] Belmonte-Sánchez, J., R., Rum classification using fingerprint analysis of volatile fraction by headspace solid microextraction coupled to gas chromatography-mass spectrometry, Talanta, 187, 348-356, 2018
[3] Horiuchi, Yu et al, Identifying novel phenotypes of acute heart failure using cluster analysis of clinical variables, Int. J. Cardiol., 262, 57-63, 2018
[4] Reh, E., Chemometrie, Grundlagen der Statistik, numerischen Mathematik und Software-Anwendung in der Chemie, de Gruyter, 2017
[5] Brereton, R.G., Chemometrics, Data Analysis for Laboratory and Chemical Plant, Wiley, 2006
[6] Ferández-Torres, R., et al, Mineral content and botanical origin of Spanish honeys, Talanta, 65, 686-691, 2005
[7] Reh, E., Validierung Clusteranalyse, www.chemometrie.info/literatur-titel.html
[8] Gao, F. et al., A comprehensive strategy using chromatographic profiles combined with chemometric methods: Application to qualtiy control of *Polygonum cuspidatum* Sieb. et Zucc., J. Chrom. A, 1466, 67-75, 2016

Anhang

a) Davies-Bouldin-Index dbi:

$$dbi_{gh} = 0{,}5\,\frac{\frac{1}{n_g(n_g-1)/2}\sum_{i\in n_g,\, i<n_g}\sum_{k\in n_g,\, k>i} a(O_i,O_k) + \frac{1}{n_h(n_h-1)/2}\sum_{i\in n_h,\, i<n_h}\sum_{k\in n_h,\, k>i} a(O_i,O_k)}{a(Z_g,Z_h)} \quad (2)$$

mit n_g, n_h: Objektzahl in Cluster g bzw. h; $a(Z_g, Z_h)$: euklidischer Abstand der Zentroide der beiden Cluster; $a(O_i, O_k)$: euklidischer intracluster Abstand zwischen den Objekten O_i und O_k

b) Mojena-Grenze mG:

$mG = \bar{x}_{d_{ij}} + mF * s_{d_{ij}}$ (3)

mit $x_{d_{ij}}$: Vereinigungsdistanz der Elemente i und j; $\bar{x}$, s: Mittelwert, Standardabweichung Vereinigungsdistanzen; mF: Mojena-Faktor (willkürlich gewahlt) $\in$ (1,25; 1,5; 1,75)

Daten

	Zn	P	B	Mn	Mg	Cu	Ca	Sr	Ba	Na	K
H1	3,211	154,3	6,988	4,539	61,34	0,809	263,00	0,271	0,225	97,66	1771,0
H3	3,101	144,2	6,878	4,429	55,84	0,699	252,00	0,260	0,214	86,66	1760,0
H10	2,632	142,6	5,150	3,485	66,26	2,117	201,60	0,354	0,215	47,09	1845,0
H7	2,422	121,6	4,940	3,275	61,16	1,907	195,60	0,333	0,194	48,90	1824,0
H6	2,275	130,8	2,962	2,438	71,18	0,554	326,00	0,461	0,217	63,60	1920,0
H4	2,425	145,8	3,112	2,588	74,38	0,704	341,00	0,476	0,232	78,60	1935,0
H8	2,921	148,5	6,069	4,012	63,80	1,463	232,30	0,313	0,220	72,38	1808,0
H9	2,871	143,5	6,019	3,962	61,80	1,413	227,30	0,308	0,215	67,38	1803,0
E1	5,438	113,0	7,065	7,427	36,61	0,578	184,40	0,408	0,444	168,60	1094,0
E4	5,327	101,9	4,967	7,316	35,51	0,542	173,30	0,597	0,555	157,50	1083,0
E2	5,821	140,3	3,163	6,600	47,64	0,645	177,70	0,823	0,493	116,20	1425,0
E5	5,611	119,3	2,953	6,390	45,54	0,533	156,70	0,613	0,393	137,20	1404,0
E3	7,672	96,6	2,314	9,318	36,88	0,517	269,30	0,635	0,522	218,50	1217,0
E8	7,825	111,9	2,467	9,471	38,38	0,670	254,00	0,788	0,675	203,20	1232,0
E6	5,426	116,0	4,842	7,950	41,50	0,636	203,00	1,175	0,380	150,00	1581,0
E9	5,375	111,0	4,792	7,900	41,00	0,586	198,00	1,125	0,330	145,00	1576,0
O1	5,297	76,7	8,120	0,456	19,22	0,531	42,70	0,691	0,106	43,44	618,9
O6	3,943	75,6	8,010	0,350	18,11	0,531	42,59	0,580	0,106	42,34	608,8
O3	4,054	53,3	6,926	0,133	15,35	0,531	51,52	0,665	0,106	11,69	455,0
O7	3,845	51,2	6,716	0,200	13,26	0,531	49,43	0,456	0,106	23,59	434,1
O4	3,268	63,2	6,506	0,286	20,86	0,531	59,20	1,462	0,106	39,39	634,9
O5	3,217	58,1	6,454	0,234	20,36	0,531	54,00	1,410	0,106	38,87	629,7
O9	4,272	68,3	5,723	0,491	19,41	0,548	89,60	0,916	0,106	27,41	647,3
O10	4,375	69,4	5,826	0,388	20,44	0,531	90,90	1,019	0,106	28,44	657,6

Tab. 1: Messwerte Elementkonzentrationen Honigproben

	gaa	epi	pol	epg	r2g	r6g	emg	res	tor	em8	ph8	sty	emo	tor	phy
An1	1,2	0,05	11,1	0,60	0,6	2,0	3,9	3,7	1,6	10,8	2,8	1,3	5,1	0,2	4,8
An2	1,2	0,05	8,8	0,05	0,7	1,5	1,1	3,5	0,3	3,5	0,9	1,4	10,6	0,2	8,3
Fu	1,9	0,05	6,2	0,60	0,7	2,2	3,2	6,8	1,4	9,3	2,5	1,1	9,8	0,4	9,7
Hu1	1,6	0,05	8,0	0,05	0,6	1,7	2,3	4,4	0,8	7,7	3,1	2,6	8,8	0,2	10,1
Hu2	0,9	0,40	19,4	1,80	1,0	2,2	4,4	3,1	2,3	12,0	3,5	1,8	4,6	0,2	6,0
Hu3	1,3	0,60	26,1	1,80	1,2	5,2	7,2	2,9	4,2	17,4	4,1	2,3	5,4	0,3	5,5
Js	1,9	0,05	17,8	0,50	0,8	1,7	2,8	2,0	1,1	10,2	2,1	1,3	8,3	0,3	7,0
Jx	1,2	0,05	16,9	1,10	0,7	3,5	4,0	3,4	1,1	9,9	1,7	2,0	7,3	0,3	5,4
Ni	1,2	0,05	11,8	0,80	0,7	2,2	2,9	3,8	1,0	8,5	2,1	1,4	7,3	0,2	6,0
He	1,1	0,70	21,8	1,80	0,9	4,4	3,7	4,7	1,6	12,6	3,6	1,5	4,6	0,2	6,7
Si	1,3	0,05	10,5	0,60	0,7	2,7	1,9	5,4	0,7	6,7	1,8	1,3	7,9	0,2	8,2
Yu1	0,9	0,05	19,0	1,90	0,7	3,0	6,8	3,3	2,0	13,0	2,3	1,5	5,7	0,2	5,4
Yu2	1,7	0,70	14,1	0,90	0,7	2,0	3,3	2,3	1,1	8,9	2,1	1,3	6,2	0,3	5,9
Zh	1,8	0,05	14,9	0,70	0,6	1,7	2,9	4,0	0,8	7,8	1,7	1,1	8,2	0,2	5,5

Tab. 2: Messwerte Konzentrationen Pflanzeninhaltsstoffe Polygonum cuspidatum