G. Rohwer, C. Dudel Wintersemester 2010/11

Methoden der Datenrepräsentation und Klassifikation

Kapitel 5: Ansätze der Clusteranalyse

5 Ansätze der Clusteranalyse

In diesem Kapitel beginnen wir mit der Diskussion von Methoden der Clusteranalyse bzw. Klassifikation. Zunächst beschäftigen wir uns mit der Frage, wie Cluster definiert werden können, und geben eine kurze Übersicht über unterschiedliche Verfahren der Clusteranalyse. Dann werden partitionierende Verfahren besprochen.

Folgende Notationen werden verwendet: $\mathcal{N} = \{1, \ldots, n\}$ repräsentiert die Menge der Objekte, die von beliebiger Art sein können. Es wird angenommen, dass eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist, die für jeweils zwei Elemente $i, j \in \mathcal{N}$ einen Abstand d_{ij} angibt. Für Teilmengen von $\mathcal{N}$, die als Cluster betrachtet werden können, wird meistens der Buchstabe C, für Partitionen wird der Buchstabe P verwendet.

5.1 Unterschiedliche Ansätze

1. Wie können Cluster definiert werden?

Clusteranalyse dient hier als Sammelbegriff für eine breite Palette von Verfahren, deren Gemeinsamkeit im Wesentlichen nur darin besteht, dass sie die Objektmenge $\mathcal{N}$ irgendwie in Cluster einteilen. Somit stellt sich zunächst die Frage, wie Cluster definiert werden können. Eine gelegentlich verfolgte Idee für die Definition von Clustern wurde von K. D. Bailey (1983: 255) so formuliert:

„It is axiomatic in cluster analysis, as in all classification, that individuals or variables in a class are considered to be more similar to each other than to individuals or variables not in the class."

Tatsächlich führt diese Idee zu einer sehr engen Clusterdefinition, die sich nur selten realisieren lässt. Um das deutlich zu machen, präzisieren wir zunächst die Formulierung. Eine echte Teilmenge $C \subset \mathcal{N}$ wird ein *separierbares* oder kurz ein *S-Cluster* genannt, wenn C mindestens zwei Elemente hat und folgende Bedingung erfüllt ist:

$$\max\{d_{ij} \,|\, i,j \in C\} < \min\{d_{ij} \,|\, i \in C, j \notin C\} \tag{5.1}$$

$\max\{d_{ij} \,|\, i,j \in C\}$ wird auch als *Durchmesser* des Clusters C bezeichnet. Die Bedingung sagt, dass der Clusterdurchmesser kleiner sein sollte als der kleinste Abstand zu einem Objekt außerhalb des Clusters.

Verwendet man das so formulierte Kriterium, lautet die Frage, ob sich $\mathcal{N}$ in zwei oder mehr S-Cluster einteilen lässt. Wie man sehen wird, ist das oft nicht der Fall.

Als Beispiel betrachten wir eine Menge $\mathcal{N} = \{1, \ldots, 5\}$, die fünf Schulabschlüsse repräsentiert: 1 = ohne Hauptschulabschluss, 2 = Hauptschulabschluss, 3 = Realschulabschluss, 4 = Fachhochschulreife, 5 = Abitur. Es wird folgende Abstandsmatrix angenommen:

$$\begin{pmatrix} 0.0 & 2.0 & 3.0 & 4.5 & 5.5 \\ 2.0 & 0.0 & 1.0 & 2.5 & 3.5 \\ 3.0 & 1.0 & 0.0 & 1.5 & 2.5 \\ 4.5 & 2.5 & 1.5 & 0.0 & 1.0 \\ 5.5 & 3.5 & 2.5 & 1.0 & 0.0 \end{pmatrix} \tag{5.2}$$

Man erkennt, dass es kein S-Cluster gibt, das den Schulabschluss 1 als Element enthält (denn wollte man den Abschluss 2 hinzufügen, müsste man auch alle anderen Abschlüsse mit aufnehmen). Es liegt also nahe, vor der Bildung von S-Clustern alle Elemente aus $\mathcal{N}$ zu entfernen, die keine Elemente von S-Clustern sein können. In unserem Beispiel ist das das Element 1. Es bleibt die reduzierte Menge $\{2,3,4,5\}$, die sich in die beiden S-Cluster $C_1 = \{2,3\}$ und $C_2 = \{4,5\}$ zerlegen lässt.

Als ein weiteres Beispiel betrachten wir die Berufsstrukturdaten aus Abschnitt 2.3. Geht man von der Abstandsmatrix für die acht Länder aus (Tabelle 2.3-2), findet man, dass die Länder 6 (Schweden) und 8 (Japan) nicht für S-Cluster verwendet werden können. Die restlichen Länder können in drei S-Cluster partitioniert werden: (Griechenland, Türkei), (Deutschland, Schweiz), (Grossbritannien, USA).

In der Literatur findet man eine Fülle unterschiedlicher Clusterkonzeptionen, die im Vergleich zur Idee der S-Cluster oft nur sehr schwache Anforderungen stellen. Hier sind einige Beispiele.

„Classification can be described as the activity of dividing a set of objects into a smaller number of classes in such a way that objects in the same class are similar to one another and dissimilar to objects in other classes.“ (Gordon 1987: 119)

„[...] an investigator would like to group together variables so that they are as homogenous as possible within subsets, and as different as possible between subsets.“ (Cliff et al. 1986: 201)

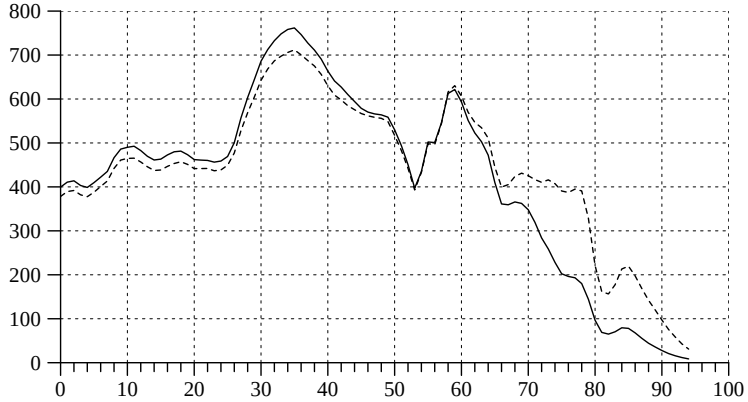

Abb. 5.1-1 Altersverteilung (absolute Häufigkeiten in 1000) der Männer (durchgezogene Linie) und Frauen (gestrichelte Linie) in Deutschland 1999.

„*Cluster analysis* refers to a wide variety of techniques used to group entities into homogeneous subgroups on the basis of their similarities.“ (Lorr 1983: 1)

„Basically, one wants to form groups in such a way that objects in the same group are similar to each other, whereas objects in different groups are as dissimilar as possible.“ (Kaufman und Leonard 1990: 1)

„Roughly speaking, the goal of a clustering algorithm is to group the objects of a database into a set of meaningful subclasses.“ (Ankerst et al. 1999: 49-60)

Offenbar liefern diese Zitate nur relativ vage Hinweise, wie Cluster gebildet werden sollten, jedoch keine Definitionen des Clusterbegriffs.

2. Abstände und Häufigkeiten

Bei vielen Überlegungen zur Clusteranalyse vermischen sich zwei Ideen: Einerseits die Idee, dass Objekte innerhalb desselben Clusters ähnlich sein sollten; und andererseits eine Idee, die mit Häufigkeiten operiert: dass Cluster aus denjenigen Objekten gebildet werden sollten, die „gehäuft“ vorkommen. Zum Beispiel:

„The goal of clustering, in general, is to discover dense and sparse regions in a dataset.“ (Ganti, Gehrke und Ramakrishnan 1999: 73)

Es ist bemerkenswert, dass es keinen wesentlichen Zusammenhang zwischen den beiden Ideen gibt. Überlegungen, die mit Ähnlichkeiten bzw. Abständen argumentieren, sind zunächst von ganz anderer Art als Überlegungen, die mit Häufigkeiten argumentieren. Während Häufigkeiten eine Bezugnahme auf Daten voraussetzen, ist das bei einer Betrachtung von Abständen nicht erforderlich. Überlegungen, die mit Abständen argumentieren, können sich beispielsweise auf Merkmalsräume beziehen, ohne dass Daten erforderlich sind. Ein gutes Beispiel ist die Kemeny-Metrik für Rangordnungen (vgl. Abschnitt 2.1, § 3).

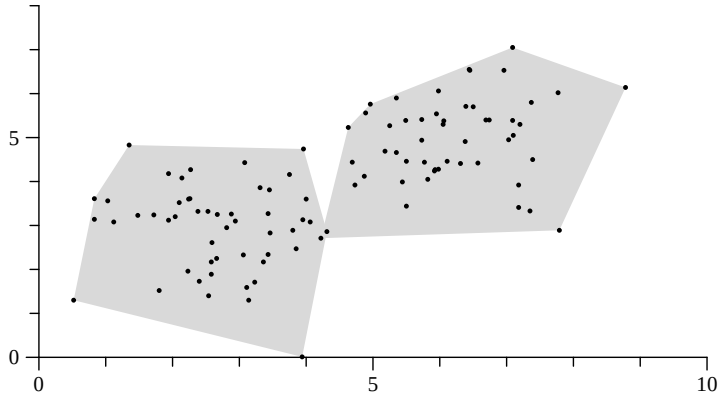

Abb. 5.1-2 100 mit einer zweidimensionalen Normalverteilung erzeugte Punkte; 50 mit dem Mittelwert $(3, 3)$, 50 dem Mittelwert $(6, 5)$.

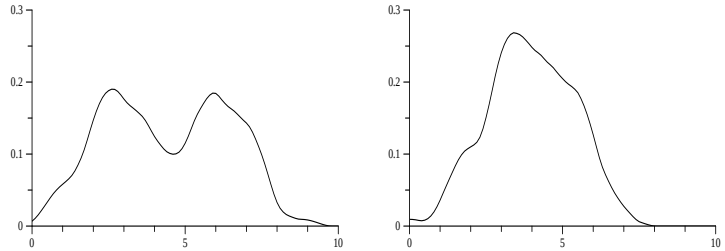

Abb. 5.1-3 Häufigkeitsverteilungen für die X- und Y-Koordinaten der 100 Punkte in Abbildung 5.1-2.

Um die Problematik von an Häufigkeiten orientierten Clusteranalysen zu verdeutlichen, genügt bereits eine Betrachtung eindimensionaler Häufigkeitsfunktionen. Als Beispiel betrachte man die Altersverteilungen in Abbildung 5.1-1. Kann man anhand dieser Häufigkeitsfunktionen sinnvolle Cluster abgrenzen?

Es ist nützlich, sich die unterschiedlichen Ansätze auch anhand eines zweidimensionalen Beispiels zu verdeutlichen. Dafür verwenden wir 100 Werte einer zweidimensionalen Normalverteilung, die ersten 50 mit dem Mittelwert $(3, 3)$ die anderen 50 mit dem Mittelwert $(6, 5)$. Abbildung 5.1-2 zeigt die erzeugten Punkte.[1]

Wie könnte in diesem Beispiel ein Ansatz verfolgt werden, der sich an Häufigkeiten orientiert. Abbildung 5.1-3 macht deutlich, dass es jedenfalls

[1] Die Daten wurden mit dem Skript `cl1.cf`, die Abbildung mit `clplot1.cf` erzeugt.

nicht genügen würde, nur die eindimensionalen Häufigkeitsprojektionen zu betrachten.

Dagegen führt eine Orientierung an Abständen zu einer anderen Idee. Sie besteht darin, einige Punkte als Clusterzentren auszuwählen und dann alle übrigen Punkte demjenigen Clusterzentrum zuzuordnen, zu dem ihr Abstand am kleinsten ist. Dies ist die Grundidee der sogenannten partitionierenden Verfahren der Clusteranalyse.

3. Eine unvollständige Übersicht

Zum Abschluss dieses Abschnitts geben wir eine kurze Übersicht über die in der Literatur hauptsächlich verfolgten und verwendeten Ansätze der Clusteranalyse.

- Partitionierende Verfahren, bei denen die Anzahl der Cluster vorgegeben werden muss und dann versucht wird, optimale Zuordnungen der Objekte zu Clustern zu finden; dabei werden optimale Zuordnungen durch unterschiedliche Kriterien definiert. Einige dieser Verfahren werden im nächsten Abschnitt besprochen.
- Hierarchische Verfahren, die nicht unmittelbar Cluster erzeugen, sondern eine hierarchische Repräsentation der Struktur einer Abstandsmatrix liefern. Es gibt hauptsächlich zwei Arten: agglomerative und divisive Verfahren. Einige dieser Verfahren werden im nächsten Kapitel besprochen.
- Verfahren, die durch Dichotomisierungen einer Abstandsmatrix erzeugte Graphen verwenden. Solche Verfahren werden in Abschnitt **??** besprochen.
- Verfahren, die sich explizit an Häufigkeiten orientieren. Solche Verfahren werden in diesem Text nicht besprochen.[2]
- Schließlich kann hier auch noch auf eine weitere Vorgehensweise hingewiesen werden, die darin besteht, zunächst räumliche Bilder (einer Abstandsmatrix) zu erzeugen (z.B. mit Verfahren der multidimensionalen Skalierung oder Korrespondenzanalyse) und dann Cluster durch visuelle Anschauung zu bestimmen.[3]

5.2 Partitionierende Verfahren

In diesem Abschnitt besprechen wir einige Ansätze der partitionierenden Clusteranalyse. Hierbei muss die Anzahl der zu bildenden Cluster, im

[2] Vgl. Everitt (1993: Kap. 6); Ankerst et al. (1999).

[3] Diese Vorgehensweise wird oft vorgeschlagen, man vgl. beispielsweise Kruskal und Wish (1978: 43ff.), Lorr (1983: 45). Es gibt jedoch auch Kritik, vgl. die Hinweise bei Bailey (1994: 73).

Folgenden k genannt, vorgegeben werden. Gesucht ist dann eine Partitionierung der Objektmenge in k disjunkte Teilmengen $C_1, \ldots, C_k$, so dass es sich um gut definierte Cluster handelt. Dafür können unterschiedliche Kriterien verwendet werden. Man kann in erster Linie zwei Ansätze unterscheiden. Einerseits Ansätze, die für jedes Cluster ein Clusterzentrum suchen und dann alle Objekte ihrem nächstgelegenen Clusterzentrum zuordnen; andererseits Ansätze, die ohne Clusterzentren auskommen. Wie sich unterschiedliche Partitionierungen vergleichen lassen, wird im nächsten Abschnitt besprochen.

1. Ansätze mit Clusterzentren

a) Ein erstes Kriterium setzt voraus, dass die Objekte durch Zeilen (oder Spalten) einer Datenmatrix gegeben sind:

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{im})' \qquad (i = 1, \ldots, n)$$

und dass euklidische Abstände verwendet werden. Dann kann für jedes Cluster C_l ein Mittelwert

$$\bar{\mathbf{x}}_l := \frac{1}{n_l} \sum\nolimits_{i \in C_l} \mathbf{x}_i$$

definiert werden (n_l bezeichnet die Anzahl der Elemente in C_l), und es wird möglich, folgendes Kriterium zu verwenden:

$$\sum_{l=1}^{k} \sum_{i \in C_l} \| \mathbf{x}_i - \bar{\mathbf{x}}_l \|^2 \longrightarrow \min \tag{5.3}$$

Es wird auch als *Abstandsquadratsummenkriterium* bezeichnet. Jeder Vektor $\mathbf{x}_i$ wird dem nächstgelegenen Clusterzentrum $\bar{\mathbf{x}}_l$ zugeordnet, und die Clusterzentren sollen so bestimmt werden, dass die gesamten quadrierten Abstände zu den Clusterzentren möglichst klein werden.

b) Wenn zunächst eine beliebige Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist, kann das Abstandsquadratsummenkriterium nicht verwendet werden. Man kann stattdessen Objekte – zunächst willkürlich – bestimmen, die als Clusterzentren dienen sollen. Hat man beispielsweise m Objekte als Zentrumsobjekte für m Cluster festgelegt, kann man alle übrigen Objekte jeweils demjenigen Cluster zuordnen, zu dessen Zentrumsobjekte sie den geringsten Abstand aufweisen. Ein weiter (rechentechnisch sehr aufwendiges) Problem besteht dann allerdings darin, diejenigen Zentrumsobjekte zu finden, die eine global optimale Lösung liefern.

2. Ansätze ohne Clusterzentren

Die beiden in § 1 besprochenen Kriterien verwenden Clusterzentren. Man kann auch versuchen, ohne Clusterzentren auszukommen.

a) Man kann beispielsweise folgendes Kriterium betrachten:

$$\sum_{l=1}^{k} \frac{1}{n_l} \sum_{i,j \in C_l : j < i} d_{ij} \longrightarrow \min \tag{5.4}$$

Unterschiedliche Varianten des Kriteriums entstehen, wenn man nicht durch n_j, sondern beispielsweise durch $n_j(n_j - 1)$ dividiert; H. Späth, der mit unterschiedlichen Varianten ausführlich experimentiert hat, hält die in (5.4) angegebene Variante für die praktisch brauchbarste.[4]

b) Anstatt sich auf eine Art von durchschnittlicher Abstandsgröße in den Clustern zu beziehen, wie bei den Kriterien der Art (5.4), kann man auch den durch $\max\{d_{ij} \,|\, i, j \in C_l\}$ definierten Clusterdurchmesser verwenden. Dann entsteht folgendes Kriterium:

$$\sum_{l=1}^{k} \max\{d_{ij} \,|\, i, j \in C_l\} \longrightarrow \min \tag{5.5}$$

In diesem Fall sollen also die Cluster so gebildet werden, dass die Summe ihrer Durchmesser minimal wird.

3. Rechentechnische Probleme

Die Hauptschwierigkeit resultiert daraus, dass die Anzahl der Möglichkeiten zur Einteilung einer Menge von N Objekten in k Cluster schnell außerordentlich groß wird. Zum Beispiel kann man 10 Objekte auf 34105 verschiedene Weisen in vier Cluster einteilen; aber bei 19 Objekten gibt es bereits 11,259,666,000 Möglichkeiten.[5] Es ist deshalb in den meisten Fällen praktisch nicht möglich, Cluster zu finden, die den globalen Minima der oben angegebenen Kriterien entsprechen.

Die normalerweise verwendeten Verfahren können nur lokale Minima der Kriterien finden. Oft handelt es sich um sog. Austauschverfahren. D.h. ausgehend von einer (irgendwie, zufällig) gebildeten Anfangspartition werden solange Objekte zwischen den Partitionen ausgetauscht, bis sich das Kriterium nicht weiter verkleinern lässt. Für das Kriterium (5.3) wird eine besonders oft verwendete Variante dieses Austauschverfahrens als *k-means Algorithmus* bezeichnet.[6] Für das Kriterium (5.4) wurde ein

[4] Vgl. Späth (1983: 92ff.; 1988).

[5] Vgl. Jain und Dubes (1988: 91).

[6] Vgl. Hartigan (1975: Kap. 4); Bacher (1994: 308ff.).

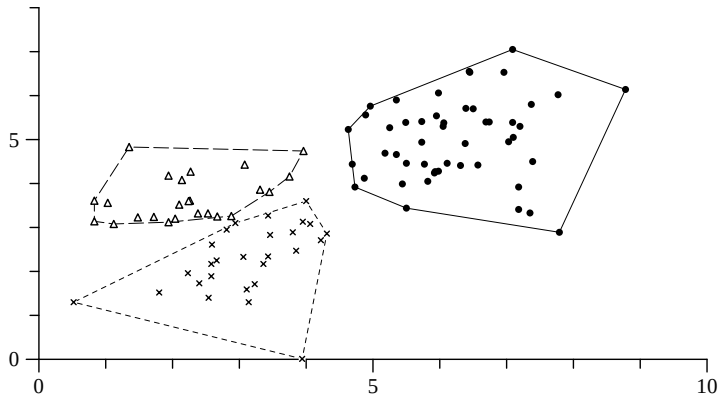

Abb. 5.2-1 Einteilung der 100 Punkte aus Abbildung 5.1-2 in drei Cluster unter Verwendung des Kriteriums (5.4). Die Kreise deuten die Clustermittelpunkte an.

Austauschverfahren von H. Späth entwickelt.[7] Teilweise andere Verfahren wurden für das Kriterium (5.5) vorgeschlagen.[8]

Bei der Verwendung partitonierender Verfahren sollte man also daran denken, dass die normalerweise verfügbaren Programme nur lokale Minima der Kriterien finden können. Es ist deshalb sinnvoll, die Berechnungen mit unterschiedlichen Anfangspartitionen zu wiederholen, um einen gewissen Einblick in das Auftreten unterschiedlicher lokaler Minima zu gewinnen.

4. Illustration mit artifiziellen Daten

Zur Illustration partitionierender Verfahren verwenden wir das Kriterium (5.4). Wir beginnen mit den in Abschnitt 5.1 (§ 5) beschriebenen artifiziellen Daten: 100 Realisierungen einer zweidimensionalen Normalverteilung (Abbildung 5.1-2). Daraus wird eine $(100, 100)$-Matrix mit euklidischen Abständen gebildet und als Eingabe für ein partitionierendes Verfahren verwendet.[9]

Versucht man, durch Minimierung des Kriteriums (5.4) zwei Cluster zu bilden, erhält man bei 100 Wiederholungen mit zufällig gebildeten Anfangspartitionen als beste Lösung folgende Einteilung: Die Punkte 1–51 (ohne Nr. 32) gehören zum ersten, die Punkte 32 und 52–100 zum zweiten Cluster.[10] Bis auf den Punkte Nr. 32 (der mit den Koordinaten $(3.96, 4.74)$ einen Grenzfall bildet) entspricht dies Ergebnis dem datenerzeugenden Prozess.

Aber warum zwei Cluster? Bildet man drei Cluster, entsteht sogleich ein vollständig anderes Bild, s. Abbildung 5.2-1.[11] In diesem Fall wird auch bei 100 Wiederholungen ein optimales Ergebnis nur in acht Fällen erreicht.

[7] Vgl. Späth (1983: 143ff.). Dieses Verfahren liegt auch der TDA-Prozedur `clp` zugrunde, die für die späteren Illustrationen verwendet wird.

[8] Vgl. Hansen und Jaumard (1987); Charikar und Panigrahy (2001).

[9] Das Datenfile mit der Abstandsmatrix wurde mit dem Skript `cl1a.cf` erzeugt und `cl1a.dat` genannt.

Box 5.2-1 Verwendung der Berufsstrukturdaten aus Tabelle 2.3-3 für eine Einteilung der Länder in zwei, drei bzw. vier Cluster.

- Türkei
- Griechenland
- Deutschland
- Japan
- Schweiz
- Schweden
- Grossbritannien
- USA

5. Beispiele mit Berufsstrukturdaten

Jetzt verwenden wir die Berufsstrukturdaten aus Abschnitt 2.3. Wir beginnen mit der Abstandsmatrix in Tabelle 2.3-3 für die acht Länder. Folgende Tabelle zeigt das Ergebnis, wenn man zwei, drei bzw. vier Cluster bildet:[12]

Land	1	2	3	4	5	6	7	8
2 Cluster	1	1	2	2	2	2	2	2
3 Cluster	1	1	3	2	3	2	2	3
4 Cluster	1	1	2	4	2	3	4	2

In diesem Beispiel entsteht auch eine hierarchische Struktur, wie sie in Box 5.2-1 noch einmal verdeutlicht wird. (Das ist bei partitionierenden Verfahren im Allgemeinen nicht der Fall.)

Verwendet man die Abstandsmatrix für die geschlechtsspezifischen Berufsverteilungen aus Abschnitt 2.3 (§ 6), findet man wiederum eine hierarchische Struktur:[13]

$$\Big(\mathrm{M1}, \ldots, \mathrm{M8}\Big)\ \Big(\big(\mathrm{F1}, \mathrm{F2}, \mathrm{F8}\big)\big((\mathrm{F3}, \mathrm{F4}, \mathrm{F5}, \mathrm{F7})(\mathrm{F6})\big)\Big)$$

[10] Verwendet wurde das Skript `clp1.cf`.

[11] Verwendet wurde das Skript `clp1a.cf`.

[12] Verwendet wurde das Skript `clp2.cf`.

[13] Erzeugt mit dem Skript `clp3.cf`.

5.3 Vergleiche von Partitionen

Manchmal möchte man zwei oder mehr Partitionen derselben Objektmenge daraufhin vergleichen, wie ähnlich oder unähnlich die durch sie vorgenommenen Klassifikationen sind. Erforderlich ist dann eine Abstandsfunktion, die jeweils zwei Partitionen für dieselbe Objektmenge eine Zahl zuordnet, die als ihr Abstand interpretiert werden kann. In der Literatur findet man viele unterschiedliche Vorschläge.[14] In diesem Abschnitt besprechen wir zwei Ideen.

1. Ein einfacher Index

Bei der ersten Idee werden alle Paare von Objekten betrachtet, und bei jedem Paar wird festgestellt, ob die beiden Objekte von einer Partition in dieselbe oder in eine unterschiedliche Klasse eingeordnet werden. Daraus wird dann ein Index gebildet.[15]

Um eine genaue Definition zu geben, beziehen wir uns auf zwei Partitionen P und P' für die Objektmenge $\mathcal{N} = \{1, \ldots, n\}$. Für die Partition P wird eine (n, n)-Matrix $\mathbf{P} = (p_{ij})$ gebildet: $p_{ij} = 1$, wenn die Objekte i und j durch die Partition P demselben Cluster zugeordnet werden; $p_{ij} = 0$ andernfalls. Analog wird eine Matrix $\mathbf{P}' = (p'_{ij})$ für die Partition P' gebildet. Dann wird ein Abstand zwischen den Partitionen durch

$$d(P, P') := \frac{2}{n\,(n-1)} \sum_{j<i} |\, p_{ij} - p'_{ij} \,| \tag{5.6}$$

definiert. Es wird also erfasst, welchen Anteil an den insgesamt $n(n-1)/2$ Paaren diejenigen Paare haben, bei denen die Objekte unterschiedlichen Clustern zugeordnet werden. Der maximale Wert des Index ist offenbar 1.

Zur Illustration verwenden wir die drei Partitionen für die Berufsstrukturdaten, die in § 5 des vorangegangenen Abschnitts gebildet wurden (2 Cluster: P, 3 Cluster: P', vier Cluster P"). Die zu P' gehörende Matrix sieht beispielsweise so aus:

$$\mathbf{P}' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

[14] Man vgl. etwa Hubert und Arabie (1985).

[15] Eine axiomatische Begründing für diesen Index haben Mirkin und Chernyi (1970) gegeben.

Bildet man analog die Matrizen $\mathbf{P}$ und $\mathbf{P}''$, findet man folgende Abstände: $d(P, P') = 9/28 = 0.32$, $d(P, P'') = 11/28 = 0.39$ und $d(P', P'') = 2/28 = 0.07$. Offenbar sind sich P' und P'' sehr ähnlich.

2. Substitutionsmetriken

Ein anderer Ansatz orientiert sich an der Idee einer Substitutionsmetrik.[16] Als Leitfaden dient die Frage, wieviele Operationen erforderlich sind, um zwei Partitionen einer Objetmenge in Übereinstimmung zu bringen. Dabei ist zu berücksichtigen, dass die Partitionen eine unterschiedliche Anzahl von Clustern aufweisen können.

Angenommen, die Partitionen P und P' bestehen aus den Clustern $\{C_1, \ldots, C_m\}$ bzw. Clustern $\{C'_1, \ldots, C'_{m'}\}$ und $m \geq m'$. Dann werden zur Partition P' $m - m'$ leere Cluster hinzugefügt, so dass P' ebenfalls aus m Clustern besteht. Dann kann man alle möglichen Zurordnungen zwischen den Clustern $C_1, \ldots, C_m$ und $C'_1, \ldots, C'_m$ betrachten und für jede Zuordnung die Anzahl der Objekte bestimmen, die einem anderen Cluster zugeordnet werden müssen, um die Partitionen in Übereinstimmung zu bringen. Schließlich verwendet man diejenige Zuordnung der Cluster, bei denen die Anzahl der erforderlichen Vertauschungsoperationen minimal ist.

Verwendet man zur Illustration wieder die Partitionen P, P' und P'' für die Berufsstrukturdaten, findet man, dass drei Operationen erforderlich sind, um P und P' in Übereinstimmung zu bringen, dass man für P' und P'' jedoch nur eine Operation benötigt.

[16] Day ...; Charon et al. (2006).