Optimierung und Validierung in der Clusteranalyse

Prof. Dr. E. Reh, Technische Hochschule Bingen, Email: e.reh@th-bingen.de

1 Einleitung

Bei der chemischen Clusteranalyse ist folgende Fragestellung gegeben:
Gibt es unterscheidbare Gruppen (Cluster) denen diverse Proben an Hand ihrer chemischen Merkmale sicher zugeordnet werden können?
Es ist a priori nicht bekannt, ob bzw. wie viele Cluster vorliegen und ob die Objekte (Proben) mit Hilfe welcher Messwerte (z.B. Signalhöhen, Konzentrationen) diesen sicher zugeordnet werden könnten (unüberwachtes Lernen).
Seit mit routinemäßiger Anwendung gekoppelter Analysemethoden (z.B. LC-MS, ICP-MS, ...) ausreichend große Daten-Pools schnell und kostengünstig zur Verfügung stehen, können solche Aufgabenstellungen erfolgreich behandelt werden.
Ein Beispiel aus der Lebensmittelchemie ist die Zuordnung von Honigproben auf Grund ihrer Elementzusammensetzung in definierte Cluster entsprechend ihrem botanischen Ursprung (Eukalyptus-, Heide-, Orangenblüten-, Rosmarien-, Akazien-Honig ...) [1] oder in der Forensik die Zuweisung von Arzneimittelfälschungen (Original, Generikum, Formulierungen mit falscher Konzentration, ohne Wirkstoff, falscher Galenik ...) an Hand von Raman-Banden [2].
Die wichtigsten Methoden für die Ermittlung der möglichen Cluster sind die k-Means- und die MASLOC-Methode, die bereits vielfach beschrieben wurden, oftmals wird auch die hierarchische Clusteranalyse eingesetzt [3].

Allen Fällen ist gemein, dass vorab nur sehr wenig Informationen vorliegen (welche Merkmale / Hauptkomponenten, wie viele Cluster, ...). Hier unterschiedet sich die Clusteranalyse von der Klassifizierung (mit definierten Clustern und einem definierten Proben-Set mit vorgegebener Zugehörigkeit zu den Clustern). Daher können bei der Clusteranalyse nur wenige Methoden für die Optimierung bzw. Validierung genutzt werden.
Generell kommt der Validierung von Merkmalen / Methoden eine besondere Bedeutung zu, um keine zufällige Aussage z.B. auf Grund eines zu kleinen oder nicht repräsentativen Probe-Set zu erhalten.

2 Optimierung

2.1 Daten

2.1.1 Irrelevante Messwerte
Durch das Analysenverfahren bedingt treten immer wieder Messartefakte auf. Dazu können z.B. Kontaminationen aus der Probenvorbereitung oder auch Leer-Signale z.B. aus dem Chromatographie-Gradienten gehören. Solche Signale müssen durch den analytischen Chemiker erkannt und aus dem Datensatz eliminiert werden.

Es ist nicht ungewöhnlich, das nicht bei allen Proben alle Messwerte vorliegen.
Treten solche fehlenden Messpunkte selten auf, können fehlende Werte durch den Mittelwert des Merkmals / Variable aller Proben ergänzt werden.
Messwerte unterhalb der Bestimmungsgrenze (LOQ) werden zumeist als 0,5*LOQ eingesetzt.
Wurde ein Merkmal nur selten ermittelt (z.B. bei < 10 % der Proben), z.B. weil die Analyse zu teuer oder langwierig ist, sollte das Merkmal insgesamt aus den Daten entfernt werden.

Mit Hilfe der Homogenitäts-Prüfung einer Variablen (z.B. Grubbs-Test [3]) können Proben als ausreißerverdächtig ermittelt werden. Empfehlenswert ist, auch die anderen Merkmale dieser ausreißerverdächtigen Probe zu betrachten. Werden mehrere Variablen der gleichen Probe als Ausreißer deklariert, könnte dies auf eine systematische Abweichung oder Verwechslung hinweisen.

2.1.2 Korrelationen
Die Kovarianzen, z.B. zusammengefasst in der Korrelations-Matrix [3], beschreiben den Zusammenhang zwischen 2 Variablen. Liegt ein hoher Korrelations-Koeffizient vor, zeigt dies eine signifikante Abhängigkeit an (diese sind nicht orthogonal), so das bei den Untersuchungen vermutlich auf eines dieser beiden Merkmale verzichtet werden kann.

Auch wenn die Daten unverändert in die Modell-Berechnung eingehen, kann eine Hauptkomponenten-Analyse zur Beurteilung der Messwerte hilfreich sein. Insbesondere die resultierenden Ladungen können zur Bewertung der Variablen wichtige Aspekte aufzeigen. Der Ladungs-Vektor beschreibt die Bedeutung der Merkmale für eine Hauptkomponente. Eine große k-te Ladung (loading) zeigt den großen Stellenwert des Merkmals k für die Hauptkomponente an. Ist die Ladung eines Merkmals bei allen Hauptkomponenten gering, weist dies darauf hin, dass das Merkmal eine geringe Bedeutung für die Beschreibung der Objekte hat, dieses Merkmal kann evtl. entfallen [3].
Die Korrelation zweier Merkmale wird auch durch den Cosinus ihres Winkels α im Ladungsplot indiziert, senkrecht stehende Merkmals-Ladungen sind unkorreliert (cos α = 0).
Daher bietet auch die visuelle Beurteilung, z.B. in einem interaktiven 3D-Biplot wichtige Aussagen über die verwendeten Merkmale.

2.2 Datenvorbehandlung

2.2.1 Skalierung Merkmale
Ist der Mittelwert / Median zweier Merkmale / Variablen mehr als eine Größenordnung unterschiedlich, sollten diese auf vergleichbare Größen skaliert werden (Skalierung, Autoskalierung, Normierung, Standardisierung [3]), damit auch kleine Merkmale einen relevanten Einfluss haben. Empfehlenswert ist die Autoskalierung bzw. Standardisierung, da hier gleichzeitig zentriert wird. Es gilt:

$$x_{ik}^{a} = \frac{x_{ik} - \bar{x}_k}{s_k} \quad (1)$$

mit x_{ik}: Original-Merkmal k des Objekts i; $\bar{x}_k$: Mittelwert des Merkmals k. s_k: Standardabweíchung des Merkmals k

Auf die Autoskalierung / Standardisierung kann verzichtet werden, wenn für die Objekt-Abstände bzw. Cluster-Distanzen Mahalanobis-Abstand bzw. -Distanz verwendet werden.

2.2.2 Transformierung Objekte
Manche Autoren transformieren die Objekte, wenn z.B. durch Variation der Probenmenge die Signalintensitäten insgesamt stark schwanken. Dabei werden alle Merkmale eines Objekts mit einem Faktor versehen, dass die Summe der Merkmale jedes Objekts gleich 1 wird. Es gilt:

$$x_{ik}^{t} = \frac{x_{ik}}{\sum_{k=1}^{p} x_{ik}} \quad (2)$$

Im Bereich der Analytischen Chemie ist davon abzuraten, da durch Einsatz eines internen Standards diese Problematik schon bei der Datenerfassung korrigiert werden sollte.

2.2.3. Selektierung Merkmale
In vielen Messverfahren wird eine kontinuierlich Folge von Signalen registriert, z.B. in der Chromatografie oder Spektroskopie. Es konnte gezeigt werden, das es weniger effektiv ist, alle Messpunkte als Merkmale zu verwenden, sondern sinnvoller, die Peaks/Banden auszuwählen, die von Bedeutung für die Cluster-Differenzierung sind (peak-selection) [4].
Es kann z.B. folgendes Procedere eingesetzt werden:
Alle Peaks/Banden werden separat ausgewertet und deren Fläche/Höhe ermittelt. Diese sind potentielle Merkmale mit denen eine Hauptkomponenten-Analyse durchgeführt wird. Es kann konstatiert werden, das Peaks/Banden mit generell minimalen Ladungen keine relevante Varianz haben und damit keinen Beitrag zur Differenzierung der Cluster zeigen (vgl. 2.1.2). Signale mit großen Ladungen zeigen eine hohe Variabilität, haben evtl. eine große Bedeutung.

2.3 Hauptkomponenten

Die Hauptkomponentenanalyse (PCA) ist keine explizite Methode der Clusteranalyse, gibt aber wichtige Vorab-Informationen.
Mit dem Wechsel von den originären Messsignalen (z.B. hochaufgelöste MS-MS-Peaks) zu den Hauptkomponenten wird zumeist eine Reduktionen der Variablenzahl angestrebt.
Dabei nutzt man die Eigenschaft, dass die 1. Hauptkomponente die größte Variabilität der Proben beschreibt und in Folge der weiteren Hauptkomponenten fallend.
Die größte Variabilität bei den Proben muss jedoch nicht die größte Aussage für die betrachtete Clusteranalyse bedeuten. Ob der Einsatz von Hauptkomponenten an Stelle der originären Merkmale sinnvoll ist, muss im jeweiligen Fall explizit untersucht und validiert werden.
Kritisch ist die Zahl der verwendeten Hauptkomponenten. Hier wird oft die Kreuzvalidierung verwendet [3]. Wie in Kapitel 3.1. beschrieben, kann der Davies Bouldin Index DBI als Qualitätsmaß für die erfolgte Cluster-Differenzierung herangezogen werden, er kann auch eingesetzt werden, um die optimale Hauptkomponentenzahl zu ermitteln.

2.3.1 Optimierung durch Kreuzvalidierung (konventionell)

Bei der Kreuzvalidierung haben RSS-Wert und PRESS-Wert besondere Bedeutung [3].
In beiden Fällen wird berechnet, wie groß die Differenz ist zwischen den ursprünglichen Merkmalen der Objekte und den nach Hauptkomponenten-Analyse geschätzten Merkmalen.
Erfolgte eine Skalierung der Original-Daten wird nach der Hauptkomponenten-Analyse vor der Berechnung der Differenz deskaliert. Beim RSS-Wert (residual sum of squares) werden Merkmale aller Objekte direkt durch Einsatz der Hauptkomponenten und Ladungen geschätzt und im Vergleich mit den originalen Daten x_{ik} die Summe der Residuenquadrate bestimmt. Dies erfolgt bei variierender Hauptkomponentenzahl t.

$$RSS_t = \sum_{i=1}^{n} \sum_{k=1}^{p} \left(x_{ik} - \hat{x}_{ik} \right)^2 \quad (3)$$

mit n: Objektzahl; p: Merkmalszahl; $\hat{x}_{ik}$: nach Hauptkomponenten-Analyse geschätztes Merkmal k von Objekt i

Bei der Berechnung des PRESS-Wertes (predictive residual sum of squares) wird meist ein Objekt nach dem anderen aus dem Datensatz entfernt. Das entfernte Objekt i entspricht dem Test-Set. Die verbliebenen Objekte dienen als Training-Set dazu, Hauptkomponenten und Ladungen zu berechnen. Hiermit werden die Merkmale des Test-Objekts geschätzt und daraus die Residuen bzw. Summe der Residuenquadrate nacheinander für alle Test-Objekte ermittelt.

$$PRESS_t = \sum_{i=1}^{n} \sum_{k=1}^{p} \left(x_{ik} - \hat{x}_{[i]k} \right)^2 \quad (4)$$

mit $\hat{x}_{[i]k}$: geschätztes Merkmal k des Testobjekts Objekt i

Abhängig von steigenden Hauptkomponentenzahlen t wird der Quotient $PRESS_t$ / RSS_{t-1} berechnet. Dieser Quotient fällt zu Anfang deutlich und steigt dann nach Überschreiten der optimalen Hauptkomponentenzahl wieder gravierend an. Überschreitet der Quotient ein Minimum bzw. steigt über 1, war die $PRESS_t$-Hauptkomponentenzahl um eine Hauptkomponente zu hoch angesetzt.

2.3.2 Optimierung durch Kreuzvalidierung (Bootstrapping)

Die Aussagen der konventionellen Kreuzvalidierung sind nicht sehr stabil, dies wird durch die Mittlung vielfacher Wiederholungen bei der Bootstrapping-Variante verbessert [6, 7].
Das Vorgehen ist wie folgt:

a) - Wahl Hauptkomponentenzahl t=1
b) - Aufteilung in Training-, Test-Set. Hierzu werden aus dem gesamten Datensatz Objekte zufällig dem Training-Set zugewiesen (in der Regel 2/3 der Objekte),
Objekte die nicht im Training-Set enthalten sind bilden den Test-Set.[1]
c) - Skalierungs-Bildung jeweils mittels Training-Set und Anwendung auf die Objekte des Test-Set.

1 Zuweisung zum Training-Set kann auch mit Wiederholungen erfolgen, fehlende Objekte bilden den Test-Set [5]. Damit entspricht Objektzahl im Training-Set der Gesamtzahl der Modellobjekte, Objektzahl im Test-Set variiert je nach Zahl der zufälligen Wiederholungen. Aus den Wiederholungen kann z.B. Varianz abgeleitet werden.

d) - Hauptkomponenten-Analyse an Hand des Training-Set.
e) - Schätzung der Merkmale des Training-Set mit den entsprechenden Hauptkomponenten / Ladungen (autoprediction).
f) - Deskalierung und Berechnung der Residuenquadrat-Summen der p Merkmale des Training-Set.
g) - Schätzung der Hauptkomponenten und folgende Schätzung der Merkmale der Test-Set-Objekte mittels Ladungsmatrix der Hauptkomponenten-Analyse (prediction).
h) - Deskalierung und Berechnung der Residuenquadrat-Summen der Merkmale des Test-Set.
i) - Wiederholung der Schritte b) - h) durch erneute, zufällige Aufteilung in Training-, Test-Set (z=200), Mittelwertbildung nach Bootstrapping-Wiederholungen für jedes Objekt gemäß der Anzahl mit dem es im Training-Set enthalten war (autoprediction ⇒ RSS-Wert), bzw. Mittelwertbildung für jedes Objekt nach seiner Anzahl im Test-Set (prediction ⇒ PRESS-Wert).
j) - Summation der mittleren Residuenquadrat-Summen aller Training-Objekte (autoprediction) ergibt gemittelten $\overline{RSS}$-Wert, analog folgt aus Summation der mittleren Residuenquadrat-Summen aller Test-Objekte (prediction) der mittlere $\overline{PRESS}$-Wert bei vorgegebener Hauptkomponentenzahl.

k) - Wiederholung aller Schritte a) - i) bei Erhöhung der Hauptkomponentenzahl t und Berechnung des jeweiligen $\overline{RSS}_t$- und $\overline{PRESS}_t$-Werts (abhängig von Hauptkomponentenzahl t).

Der Einsatz des Quotienten $\overline{PRESS}_t$ / $\overline{RSS}_{t-1}$ erfolgt wie bei der konventionellen Variante zur Ermittlung der optimalen Hauptkomponentenzahl. Abb. 1 verdeutlicht das Procedere.

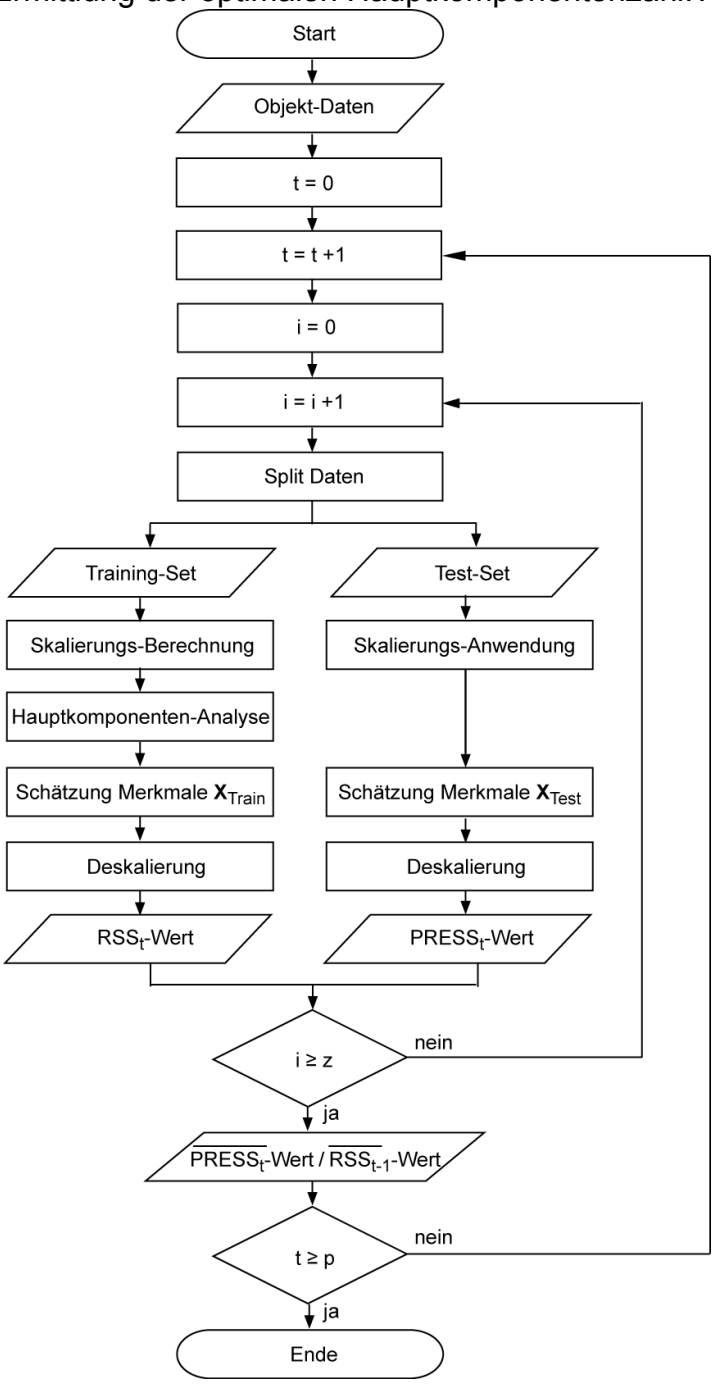

Abb. 1: Bootstrapping Kreuzvalidierung

Die Kreuzvalidierung mit Bootstrapping erbringt meist kleinere Hauptkomponentenzahlen als die konventionelle Variante. Es konnte gezeigt werden, das bei Einsatz von Hauptkomponenten (mit Bootstrapping) deutlich bessere Resultate der Clusteranalyse erbringt.

2.4. Abstands-, Distanzmaße

Allen Methoden der Clusteranalyse liegen die Abstände der betrachteten Objekte zueinander bzw. die Distanz eines Objekts zum betrachteten Cluster zugrunde.
Es sind eine Vielzahl von Abstands- (Euklidisch, Pearson, Mahalanobis ...) bzw. Distanz-Maßen (Average Linkage, Median Linkage, Single Linkage, Centroid Linkage ...) beschrieben [3], die einen Einfluss auf die ermittelten Cluster haben können. Im Hinblick auf die resultierenden Cluster bzw. deren Validierung sollten diverse Maße erprobt werden.

3. Validierung

Während bei der Klassifizierung (z.B. Diskriminanzanalyse) eine Vielzahl von Optionen zur Validierung verfügbar sind, ist dies bei der Clusteranalyse sehr limitiert.

3.1 Cluster-Differenzierung

Ein Maß für die Unterscheidung der Cluster ist wichtig z.B. zur Beurteilung unterschiedlicher Clusterzahlen bei der k-Means-Methode oder auch beim Vergleich der Clusteranalyse unterschiedlicher Daten (z.B. skaliert / standardisiert, Euklid- / Mahalanobis-Abstand).

Etabliert ist der Davies Bouldin Index DBI als Maß zum paarweisen Vergleich der Cluster [8][2].
Basis für eine Bewertung der Differenzierung der beiden Cluster g und h ist ein Quotient von mittlerem Objekt-Abstand (innerhalb Cluster g bzw. innerhalb Cluster h) und dem Zentroid-Abstand der beiden Cluster. Es gilt:

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

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

Räumlich eng begrenzte Cluster mit großer Distanz der Zentroide haben einen kleinen DBI zur Folge, ein Davies Bouldin Index < 1 steht für eine sehr gute Differenzierung, DBI < 5 ist akzeptabel. Bei mehr als 2 Clustern werden die Indices für jede Cluster-Kombination berechnet, daraus folgt der mittlere Index $\overline{DBI}$.
In der Regel basiert die Abstandsberechnung auf den Original-Variablen.
Bei Einsatz von Hauptkomponenten kann der $\overline{DBI}$ gegen die Hauptkomponenten-Zahl aufgetragen werden, die Hauptkomponenten-Zahl ab der sich der Davies Bouldin Index nicht mehr wesentlich ändert kann als optimale Anzahl Hauptkomponenten betrachtet werden.

3.2. Objektzuordnung

In manchen Fällen kann bei den Proben eine Zugehörigkeit angegeben werden, z.B. bei der Clusteranalyse der Honigproben kann zu Beginn ein Set von Honigproben eingesetzt werden, deren botanischer Ursprung (Orangenblüten-, Rosmarien-, Akazien-Honig ...) bekannt ist. Nach Optimierung der Clusteranalyse z.B. mittels k-Means-Methode kann in solchem Fall beurteilt werden, wie gut die Zuordnung der Proben erfolgte.

Ein weit verbreitetes Kriterium ist die prozentuale, korrekte Klassifizierung (%CC):
%CC: prozentualer Anteil der korrekten Klassifizierung

$$\%CC = 100 \frac{1}{n} \sum_{c=1}^{m} TP_c \quad (6)$$

mit n: Gesamtzahl Objekte, c: Cluster-Laufindex, m: Clusterzahl, TP_c: Zahl korrekt zugeordnete Objekte in Cluster c

2 Alternativ verwendet werden silhouhette width oder overlap coefficient, die numerisch aufwändiger sind, aber keine Vorteile gegenüber dem David-Bouldin-Index haben.

Es muss berücksichtigt werden, dass diese Aussage des %CC-Werts nur für die gerade vorliegenden Proben gültig ist, deren Zugehörigkeit bekannt war.

%MS-Wert (model stability): Maß für Zuordnung zum 1. bzw. 2. Cluster (2-Cluster-Fall), beschreibt, wie oft die Zuordnung eines Objekts wechselt bei Wiederholung der Clusteranalyse (z.B. k-Means-Methode mit unterschiedlichen Startparametern bzw. MASLOC-Methode mit unterschiedlichen Zentrotyp-Objekten).

$$\%MS_i = 100 \frac{|ZZ_1 - ZZ_2|}{v} \quad (7)$$

mit ZZ_1, ZZ_2: Zahl der Zuordnungen von Objekt i zu Cluster 1 bzw. 2, v: Gesamtzahl Zuordnungen Objekt i

Zum Beispiel wurde Objekt 5 insgesamt 21 x dem Cluster 1 und 7 x Cluster 2 zugeordnet, der $\%MS_5$-Wert ist 50%.
Ein $\%MS_i$-Wert um 0% steht dafür, dass das Objekt nicht sicher den gegebenen Clustern zugeordnet werden kann. Damit ist der $\%MS_i$-Wert ein Indikator für die Stabilität der Clusteranalyse, d.h. $\%MS_i$ = 100% bedeutet, das das Objekt i immer zuverlässig einem Cluster zugeordnet wird bei Wiederholung der Clusteranalyse.
Es kann auch der Mittelwert aller Objekte $\overline{\%MS}$ angegeben werden.

3.4. Null-Datenpool

Ein wesentlicher Baustein zur Validierung der Clusteranalyse ist der Einsatz eines zufallsbedingten Null-Datenpool.
Hierzu wird bei jedem Objekt für jedes verwendete Merkmal ein Wert im Bereich Minimum- und Maximum-Wert zufallsgeneriert.

Bei Einsatz des Null-Datenpool erfolgt keine relevante Zuordnung zu definierten Clustern, d.h. der Davies Bouldin Index sollte groß, der %CC-Wert etwa 50 %, der $\overline{\%MS}$-Wert etwa 0 % sein. Ist dies nicht der Fall, ist die gewählte Methode zur Clusteranalyse fragwürdig.

4 Zusammenfassung

Bei der Clusteranalyse gibt es nur wenige Optionen, die Objekt-Merkmale zu bewerten, dazu gehören u.a. die Korrelationen der Korrelationsmatrix als auch die Ladungen nach der Hauptkomponenten-Analyse.
Bei der Clusteranalyse (ohne bekannte Zugehörigkeit von Basis-Objekten) stehen nur wenig Parameter für eine Validierung zur Verfügung. Zur Bewertung der Hauptkomponenten-Zahl können Davies Bouldin Index DBI als Maß der Cluster-Differenzierung bzw. der Quotient $PRESS_t$ / RSS_{t-1} zur Beschreibung der Residuen verwendet werden, am besten in der Bootstrapping-Variante. Der Einsatz von Hauptkomponenten an Stelle der Original-Merkmale hat sich in vielen Fällen bewährt.
Ein weiterer Indikator zur Validierung ist der $\%MR_i$-Wert, der angibt, wie oft ein Objekt i bei wiederholter Clusteranalyse die Cluster-Zuordnung wechselt.
Zur Validierung kommt dem Einsatz eines Null-Datensatz besondere Bedeutung zu.

5 Literatur

[1] Fernadez-Torres, R. et al, Mineral content and botanical origin of Spanish honeys, Talanta, 65, 686-691, 2005
[2] Been, F., Profiling of conterfeit medicines by vibrational spectroscopy, Forensic Science International, 211, 83-100, 2011
[3] Reh, E., Chemometrie, Grundlagen der Statistik, Numerischen Mathematik und Software-Anwendung in der Chemie, de Gruyter, Berlin, 2017
[4] Silva,A.C., Pontes, L.F.B.L., Pimentel, M.F., Pontes, M.J.C., Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, Talanta, 93, 129– 134, 2012
[5] Brereton, R.G., Chemometrics for Pattern Recognition, Wiley-VCH, Weinheim, 2009
[6] Efro, B., Tibshirani, R.B., An Introduction to the Bootstrap, Chapman & Hall, New York, 1993
[7] Wehrens, R., Putter, H., Buydens, L.M., The Bootstrap: a tutorial, Chemometrics Intell. Lab. Systems, 54, 35-52, 2000
[8] Davies, D.L., Bouldin, D.W., A Cluster Separation Measure, IEEE Trans. Pattern Anal. Machine. Intell., 1, 224-227, 1979