Kategorisierung von Indizes zur Clustervalidierung

Studienarbeit

Philipp Plöhn
Wirtschaftsinformatik

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Kategorisierung von Indizes zur Clustervalidierung
Categorization of Cluster Validity Indices

Vorgelegte Studienarbeit von Philipp Plöhn

1. Gutachten:
2. Gutachten:

Tag der Einreichung:

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht.

Darmstadt, den 29. September 2014

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Clusteranalyse ist eine unüberwachte Methode zur Gruppierung von Daten. Im Gegensatz zur Klassifikation sind die wahren Klassenzugehörigkeiten nicht bekannt. Die Clusteranalyse wird mittels eines Clustering-Algorithmus ausgeführt, der die Daten in Cluster aufteilt. Das Ziel ist, dass die Objekte innerhalb eines Clusters sehr ähnlich sind, während sich Objekte in verschiedenen Clustern stark unterscheiden. Die „natürliche“ Struktur der Daten soll damit aufgedeckt werden. Mögliche Anwendungen der Clusteranalyse sind die Datenreduzierung sowie die Bildung von Hypothesen und Prognosen (Halkidi et al., 2001), die in vielen Bereichen eingesetzt werden: Bioinformatik (Handl et al., 2005), Machine Learning (Strehl & Ghosh, 2002), Marketing (Dimitriadou et al., 2002), Pattern Recognition (Arbelaitz et al., 2013) und Strategisches Management (Ketchen Jr. & Shook, 1996).

Die Clusteranalyse besteht aus drei Phasen: Vorbereitung, Clustering und Clustervalidierung (Halkidi et al., 2001; Handl et al., 2005). Zur Vorbereitung gehören die Auswahl der relevanten Variablen und die Normalisierung der Daten. Anschließend müssen ein Clustering-Algorithmus und dessen Parameter ausgewählt werden, damit das Clustering durchgeführt werden kann. Danach erfolgt die Clustervalidierung, denn nach der Durchführung eines Clusterings ist nicht bekannt, wie gut die Cluster zu den Daten passen. Die Cluster-Algorithmen können nicht sicherstellen, dass die „perfekte“ Partitionierung gefunden wird. Es müssen daher mehrere Algorithmen und deren Parameter ausprobiert werden. Die verwendeten Clustering-Algorithmen werden stets eine Lösung finden, auch wenn keine Struktur in den Daten vorhanden ist. Um diese fehlerhaften Clusterings aufzudecken, ist die Clustervalidierung notwendig. Dazu wird evaluiert, wie gut die ermittelten Cluster zu den zugrundeliegenden Daten passen. Auch das Aufrufen des Algorithmus mit einem falschen Parameter, wie einer zu hohen Clusteranzahl, wird zu einer suboptimalen Lösung führen. Der Grund ist, dass die meisten Algorithmen nicht die „perfekte“ Clusteranzahl ermitteln können. Die Algorithmen müssen für verschiedene Clusteranzahlen verglichen werden. Die Lösungen müssen mit Hilfe der Clustervalidierung verglichen werden, um festzustellen, welche Clusteranzahl optimal ist. Eine Darstellung der Daten um visuell die korrekte Anzahl der Cluster zu bestimmen, ist nur bis drei Dimensionen möglich. Diese Beschränkung wird in realen Anwendungen regelmäßig überschritten, daher werden Indizes zur Clustervalidierung verwendet.

Es wurden schon sehr viele Clustervalidierungsindizes vorgeschlagen und fast genauso viele Vergleiche durchgeführt. Es konnte sich allerdings kein Index durchsetzen. Meist wurde

festgestellt, dass bestimmte Indizes in bestimmten Situation besser abschneiden. Die Folge ist, dass viele Verfahren ausprobiert werden und je nach Situation andere Methoden zur Anwendung kommen. Die Auswahl der Indizes gestaltet sich schwierig, da kein umfassendes Framework zur Strukturierung der Clustervalidierungsindizes existiert. Es ist allerdings bekannt, dass einige Indizes dieselben Konzepte verwenden, um die Cluster zu validieren. Diese Indizes lassen sich möglicherweise zu einem Prototyp zusammenfassen, wenn sie Clusterings identisch bewerten. In dieser Arbeit soll diese Frage geklärt werden:

> **Forschungsfrage**: *„Wie können Clustervalidierungsindizes kategorisiert werden und gibt es innerhalb der Kategorien Prototypen von Indizes?“*

Die Beantwortung der Forschungsfrage wurde wie folgt angegangen. Zuerst wurde eine umfangreiche Literaturrecherche durchgeführt. Dazu wurden Studien berücksichtigt, die neue Clustervalidierungsindizes entwickelt haben oder die Übersichten und Vergleiche von Indizes erstellt haben. Die Arbeiten wurden systematisch aufbereitet und Informationen über die Indizes gesammelt. Wichtig waren vor allem die verwendeten Maße und die zugrundeliegenden Konzepte. Darauf aufbauend konnten die Indizes kategorisiert werden und Ähnlichkeiten festgestellt werden. Für die größte Gruppe der Indizes wurde genauer analysiert, wie Clusterings bewertet werden. Die Beurteilungen wurden untereinander und mit Prototypen verglichen, um identische Indizes zu erkennen.

Zunächst werden die möglichen Unterteilungen der Clustervalidierung erläutert und zu einer kompletten Kategorisierung zusammengefasst. Für die größte Kategorie werden die vorhandenen Konzepte vorgestellt. In den anschließenden Kapiteln erfolgt die Beschreibung der Clustervalidierungsindizes getrennt nach ihren Kategorien in chronologischer Reihenfolge. Danach werden Prototypen erstellt und mit vorhandenen Indizes verglichen. Abschließend werden die Ergebnisse zusammengefasst und ein Ausblick gegeben.

2 Clustervalidierungsindizes

In diesem Kapitel wird die Theorie hinter den Clustervalidierungsindizes erläutert. Dazu werden zunächst die Kategorisierung der Indizes vorgestellt, um die unzähligen Indizes zu strukturieren. Anschließend werden die grundlegenden Konzepte erläutert, die zur Berechnung der Indizes verwendet werden.

2.1 Kategorien der Clustervalidierungsindizes

Es wurden schon sehr viele Clustervalidierungsindizes vorgeschlagen, die alle unterschiedliche Vor- und Nachteile besitzen. Es gibt daher keinen Index, der immer am besten abschneidet. Viele Indizes sind für bestimmte Situation entwickelt worden und schneiden in diesen dann besonders gut ab. Um Clustervalidierungsindizes zu verstehen und korrekt auszuwählen, müssen die vorhandenen Indizes demnach strukturiert aufbereitet werden. Im Folgenden werden mögliche Unterteilungen der Clustervalidierungsindizes vorgestellt.

2.1.1 Verfügbare Information

Die meistverwendete Kategorisierung der Clustervalidierung unterteilt die Indizes nach der Verfügbarkeit von Informationen (Halkidi et al., 2001; Brun et al., 2007; Pfitzner et al., 2008; Arbelaitz et al., 2013). Es werden drei Arten unterschieden: interne, relative und externe Clustervalidierung (vgl. Abbildung 6).

Interne Indizes verfügen nur über Informationen, die auch für die Clusteranalyse verwendet oder dabei erstellt werden. Dies sind die Datenmatrix und das Clustering. Externen Indizes sind zudem die wahren Klassen bekannt, die mit dem Clustering verglichen werden können. In realen Anwendungen sind diese Informationen nicht vorhanden, so dass die Anwendbarkeit der externen Clustervalidierung eingeschränkt ist. Sie kann höchstens verwendet werden, um Methoden der Clusteranalyse, wie Algorithmen und interne Indizes, auf bekannten Datensätzen zu vergleichen. Die relative Clustervalidierung stellt eine Mischform der internen und externen Validierung dar.[1] Die relativen Indizes verfügen über

[1] Zu beachten ist, dass in der Literatur keine Einigkeit über die Trennlinie zwischen interner und relativer Clustervalidierung besteht. Teilweise werden die internen Indizes zur relativen Clustervalidierung zugeordnet, da sie dazu verwendet werden Clusterings mit verschiedenen Clusterzahlen zu vergleichen. In

dieselben Informationen wie interne Indizes, vergleichen aber Clusterings analog zu externen Indizes. Die Clusterings werden mit verschiedenen Datenteilen, Algorithmen und Parametern erstellt und untereinander verglichen. Ein Framework für alle drei Arten der Clustervalidierung zu erstellen ist sehr schwierig und bisher noch nicht gelungen (Arbelaitz et al., 2013). Es wurde allerdings eine **Studie** durchgeführt, die alle drei Typen vergleicht (Brun et al., 2007). Dabei erzielten externe Indizes die höchste Übereinstimmung mit der optimalen Clusterzahl und schnitten, wie erwartet, am besten ab. Die Unterscheidung zwischen den Indizes, die über keine externen Informationen verfügen, ist nicht so eindeutig. Die relativen Indizes erreichten leicht höhere Erkennungsraten als die internen Indizes. Dafür brauchen sie einen viel höheren rechnerischen Zeitaufwand, da zur relativen Clustervalidierung etliche Wiederholungen der Clusterings notwendig sind. Die nicht signifikant besseren Ergebnisse rechtfertigen daher nicht zwingend die Bevorzugung vor den größtenteils einfach zu berechnenden internen Indizes.

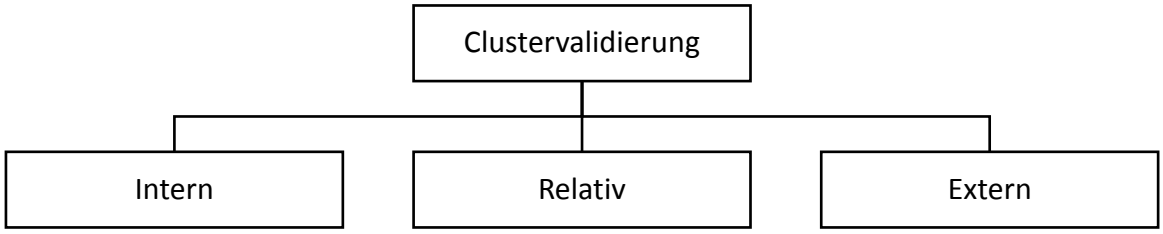

Abbildung 1: Kategorisierung der Clustervalidierung nach den verfügbaren Informationen.

2.1.2 Ergebnis des Clustering-Algorithmus

Clustervalidierungsindizes können nach dem Ergebnis des verwendeten Clustering-Algorithmus unterteilt werden (Kim & Ramakrishna, 2005). Hauptsächlich gibt es zwei Typen von Ergebnissen: harte und weiche Clusterings (vgl. Abbildung 6). Harte Clusterings, wie sie von K-Means und den Linkage-Verfahren erzeugt werden, ordnen jedes Objekt genau einem Cluster zu. Weiche Clusterings, wie von Fuzzy C-Means, ordnen jedem Objekt einen Gewichtsvektor zu, der angibt wie stark ein Objekt zu dem jeweiligen Cluster gehört. Ein hartes Clustering kann daher als Spezialfall des weichen Clusterings angesehen werden, bei dem alle Gewichte eines Objekts 0 sind außer dem Gewicht für einen Cluster, das 1 beträgt.

Die Unterteilung nach dem betrachteten Clustering-Algorithmus wird implizit in vielen **Studien** durchgeführt, die sich auf die Clustervalidierung einer der beiden Kategorien

dieser Arbeit werden aber nur Methoden, die explizit mehrere Clusterings vergleichen um die Stabilität zu evaluieren, als relative Indizes bezeichnet.

beschränken. Mit der Validierung der harten Clusterings beschäftigt sich die Forschung schon länger. Die erste große und immer noch meist zitierte Übersicht zu internen Clustervalidierungsindizes für harte Clusterings stammt von Milligan & Cooper (1985). In dieser Studie wurden 30 Indizes für vier Algorithmen und viele Datensätze mit unterschiedlichen Eigenschaften verglichen. Die Ergebnisse dieses Vergleiches hatten großen Einfluss und lange Bestand. Die Methodik wurde allerdings häufiger kritisiert (Halkidi et al., 2001; Vendramin et al., 2010; Gurrutxaga et al., 2011). Eine aktuelle und ausführliche Vergleichsstudie, die die Kritik beachtet und ein neues Verfahren verwendet hat, erstellten Arbelaitz et al. (2013). Sie verglichen ebenfalls 30 interne Indizes mit mehreren Algorithmen und verwendeten sowohl künstliche als auch reale Datensätze, die verschiedenste Eigenschaften abdecken. Im Gegensatz dazu ist die Forschung zur Clustervalidierung von weichen Clusterings noch nicht so alt. Die aktuell umfangreichste Studie haben Wang & Zhang (2007) verfasst, bei der sie 27 interne Indizes für Fuzzy C-Means vergleichen.

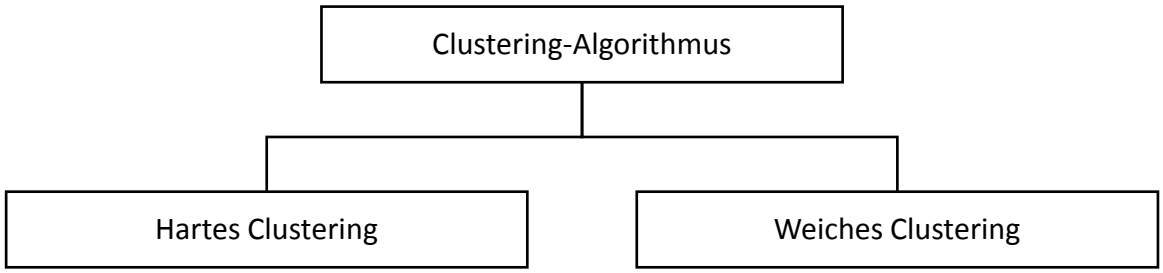

Abbildung 2: Kategorisierung der Clustervalidierung nach dem Clustering-Algorithmus.

2.1.3 Art der Optimierung

Interne Clustervalidierungsindizes können nach der Art der Optimierung unterschieden werden. Das Kriterium zur Kategorisierung wurde von Vendramin et al. (2010) in ihrer Studie über interne Indizes für harte Clustering verwendet. Sie haben in optimierende und vergleichende Indizes unterschieden (vgl. Abbildung 3). Optimierende Indizes nehmen für ein optimales Clustering einen minimalen oder maximalen Wert an (vgl. Abbildung 4A). Vergleichende Indizes hingegen bewerten den Unterschied zwischen den Werten für verschiedene Clusterzahlen (vgl. Abbildung 4B). Diese Werte fallen bei hierarchischen Clustering-Algorithmen automatisch an, daher sind vergleichende Indizes gut zur Validierung dieser Algorithmen geeignet. Das Optimum ist dann durch ein auffälliges „Knie“ oder „Ellbogen“ in der graphischen Darstellung der Werte für verschiedene Clusterzahlen gekennzeichnet. Dieses manuelle und subjektive Verfahren eignet sich nicht zur automatischen Bestimmung der Clusteranzahl, wie es eigentlich üblich ist. Es wurden daher

Transformationen vorgeschlagen, die die relativen Wertunterschiede mit Formeln messen und das „Knie“ so quantifizieren. Dadurch ist es möglich vergleichende Indizes zu optimierenden Indizes zu machen. Bei der Vergleichsstudie von 24 Indizes beider Kategorien schnitten die transformierten vergleichenden Indizes trotzdem durchweg schlechter als die optimierenden Indizes ab (Vendramin et al., 2010). Ein möglicher Grund ist, dass die vergleichenden Indizes deutlich früher entwickelt wurden und deren Konzepte nicht so ausgereift sind wie bei neueren Indizes. Es lässt sich auch in anderen Studien beobachten, dass vergleichende Indizes nur selten bei Vergleichen herangezogen werden und dass kaum neue Indizes dieser Art vorgeschlagen werden. Dies ist auch der Grund, warum keine vergleichenden Indizes sondern nur optimierende Indizes für weiches Clustering existieren.

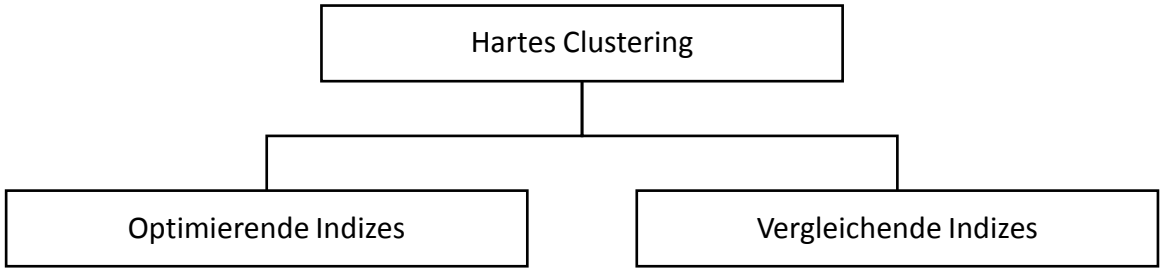

Abbildung 3: Kategorisierung der Clustervalidierungsindizes nach der Art der Optimierung.

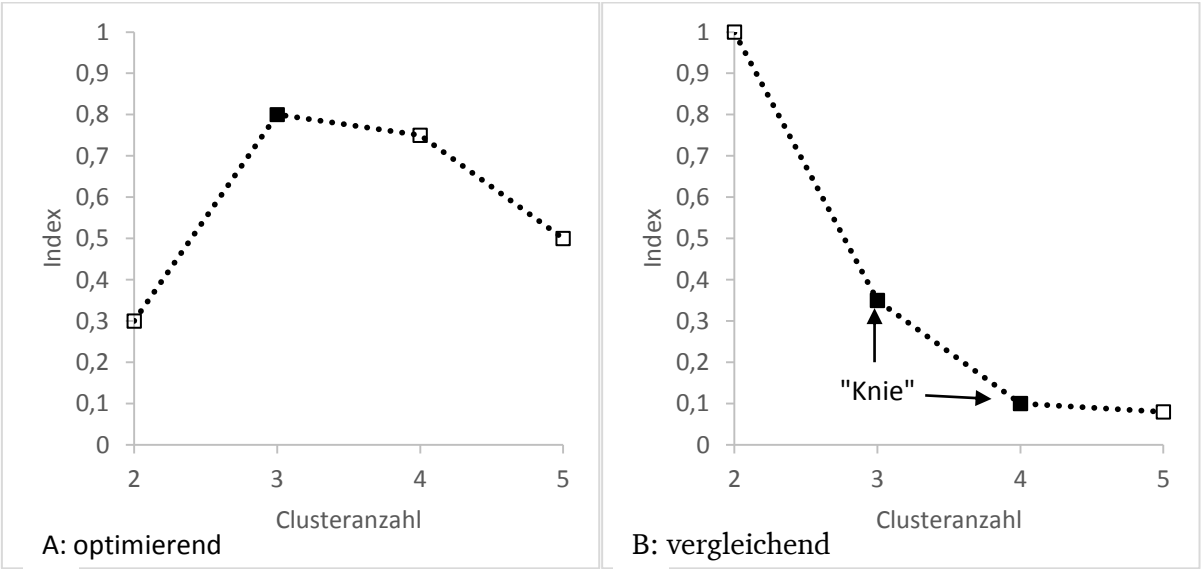

Abbildung 4: Gegenüberstellung der optimierenden und vergleichenden Indizes.

2.1.4 Verwendete Statistiken

Clustervalidierungsindizes lassen sich auch nach den verwendeten Statistiken unterteilen (vgl. Abbildung 5). Verbreitet ist die Unterteilung der internen Clustervalidierung für weiche

Clusterings in Indizes, die nur die Clusterzugehörigkeitsmatrix mit Gewichten berücksichtigen, und in Indizes, die sowohl die Clusterzugehörigkeitsmatrix als auch die Clusterzentren, also die Daten, verwenden (Halkidi et al., 2001; Xu & Brereton, 2005). Auch hier zeigt sich, dass der Zeitraum, in denen die beiden Arten von Indizes vorgeschlagen, sich stark unterscheidet. Die Indizes, die nur die Gewichte verwenden, sind größtenteils älter als die Indizes der Kategorie, die auch die Daten direkt einbezieht. Ein signifikanter Unterschied in der Erfolgsrate bei dem Ermitteln der optimalen Clusteranzahl ist aber nicht vorhanden (Wang & Zhang, 2007). Die Unterteilung kann auch auf interne, optimierende Indizes für hartes Clustering übertragen werden. Die meisten dieser Indizes verwenden ausschließlich das Clustering an sich. Eine andere Gruppe von Indizes berechnet die Korrelation zwischen der Distanzmatrix der Daten und der Clusterzugehörigkeitsmatrix. Vergleichende Indizes können ebenfalls nach den verwendeten Statistiken unterschieden werden (Dimitriadou et al., 2002). Hier gibt es Indizes, die die Varianz der Daten berechnen, und Indizes, die die Kovarianz verwenden.

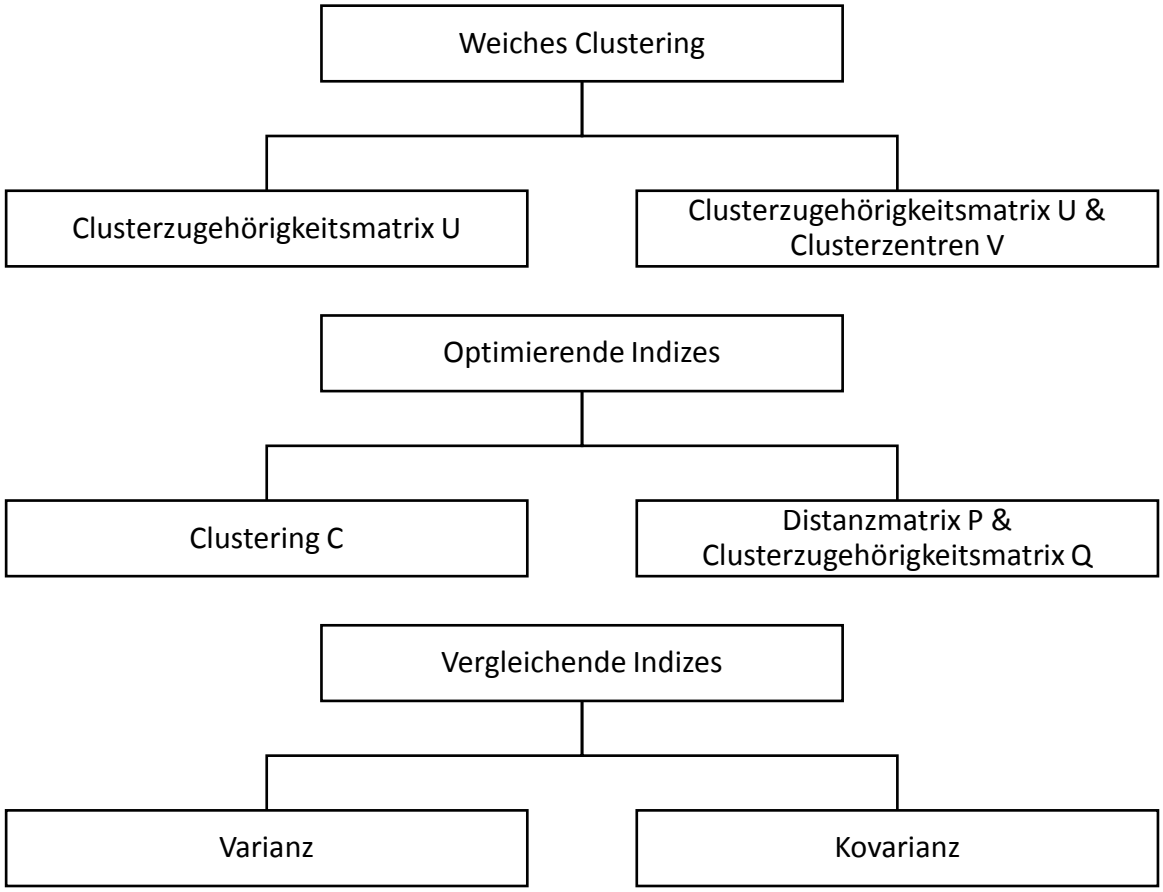

Abbildung 5: Kategorisierung der Clustervalidierungsindizes nach verwendeten Statistiken.

2.1.5 Kombination der Kategorien

Die **Kombination** der hier vorgestellten Unterteilungen der Clustervalidierungsindizes ermöglicht eine Kategorisierung aller Indizes in eine Struktur (vgl. Abbildung 6). In der obersten Ebene wird in interne, relative und externe Clustervalidierung unterschieden. Die **internen Indizes** werden nach dem Ergebnis des Clustering-Algorithmus, harte oder weiche Clusterings, unterteilt. Das nächste Kriterium ist die Art der Optimierung. Die internen Indizes für harte Clusterings werden in optimierende und vergleichende Indizes unterschieden, während es für weiche Clusterings nur optimierende Indizes gibt. Im letzten Schritt werden die internen Indizes nach den verwendeten Statistiken aufgeteilt.

Die **relative Clustervalidierung** wird nach der Vorgehensweise der Indizes unterteilt (Handl et al., 2005). Einige Indizes basieren auf dem wiederholten Ziehen von überlappenden Stichproben, die verglichen werden um die Stabilität des Clustering zu erfassen. Es gibt Indizes, die die Daten in Trainings- und Testdatensatz teilen und die darauf erstellte Prognose mit den Ergebnissen des Clusterings beurteilt. Schließlich existieren Indizes, die die Daten manipulieren und die darauf berechneten Clusterings gegenüberstellen. Übersichten und Vergleiche zu relativen Indizes haben Lange et al. (2004) und Giancarlo et al. (2008) verfasst.

Eine umfangreiche Studie zu dem aktuellen Stand der **externen Clustervalidierung** wurde von Pfitzner et al. (2008) ausgearbeitet. Sie haben die externen Indizes in zwei Gruppen geteilt, die sich in den verwendeten Methoden unterscheiden. Eine Gruppe zählt bestimmte Objektpaare in den wahren Klassen und den Clustern. Ein neuerer Ansatz verwendet Maße der Informationstheorie um die Übereinstimmung zwischen den zwei Partitionierungen zu messen.

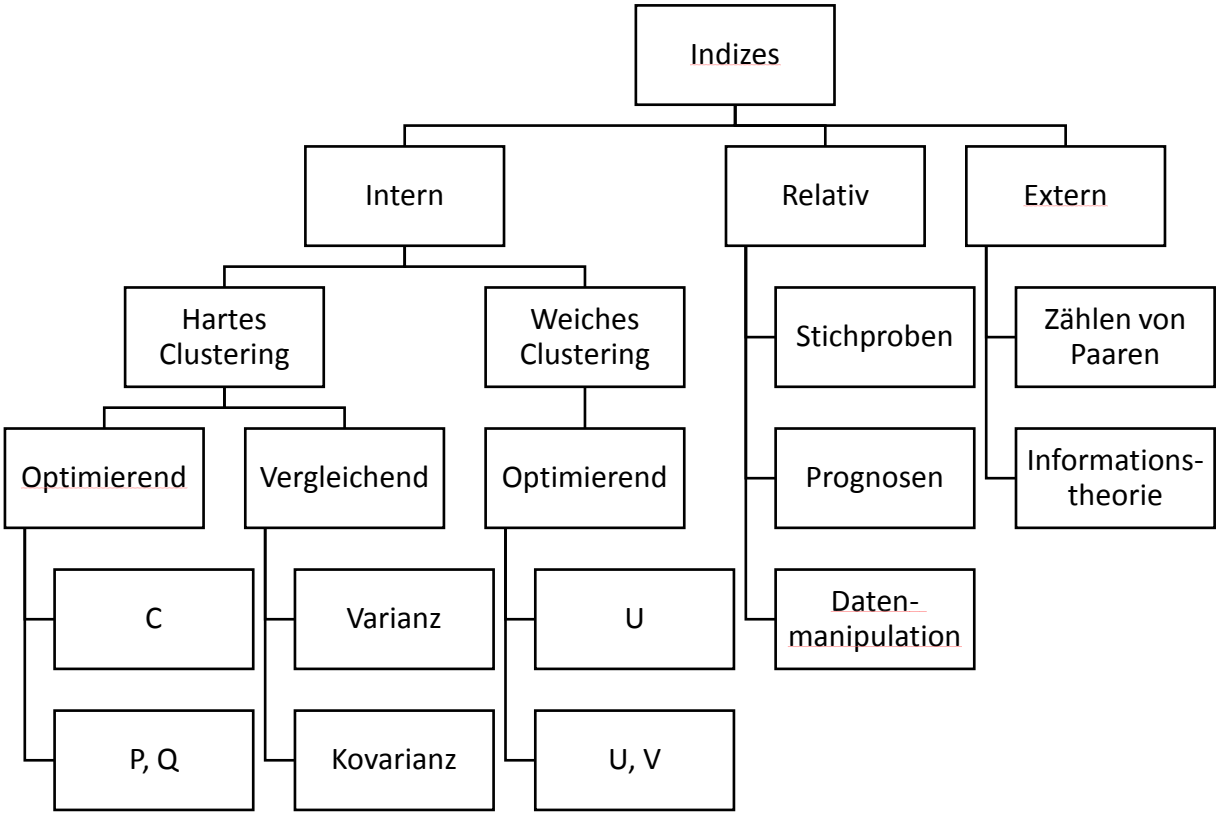

Abbildung 6: Kategorisierung der Clustervalidierungsindizes.

2.2 Konzepte der Clustervalidierungsindizes

Die Clustervalidierungsindizes, vor allem die internen Indizes, beruhen auf bestimmten Konzepten um die Cluster zu evaluieren. Eine Übersicht über diese Konzepte haben Handl et al. (2005) erstellt. Sie unterteilen in vier Gruppen: Indizes basierend auf Kompaktheit, Separierung, Dichte oder einer Kombination von den vorigen Konzepten (vgl. Abbildung 7). Diese Kategorisierung wurde ursprünglich für Clustering-Algorithmen erstellt und auf Clustervalidierungsindizes übertragen.

Zur Berechnung der Konzepte werden die Distanzen zwischen den Objekten der Cluster berechnet. Dazu wird ein Datensatz X mit N Objekten in einem F-dimensionalen Raum benötigt: $X = \{x_1, \dots, x_i, \dots, x_N\}$. Ein hartes Clustering C teilt den Datensatz X in K Gruppen auf: $C = \{c_1, \dots, c_k, \dots, c_K\}$. Das Clusterzentrum $\bar{c}_k$ wird aus dem Mittelwert der Objekte des Clusters c_k berechnet: $\bar{c}_k = \frac{1}{|c_k|} \sum_{x_i \in c_k} x_i$. Bei einem weichen Clustering entsteht eine Clusterzugehörigkeitsmatrix U der Größe $K * N$, bei der das Gewicht u_{ki} die Zugehörigkeit des

Objekts x_i zu Cluster c_k angibt. Das Clusterzentrum v_k wird hier berechnet als: $v_k = \frac{\sum_{i=1}^{N} u_{ki}^m * x_i}{\sum_{i=1}^{N} u_{ki}^m}$.

Der Mittelwert der Daten $\bar{X}$ ist definiert als: $\bar{X} = \frac{1}{N}\sum_{x_i \in X} x_i$. Die Distanz $d(x_i, x_j)$ zwischen den Objekten x_i und x_j wird mit einem Distanzmaß berechnet. Die am häufigsten verwendeten Distanzmaße sind der euklidische Abstand d_e und der quadrierte euklidische Abstand d_e^2. Es wurden allerdings auch andere Distanzmaße vorgeschlagen, die die Distanz mittels Graphen berechnen (Pal & Biswas, 1997; Saha & Bandyopadhyay, 2012) oder an relationale Daten anpassen (Sledge et al., 2010).

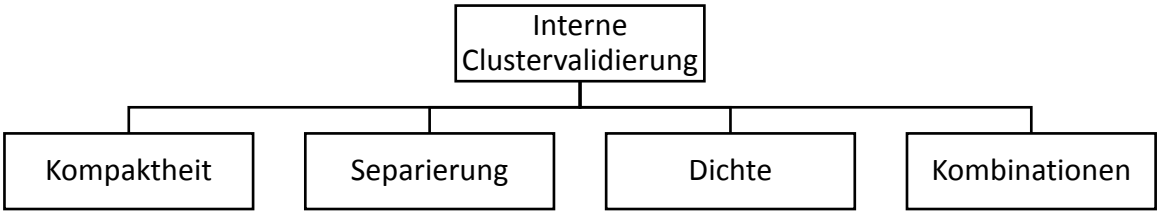

Abbildung 7: Kategorisierung der Konzepte der Clustervalidierungsindizes.

2.2.1 Kompaktheit

Unter der Kompaktheit werden die Intra-Cluster-Distanzen zusammengefasst, die erfassen, wie eng die Objekte der Cluster zusammenliegen. Die Kompaktheit wird für jeden Cluster separat gemessen und lässt sich mit verschiedenen Maßen berechnen, die als Durchmesser der Cluster interpretiert werden können (vgl. Abbildung 8A). Diese Kompaktheitsmaße wurden von Bezdek & Pal (1998) vorgeschlagen und werden von vielen Indizes verwendet. Eine Anpassung der für harte Clusterings entwickelten Maße an weiche Clusterings wurde ebenfalls erstellt (Hassar & Bensaid, 1999).

Das erste Kompaktheitsmaß ist der kompletter Durchmesser („complete diameter“) des Clusters. Es misst die maximale Distanz zwischen zwei Objekten eines Clusters.

$$Comp_1(c_k) = \max_{x_i, x_j \in c_k} \{d(x_i, x_j)\}$$

Das zweite Kompaktheitsmaß berechnet den mitterlen Durchmesser („average diameter“) des Clusters, indem es den Mittelwert der Distanzen zwischen allen Objekten eines Cluster bildet.

$$Comp_2(c_k) = \frac{1}{|c_k|(|c_k|-1)} \sum_{(x_i \neq x_j) \in c_k} d(x_i, x_j)$$

Das dritte und meist verwendete Kompaktheitsmaß ist der Abstand zum Clusterzentrum („centroid diameter“). Dabei wird die mittlere Distanz zwischen den Objekten des Cluster und

dem Clusterzentrum ermittelt. Wird dieses Maß über alle Cluster aufsummiert, entsteht die bekannte Intra-Cluster-Varianz.

$$Comp_3(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d(x_i, \bar{c}_k)$$

$$Comp_3(u_k, v_k) = \frac{1}{N} \sum_{x_i \in X} u_{ki}^m * d(x_i, v_k)$$

Für alle drei Maße gilt, dass die Distanzen minimiert werden müssen, um die optimale Kompaktheit zu erreichen. Die Kompaktheitsmaße sind effektiv bei kugelförmigen und klar getrennten Clustern, bei komplizierteren Strukturen versagen sie allerdings (Handl et al., 2005).

2.2.2 Separierung

Die Separierung erfasst die Trennung zwischen den Clustern und misst dazu die Inter-Cluster-Distanzen. Diese Distanzen werden jeweils zwischen zwei Clustern berechnet und geben deren Abstand wieder (vgl. Abbildung 8B). Die meisten Separierungsmaße wurden erneut von Bezdek & Pal (1998) vorgeschlagen und von Hassar & Bensaid (1999) an weiche Clusterings angepasst.

Das erste Separierungsmaß ist der minimale Abstand („single linkage“) zwischen zwei Clustern. Berechnet wird dieses Maß mit der minimalen Distanz zwischen einem Objekt aus dem einen Cluster und einem zweiten Objekt aus einem anderen Cluster.

$$Sep_1(c_k, c_l) = \min_{x_i \in c_k} \min_{x_j \in c_l} \{d(x_i, x_j)\}$$

Das zweite Separierungsmaß ist der maximaler Abstand („complete linkage“) zwischen zwei Clustern. Dieses Maß ermittelt die maximale Distanz zwischen zwei Objekten aus zwei verschiedenen Clustern.

$$Sep_2(c_k, c_l) = \max_{x_i \in c_k} \max_{x_j \in c_l} \{d(x_i, x_j)\}$$

Der mittlere Abstand („average linkage“) zwischen zwei Clustern ist das dritte Maß und berechnet den Mittelwert der Distanzen zwischen den Objekten des einen Clusters und den Objekten des zweiten Clusters.

$$Sep_3(c_k, c_l) = \frac{1}{|c_k||c_l|} \sum_{x_i \in c_k} \sum_{x_j \in c_l} d(x_i, x_j)$$

Das meist verwendete Separierungsmaß ist der Abstand zwischen den Clusterzentren („centroid linkage“) zweier Cluster, der sich aus der Distanz der beiden Clusterzentren berechnet.

$$Sep_4(c_k, c_l) = d(\bar{c}_k, \bar{c}_l)$$

$$Sep_4(u_k, u_l) = d(v_k, v_l)$$

Ein Separierungsmaß, das nicht zwischen zwei Clustern berechnet wird, ist der Abstand zum Mittelpunkt der Daten („grand mean distance“). Dieses Maß erfasst die Distanz zwischen dem Zentrum eines Clusters und dem Mittelpunkt des kompletten Datensatzes. Die Summe des Maßes über alle Cluster ergibt die Inter-Cluster-Varianz.

$$Sep_5(c_k) = |c_k| d(\bar{c}_k, \bar{X})$$

$$Sep_5(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m * d(v_k, \bar{X})$$

Alle fünf Separierungsmaße müssen maximiert werden, um die optimale Separierung zu erreichen. Es wurden zusätzlich zwei weitere Separierungsmaße vorgeschlagen: eine Kombination von dem mittleren Abstand und dem Abstand zwischen Clusterzentren („average to centroids linkage“) und die Hausdorff-Metrik (Bezdek & Pal, 1998). Beide Maße wurden bisher von keinem bekannten Index verwendet und werden daher nicht weiter betrachtet.

2.2.3 Dichte

Die Dichte ist ein neueres Konzept, das den Zusammenhang der Objekte betrachtet. Es basiert darauf, dass benachbarte Objekte demselben Cluster angehören sollten (vgl. Abbildung 8C). Da dieses Konzept noch relativ neu ist, gibt es nur wenige Indizes, die die Dichte der Cluster messen, und daher auch noch keine Sammlung von häufig verwendeten Dichtemaßen. Es ist allerdings möglich Kompaktheitsmaße bei Verwendung der punktsymmetrischen Distanz in Dichtemaße zu überführen. Für die Dichtemaße spricht, dass sie effektiv bei willkürlichen Cluster-Strukturen sind, die in realen Anwendungen häufig anzutreffen sind (Handl et al., 2005). Sie sind aber nicht robust bei engen Abständen zwischen Clustern.

2.2.4 Kombinationen

Die meisten Indizes verwenden Kombinationen aus den drei vorgestellten Konzepten. Dazu werden zunächst die Distanzen, die auf Clusterebene bestimmt werden, mittels Minimum, Maximum oder Mittelwert aggregiert, um Maße für das komplette Clustering zu erhalten. Die Werte für die jeweiligen Konzepte werden dann durch eine Summe, Differenz, Produkt oder Quotient kombiniert. Dabei ist zu beachten, dass nicht alle Methoden gleich gut zu verwenden sind. Das Mitteln der Distanzen und Summieren der Konzepte verwischt die Werte für die einzelnen Cluster und verhindert, dass die Indizes die Cluster korrekt evaluieren (Kim & Ramakrishna, 2005). Eine Kombination ist notwendig, da das Evaluieren eines einzigen Konzepts oft zu trivialen Lösungen führt. Wird die optimale Clusteranzahl nur anhand der Kompaktheit eines Clusterings ermittelt, wird die Anzahl der Objekte als „perfekte“ Clusterzahl vorgeschlagen. Denn in diesem Fall sind alle Intra-Cluster-Distanzen 0 und die Kompaktheit dadurch optimal. Ähnliches kann festgestellt werden, wenn nur die Separierung eines Clusterings betrachtet. Hier wird nur ein Cluster als Optimum vorgeschlagen, so dass die Inter-Cluster-Distanzen 0 betragen. Eine Kombination der beiden Konzepte löst das Problem und ermöglicht das Finden der optimalen Clusterzahl.

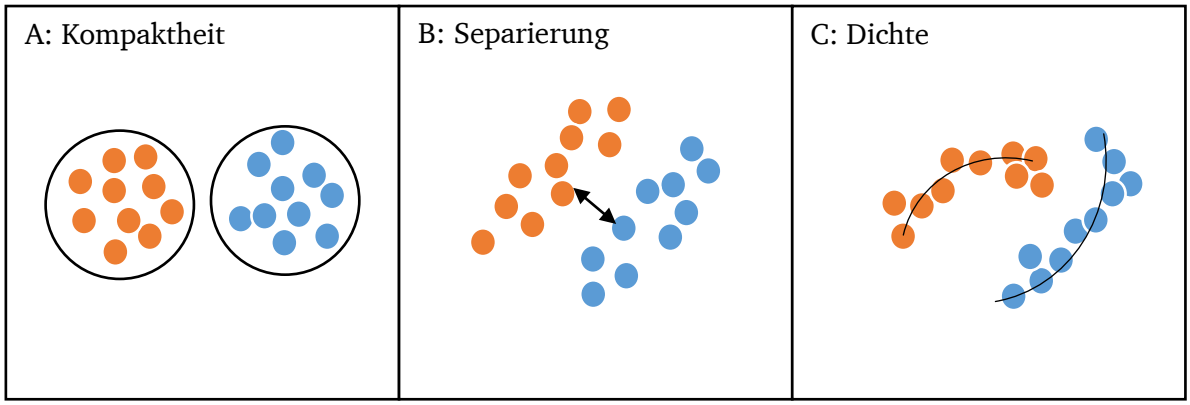

Abbildung 8: Gegenüberstellung der Konzepte der Clustervalidierungsindizes.

Die orangenen und blauen Kreise stehen für die Objekte der jeweiligen Cluster. In schwarz sind die Konzepte eingezeichnet: A - Kompaktheit der Cluster, B - Separierung zweier Cluster und C - Dichte der Cluster. In Anlehnung an Handl et al. (2005).

3 Interne Clustervalidierungsindizes

Zur Kategorie der internen Clustervalidierung gehören alle Indizes, die nur die Daten verwenden, die auch der Clusteranalyse zur Verfügung stehen oder die bei der Clusteranalyse erzeugt werden. Die internen Clustervalidierungsindizes können in zwei Gruppen unterteilt werden, die sich nach dem verwendeten Clustering-Algorithmus unterteilen. Für harte Clusterings, bei denen jedes Objekt genau einem Cluster zugeordnet wird, werden andere Indizes verwendet als für weiche Clusterings, bei denen für jedes Objekte und jeden Cluster ein Gewichte berechnet wird. Im Folgenden werden die meist verwendeten Indizes beider Gruppen vorgestellt.

3.1 Indizes für harte Clusterings

Die Clustervalidierungsindizes für harte Clusterings lassen sich nach der Art der Optimierung unterscheiden. Die meisten Indizes können optimiert werden, indem das Maximum oder Minimum ermittelt wird. Einige Indizes steigen oder fallen allerdings monoton mit der Clusteranzahl. Für diese vergleichenden Indizes müssen die Werte für alle sinnvollen Clusteranzahlen verglichen werden und nach einem signifikanten „Knie“ oder Sprung in der Verteilung gesucht werden. Zunächst werden die optimierenden Indizes besprochen, im Anschluss folgen die vergleichenden Indizes.

3.1.1 Optimierende Indizes

Die optimierenden Indizes werden nach den verwendeten Statistiken getrennt. Die meisten Indizes verwenden das Clustering, um die Evaluierung vorzunehmen (vgl. Tabelle 1). Einige Indizes berechnen allerdings auch die Korrelation zwischen der Distanzmatrix der Daten und der Clusterzugehörigkeitsmatrix (vgl. Tabelle 2). Beide Gruppen werden im Folgenden vorgestellt.

Es existieren auch einige weitere optimierende Indizes, die sich keiner der beiden Gruppen zuordnen lassen. Dazu gehören der G(+) Index (Rohlf, 1974) und der Gamma Index (Baker & Hubert, 1975), die die Cluster beurteilen, indem sie Objektpaare mit bestimmten Eigenschaften zählen. Akaike Information Criterion (Akaike, 1974) und Bayesian Information Criterion (Schwarz, 1978) werden häufig bei modellbasierten Clusterings verwendet, wo sie die Güte eines statistischen Modells erfassen. Der Negentropy Increment Index (Lago-

Fernández & Corbacho, 2010) misst die Abweichung der Verteilung der Cluster von der Normalverteilung und versucht auf diese Weise die Qualität der Cluster zu beurteilen.

3.1.1.1 Indizes basierend auf dem Clustering

Der **Calinski-Harabasz Index** (Calinski & Harabasz, 1974), auch bekannt als Variance Ratio Criterion, ist ein Quotient aus Separierung und Kompaktheit. Der Index beinhaltet zudem einen Normalisierungsfaktor $\frac{N-K}{K-1}$, der verhindert, dass der Quotient monoton mit der Clusteranzahl steigt (Vendramin et al., 2010).

$$CH(C) = \frac{N-K}{K-1}\frac{Sep(C)}{Comp(C)}$$

Das Kompaktheitsmaß, das von dem Calinski-Harabasz Index verwendet wird, ist die Intra-Cluster-Varianz. Diese Varianz wird durch die Quadratsumme der Distanz zwischen den Objekten und deren Clusterzentren berechnet.

$$Comp(C) = \sum_{c_k \in C} Comp(c_k)$$

$$Comp_3(c_k) = \sum_{x_i \in c_k} d_e^2(x_i, \bar{c}_k)$$

Das verwendete Separierungsmaß ist die Inter-Cluster-Varianz, die sich aus der Quadratsumme der Distanz zwischen den Clusterzentren und dem Mittelpunkt der Daten berechnet. Der Abstand der Clusterzentren wird dabei mit der Clustergröße gewichtet, die die Anzahl der Objekte in einem Cluster angibt.

$$Sep(C) = \sum_{c_k \in C} Sep(c_k)$$

$$Sep_5(c_k) = |c_k| d_e^2(\bar{c}_k, \bar{X})$$

Für ein optimales Clustering wird der Calinski-Harabasz Index den maximalen Wert erreichen. Dies liegt daran, dass die zu maximierende Separierung im Zähler steht und sich die zu minimierende Kompaktheit im Nenner befindet.

Dieser Index wurde oft verwendet, u.a. weil er in der bekannten Studie von Milligan & Cooper (1985) am häufigsten die korrekte Clusteranzahl erkannte und damit den Vergleich gegen 30 Indizes gewann. Der Index birgt allerdings auch Nachteile, da er anfällig gegen Rauschen in den Daten ist (Liu et al., 2010). Wenn einem Clustering Rauschen hinzugefügt wird, steigt die

Intra-Cluster-Varianz stärker an als die Inter-Cluster-Varianz. Für die gleiche Anzahl an Clustern sinkt der Calinski-Harabasz Index unter dem Einfluss von Rauschen, was den Index instabil macht. Dies kann dazu führen, dass die optimale Anzahl an Clustern nicht mehr gefunden werden kann. Ein weiterer Nachteil ist, dass die Intra-Cluster-Varianz voraussetzt, dass die Objekte innerhalb der Cluster kugelförmig verteilt sind. Liegt eine schiefe Verteilung vor, versagt das Kompaktheitsmaß und der Calinski-Harabasz Index kann nicht mehr die korrekte Anzahl an Clustern bestimmen. Dies gilt auch für den Clustering-Algorithmus K-Means und andere Indizes, die auf der Intra-Cluster-Varianz basieren.

Der **Dunn Index** (Dunn, 1974) erfasst die Qualität eines Clusterings mit einem Quotienten aus Separierung und Kompaktheit.

$$Dunn(C) = \frac{Sep(C)}{Comp(C)}$$

Die Kompaktheit wird durch den maximalen Durchmesser eines Clusters erfasst, der sich aus der größten Distanz zwischen zwei Objekten des Clusters berechnet. Als Kompaktheitsmaß für das Clustering wird der Cluster mit dem größten Durchmesser gewählt.

$$Comp(C) = \max_{c_k \in C} \{Comp(c_k)\}$$

$$Comp_1(c_k) = \max_{x_i, x_j \in c_k} \{d_e(x_i, x_j)\}$$

Das Separierungsmaß berechnet sich aus der minimalen Separierung zwischen zwei Clustern im Clustering. Die Separierung wird durch die minimale Distanz zwischen einem Objekt des einen Clusters und einem Objekt des anderen Clusters gemessen.

$$Sep(C) = \min_{c_k \in C} \min_{c_l \in C \setminus c_k} \{Sep(c_k, c_l)\}$$

$$Sep_1(c_k, c_l) = \min_{x_i \in c_k} \min_{x_j \in c_i} \{d_e(x_i, x_j)\}$$

Um die optimale Clusteranzahl zu ermitteln, muss der Dunn Index maximiert werden.

Viele Studien haben festgestellt, dass der Dunn Index sehr stark auf Rauschen reagiert (Halkidi et al., 2001; Handl et al., 2005; Liu et al., 2010). Durch das Hinzufügen von Rauschen verändern sich sowohl das Kompaktheits- als auch das Separierungsmaß sehr schnell, da die maximale bzw. minimale Distanz verwendet wird. Das Mitteln der Distanzen könnte die Abhängigkeit reduzieren. Der Dunn Index ist zudem rechenintensiv, da zur Berechnung der Kompaktheit über alle Paare innerhalb der Cluster iteriert werden muss und für die Separierung sogar die Distanz zwischen allen Objektpaaren benötigt wird. Um diese beiden Nachteile zu verringern wurden die generalisierten Dunn Indizes (Bezdek & Pal, 1998)

vorgeschlagen. Dazu gehörten zwei Kompaktheitsmaße (vgl. Kapitel 2.2.1) und fünf Separierungsmaße (vgl. Kapitel 2.2.2) als Alternative zu den verwendeten Distanzen im Dunn Index. Ziel aller alternativen Maße war es die Rechenzeit und die Abhängigkeit von Rauschen zu vermindern.

Der **C Index** (Hubert & Levin, 1976) ist ein normalisierter Kompaktheitsindex. Die Normalisierung erfolgt dadurch, dass von dem Kompaktheitsmaß der minimale Wert abgezogen wird und diese Differenz durch die Länge des Wertebereiches des Maßes geteilt wird. Der Index bewegt sich daher zwischen 0 und 1 für alle Clusterings und Clusteranzahlen.

$$CI(C) = \frac{Comp(C) - Comp_{min}(C)}{Comp_{max}(C) - Comp_{min}(C)}$$

Die Kompaktheit wird bei diesem Index mit der mittleren Distanz zwischen allen Objekten eines Clusters erfasst, die über alle Cluster aufsummiert wird.

$$Comp(C) = \sum_{c_k \in C} Comp(c_k)$$

$$Comp_2(c_k) = \sum_{x_i, x_j \in c_k} d_e(x_i, x_j)$$

Die minimal und maximal mögliche Kompaktheit wird durch die Summe der n_w kleinsten bzw. größten paarweisen Distanzen im Datensatz berechnet. Der Wert n_w stellt dabei die Anzahl der Objektpaare dar, die sich im selben Cluster befinden: $n_w = \sum_{c_k \in C} \binom{|c_k|}{2}$.

$$Comp_{min}(C) = \sum \min_{x_i, x_j \in X}(n_w)\{d_e(x_i, x_j)\}$$

$$Comp_{max}(C) = \sum \max_{x_i, x_j \in X}(n_w)\{d_e(x_i, x_j)\}$$

Die Kompaktheit des Clusterings sollte nahe an der minimalen Kompaktheit liegen, so dass der Index, der nur die Kompaktheit misst, minimiert werden muss.

Der **Ratkowsky-Lance Index** (Ratkowsky & Lance, 1978), auch bekannt als $C/\sqrt{K}$ Index, berücksichtigt nur die Separierung der Cluster. Der Index summiert über alle Variablen die Wurzel des Quotienten aus der Inter-Cluster-Varianz und der Varianz des Datensatzes in der jeweiligen Dimension. Normalisiert wird der Index durch $\frac{1}{N\sqrt{K}}$ um ein monotones Steigen mit der Clusteranzahl zu verhindern.

$$RL(C) = \frac{1}{N\sqrt{K}} \sum_{f=1}^{F} \sqrt{\frac{Sep_f(C)}{Var_f(X)}}$$

Die Separierung wird pro Variable erfasst und berechnet sich aus der Quadratsumme der Distanz zwischen den Clusterzentren und dem Mittelpunkt des Datensatzes. Die Inter-Cluster-Varianz wird durch die komplette Varianz geteilt, die durch die Quadratsumme zwischen den Objekten und dem Datenmittelpunkt ermittelt wird.

$$Sep_f(C) = \sum_{c_k \in c} \left(\bar{c}_{k_f} - \bar{X}_f\right)^2 = SSB$$

$$Var_f(X) = \sum_{x_i \in X} \left(x_{i_f} - \bar{X}_f\right)^2 = SST$$

Gut separierte Cluster erzeugen hohe Werte des Index, der daher maximiert werden sollte.

Der **Davies-Bouldin Index** (Davies & Bouldin, 1979) besteht aus dem Quotienten von Kompaktheit und Separierung mit der Besonderheit, dass der Index pro Clusterpaar berechnet wird. Dazu wird die Kompaktheit beider Cluster addiert und durch die Separierung zwischen den Cluster geteilt. Über alle Cluster wird der jeweilige maximale Wert des Quotienten aufsummiert. Damit wird für jeden Cluster nur der Cluster betrachtet, der in der Nähe liegt und nicht sehr kompakt ist.

$$DB(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{Comp(c_k) + Comp(c_l)}{Sep(c_k, c_l)} \right\}$$

Die Kompaktheit der Cluster wird durch die mittlere Distanz der Objekte in dem Cluster zu dem Clusterzentrum ermittelt.

$$Comp_3(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, \bar{c}_k)$$

Die Separierung zweier Cluster wird durch die Distanz zwischen den beiden Clusterzentren berechnet.

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Der Davies-Bouldin Index muss minimiert werden, um kompakte und separierte Clustering zu erhalten.

Der **Coggins-Jain Index** (Coggins & Jain, 1985) verwendet den Quotienten aus Separierung und Kompaktheit, der für jeden Cluster berechnet wird. Allerdings wird nur der Cluster mit

dem minimalen Wert betrachtet, der anzeigt, dass es sich um den Cluster mit der geringsten Kompaktheit und Separierung handelt.

$$CJ(C) = \min_{c_k \in C} \left\{ \frac{Sep(c_k)}{Comp(c_k)} \right\}$$

Die Kompaktheit des Clusters wird durch den mittleren Abstand der Objekte im Cluster zu dem Clusterzentrum berechnet.

$$Comp_3(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, \bar{c}_k)$$

Die Separierung des Clusters wird durch die Distanz von dem Clusterzentrum zu dem nächstgelegenen Clusterzentrum ermittelt.

$$Sep(c_k) = \min_{c_l \in C \setminus c_k} \{Sep(c_k, c_l)\}$$

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Der Coggins-Jain Index muss maximiert werden, um die optimale Clusterzahl zu finden. Die Autoren merken zudem an, dass der Index für jeden Cluster den Wert 1,7 übersteigen sollte. Dies bedeutet für jeden Cluster, dass der Abstand zum nächsten Clusterzentrum 1,7-fach größer als der mittlere Abstand zum eigenen Clusterzentrum sein sollte.

Der **Silhouette Index** (Rousseeuw, 1987), auch bekannt als Silhouette Width Criterion, berechnet die normalisierte Differenz aus Separierung und Kompaktheit, um Clusterings zu bewerten. Die Besonderheit ist, dass die Differenz für jedes Objekt berechnet wird und dann über alle Objekte gemittelt wird. Die Normalisierung wird erreicht, indem durch den größeren Wert des Kompaktheits- oder Separierungsmaß geteilt wird.

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{Sep(x_i, c_k) - Comp(x_i, c_k)}{\max\{Comp(x_i, c_k), Sep(x_i, c_k)\}}$$

Die Kompaktheit für ein Objekt berechnet sich aus der mittleren Distanz zu den Objekten desselben Clusters.

$$Comp(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, x_j)$$

Die Separierung für ein Objekt wird ermittelt, indem die mittlere Distanz zu den Objekten eines anderen Clusters berechnet wird. Diese Rechnung wird für alle vorhandenen Cluster wiederholt und der Cluster mit der geringsten mittleren Distanz wird als Separierungsmaß verwendet.

$$Sep(x_i, c_k) = \min_{c_l \in C \setminus c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} d_e(x_i, x_j) \right\}$$

Der Wertebereich des Silhouette Index liegt zwischen -1 und 1. Der minimale Wert wird angenommen, wenn der Abstand zum nächsten Cluster sehr gering ist und die Distanzen innerhalb des Cluster viel größer sind. Der maximale Wert steht für kompakte und separierte Cluster und ist demnach anzustreben.

Eine Variation des Index ist der **alternative Silhouette Index** (Hruschka et al., 2006), bei dem statt der normalisierten Differenz der Quotient aus Separierung und Kompaktheit gebildet wird. Der Unterschied besteht darin, dass die Konzepte nicht mehr linear sondern nichtlinear kombiniert werden. Dem Nenner wird eine kleine Konstante ε hinzugefügt, für den Fall, dass die Kompaktheit 0 beträgt.

$$Sil_{alt}(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in C} \frac{Sep(x_i, c_k)}{Comp(x_i, c_k) + \varepsilon}$$

Eine weitere Variation ist der **Simplified Silhouette Index** (Vendramin et al., 2010), der die Berechnung der Kompaktheits- und Separierungsmaße vereinfacht. Dazu wird die mittlere Distanz zu den Objekten eines Clusters durch die Distanz zu dem Clusterzentrum ersetzt. Mit den neuen Maßen muss nur eine Distanz pro Cluster berechnet werden, anstatt die Distanzen zu allen Objekten des Clusters zu ermitteln. Ziel ist die zeitintensive Berechnung des Silhouette Index zu verkürzen.

$$Comp(x_i, c_k) = d_e(x_i, \bar{c}_k)$$

$$Sep(x_i, c_k) = \min_{c_l \in C \setminus c_k} \{d_e(x_i, \bar{c}_l)\}$$

Der **Boudraa Index** (Boudraa, 1999) bildet die gewichtete Summe aus Kompaktheit und Separierung. Ein Gewichtungsfaktor α dient dazu die Summe zu normalisieren. Der Faktor berechnet sich aus dem Quotienten zwischen der Separierung für ein Clustering mit N Clustern und der Kompaktheit für ein Clustering mit 2 Clustern: $\alpha = \frac{Sep(C^{(N)})}{Comp(C^{(2)})}$. Die Clusteranzahlen wurden so vorgegeben, dass die Maße jeweils die schlechtesten Werte erreichen. Die Separierung bei Clusterings mit vielen Clustern wird gering und damit weit von dem Optimum entfernt sein. Ein Clustering mit nur zwei Clustern ist nicht sehr kompakt, so dass das Kompaktheitsmaß relativ groß sein wird.

$$B_{crit}(C) = \alpha * Comp(C) + Sep(C)$$

Die Kompaktheit wird aus dem Quotienten aus der Intra-Cluster-Varianz und der Varianz des Datensatzes berechnet. Die Varianzen werden für jede Variable einzeln ermittelt und aufsummiert.

$$Comp(C) = \frac{1}{K}\frac{\sum_{f=1}^{F}\sum_{c_k \in C} Var_f(c_k)}{\sum_{f=1}^{F} Var_f\,(X)}$$

Das Separierungsmaß ergibt sich aus dem Verhältnis der maximalen Separierung zu der minimalen Separierung zwischen zwei Clustern des Datensatzes. Die Separierung berechnet sich aus der Distanz zwischen den Clusterzentren der jeweiligen Cluster.

$$Sep(C) = \frac{\max_{c_k, c_l \in C}\{Sep(c_k, c_l)\}}{\min_{(c_k \neq c_l) \in C}\{Sep(c_k, c_l)\}}$$

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Das Minimum des Boudraa Index zeigt, dass ein kompaktes und separiertes Clustering gefunden wurde.

Der **SD Index** (Halkidi et al., 2000) verwendet die gewichtete Summe aus Kompaktheit und Separierung, um die Güte eines Clustering zu erfassen. Der Index basiert auf dem CWB Index (Rezaee et al., 1998), der für weiches Clusterings vorgeschlagen wurde. In diesem Index wird auch ein Gewichtungsfaktor α verwendet, der die Separierung für das Clustering mit der größtmöglichen Clusteranzahl berechnet: $\alpha = Sep\left(C^{(K_{max})}\right)$. Der Faktor wird benötigt, da die Maße der beiden Konzepte nicht zwingend über den gleichen Wertebereiche verfügen.

$$SD(C) = \alpha * Comp(C) + Sep(C)$$

Die Kompaktheit des Clusterings wird aus dem Quotienten der Intra-Cluster-Varianz und der Varianz des Datensatzes berechnet, der über alle Cluster gemittelt wird.

$$Comp(C) = \frac{1}{K}\sum_{c_k \in C}\frac{\|Comp(c_k)\|}{\|Var(X)\|}$$

$$Comp_3(c_k) = \frac{1}{|c_k|}\sum_{x_i \in c_k} d_e^2(x_i, \bar{c}_k)$$

$$Var(X) = \frac{1}{N}\sum_{x_i \in X} d_e^2(x_i, \bar{X})$$

Das Separierungsmaß für das Clustering wird aus dem Verhältnis der maximalen zu der minimalen Separierung multipliziert mit der invertierten Summe aller Separierungen

berechnet. Die Separierung wird jeweils durch die Distanz zwischen den Clusterzentren beurteilt.

$$Sep(C) = \frac{\max_{c_k,c_l \in C}\{Sep(c_k,c_l)\}}{\min_{(c_k \neq c_l) \in C}\{Sep(c_k,c_l)\}} \sum_{c_k \in C} \frac{1}{\sum_{c_l \in C} Sep(c_k,c_l)}$$

$$Sep_4(c_k,c_l) = d_e(\bar{c}_k,\bar{c}_l)$$

Durch die verwendeten Gewichtungen und Normalisierungen sind beide Maße so angepasst, dass sie mit geringen Werte gute Clusterings anzeigen. Der SD Index muss demnach minimiert werden, um die optimale Clusteranzahl zu finden.

Der **S_{Dbw} Index** (Halkidi & Vazirgiannis, 2001) ist eine Variante des SD Index, bei der die Separierung durch ein Dichtemaß ersetzt wird. Es wird analog zu dem SD Index die Summe aus der Kompaktheit und der Dichte eines Clusterings gebildet. Die Kompaktheit wird identisch wie beim SD Index aus dem Quotienten aus der Intra-Cluster-Varianz und der Varianz des Datensatzes berechnet.

$$S_{Dbw}(C) = Comp(C) + Dens(C)$$

Die Dichte eines Clusterings wird für jedes Clusterpaar separat erfasst und gemittelt. Zur Evaluierung von zwei Clustern c_k und c_l wird die Dichte am Mittelpunkt u_{kl} zwischen beiden Clusterzentren ermittelt und durch den Wert des dichteren Clusterzentrums geteilt. Die Dichte eines Punktes wird mit der Dichtefunktion f erfasst, die die Anzahl der Objekte zählt, die mit einem Abstand von weniger als der Standardabweichung von dem Punkt entfernt sind.

$$Dens(C) = \frac{1}{K(K-1)} \sum_{c_k \in C} \sum_{c_l \in C \setminus c_k} Dens(c_k,c_l)$$

$$Dens(c_k,c_l) = \frac{\sum_{x_i \in c_k,c_l} f(x_i,u_{kl})}{\max\left\{\sum_{x_i \in c_k} f(x_i,\bar{c}_k), \sum_{x_j \in c_l} f(x_j,\bar{c}_l)\right\}}$$

$$f(x_i,u_{kl}) = \begin{cases} 0, wenn\ d(x_i,u_{kl}) > stdev \\ 1, sonst \end{cases}$$

Das Dichtemaß muss minimiert werden, da die Dichte in der Mitte zwischen zwei Clustern geringer sein sollte als die Dichte in den Clusterzentren. Hohe Werte sind ein Indikator dafür, dass sich die meisten Werte zwischen den Clustern befinden, während niedrige Werte für dichte Cluster stehen. Das Kompaktheitsmaß muss auch minimiert werden, so dass ein Minimum des S_{Dbw} Index für ein optimales Clustering steht.

Der **PS Index** (Chou et al., 2002) berechnet das Verhältnis von Dichte zu Separierung. Die Dichte wird durch die punktsymmetrische Distanz für jedes Objekt erfasst und durch die Separierung des Clusterings geteilt. Dieser Quotient wird für alle Objekte berechnet und gemittelt.

$$PS(C) = \frac{1}{K} \sum_{c_k \in C} \left[\frac{1}{|c_k|} \sum_{x_i \in c_k} \frac{Dens(x_i, c_k)}{Sep(C)} \right]$$

Das Dichtemaß besteht aus der punktsymmetrischen Distanz zwischen einem Objekt und dem zugehörigen Clusterzentrum und wird mit der euklidischen Distanz multipliziert. Bei eng zusammenliegenden Punkten dominiert die Punktsymmetrie, während die euklidische Distanz verhindert, dass der Index nur die Symmetrie betrachtet. Die punktsymmetrische Distanz d_{sym} berechnet, ob für ein Objekt ein punktsymmetrisches Pendant existiert. Als Symmetriepunkt wird das jeweilige Clusterzentrum ausgewählt. Der Wert wird durch die einzelnen Distanzen normalisiert.

$$Dens(x_i, c_k) = d_{sym}(x_i, \bar{c}_k) * d_e(x_i, \bar{c}_k)$$

$$d_{sym}(x_i, \bar{c}_k) = \min_{i \neq j} \left\{ \frac{d_e\left((x_i, \bar{c}_k), (x_j, \bar{c}_k)\right)}{d_e(x_i, \bar{c}_k) + d_e(x_j, \bar{c}_k)} \right\}$$

Die Separierung des Clusterings wird durch die minimale Distanz zwischen zwei Clusterzentren im Datensatz beurteilt.

$$Sep(C) = \min_{k \neq l} \{ d_e(\bar{c}_k, \bar{c}_l) \}$$

Das Minimum des PS Index gibt an, wann ein optimales Clustering gefunden wurde.

Der **CS Index** (Chou et al., 2004) berechnet den Quotienten aus Kompaktheit und Separierung. Die Besonderheit liegt in den verwendeten Maßen, die mit Clustern verschiedener Dichten und Größen umgehen können.

$$CS(C) = \frac{Comp(C)}{Sep(C)}$$

Zur Ermittlung der Kompaktheit wird die maximale Intra-Cluster-Distanz für jedes Objekt berechnet. Diese Distanzen werden über alle Cluster gemittelt, um das Kompaktheitsmaß zu erhalten.

$$Comp(C) = \sum_{c_k \in C} Comp(c_k)$$

$$Comp(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \max_{x_j \in c_k} \{d_e(x_i, x_j)\}$$

Die Separierung wird durch die Distanzen zwischen den Clusterzentren beurteilt. Das Separierungsmaß ergibt sich aus der Summe der Separierungen zu dem jeweils nächstgelegenen Cluster.

$$Sep(C) = \sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \{Sep(c_k, c_l)\}$$

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Bei dem CS Index wird die Kompaktheit durch die Separierung geteilt, daher ist das Minimum ein Indikator für ein kompaktes und separiertes Clustering.

K-Nearest-Neighbor-Consistency (Ding & He, 2004) ist ein Index, der ausschließlich die Dichte eines Clusterings betrachtet. Dazu wird für jeden Cluster der Anteil der kNN-konsistenten Objekte berechnet. Ein Objekt x innerhalb eines Clusters c ist kNN-konsistent in Bezug auf Cluster c, wenn die k nächsten Nachbarn von x auch zu Cluster c gehören. Die Anzahl der kNN-konsistenten Punkte ist durch n_{cons} angegeben.

$$Cons(C) = Dens(C) = \frac{1}{K} \sum_{c_k \in C} \frac{n_{cons}}{|c_k|}$$

Ein hoher Anteil von kNN-konsistenten Objekten deutet auf dichte Cluster hin, daher gilt es den Index zu maximieren.

Der **PBM Index** (Pakhira et al., 2004), auch bekannt als I Index (Maulik & Bandyopadhyay, 2002), ermittelt die Güte der Cluster mit einem quadratischen Produkt aus drei Termen. Der erste Term ist ein Gewichtungsfaktor, um ein monotones Steigen des Index zu verhindern. Der zweite Term erfasst die Kompaktheit und bildet das Verhältnis aus der Distanz der Objekte zu dem Mittelpunkt der Daten und der Distanz der Objekte zu dem jeweiligen Clusterzentrum. Der dritte Term besteht aus der Separierung des Clusterings.

$$PBM(C) = \left(\frac{1}{K} \frac{Var(X)}{Comp(C)} Sep(C) \right)^2$$

Die Kompaktheit besteht aus dem Verhältnis von der Summe der Intra-Cluster-Distanzen und der Distanz zwischen den Objekten und dem Mittelpunkt der Daten.

$$Comp(C) = \sum_{c_k \in C} Comp(c_k)$$

$$Comp_3(c_k) = \sum_{x_i \in c_k} d_e(x_i, \bar{c}_k)$$

$$Var(X) = \sum_{x_i \in X} d_e(x_i, \bar{X})$$

Als Separierungsmaß wird die maximale Distanz zwischen zwei Clusterzentren im Datensatz verwendet.

$$Sep(C) = \max_{c_k \in C} \max_{c_l \in C} \{Sep(c_k, c_l)\}$$

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Ein maximaler Wert des PBM Index gibt an, dass ein kompaktes und separiertes Clustering vorliegt. Der Index konnte in Vergleichen erstaunlich oft die optimale Clusterzahl vorhersagen. In zwei aussagekräftigen Vergleichsstudien schnitt der PBM Index sogar besser als alle anderen getesteten Indizes ab (Wang & Zhang, 2007; Vendramin et al., 2010).

Connectivity (Handl & Knowles, 2005) ist ein Index, der die Dichte der Clusterings beurteilt. Dazu wird für jedes Objekt der Anteil der benachbarten Objekte betrachtet, die sich im selben Cluster befinden. Es werden die L nächsten Nachbarn eines Objekts ausgewählt und gemäß ihrer Distanz zu dem Objekt eine Reihenfolge gebildet. Befinden sich Nachbarn in einem anderen Cluster, so wird das Inverse ihres Ranges aufsummiert. Nachbarn aus demselben Cluster werden mit 0 bewertet.

$$Conn(C) = Dens(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

$$x_{i,nn_{i(j)}} = \begin{cases} 1/j\,, wenn\ sich\ j\ in\ einem\ anderen\ Cluster\ befindet \\ 0, sonst \end{cases}$$

Insgesamt soll Connectivity minimiert werden, da nur fehlerhafte Zuordnungen gemäß ihrer Schwere aufsummiert werden.

Score Function (Saitta et al., 2007) beurteilt Clusterings mittels der Differenz aus der Separierung und der Kompaktheit des Clusterings. Diese Differenz wird normalisiert, so dass die Werte des Index zwischen 0 und 1 liegen. Dies ermöglicht eine Aussage darüber zu treffen, wie nah ein Clustering am Optimum ist.

$$SF(C) = 1 - \frac{1}{e^{e^{Sep(C)-Comp(C)}}}$$

Die Kompaktheit des Clusterings wird durch die mittlere Distanz der Objekte zu den jeweiligen Clusterzentren erfasst.

$$Comp(C) = \sum_{c_k \in C} Comp(c_k)$$

$$Comp_3(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, \bar{c}_k)$$

Die Separierung berechnet sich aus der mittleren Distanz von den Clusterzentren zu dem Mittelpunkt der Daten.

$$Sep(C) = \frac{1}{N * K} \sum_{c_k \in C} Sep(c_k)$$

$$Sep_4(c_k) = |c_k| d_e(\bar{c}_k, \bar{X})$$

Ein kompaktes und separiertes Clustering wird einen Wert nahe 1 einnehmen, daher muss Score Function maximiert werden, um die optimale Clusteranzahl zu ermitteln.

Der **C_{Dbw} Index** (Halkidi & Vazirgiannis, 2008) berechnet das Produkt von Kompaktheit, Separierung und Kohäsion mit Hilfe von Dichtemaßen, um die Qualität von Clusterings zu beurteilen.

$$C_{Dbw}(C) = Cohesion(C) * Sep_{Dens}(C) * Comp_{Dens}(C)$$

Die Kompaktheit des Clusterings wird mit der mittleren Intra-Cluster-Dichte erfasst. Diese ermittelt den Anteil der Objekte pro Cluster, die zur Nachbarschaft des Clusterrepräsentanten gehören. Die Nachbarschaft ist eine Kugel rund um den Clusterrepräsentanten mit der Standardabweichung der Daten als Radius. Es werden also alle Objekte eines Clusters erfasst, deren Distanz zu dem Clusterrepräsentanten geringer als die Standardabweichung ist. Die Separierung des Clusterings wird durch den Quotienten von zwei Maßen erfasst. Im Zähler befindet sich die Summe über die Distanzen aller Cluster zu dem jeweils nächstgelegenen Clusterzentrum. Die Summe wird geteilt durch die Inter-Cluster-Dichte, die über alle Cluster die maximale Dichte im Zwischenraum aufsummiert. Der dritte Faktor ist die Kohäsion, die Änderungen in der Intra-Cluster-Dichte erfasst. Diese sollten möglichst gering sein. Berechnet wird die Kohäsion aus dem Verhältnis der Kompaktheit des Clusterings und der Größe der Änderungen.

Alle drei Faktoren nehmen große Werte an, wenn ein optimales Clustering vorliegt. Der C_{Dbw} Index muss demnach maximiert werden, um die korrekte Clusteranzahl zu finden. Gegen die Verwendung des Index spricht, dass die Berechnung sehr rechenaufwändig ist. Zudem ist es schwierig adäquate Repräsentanten für jeden Cluster zu finden, daher wird dieser Index instabil (Liu et al., 2010).

Der **NIVA Index** (Rendón et al., 2008) verwendet einen Quotienten aus Kompaktheit und Separierung, um das Clustering zu validieren. Die Besonderheit ist, dass das Kompaktheitsmaß auf einem erneuten Clustering der Cluster basiert.

$$NIVA(C) = \frac{CompC(C)}{Sep(C)}$$

Die Kompaktheit des Clusterings wird ermittelt, indem für jeden Cluster das Produkt aus Kompaktheit und Separierung der Sub-Cluster bewertet wird. Ein Sub-Clustering SC_k stellt das Clustering dar, das Cluster c_k in L_k Sub-Cluster aufteilt: $SC = \{sc_1, \dots, sc_L\}$.

$$CompC(C) = \frac{1}{K} \sum_{c_k \in C} CompC(c_k) * SepC(c_k)$$

Die Kompaktheit der Sub-Cluster errechnet sich aus der mittleren Distanz der Objekte x_i zu ihrem nächsten Nachbarn $x_{i,nn}$.

$$CompC(c_k) = \frac{1}{L_k} \sum_{l=1}^{L_k} CompSC(sc_l)$$

$$CompSC(sc_l) = \frac{1}{|sc_l|} \sum_{i=1}^{|sc_l|} d_e(x_i, x_{i,nn})$$

Die Separierung der Sub-Cluster wird durch die mittlere Distanz zu dem jeweils entferntesten Sub-Clusterzentrum erfasst.

$$SepC(c_k) = \frac{1}{L_k} \sum_{l=1}^{L_k} \max_{p \in SC_k} \{SepSC(sc_l, sc_p)\}$$

$$SepSC(sc_l, sc_p) = d_e(\bar{sc}_l, \bar{sc}_p)$$

Die Separierung des Clusterings verwendet keine Sub-Cluster, sondern die minimale Distanz zwischen Clusterzentren. Für jeden Cluster wird so die minimale Separierung zu dem nächstgelegenen Cluster berechnet und über alle Cluster gemittelt.

$$Sep(C) = \frac{1}{K} \sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \{Sep(c_k, c_l)\}$$

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Der NIVA Index nimmt minimale Werte an, wenn ein kompaktes und separiertes Clustering gefunden wurde.

Der **COP Index** (Gurrutxaga et al., 2010) verwendet einen Quotienten aus Kompaktheit und Separierung mit der Besonderheit, dass das Verhältnis pro Cluster berechnet und über alle Cluster gemittelt wird.

$$COP(C) = \frac{1}{N} \sum_{c_k \in C} |c_k| \frac{Comp(c_k)}{Sep(c_k)}$$

Die Kompaktheit der Cluster wird mit der mittleren Distanz zwischen den Objekten eines Clusters und dem Clusterzentrum beurteilt.

$$Comp_3(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, \bar{c}_k)$$

Zur Ermittlung der Separierung der Cluster wird die maximale Distanz zwischen zwei Objekten von zwei verschiedenen Clustern betrachtet. Die Separierung des Clusterpaars, für das die Distanz minimal ist, wird als Separierungsmaß verwendet.

$$Sep(c_k) = \min_{c_l \in C \setminus c_k} Sep(c_k, c_l)$$

$$Sep_2(c_k, c_l) = \max_{x_i \in c_k} \max_{x_j \in c_l} \{d_e(x_i, x_j)\}$$

Typischerweise für einen Index, bei dem ein Kompaktheitsmaß im Zähler und ein Separierungsmaß im Nenner stehen, muss der COP Index minimiert werden.

Der **SV Index** (Žalik & Žalik, 2011) berechnet den Quotienten aus Separierung und Kompaktheit zur Ermittlung der Güte des Clusterings.

$$SV(C) = \frac{Sep(C)}{Comp(C)}$$

Das Kompaktheitsmaß wird durch die Summe der Distanzen der Objekte ermittelt, die sich am weitesten entfernt von ihrem Clusterzentrum befinden.

$$Comp(C) = \sum_{c_k \in C} Comp(c_k)$$

$$Comp(c_k) = \max_{x_i \in c_k} \{d_e(x_i, \bar{c}_k)\}$$

Die Separierung wird über alle Cluster summiert und berechnet sich jeweils aus der Distanz zu dem nächstgelegenen Clusterzentrum.

$$Sep(C) = \sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \{Sep(c_k, c_l)\}$$

$$Sep_4(c_k, c_l) = d_e(\bar{c}_k, \bar{c}_l)$$

Folglich muss der SV Index maximiert werden, um gute Clusterings zu erhalten.

Der **OS Index** (Žalik & Žalik, 2011) stellt eine Variante des SV Index dar, bei der das Separierungsmaß durch eine komplexere Version ersetzt wird, die die Überlappung der Cluster misst. Das Kompaktheitsmaß wird nicht verändert. Die Überlappung wird pro Objekt eines Clusters berechnet. Dazu wird die Distanz a verwendet, die die mittlere Distanz zu den Objekten im selben Cluster angibt: $a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, x_j)$. Diese Distanz wird verglichen mit b, die die mittlere Distanz zu den nächsten $|c_k|$ Objekten in anderen Clustern erfasst: $b(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_i \notin c_k} \min(|c_k|)\{d_e(x_i, x_j)\}$. Wenn für ein Objekt die Distanz a nicht bedeutend kleiner als b ist, dann wird der Strafterm a/b addiert. Ist a klein genug, wird 0 addiert. Die Strafterme werden über alle Cluster und deren Objekte aufsummiert und ergeben das Separierungsmaß.

$$Sep(C) = \sum_{c_k \in C} \sum_{x_i \in c_k} ov(x_i, c_k)$$

$$ov(x_i, c_k) = \begin{cases} a/b\,, wenn\ (b-a)/(b+a) < 0.4 \\ 0, sonst \end{cases}$$

Der **Gaussian Fuzzy Index** (Ghosh & De, 2013) berechnet den Quotienten aus Separierung und Kompaktheit. Die Distanzen werden im Gegensatz zu den meisten anderen Indizes nicht mit einer euklidischen Distanzfunktion berechnet, sondern mit einer Zugehörigkeitsfunktion μ nach Gauß beurteilt.

$$GF(C) = \frac{Sep(C)}{1 + Comp(C)}$$

Die Kompaktheit des Clusterings wird mit der über alle Cluster und Objekte gemittelten Zugehörigkeit $\mu_k(x_i)$ erfasst. Der Term $\mu_k(x_i)$ gibt die Zugehörigkeit von Objekt x_i zu Cluster c_k an.

$$Comp(C) = \frac{1}{K} \sum_{c_k \in C} Comp(c_k)$$

$$Comp(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \mu_k(x_i)$$

Das Separierungsmaß verwendet ebenfalls das Mittel der Zugehörigkeit $\mu_k(c_l)$, das die Zugehörigkeit von Cluster c_l zu Cluster c_k angibt.

$$Sep(C) = \frac{2}{K(K-1)} \sum_{c_k \in C} Sep(c_k)$$

$$Sep(c_k) = \sum_{c_l \in C \setminus c_k} \mu_k(c_l)$$

Das Minimum des GF Index zeigt an, dass ein optimales Clustering mit kompakten und separierten Clustern gefunden wurde.

Der **HS Index** (Satapathy et al., 2014) berechnet den Quotienten aus Kompaktheit und Separierung.

$$HS(C) = \frac{Comp(C)}{Sep(C)}$$

Die Kompaktheit des Clusterings wird mit Hilfe eines minimalen Spannbaums (MST) berechnet. Dazu wird ein Graph für jeden Cluster erstellt, bei dem die Objekte des Clusters die Knoten darstellen und die Kanten die euklidischen Distanzen zwischen den Objekten repräsentieren. Der minimale Spannbaum umfasst alle Objekte und minimalisiert den Wert der Kanten. Das Gewicht des Baums wird als Kompaktheitsmaß verwendet und über alle Cluster gemittelt.

$$Comp(C) = \frac{1}{K} \sum_{c_k \in C} Comp(c_k)$$

$$Comp(c_k) = MST_k$$

Die mittlere Separierung der Cluster erfasst die kleinste Separierung von einem anderen Cluster, berechnet durch die minimale Distanz zwischen Objekten der beiden Cluster.

$$Sep(C) = \frac{1}{K} \sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \{Sep(c_k, c_l)\}$$

$$Sep_1(c_k, c_l) = \min_{x_i \in c_k} \min_{x_j \in c_l} \{d_e(x_i, x_j)\}$$

Der HS Index nimmt minimale Werte an, wenn das Clustering kompakt und separiert ist.

3.1.1.2 Indizes basierend auf dem Clustering und der Distanzmatrix

In diesem Abschnitt wird auf die Indizes eingegangen, die das Clustering, repräsentiert durch die binäre Clusterzugehörigkeitsmatrix Q, mit der Distanzmatrix P vergleichen. Alle Indizes dieser Kategorie messen die Korrelation zwischen den beiden Matrizen und erfassen damit, wie gut die Cluster die Datenstruktur abbilden.

Der **Tau Index** (Rohlf, 1974) berechnet die Tau Korrelation zwischen den Matrizen, auch bekannt als Kendalls Tau,. Der Index misst die Differenz zwischen der Anzahl der übereinstimmenden und uneinigen Objektpaare. Der Wertebereich befindet sich zwischen -1 und 1, da sich im Nenner ein Normalisierungsterm befindet.

$$\tau(P,Q) = \frac{S_+ - S_-}{\sqrt{(t(t-1)/2 - t_{ie})(t(t-1)/2)}}$$

Die Anzahl an übereinstimmenden Paaren S_+ gibt die Anzahl an Distanzen zwischen Objektpaaren desselben Clusters wider, die kleiner sind als die Distanz zwischen Objektpaaren aus verschiedenen Clustern. Die Anzahl an uneinigen Paaren S_- steht für die Anzahl an Distanzen zwischen Objektpaaren desselben Clusters, die größer sind als die Distanz zwischen Objektpaaren aus verschiedenen Clustern. t ist die Anzahl der Objektpaare.

Ist die Anzahl der uneinigen Paare größer als die der übereinstimmenden Paare, so nimmt der Tau Index negative Werte an. Positive Werte ergeben sich, wenn die Anzahl der übereinstimmenden Paare überwiegt. Das Maximum erreicht der Index bei der optimalen Clusterzahl.

Der **punktbiseriale Korrelationskoeffizient** (Milligan, 1981) berechnet die Korrelation zwischen den Matrizen, indem die mittleren Distanzen verglichen werden. Dazu wird die Differenz aus der mittleren Inter-Cluster-Distanz d_b und der mittleren Intra-Cluster-Distanz d_w gebildet. Die Differenz wird durch die Standardabweichung aller Distanzen s_d geteilt und der Quotient mit einem Normalisierungsterm, bestehend aus der Anzahl der Intra-Cluster-Distanzen w_d und der Inter-Cluster-Distanzen b_d, multipliziert.

$$PB(P,Q) = \frac{(d_b - d_w)}{s_d}\sqrt{w_d * b_d / t^2}$$

Positive Werte des Index geben an, dass die Distanzen zwischen den Clustern größer sind als innerhalb der Cluster. Zur Ermittlung der optimalen Clusteranzahl muss der Index demnach maximiert werden.

Huberts Γ Statistik (Hubert & Arabie, 1985) berechnet die punktseriale Korrelation zwischen den Matrizen, die die Distanzen von Objekten, die sich nicht im selben Cluster befinden, aufsummiert.

$$\Gamma(P,Q) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij} q_{ij}$$

Die **normalisierte Huberts Γ Statistik** misst die lineare Korrelation zwischen den beiden Matrizen und ist äquivalent zu Pearsons Korrelationskoeffizienten. Der Vorteil der Normalisierung liegt darin, dass sich der Wertebereich nun zwischen -1 und 1 befindet. Die Anwendung der Γ Statistiken benötigt allerdings Wissen über die Verteilung der Daten, das beim Clustering nicht vorhanden ist. Die Statistik müsste daher für alle $n!$ Kombinationsmöglichkeiten berechnet werden. Um den aufwändigen Prozess zu vermeiden wurde die **modifizierte Huberts Γ Statistik (Hubert & Arabie, 1985)** vorgeschlagen. Bei der modifizierten Version wird die binäre Clusterzugehörigkeitsmatrix durch eine Matrix ersetzt, die die Distanz zwischen den Clusterzentren der beiden Objekte als Einträge besitzt. Die Statistik errechnet sich folglich aus dem Produkt der Distanz zwischen zwei Objekten und der Distanz zwischen deren Clusterzentren. Der Mittelwert des Produkts wird für alle Objektpaare berechnet und kann als Clustervalidierungsmaß verwendet werden.

$$MH\Gamma(C) = \frac{1}{t} \sum_{x_i, x_j \in X} d_e(x_i, x_j) d_{x_i \in c_k, x_j \in c_l}(\bar{c}_k, \bar{c}_l)$$

Der Index steigt monoton mit der Anzahl der Cluster an, daher muss ein markantes „Knie“ in der graphischen Darstellung der Werte gesucht werden, um die optimale Clusteranzahl festzustellen (Liu et al., 2010).

Der **kophenetische Korrelationskoeffizient** (Halkidi et al., 2001) berechnet die Korrelation zwischen der Distanzmatrix und der kophenetischen Distanzmatrix. Die Einträge der kophenetischen Distanzmatrix bestehen aus den kleinsten Abständen, bei denen Objektpaare erstmalig im selben Cluster liegen. Die Werte entsprichen der Ordinate im Dendrogramm, wie sie bei hierarchischen Clustering-Algorithmen entsteht. Daher ist der kophenetische Korrelationskoeffizient auf die Anwendung bei solchen Algorithmen beschränkt.

Eine starke Korrelation der Matrizen und damit hohe Werte des Index zeigen an, dass die Cluster die Struktur der Daten widergeben. Um die optimale Clusteranzahl zu bestimmen, muss der Index demnach maximiert werden.

3.1.2 Vergleichende Indizes

Die vergleichenden Indizes zur internen Clustervalidierung von harten Clusterings lassen sich nach den verwendeten Statistiken unterscheiden. Eine Gruppe verwendet die Varianzen, während die Indizes der anderen Gruppe Kovarianzen berechnen. Einige Indizes, wie der McClain-Rao Index (McClain & Rao, 1975) und der S Index (Starczewski, 2012), lassen sich keiner Gruppe zuordnen und werden daher nicht weiter erläutert.

3.1.2.1 Indizes basierend auf Varianzen

Indizes, die auf Varianzen basieren, verwenden zur Clustervalidierung drei verschiedene Varianzen, um Distanzen zu beurteilen (vgl. Tabelle 3). Die Intra-Cluster-Varianz (SSW) berechnet die Quadratsumme der Distanz zwischen den Objekten eines Cluster und dem Clusterzentrum. Gemittelt über alle Cluster stellt dies ein Kompaktheitsmaß dar.

$$SSW = \sum_{c_k \in C} \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e^2(x_i, \bar{c}_k) = Comp(C)$$

Die Inter-Cluster-Varianz (SSB) ist ein Separierungsmaß, bei dem die Quadratsumme der Distanz zwischen den Clusterzentren und dem Mittelpunkt der Daten erfasst wird.

$$SSB = \sum_{c_k \in C} d_e^2(\bar{c}_k, \bar{X}) = Sep(C)$$

Die Varianz der kompletten Daten (SST) ist die Quadratsumme der Distanz zwischen allen Objekten des Datensatzes und dem Mittelpunkt der Daten. Ein alternativer Berechnungsweg der Varianz des Datensatzes ist die Addition der Intra- und der Inter-Cluster-Varianz.

$$SST = \sum_{x_i \in X} d_e^2(x_i, \bar{X}) = SSW + SSB = Var(X)$$

Der **Ball-Hall Index** (Ball & Hall, 1965) verwendet nur die Intra-Cluster-Varianz. Es ist daher ein reines Kompaktheitsmaß, so dass das Optimum durch ein „Knie“ angezeigt wird.

$$BH(C) = SSW$$

Der **Hartigan Index** (Hartigan, 1975) berechnet den logarithmierten Quotienten aus der Inter- und der Intra-Cluster-Varianz. Der Index ist somit ähnlich wie der Calinksi-Harabasz Index aufgebaut, der zusätzlich einen Normalisierungsfaktor besitzt. Der Hartigan Index wird nicht normalisiert, so dass das Optimum nur durch ein „Knie“ in der Verteilung gefunden werden kann.

$$H(C) = N \log_{10}\left(\frac{SSB}{SSW}\right)$$

Der **Krzanowski-Lai Index** (Krzanowski & Lai, 1988) berechnet den Unterschied zwischen den Intra-Cluster-Varianzen von Clusterings mit verschiedenen Clusteranzahlen. Es ist daher ein Kompaktheitsmaß, das schon transformiert wurde. Daher muss bei diesen Index kein „Knie“ ermittelt werden, sondern der maximale Wert wird gesucht, der die optimale Clusteranzahl anzeigt.

$$KL(C) = \frac{\left|Diff^{(k)}\right|}{\left|Diff^{(k+1)}\right|}$$

$$Diff^{(k)} = (k-1)^{2/F} * SSW^{(k-1)}(C) - N^{2/F} * SSW^{(k)}(C)$$

Der **RS-RMSSTD Index** (Sharma, 1996) kombiniert zwei Maße um die Güte der Cluster zu messen. Zum einen wird RMMSTD, ein Kompaktheitsmaß, verwendet, das die Wurzel der normierten Intra-Cluster-Varianz berechnet. Zum anderen misst RS den Anteil der Inter-Cluster-Varianz an der Varianz des Datensatzes und ermittelt so die Separierung des Clustering. Durch die Kombination der Maße wird der Nachteil, dass jeweils nur ein Konzept berücksichtigt wird, aufgehoben. Allerdings wird das Optimum für beide Maß nur durch ein „Knie“ angezeigt. Die Kombination der Vorhersagen ist auch nicht eindeutig.

$$RMSSTD(C) = \sqrt{\frac{SSW}{F(N-K)}}$$

$$RS(C) = \frac{SSB}{SST}$$

3.1.2.2 Indizes basierend auf Kovarianzen

Die zweite Gruppe der vergleichenden Indizes verwendet Kovarianzmatrizen zu Validierung der Cluster (vgl. Tabelle 4). Es gibt drei mögliche Kovarianzen:

Die Intra-Cluster-Kovarianz (W) ist die Summe der Kovarianzmatrizen aller Cluster, die sich die Abstände der Objekte von den Clusterzentren berechnen.

$$W = \sum_{c_k \in C} \sum_{x_i \in c_k} (x_i - \bar{c}_k)(x_i - \bar{c}_k)^T = Comp(C)$$

Die Inter-Cluster-Kovarianz (B) berechnet die Abstände der Clusterzentren von dem Mittelpunkt der Daten.

$$B = \sum_{c_k \in C} |c_k|(\bar{c}_k - \bar{X})(\bar{c}_k - \bar{X})^T = Sep(C)$$

Die Kovarianz der kompletten Daten (T) wird durch den Abstand der Objekte von dem Mittelpunkt der Daten erfasst. Analog zu den Quadratsummen ergibt hier auch die Addition der Intra- und Inter-Cluster-Kovarianz die Kovarianzmatrix des Datensatzes.

$$T = W + B = \sum_{x_i \in X} (x_i - \bar{X})(x_i - \bar{X})^T = Var(X)$$

Der **Rubin Index** (Friedman & Rubin, 1967) berechnet den Quotienten aus der Determinante der Kovarianz des kompletten Datensatzes und der Determinante der Intra-Cluster-Kovarianz. Das Optimum des Index wird durch ein „Knie“ angezeigt.

$$RI(C) = \frac{|T|}{|W|}$$

Der **Trace(W) Index** (Friedman & Rubin, 1967) berechnet die Spur der Intra-Cluster-Kovarianz. Hier gilt ebenfalls ein „Knie“ als Optimum.

$$TrW(C) = trace(W)$$

Der **Friedman Index** (Friedman & Rubin, 1967) verwendet die Spur des Produkts der invertierten Intra-Cluster-Kovarianz und der Inter-Cluster-Kovarianz. Der größte Sprung in den Werten des Index gibt die optimale Clusteranzahl an.

$$FI(C) = trace(W^{-1}B)$$

Der **Marriot Index** (Marriott, 1971) berechnet die Determinante der Intra-Cluster-Kovarianz und multiplizierte diese mit der quadrierten Clusteranzahl. Ein „Knie“ gibt das Optimum des Index wider.

$$MI(C) = K^2|W|$$

Der **Scott-Symons Index** (Scott & Symons, 1971) stellt eine Variante des Rubin Index dar, bei der das Verhältnis der Kovarianz zur Intra-Cluster-Kovarianz logarithmiert wird. Der

maximale Anstieg der resultierenden Kurve zeigt an, für welche Clusteranzahl das Optimum erreicht wird.

$$SS(C) = N \log_{10}\left(\frac{|T|}{|W|}\right)$$

Der **Trace(CovW) Index** (Milligan & Cooper, 1985) berechnet die Spur der Kovarianz der Kovarianzmatrizen für die einzelnen Cluster. Das Optimum des Index wird durch ein „Knie“ in der Kurve festgestellt.

$$TrCovW(C) = trace\big(cov(W)\big)$$

3.2 Indizes für weiche Clusterings

Die internen Indizes für weiche Clusterings werden nach den verwendeten Statistiken unterteilt. Die erste Gruppe benutzt nur die Clusterzugehörigkeitsmatrix U, deren Einträge angeben, wie stark die Objekte zu den Clustern gehören (vgl. Tabelle 5). Die Indizes dieser Gruppe basieren ausschließlich auf den Gewichten, die bei weichen Clustering-Algorithmen entstehen. Die zweite Gruppe von Indizes bezieht zusätzlich die Geometrie der Daten ein, repräsentiert durch den Vektor mit den Clusterzentren (vgl. Tabelle 6).

Es gibt noch weitere Indizes, die sich in keine der beiden Gruppen einsortieren lassen. Der Merging Index (Chong et al., 2002) ist ein hybrider Ansatz, der zunächst einen herkömmlichen Index verwendet und anschließend prüft, ob es sinnvoll ist Cluster zusammenzulegen. Der Bayesian Score (Cho & Yoo, 2006) wendet den Satz von Bayes auf Clusterings an, der berechnet, wie gut die Cluster die Daten repräsentieren. Der Stability Index (Yu & Li, 2006) misst die Stabilität der Cluster, indem die Konditionszahl der Hesse-Matrix der Zielfunktion von Fuzzy C-Means berechnet wird.

3.2.1 Indizes basierend auf Clusterzugehörigkeitsgraden

Der **Partition Coefficient** (Bezdek, 1974) misst die Kompaktheit eines Clusterings. Der Index berechnet den Mittelwert der quadrierten Gewichte über alle Objekte und Cluster. Dadurch wird erfasst, wie stark sich die Gewichte der Objekte auf mehrere Cluster verteilen. Im Optimalfall sind alle Gewichte entweder 0 oder 1 und der Index damit maximal. Im schlechtesten Fall sind die Gewichte der Objekte für jeden Cluster identisch und der Index

erreicht das Minimum von $\frac{1}{K}$. Zur Bestimmung der korrekten Clusteranzahl muss der Partition Coefficient also maximiert werden.

$$PC(U) = Comp(U) = \frac{1}{N}\sum_{k=1}^{K} Comp(u_k)$$

$$Comp(u_k) = \sum_{i=1}^{N} u_{ki}^{m}$$

Es konnte gezeigt werden, dass der Partition Coefficient robust gegen Rauschen und Ausreißer in den Daten ist (Wu et al., 2009). Er hat allerdings auch einige Nachteile (Halkidi et al., 2001). Der Index steigt monoton mit der Anzahl der Cluster an, so dass das Optimum teilweise nur durch eine „Knie“ in der grafischen Darstellung zu erkennen ist. Der Index ist auch abhängig von dem Fuzzifier m, ein Parameter des Fuzzy C-Means. Wenn $m \to 1$, nimmt der Index für jede Anzahl an Clustern den gleichen Wert an. Wenn $m \to \infty$, hat der Index immer ein signifikante „Knie“ bei zwei Clustern. Oft wird der $m = 2$ gewählt oder Werte in der Nähe.

Es wurden zwei Alternativen vorgeschlagen, die die monotone Abhängigkeit verringern sollen: der Fuzziness Performace Index (Roubens, 1982) und der modifizierte Partition Coefficient (Dave, 1996). Beide Indizes normalisieren auf den Wertebereich 0 bis 1, indem sie den Partition Coefficient auf bestimmte Weise mit der Clusteranzahl kombinieren. Der Fuzziness Performace Index muss minimiert werden, um die optimale Clusterzahl zu bestimmen, während der modifizierte Partition Coefficient nach wie vor maximiert wird.

$$FPI(U) = 1 - \frac{K * PC(U) - 1}{K - 1}$$

$$MPC(U) = 1 - \frac{K}{K - 1}(1 - PC(U))$$

Die **Partition Entropy** (Bezdek, 1973) misst die Unschärfe des Clusterings, indem der Mittelwert der Entropien der Gewichte berechnet wird. Sind die Gewichte eindeutig (0 und 1), ist die Entropie 0. Sind alle Gewichte gleich groß, ist die Entropie $log\ K$. Folglich muss der Index minimiert werden, um das Clustering mit den eindeutigsten Clustern zu erhalten.

$$PE(U) = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} u_{ki} \log_a u_{ki}$$

Für die Partition Entropy gelten dieselben Vor- und Nachteile wie für den Partition Coefficient. Um die monotone Abhängigkeit zu verhindern, wurde die modifizierte Partition

Entropy (Roubens, 1982) vorgeschlagen, die den Index durch den Logarithmus der Clusteranzahl dividiert. Dadurch wird die Partition Entropy auf den Wertebereich 0 bis 1 normiert. Das Optimum wird weiterhin durch Minimieren erreicht.

$$MPE(U) = \frac{1}{\log_a K} PE(U)$$

Der **P Index** (Chen & Linkens, 2001) beurteilt die Güte der Clusterings, indem die Differenz aus einem Kompaktheits- und einem Separierungsmaß gebildet wird.

$$P(U) = Comp(U) - Sep(U)$$

Das Kompaktheitsmaß erfasst für jedes Objekt das maximale Gewicht und mittelt diese. Dadurch wird gemessen, wie stark die Objekte zu den Clustern zugeordnet werden. Das Maß muss demnach maximiert werden, um möglichst eindeutige Cluster zu erhalten.

$$Comp(U) = \frac{1}{N} \sum_{i=1}^{N} \max_k \{u_{ki}\}$$

Die Separierung wird für jedes Clusterpaar berechnet und gemittelt. Dazu wird für jedes Objekt das kleinere Gewicht der Cluster ermittelt. Dadurch kann die Überlappung der zwei Cluster gemessen werden. Ist die Separierung hoch, dann ist selbst das kleinere Gewicht eines Objekts für beide Cluster noch groß. Sind die Cluster klar getrennt, wird der Wert minimal.

$$Sep(U) = \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} Sep(u_k, u_l)$$

$$Sep(u_k, u_l) = \frac{1}{N} \sum_{i=1}^{N} \min\{u_{ki}, u_{li}\}$$

Der P Index wird maximiert, da untypischerweise das Kompaktheitsmaß maximiert und das Separierungsmaß minimiert werden muss.

Der **KYI Index** (Kim et al., 2004b) ist ein Separierungsmaß, das den Überlappungsgrad der Cluster misst. Die Überlappung wird für alle Clusterpaare errechnet und gemittelt. Dazu werden für jedes Objekt die Gewichte von zwei Clustern betrachtet und deren Überlappung ermittelt, die mit der Entropie der Gewichte des Objekts sowie der Clusteranzahl multipliziert wird.

$$KYI(U) = Sep(U) = \frac{2}{K(K-1)} \sum_{k \neq l}^{K} \sum_{i=1}^{N} [K * (u_{ki} \wedge u_{li}) * h_i]$$

Sind die Werte des KYI Index gering, so sind nur wenige Überlappungen zwischen den Clustern vorhanden. Um ein Clustering mit klar getrennten Cluster zu erhalten, muss der Index demnach minimiert werden.

Der **OS Index** (Kim et al., 2004a) berechnet den Quotienten aus Überlappung und Separierung, wobei beide Terme jeweils durch die maximalen Werte normiert werden. Diese Normierung ist notwendig, da beide Maße über verschiedene Wertebereiche verfügen.

$$OS(U) = \frac{Overlap(U)/Overlap_{max}}{Sep(U)/Sep_{max}}$$

Die Überlappung P wird jeweils für zwei Cluster berechnet und über alle möglichen Paare gemittelt. Dazu werden die Gewichte der Objekte für jeweils zwei Cluster betrachtet. Überschreiten beide Gewichte einen gewissen Schwellwert μ wird dies als Überlappung δ gezählt, die mit einem Gewichtungsfaktor ω multipliziert wird, der angibt wie unscharf ein Objekt ist. Ist ein Objekt relativ eindeutig einem Cluster zugeordnet, wird eine Überlappung nicht so stark gewichtet. Handelt es sich um ein Objekt, das nur über geringe Gewichte verfügt, so wird die Überlappung stärker bestraft. Das Überlappungsmaß sollte demnach minimiert werden.

$$Overlap(U) = \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} P(c_k, c_l)$$

$$P(c_k, c_l) = \sum_{\mu} \sum_{i=1}^{N} \delta(x_i, \mu: c_k, c_l) * \omega(x_i)$$

Zur Ermittlung der Separierung wird für jedes Clusterpaar das Maximum des jeweils kleineren Gewichts ermittelt. Ist der Wert groß, zeigt das an, dass die Cluster nicht stark getrennt sind. Das Clusterpaar, für den dieses Gewicht minimal ist und das daher die größte Separierung aufweist, wird für das Separierungsmaß verwendet. Das Gewicht wird von 1 subtrahiert, so dass maximale Werte der Separierung zu bevorzugen sind.

$$Sep(U) = 1 - \min_{k \neq l} \left\{ \max_{x_i \in X} \{ \min\{u_{ki}, u_{li}\} \} \right\}$$

Zur Ermittlung der optimalen Clusteranzahl muss der OS Index demnach minimiert werden.

3.2.2 Indizes basierend auf Clusterzugehörigkeitsgraden und den Daten

Der **Fukuyama-Sugeno Index** (Fukuyama & Sugeno, 1989) bildet die Differenz aus Kompaktheit und Separierung zur Validierung der Cluster.

$$FS(U,V) = Comp(U,V) - Sep(U,V)$$

Die Kompaktheit des Clusterings ist die Summe der Distanzen von den Objekten zu den Clusterzentren. Die Distanzen werden mit den Clusterzugehörigkeiten multipliziert und dadurch gewichtet. Dieses Maß entspricht der Zielfunktion des meist verwendeten, weichen Clustering-Algorithmus Fuzzy C-Means. Wird dieses Maß für ein hartes Clustering berechnet, so entspricht es der Intra-Cluster-Varianz und damit der Zielfunktion von K-Means.

$$Comp(U,V) = \sum_{k=1}^{K} Comp(u_k, v_k)$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m d_e^2(x_i, v_k)$$

Das Separierungsmaß erfasst die Summe der Distanzen zwischen den Clusterzentren und dem Mittelpunkt der Daten. Dieses Maß entspricht der Inter-Cluster-Varianz.

$$Sep(U,V) = \sum_{k=1}^{K} Sep(u_k, v_k)$$

$$Sep_5(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m d_e^2(v_k, \bar{v})$$

Das Minimum des Fukuyama-Sugeno Index zeigt an, dass es sich um ein kompaktes und separiertes Clustering handelt. Bei der Ermittlung der optimalen Clusteranzahl wird der Index stark von Rauschen und Ausreißern in den Daten beeinflusst (Wu et al., 2009).

Gath & Geva (1989) haben mehrere Indizes vorgeschlagen, die die Kompaktheit und Dichte der Cluster evaluieren. **Fuzzy Hyper Volume** berechnet für jeden Cluster die Kovarianzmatrix Σ_k. Die Wurzel der Determinanten der Matrix wird über alle Cluster aufsummiert und erfasst die Kompaktheit des Clusterings. Minimale Werte des Index sind ein Indikator für kompakte Cluster.

$$FHV(U,V) = Comp(U,V) = \sum_{k=1}^{K} \sqrt{|\Sigma_k|}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} u_{ki}^m (x_i - v_i)(x_i - v_i)^T}{\sum_{i=1}^{N} u_{ki}^m}$$

Average Partition Density errechnet für jeden Cluster den Quotienten aus Kompaktheit und Dichte und mittelt diesen.

$$APD(U,V) = \frac{1}{K} \sum_{k=1}^{K} \frac{Dens(u_k)}{Comp(u_k, v_k)} = \frac{1}{K} \sum_{k=1}^{K} \frac{S_k}{\sqrt{|\Sigma_k|}}$$

Die Kompaktheit der Cluster wird, wie bei Fuzzy Hyper Volume, mit der Wurzel der Determinanten der Kovarianzmatrix beurteilt.

$$Comp(u_k, v_k) = \sqrt{|\Sigma_k|}$$

Das Dichtemaß summiert die Gewichte der zentralen Mitglieder X_k eines Clusters. X_k stellt eine Menge von Objekten dar, die sich in einer vorgegebenen Region rund um das Clusterzentrum befinden.

$$Dens(u_k) = S_k = \sum_{x_i \in X_k} u_{ki}$$

Das Maximum von Average Partition Density zeigt an, dass ein Clustering mit kompakten und dichten Clustern gefunden wurde.

Partition Density kombiniert die beiden vorigen Indizes, indem es den Quotienten aus der Dichte und der Kompaktheit des Clusterings bildet. Das Dichtemaß S_k wird über alle Cluster summiert und durch Fuzzy Hyper Volume geteilt. Maximale Werte der Partition Density geben die optimale Clusteranzahl wider.

$$PD(U,V) = \frac{Dens(U)}{Comp(U,V)} = \frac{\sum_{k=1}^{K} S_k}{FHV(U,V)}$$

Der **Xie-Beni Index** (Xie & Beni, 1991) berechnet den Quotienten aus Kompaktheit und Separierung und ist einer der meist verwendeten Indizes für weiche Clusterings.

$$XB(U,V) = \frac{1}{N} \frac{Comp(U,V)}{Sep(V)}$$

Das Kompaktheitsmaß verwendet die Intra-Cluster-Varianz und berechnet demnach die gewichtete Distanz der Objekte zu den Clusterzentren.

$$Comp(U,V) = \sum_{k=1}^{K} Comp(u_k, v_k)$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m d_e^2(x_i, v_k)$$

Die Separierung erfasst die minimale Distanz zwischen zwei Clusterzentren im Datensatz.

$$Sep(V) = \min_k \min_l \{Sep(v_k, v_l)\}$$

$$Sep_4(v_k, v_l) = d_e^2(v_k, v_l)$$

Zur Ermittlung der optimalen Clusteranzahl muss der Xie-Beni Index minimiert werden. Dabei wird der Index stark von Rauschen und Ausreißern in den Daten beeinflusst (Wu et al., 2009).

Weitere Nachteile sind, dass der Index monoton fällt, wenn die Anzahl der Cluster sehr groß ist und dass der Index abhängig vom Fuzzifier m ist, da gilt: wenn $m \to \infty$, dann $XB \to \infty$ (Halkidi et al., 2001). Es wurde daher einige Alternative vorgestellt, die vor allem versuchen die Monotonität zu verhindern.

Der **Partition Index** (Bensaid et al., 1996) nimmt eine Anpassung des Xie-Beni Index an die Eigenschaften des Fuzzy C-Means vor. Das Separierungsmaß wird dahingehend verändert, dass nicht die minimale Separierung, sondern die mittlere Separierung der Cluster betrachtet wird. Die minimale Distanz ist gut zum Finden der optimalen Anzahl an Clustern, ist aber nicht geeignet, um verschiedene Clusterings zu vergleichen. Eine weitere Änderung ist, dass der Quotient aus Kompaktheit und Separierung pro Cluster ermittelt wird. Dabei findet auch eine Normalisierung der Werte durch die Summe der Gewichte für den Cluster statt (Clustergröße), die das monotone Fallen verhindern soll. Die Zielfunktion des Fuzzy C-Means erfüllt diese Anforderung nicht.

$$SC(U,V) = \sum_{k=1}^{K} \frac{1}{\sum_{i=1}^{N} u_{ki}} \frac{Comp(u_k, v_k)}{\sum_{l=1}^{K} Sep(v_k, v_l)}$$

Der **K Index** (Kwon, 1998) stellt eine Anpassung des Xie-Beni Index dar, um das monotone Fallen bei hohen Clusterzahlen zu vermeiden. Dazu wird ein Strafterm hinzugefügt, der die mittlere Distanz der Clusterzentren zu dem Mittelpunkt der Daten erfasst.

$$K(U,V) = \frac{Comp(U,V) + 1/K \sum_{k=1}^{K} d_e^2(v_k, \bar{X})}{Sep(V)}$$

Der **T Index** (Tang et al., 2005) verfolgt die gleiche Idee wie Kwon (1998) und fügt zwei Strafterme hinzu. Im Zähler wird die mittlere Distanz zwischen den Clusterzentren addiert, um das monotone Fallen des Index zu verhindern. Im Nenner wird $1/K$ addiert, um den Index insgesamt stabiler zu machen.

$$T(U,V) = \frac{Comp(U,V) + 1/(K(K-1)) \sum_{k=1}^{K} \sum_{l=1 \neq k}^{K} d_e^2(v_k, v_l)}{Sep(V) + 1/K}$$

Der **CV Index** (Rhee & Oh, 1996) erfasst die Qualität der Clusterings mit dem Quotienten aus Separierung und Kompaktheit. Der Index ist recht einfach gehalten, da das Hauptaugenmerk der Autoren auf einer neuen, komplexen Methode zur Ermittlung der optimalen Clusteranzahl lag.

$$CV(U,X) = \frac{Sep(U,X)}{Comp(U,X)}$$

Das Kompaktheitsmaß besteht aus der mittleren Distanz zwischen allen Objekte multipliziert mit dem jeweils kleineren Gewicht des Objektpaares für den Cluster.

$$Comp(U,X) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{K} \left[\min\{u_{ki}, u_{kj}\} * d_e^2(x_i, x_j)\right]$$

Das Separierungsmaß wird ebenfalls für alle Objektpaare ermittelt. Die Separierung der Objekte wird berechnet durch die Distanz multipliziert mit dem maximalen Gewicht des unschärferen Objekts.

$$Sep(U,X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\min\left\{\max_k\{u_{ki}\}, \max_l\{u_{li}\}\right\} * d_e^2(x_i, x_j)\right]$$

Das Maximum des CV Index zeigt an, dass ein kompaktes und separiertes Clustering gefunden wurde.

Der **CWB Index** (Rezaee et al., 1998) bestimmt die Summe aus Kompaktheit und Separierung, wobei ein Gewichtungsfaktor verwendet wird, da die Maße verschiedene Wertebereiche besitzen. Der SD Index verwendet das identische Konzept und kann für die genauen Berechnungen nachgeschaut werden (vgl. Kap. 3.1.1.1).

Sun et al. (2004) haben den WSJ Index vorgeschlagen, der nahezu identisch zu dem CWB Index ist. Der einzige Unterschied besteht in dem verwendeten Distanzmaß zur Ermittlung

der Separierung des Clusterings. Bei dem CWB Index wird die euklidische Distanz verwendet, während bei dem WSJ Index die quadrierte euklidische Distanz angewandt wurde.

Der **SC_{ZLE} Index** (Zahid et al., 1999) besteht aus der Differenz von zwei Quotienten, die jeweils Separierung und Kompaktheit des Clusterings evaluieren. Im ersten Quotienten werden dazu die Clusterzugehörigkeiten und die Geometrie der Daten verwendet, während im zweiten Quotienten nur die Clusterzugehörigkeiten berücksichtigt werden.

$$SC_{ZLE}(U,V) = SC_1(U,V) - SC_2(U)$$

Der erste Term berechnet das Verhältnis von Separierung zu Kompaktheit. Das Kompaktheitsmaß berechnet die mittlere Distanz der Objekte zu den Clusterzentren, gewichtet durch die Summe der Gewichte für jeden Cluster.

$$SC_1(U,V) = \frac{Sep(V)}{Comp(U,V)}$$

$$Comp(U,V) = \sum_{k=1}^{K} \frac{Comp(u_k, v_k)}{\sum_{i=1}^{N} u_{ki}}$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m \, d_e^2(x_i, v_k)$$

Die Separierung wird durch die mittlere Distanz der Clusterzentren zu dem Mittelpunkt der Daten erfasst.

$$Sep(V) = \frac{1}{K} \sum_{k=1}^{K} Sep(v_k)$$

$$Sep_5(v_k) = d_e^2(v_k, \bar{X})$$

Der zweite Term berechnet ebenfalls das Verhältnis von Separierung zu Kompaktheit, verwendet aber nur die Clusterzugehörigkeitsmatrix. Die Kompaktheit wird erfasst, indem die Summe der quadrierten, maximalen Gewichte pro Objekt berechnet wird und durch die einfache Summe gewichtet wird. Dieses Kompaktheitsmaß nimmt maximale Werte an, wenn die Objekte klar den Clustern zugeordnet sind.

$$SC_2(U) = \frac{Sep(U)}{Comp(U)}$$

$$Comp(U) = \frac{\sum_{i=1}^{N} (\max_k \{u_{ki}\})^2}{\sum_{i=1}^{N} \max_k \{u_{ki}\}}$$

Die Separierung ermittelt die Überlappung der Cluster. Dazu wird die Summe der quadrierten, minimalen Gewichte pro Objektpaar berechnet und ebenfalls durch die einfache Summe gewichtet. Dieses Maß gilt es zu minimieren.

$$Sep(U) = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \frac{\sum_{i=1}^{N} (\min\{u_{ki}, u_{li}\})^2}{\sum_{i=1}^{N} \min\{u_{ki}, u_{li}\}}$$

Die optimale Clusteranzahl kann ermittelt werden, indem der SC_{ZLE} Index maximiert wird. Die beiden Terme befinden sich allerdings nicht zwingend im selben Wertebereich. Der erste Term SC_1 dominiert den Index, wenn die Cluster kompakt und separiert sind.

Die **Scattering Criteria** (Geva et al., 2000) basieren auf den Streumatrizen, die Kovarianz innerhalb und zwischen den Clustern erfassen. Die Intra-Cluster-Streumatrix S_w misst die Differenz zwischen den Objekten und den Clusterzentren. Die Inter-Cluster-Streumatrix S_B erfasst die Distanz zwischen den Clusterzentren und dem Mittelpunkt der Daten.

$$S_w = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ki}(x_i - v_k)(x_i - v_k)^T$$

$$S_B = \sum_{k=1}^{K} \left(\sum_{i=1}^{N} u_{ki} \right) (v_k - \bar{X})(v_k - \bar{X})^T$$

Insgesamt wurden vier Indizes vorgeschlagen, die die Spur oder Determinante der Streumatrizen verwenden. Allerdings werden nur bei einem Index beide Matrizen kombiniert, so dass ein Maß entsteht, das sich mit anderen Separierungs-und Kompaktheitsmaßen vergleichen lässt.

Das **Invariant Criterion** berechnet die Spur der Produktmatrix aus der invertierten Intra-Cluster-Streumatrix und der Inter-Cluster-Streumatrix. Der Index wird durch die quadrierte Clusteranzahl normalisiert, um ein monotones Verhalten zu verhindern. Das Maximum des Index zeigt die optimale Clusterzahl an.

$$Inv(U, V) = \frac{1}{K^2} trace\left({S_w}^{-1} S_B\right)$$

Der **SV Index** (Kim et al., 2001) berechnet die Summe aus Kompaktheit und der invertierten Separierung zur Validierung der Clusterings. Die Besonderheit bei dem Index liegt darin, dass die Clusterzugehörigkeiten und damit die Gewichte ignoriert werden und nur die Clusterzentren zur Berechnung verwendet werden. Zusätzlich werden beide Maße normalisiert, indem der minimale Wert abgezogen wird und durch den Wertebereich geteilt wird.

$$SV(V) = Comp_N(V) + \frac{1}{Sep_N(V)}$$

$$Comp_N(V) = \frac{Comp(V) - Comp_{min}}{Comp_{max} - Comp_{min}}$$

Die Kompaktheit des Clusterings wird mit der mittleren Distanzen zwischen den Objekten und den Clusterzentren erfasst.

$$Comp(V) = \frac{1}{K}\sum_{k=1}^{K} Comp(v_k)$$

$$Comp_3(v_k) = \sum_{x_i \in c_k} \frac{d_e(x_i, v_k)}{|c_k|}$$

Das Separierungsmaß ist die minimale Distanz zwischen zwei Clusterzentren im Datensatz.

$$Sep(V) = \frac{1}{K} \min_k \min_l \{Sep(v_k, v_l)\}$$

$$Sep_4(v_k, v_l) = d_e(v_k, v_l)$$

Das Minimum des SV Index zeigt die optimale Clusterzahl an, da die zu maximierende Separierung invertiert wird.

Der **SVI Index** (Tsekouras & Sarimveis, 2004) wird ermittelt, indem der Quotient aus Kompaktheit und Separierung berechnet wird.

$$SVI(U, V) = \frac{Comp(U, V)}{Sep(V)}$$

Das Kompaktheitsmaß ist identisch wie bei dem SC_{ZLE} Index. Es wird die mittlere Distanz zwischen den Objekten und den Clusterzentren berechnet und durch die Summe der Gewichte des Clusters gewichtet.

$$Comp(U, V) = \sum_{k=1}^{K} \frac{Comp(u_k, v_k)}{\sum_{i=1}^{N} u_{ki}}$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m \, d_e^2(x_i, v_k)$$

Das Separierungsmaß misst die mittlere Distanz zwischen den Clusterzentren und multipliziert diese mit einem komplexen Gewichtungsfaktor. Dazu wird ein neuer Vektor z benötigt, der die Clusterzentren und den Mittelpunkt des Datensatzes beinhaltet:

$[z_1, z_2, \dots, z_K, z_{K+1}] = [v_1, v_2, \dots, v_K, \bar{X}]$. Darauf aufbauend wird eine Clusterzugehörigkeitsfunktion μ und ein Gewichtungsfaktor $\omega \in (0, \infty)$ verwendet.

$$Sep(V) = \sum_{i=1}^{K+1} \sum_{\substack{j=1 \\ j \neq i}}^{K+1} (\mu_{ij})^{\frac{2+\omega}{\omega}} d_e(z_i, z_j)$$

Der SV Index nimmt minimale Werte an, wenn das Clustering kompakt und separiert ist.

Der **GD Index**, auch bekannt als Separation-Compactness Index (Xie et al., 2002) und als Granularity-Dissimilarity Index (Xie et al., 2005), berechnet die quadrierte Separierung und teilt diese durch die Kompaktheit.

$$GD(U, V) = \frac{\big(Sep(V)\big)^2}{Comp(U, V, X)}$$

Das Kompaktheitsmaß besteht aus der mittleren Distanz der Objekte zu den Clusterzentren. Zur Berechnung sollten laut den Autoren dieses Index nur Cluster einbezogen werden, die mehr als ein Objekt beinhalten.

$$Comp(U, V) = \sum_{k=1}^{K} \frac{Comp(u_k, v_k)}{\sum_{i=1}^{N} u_{ki}^2}$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^2 \, d_e^2(x_i, v_k)$$

Das Separierungsmaß besteht aus dem Mittlerwert der minimalen Separierung zwischen zwei Clustern. Die Separierung wird durch die Distanz zwischen den Clusterzentren erfasst.

$$Sep(V) = \frac{1}{K} \sum_{k=1}^{K} \min_{l \neq k} \{Sep(v_k, v_l)\}$$

$$Sep_4(v_k, v_l) = d_e^2(v_k, v_l)$$

Der GD Index muss maximiert werden, um die optimale Clusteranzahl zu ermitteln.

Der **PCAES Index** (Wu & Yang, 2005) erfasst die Qualität der Clusterings durch die Differenz aus Kompaktheit und Separierung.

$$PCAES(U, V) = Comp(U) - Sep(V)$$

Die Kompaktheit eines Clusters wird dem Partition Coefficient erfasst, d.h. die quadratische Summe der Gewichte wird berechnet. Die Werte der Cluster werden durch die Summe der Gewichte für den kompaktesten Cluster geteilt und summiert.

$$Comp(U) = \sum_{k=1}^{K} \frac{Comp(u_k)}{\min_{1 \le l \le K} \{Comp(u_l)\}}$$

$$Comp(u_k) = \sum_{i=1}^{N} u_{ki}^2$$

Die Separierung wird für jeden Cluster einzeln ermittelt. Dazu wird die minimale Distanz zu dem nächsten Clusterzentrum durch die mittlere Distanz der Clusterzentren zu dem Mittelpunkt der Daten geteilt. Dieser Wert wird auf die Exponentialfunktion angewandt und über alle Cluster aufsummiert.

$$Sep(V) = \sum_{k=1}^{K} \exp\left(-\frac{Sep(v_k)}{Sep'(V)}\right)$$

$$Sep(v_k) = \min_{l \neq k} \{d_e^2(v_k, v_l)\}$$

$$Sep_4(v_k, v_l) = d_e^2(v_k, v_l)$$

$$Sep'(V) = Sep_5(V) = \frac{1}{K} \sum_{k=1}^{K} d_e^2(v_k, \bar{X})$$

Erreicht der PCAES Index maximale Werte, so wurden ein kompaktes und separiertes Clustering gefunden.

Der **SC_{BWS} Index** (Bouguessa et al., 2006) berechnet den Quotienten aus Separierung und Kompaktheit.

$$SC_{BWS}(U, V) = \frac{Sep(U, V)}{Comp(U, V)}$$

Das Kompaktheitsmaß erfasst die Summe der Distanzen zwischen den Objekten und den Clusterzentren, gewichtet mit der Summe der Gewichte für den jeweiligen Cluster. Diese Normalisierung verhindert ein monotones Fallen des Kompaktheitsmaßes und das Maß reagiert dadurch sensibler in Bezug auf Änderungen in der Kovarianzstruktur der Cluster (Bouguessa et al., 2006).

$$Comp(U, V) = \sum_{k=1}^{K} \frac{Comp(u_k, v_k)}{\sum_{i=1}^{N} u_{ki}}$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m \, d_e^2(x_i, v_k)$$

Die Separierung des Clustering errechnet sich aus der Summe der Distanzen zwischen den Clusterzentren und dem Mittelpunkt der Daten.

$$Sep(U,V) = \sum_{k=1}^{K} Sep(u_k, v_k)$$

$$Sep_5(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^m d_e^2(v_k, \bar{X})$$

Das Maximum des SC_{BWS} Index gibt an, wann ein kompaktes und separiertes Clustering vorliegt.

Der **W Index** (Zhang et al., 2008) berechnet den Quotienten aus Kompaktheit und Separierung, wobei die Maße mit dem jeweiligen maximalen Wert normalisiert werden.

$$W(U,V) = \frac{Comp(U,V)/Comp_{max}}{Sep(U)/Sep_{max}}$$

Die Kompaktheit des Clusterings wird pro Cluster berechnet, aufsummiert und mit dem Faktor $\sqrt{\frac{K+1}{K-1}}$ gewichtet. Die Kompaktheit der Cluster berechnet sich aus der mittleren Distanz zwischen den Objekten und den Clusterzentren. Das Besondere dabei ist die verwendete Distanzmetrik d_{exp}^2, die die Exponentialfunktion verwendet, da diese robuster als die euklidische Distanz ist und gut für die Anwendung in der Clusteranalyse geeignet ist. Die Distanzmetrik berechnet sich aus der Wurzel der Exponentialfunktion von dem Produkt der Distanz und der invertierten Stichprobenkovarianz. Zur Berechnung der Kompaktheit wird auch der Gewichtungsfaktor n_k verwendet, der die Anzahl der Objekte in Cluster k misst. Dieser Faktor sorgt dafür, dass das Kompaktheitsmaß nicht monoton fällt, wenn sich die Anzahl der Cluster der der Anzahl der Objekte annähert.

$$Comp(U,V) = \sqrt{\frac{K+1}{K-1}} * \sum_{k=1}^{K} Comp(u_k, v_k)$$

$$Comp(u_k, v_k) = \sum_{i=1}^{N} \frac{u_{ki}}{n_k} d_{exp}^2(x_i, v_k)$$

$$d_{exp}^2(x_i, v_k) = \sqrt{1 - \exp\left(\beta * d_e^2(x_i, v_k)\right)}$$

Das Separierungsmaß ist von dem OS Index bekannt (vgl. Kapitel 3.2.1). Es ermittelt das Clusterpaar mit der größten Separierung, gemessen an dem Gewicht für den jeweils schwächeren Cluster.

$$Sep(U) = 1 - \min_{k \neq l} \left\{ \max_{x_i \in X} \{ \min\{u_{ki}, u_{li}\} \} \right\}$$

Zur Berechnung der optimalen Clusterzahl muss der W Index minimiert werden.

Der **SC_R Index** (Rezaee, 2010) evaluiert die Summe aus Kompaktheit und Separierung, wobei beide Maße jeweils mit dem maximalen Wert gewichtet werden.

$$SC_R(U,V) = \frac{Sep(U)}{Sep_{max}} + \frac{Comp(U,V)}{Comp_{max}}$$

Das Kompaktheitsmaß besteht aus der Summe der Distanzen zwischen den Objekten und den jeweiligen Clusterzentren.

$$Comp(U,V) = \sum_{k=1}^{K} Comp(u_k, v_k)$$

$$Comp_3(u_k, v_k) = \sum_{i=1}^{N} u_{ki}^2 d_e^2(x_i, v_k)$$

Das Separierungsmaß, auch bekannt als KYI Index (vgl. Kapitel 3.2.1), misst den Überlappungsgrad der Cluster.

$$Sep(U) = \frac{2}{K(K-1)} \sum_{k \neq l}^{K} \sum_{i=1}^{N} [K * (u_{ki} \wedge u_{li}) * h_i]$$

Minimale Werte des SC_R Index sind ein Indikator für ein kompaktes und separiertes Clustering.

4 Relative Clustervalidierungsindizes

Die relativen Clustervalidierungsindizes werden in drei Gruppen unterteilt, die sich in dem Vorgehen zur Ermittlung der Clusterstabilität unterscheiden (vgl. Tabelle 7). Die Indizes der ersten Gruppen erstellen überlappende Stichproben von dem Datensatz und führen jeweils eine Clusteranalyse durch. Die Ergebnisse werden verglichen und damit die Stabilität erfasst. Die zweite Gruppe von Indizes teilt die Daten in zwei Hälften und lernt einen Klassifizierer, dessen Prognosen einem Clustering gegenübergestellt werden. Die Indizes der dritten Gruppen verändern die Daten, indem sie Störterme addieren oder Variablen weglassen. Die Indizes erfassen die Stabilität der Clusterings auf den manipulierten Daten.

4.1 Indizes basierend auf überlappenden Stichproben

Figure of Merit (Levine & Domany, 2001) basiert auf der Erstellung und Clustering von vielen Stichproben und misst deren Übereinstimmung mit dem Clustering auf allen Daten. Ist die Übereinstimmung groß, dann ist das ein Indikator dafür, dass die Cluster stabil bleiben auch wenn ein Teil der Daten fehlt. Die Übereinstimmung zwischen den Stichproben und den kompletten Daten wird durch die Figure of Merit berechnet, die die Konnektivitätsmatrizen der Stichproben mit der Matrix des kompletten Datensatzes vergleicht:

$$\mathcal{M}(K) = \left\langle \left\langle \delta_{M(i,j),M^{(h)}(i,j)} \right\rangle \right\rangle_H$$

$\langle\langle \ . \ \rangle\rangle_H$ steht für das Mitteln der Werte über alle Stichproben. M ist die Konnektivitätsmatrix des kompletten Datensatzes, während $M^{(h)}$ die Matrix der $h$ten Stichprobe darstellt. Die Konnektivitätsmatrix ist eine NxN Matrix, die die die Clusterzugehörigkeiten der Objekte widergibt. Befinden sich zwei Objekte im selben Cluster, so ist der Eintrag 1, ansonsten 0. Das Verfahren der Methode läuft folgendermaßen ab:

1. Clustering auf dem kompletten Datensatz ausführen.
2. H Stichproben erstellen, indem zufällig fN Objekte ausgewählt werden. Der Faktor f gibt an, wie groß die Stichproben sind.
3. Clustering auf jeder Stichprobe ausführen.
4. Berechnung der Figure of Merit $\mathcal{M}$.

5. Variation von der Clusteranzahl K um stabile Clusterings zu identifizieren, bei denen der Index ein lokales Maximum erreicht.

Model Explorer (Ben-Hur et al., 2002) erstellt ebenfalls Stichproben und vergleicht diese mittels externer Indizes, die in dieser Anwendung auch Ähnlichkeitsmaße genannt werden. Bei dem Verfahren werden immer zwei Stichproben gleichzeitig erstellt, deren Übereinstimmung dann gemessen wird:

1. Erstellung von zwei Stichproben des kompletten Datensatzes mit Faktor f.

$$sub_1 = subsamp(X, f)\ ;\ sub_2 = subsamp(X, f)$$

2. Clustering auf beiden Stichproben ausführen, um zwei Sätze von Labeln zu erhalten.

$$L_1 = cluster(sub_1, K)\ ;\ L_2 = cluster(sub_2, K)$$

3. Berechnung der Schnittmenge der beiden Stichproben.

$$intersect = sub_1 \cap sub_2$$

4. Berechnung der Ähnlichkeit der Objekte, die in beiden Proben enthalten sind.

$$S(h, K) = s(L_1(intersect), L_2(intersect))$$

Das Verfahren wird für viele Stichprobenpaare H wiederholt und berechnet für wie viele Stichprobenpaare das Ähnlichkeitsmaß über einem gewissen Schwellwert liegt. Dies wird für verschiedene Clusteranzahlen wiederholt und grafisch dargestellt. In der resultierenden Kurve wird dann nach einem signifikanten Sprung gesucht, der die optimale Clusterzahl angibt. Diese Methodik wurde kritisiert, da sie nicht quantifiziert ist und dementsprechend subjektiv ist (Lange et al., 2004).

Consensus (Monti et al., 2003) kombiniert Teile der beiden vorigen Verfahren. Analog zu Model Explorer wird die Übereinstimmung nur zwischen den Stichproben ermittelt. Dazu werden die Konnektivitätsmatrizen verwendet, die auch bei Figure of Merit berechnet wurden. Der Unterschied besteht darin, dass die Konnektivitätsmatrizen aufsummiert und normalisiert werden. Die resultierende Consensus-Matrix gibt an, wie oft ein Objektpaar demselben Cluster zugeordnet wurde geteilt durch die Anzahl der Stichproben, in denen beide Objekte enthalten waren. Ziel ist es, dass die Einträge der Matrix nur aus Nullen und Einsen bestehen. Das formalisierte Verfahren: Wiederholung der folgenden Schritte für alle möglichen Clusteranzahlen.

1. Erstelle H Stichproben und Indikatormatrizen $I^{(h)}$, die angeben, ob sich Objektpaare in einer Stichprobe befinden.
2. Clustering auf allen Stichproben ausführen.
3. Erstellung der Konnektivitätsmatrizen $M^{(h)}$.
4. Berechnung der Consensus-Matrix: $\mathcal{M}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)}$.

Zur Bestimmung der optimalen Clusterzahl wird die Fläche unterhalb der Verteilungsfunktion für die Consensus-Matrix berechnet. Mit Hilfe eines quantifizierten Verfahrens wird der größte Sprung in der Größe der Fläche ermittelt und daraus die optimale Anzahl an Clustern abgeleitet.

4.2 Indizes basierend auf dem Vergleich mit einer Prognose

Die **Gap Statistik** (Tibshirani et al., 2001) berechnet die Kompaktheit der Cluster des Datensatzes und einer Referenzverteilung, die keine Clusterstrukturen besitzt. Werden die Werte des Kompaktheitsmaßes verglichen, zeigt sich, ob zwischen dem Datensatz und der Referenzverteilung ein signifikanter Abstand besteht. Ist dies nicht der Fall, bedeutet das, dass in den Daten ebenfalls keine Clusterstrukturen enthalten sind und ein Clusteranalyse keine sinnvollen Cluster finden kann. Die Gap Statistik ermöglicht es herauszufinden, ob ein Clustering überhaupt angewandt werden sollte, was mit den meisten internen Indizes nicht möglich ist. Zur Ermittlung der optimalen Clusteranzahl wird ein signifikanter Abstand zwischen der Kompaktheit des Datensatzes und der Referenzverteilung gesucht.

1. Clustering auf dem originalen Datensatz ausführen.
2. Generation und Clustering von H Referenzdatensätzen.
3. Berechnung der Gap Statistik: $Gap(K) = \frac{1}{H}\sum_h \log\left(W_{k,h}^*\right) - \log(W_k)$. W_K ist das Kompaktheitsmaß und berechnet die Summe der Intra-Cluster-Distanzen: $W_K = \sum_{c_k \in C} \frac{1}{2|c_k|} \sum_{x_i, x_j \in c_k} d\left(x_i, x_j\right)$.
4. Berechnung der Standardabweichung: $sd_K = \sqrt{\frac{1}{H}\sum_h \left(\log\left(W_{k,h}^*\right) - \frac{1}{H}\sum_h \log\left(W_{k,h}^*\right)\right)^2}$ und $s_K = sd_K\sqrt{1 + 1/H}$.
5. Wahl der kleinsten Clusteranzahl K, die folgende Bedinung erfüllt: $Gap(K) \geq Gap(K+1) - s_{K+1}$.

Clest (Dudoit & Fridlyand, 2002) teilt den Datensatz in zwei Teile auf, die dazu verwendet einen Klassifizierer zu lernen und die Ergebnisse mit einem Clustering zu vergleichen. Zusätzlich wird das Verfahren mit Referenzverteilungen ausgeführt, die über keine Clusterstrukturen verfügen. Dadurch ist es möglich eine Aussage zu treffen, ob überhaupt eine Clusteranalyse auf den Daten durchgeführt werden sollte. Clest ist folgendermaßen aufgebaut:

1. Ausführung der folgende Schritte für jede mögliche Anzahl an Clustern
    a. Wiederholung für H Iterationen
        i. Teilung des Datensatzes in zwei nicht überlappende Teile: Trainingsdatensatz und Testdatensatz.
        ii. Clustering auf dem Trainingsdatensatz ausführen.
        iii. Lernen eines Klassifizierers auf dem Trainingsdatensatz.
        iv. Klassifikation auf dem Testdatensatz anwenden.
        v. Clustering auf dem Testdatensatz ausführen.
        vi. Berechnung der Übereinstimmung zwischen den Klassen und den Clustern mit einem externen Index $s_{k,h}$.
    b. Berechnung der Ähnlichkeitsstatistik: $t_k = median(s_{k,1}, \dots, s_{k,H})$.
    c. Generation von H_0 Referenzverteilungen unter einer geeigneten Nullhypothese. Wiederholung der Schritte a und b für die Referenzverteilungen.
2. Ermittlung von allen Clusterzahlen, die folgende Bedingungen erfüllen:
    a. Der Anteil der Ähnlichkeitsstatistiken, die mindestens so groß sind wie die Ähnlichkeitsstatistik der originalen Daten, muss unter einem gewissen Schwellwert liegen: $p_k = P(t_{k,h} \geq t_k) \leq p_{max}$.
    b. Zudem muss die Differenz zwischen der Ähnlichkeitsstatistik der originalen Daten und der mittleren Ähnlichkeitsstatistik der Referenzverteilungen über einem bestimmten Schwellwert liegen: $d_k = t_k - \frac{1}{H_0}\sum_{h=1}^{H_0} t_{k,h} \geq d_{min}$.

3. Auswahl der optimalen Clusteranzahl, die beide Bedingungen erfüllt und die maximale Differenz d_k aufweist. Falls keine Clusterzahl die Bedingungen erfüllt, unterscheidet sich der Datensatz nicht von den Referenzverteilungen und demnach liegen Clusterstrukturen in den Daten vor.

Das **Stabilitätsmaß** (Lange et al., 2004) misst die Fähigkeit eines Clusterings das Ergebnis eines weiteren Clusterings zu prognostizieren, das auf anderen Daten derselben Quelle basiert. Dazu wird der Datensatz in einen Trainings- und einen Testdatensatz unterteilt. Auf beiden Datensätzen wird ein Clustering durchgeführt. Der Trainingsdatensatz wird zudem genutzt, um einen Klassifizierer zu erlernen, der den Testdatensatz klassifiziert. Die Übereinstimmung zwischen den Klassen und den Clustern wird mit einem externen Index erfasst. Diese Prozedur wird mehrfach wiederholt und durch die Ergebnisse von zufälligen Klassifizierungen normalisiert. Die Clusterzahl, bei der der Abstand zur zufälligen Verteilung am größten ist, wird als Optimum vorgeschlagen.

Prediction Strength (Tibshirani & Walther, 2005) ist ähnlich zu Clest und dem Stabilitätsmaß. Es teilt die Daten in einen Trainings- und Testdatensatz und vergleicht die Ergebnisse der Klassifikation mit der des Clusterings. Die Güte des Clusterings wird daran gemessen, wie gut der Trainingsdatensatz vorhersagen kann, welche Objektpaare sich im selben Cluster im Testdatensatz befinden. Zur Ermittlung der optimalen Clusteranzahl wird nur der Cluster betrachtet, der am schlechtesten vorhergesagt wurde. Liegt der Anteil der korrekten Prognose über einem Schwellwert (z.B. 0,9), dann liegt ein gutes Clustering vor. Die maximale Clusteranzahl, für die diese Bedingung erfüllt ist, wird als Optimum empfohlen.

Der **Stabilitätsindex** (Pascual et al., 2010) definiert die Clusterstabilität als ein Informationsproblem. Der Index verwendet statt dem Zählen der Übereinstimmungen Maße aus der Informationstheorie. Konkret wird die Standardabweichung der Transinformation verwendet, um den Zusammenhang zwischen dem Trainings- und Testdatensatz zu ermitteln. Die Clusterstabilität ist am größten, wenn die Standardabweichung minimal ist und demnach das Kriterium zum Finden der optimalen Clusterzahl.

4.3 Indizes basierend auf Datenmanipulation

Weighted Average Discrepant Pairs (WADP) (Bittner et al., 2000) vergleicht das Clustering auf den originalen Daten mit Clusterings auf manipulierten Daten. Die Daten werden

verändert, indem auf den Wert etwas Rauschen hinzuaddiert wird. Die Methode misst also, wie Reproduzierbarkeit von Clustern unter der Bedingung von zusätzlichem Rauschen. Der Vergleich der Clusterings erfolgt durch das Zählen von Objektpaaren, die sich durch die Datenmanipulation nicht mehr im selben Cluster befinden. Diese Kennzahl wird durch die gesamte Anzahl aller Objektpaare gewichtet. Das Verfahren schematisch dargestellt:

1. Clustering auf dem originalen Datensatz ausführen. Anzahl der Objektpaare M_k in jedem Cluster c_k zählen.
2. Manipulation der Daten indem auf jeden Wert ein zufälliger Störterm der $N(0, \sigma)$-Verteilung addiert wird. σ muss gewählt werden (z.B. Median der Varianzen der Variable).
3. Clustering auf den manipulierten Daten ausführen.
4. Für alle Elemente aus M_k zählen, wie viele Paare nicht mehr zusammen sind.
5. Berechnung der gesamten Abweichung: $DR = \frac{\sum_{k=1}^{K} D_k}{\sum_{k=1}^{K} M_k}$.

Diese Schritte werden für viele Iterationen wiederholt und die jeweiligen Abweichungen werden gemittelt: $WADP_k$. Das Verfahren wird für alle möglichen Clusteranzahlen durchgeführt und die Clusterzahl ausgewählt, für die $WADP_k$ minimal ist. Es werden also die Clusterings ausgewählt, bei denen die wenigstens Objektpaare durch die Datenmanipulation getrennt werden.

Bootstrap Clustering (Kerr & Churchill, 2001) verwendet das gleich Vorgehen wie WADP. Ein wichtiger Unterschied besteht in der Wahl der Störterme. Diese Methode verwendet ein ANOVA-Modell, um die manipulierten Datensätze zu erstellen. Dazu werden die Residuen des angepassten Modells verwendet und als Schätzer der Fehlerverteilungen auf die Daten addiert. Zur Messung der Übereinstimmung zwischen dem originalen Clustering und den manipulierten Clusterings kommen auch andere Maße zum Einsatz. Zum einen wurde vorgeschlagen, alle Objekte zu zählen, die zu 95% demselben Cluster zugeordnet werden. Eine andere Möglichkeit zur Ermittlung der Übereinstimmung besteht in dem Messen der Häufigkeit, dass sich Objektpaare im selben Cluster befinden. Die letzte Methode ähnelt der des WADP, wobei hier die Clusteranzahl gesucht wird, die die Kennzahl maximiert.

Figure of Merit (Yeung et al., 2001) vergleicht das Clustering auf den originalen Daten mit Clusterings, bei denen eine Variable weggelassen wurde. Diese Methode funktioniert nur, wenn die abhängigen Variablen korreliert sind wie es bei Microarray-Daten in der Bioinformatik der Fall ist (Handl et al., 2005). Bei diesem Index wird berechnet, wie groß die

Intra-Cluster-Varianz der fehlenden Variablen ist. Figure of Merit misst also die Kompaktheit einer Variablen in einem Cluster, der ohne diese Variable erstellt wurde:

$$FOM(K) = \sum_{j=1}^{F} \sqrt{\frac{1}{N} \sum_{k=1}^{K} \sum_{x_i^j \in C_k^j} d_e(x_i^j, \bar{c}_k^j)}$$

C^j steht dabei für ein Clustering ohne die $j$te Variable. Die Clusterzahl, für die FOM minimal ist, hat die geringste Varianz in den fehlenden Variablen, ist demnach am stabilsten und optimal.

Ein Nachteil der Figure of Merit ist, dass die Funktion mit der Anzahl der Cluster sinkt. Dies liegt daran, dass bei mehr Clustern die Intra-Cluster-Distanzen kleiner werden. Um diesen Effekt zu verhindern haben Brun et al. (2007) den Korrekturfaktor $\sqrt{(N-K)/N}$ hinzugefügt. Die **angepasste Figure of Merit** ist definiert als:

$$FOM_{adj}(K) = \frac{1}{\sqrt{(N-K)/N}} FOM(K).$$

Basierend auf dem Weglassen von Variablen wurden weitere Indizes von Datta & Datta (2003) vorgeschlagen. **Average Proportion of Non-Overlap** misst den durchschnittlichen Anteil der Objekte, die sich nicht mehr im selben Cluster befinden, wenn eine Variable weggelassen wird.

$$APN(C) = \frac{1}{FN} \sum_{i=1}^{N} \sum_{j=1}^{F} \left(1 - \frac{n\left(C^{i,j} \cap C^{i,0}\right)}{n(C^{i,0})}\right)$$

$C^{i,0}$ repräsentiert den Cluster mit Objekt i auf den originalen Daten, während $C^{i,j}$ den Cluster mit Objekt i darstellt, der ohne Variable j erstellt wurde. Der Wertebereich von APN liegt zwischen 0 und 1, wobei kleinere Werte stabilere Clusterings bedeuten.

Average Distance misst die mittlere Distanz zwischen Objekten des originalen Clusterings und der angepassten Clusterings, die sich im selben Cluster befinden.

$$AD(C) = \frac{1}{FN} \sum_{i=1}^{N} \sum_{j=1}^{F} \frac{1}{n(C^{i,0}) n(C^{i,j})} \left[\sum_{x_i \in C^{i,0}, x_j \in C^{i,j}} d_e(x_i, x_j) \right]$$

Die Clusterzahl, die AD minimiert, besitzt die geringste mittlere Distanz und ist demnach am stabilsten.

Average Distance Between Means erfasst die mittlere Distanz zwischen den Clusterzentren des originalen Clusterings und der angepasssten Clustering.

$$ADM(C) = \frac{1}{FN}\sum_{i=1}^{N}\sum_{j=1}^{F} d_e(\bar{c}^{i,j}, \bar{c}^{i,0})$$

Auch dieser Index muss minimiert werden, um die optimale / stabilste Clusterzahl zu finden.

5 Externe Clustervalidierungsindizes

Die externen Indizes der Clustervalidierung sind auch als Ähnlichkeitsmaße bekannt, da sie die Ähnlichkeit von zwei Matrizen ermitteln (vgl. Tabelle 8). Bei der externen Clustervalidierung stehen die wahren Klassen als Referenzpartitionierung R zur Verfügung, so dass die Übereinstimmung zwischen den wahren Klassen und den gefundenen Clustern C oder den Clusterzugehörigkeiten Q berechnet werden kann. Zur Ermittlung der Übereinstimmung gibt zwei Ansätze, denen die meisten externen Indizes zugeordnet werden können: das Zählen von Objektpaaren und Maße aus der Informationstheorie. Beide Ansätze werden erläutert und jeweils die wichtigsten Indizes vorgestellt. Es gibt nur wenige Indizes, die sich nicht einordnen lassen. Dazu gehören der Minkowski Score (Jardine & Sibson, 1971), der die kophenetischen Matrizen von R und Q betrachtet, und Huberts Γ Statistik (Hubert & Arabie, 1985), die in Kapitel 3.1.2 vorgestellt wurde.

5.1 Indizes basierend auf dem Zählen von Paaren

Die externen Indizes, die mit dem Zählen von Paaren die Übereinstimmung zwischen zwei Clusterings messen, nutzen größtenteils die Kontingenztabelle. Die Einträge der Tabelle sind:

- a: Anzahl an Objektpaaren, die zur selben Klasse in R und demselben Cluster in Q gehören.
- b: Anzahl an Objektpaaren, die zur selben Klasse in R und verschiedenen Clustern in Q gehören.
- c: Anzahl an Objektpaaren, die zu verschiedenen Klassen in R und demselben Cluster in Q gehören.
- d: Anzahl an Objektpaaren, die zu verschiedenen Klassen in R und verschiedenen Clustern in Q gehören.
- M: Anzahl an Objektpaaren im Datensatz: $M = \frac{N(N-1)}{2} = a + b + c + d$.

Der **Rand Index** (Rand, 1971) misst den Anteil der übereinstimmenden Objektepaare a und d an allen Objektpaaren:

$$RI(R, Q) = \frac{a + d}{a + b + c + d}$$

Ein größerer Wert des Rand Index spricht für eine größere Überstimmung der beiden Partitionierungen.

Am Rand Index gibt es Kritik, da der Wert für eine zufällige Partitionierung nicht 0 ist, daher wurde der **Adjusted Rand Index** (Morey & Agresti, 1984; Hubert & Arabie, 1985) eingeführt, der den Index normalisiert:

$$ARI(R,C) = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{M}}$$

Der angepasste Index ist 0, wenn die Partitionierung zufällig sind, und nimmt den Wert 1 an, wenn die Partitionierung identisch sind.

Ein weiterer Kritikpunkt am Rand Index ist die gleiche Gewichtung von den Übereinstimmungen a und d. Der Grund dafür ist, dass d den Index dominiert, wenn die Anzahl an Klassen groß wird. Der **Jaccard Koeffizient** (Jaccard, 1901) löst das Problem, indem er d nicht in die Berechnung aufnimmt. Der Index berechnet den Anteil der übereinstimmenden Objektpaare an allen Objektpaaren, die derselben Klasse oder demselben Cluster angehören:

$$JC(R,Q) = \frac{a}{a+b+c}$$

Für den Index gilt weiterhin, dass hohe Werte nahe 1 für eine große Übereinstimmung zwischen zwei Partitionierungen sprechen.

Einer weiterer, häufig verwendeter externer Index ist der **Fowlkes-Mallows Index** (Fowlkes & Mallows, 1983), der das geometrische Mittel nutzt und auch d nicht verwendet. Der Index misst den Anteil der übereinstimmenden Objektpaare am geometrischen Mittel der Objektpaare, die sich in derselben Klasse befinden, und den Objektpaaren, die zu denselben Cluster zugeordnet wurden:

$$FM(R,Q) = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Diesen Index gilt es ebenfalls im Wertebereich 0 bis 1 zu maximieren.

Ein Index, der die bekannte Idee von Precision und Recall verwendet, ist das **F-Maß** (van Rijsbergen, 1979). Es verwendet das harmonische Mittel zur Kombination von Precision und Recall, um die Übereinstimmung zwischen einer Klasse und dem dazugehörigen Cluster zu ermitteln. Der Mittelwert über alle Klassen ergibt das F-Maß:

$$F(R,C) = \sum_{r \in R} \frac{n_c}{N} * \max_{c \in C} \{F_\beta(r,c)\}$$

$$F_\beta(r,c) = \frac{(\beta^2 + 1) * P(r,c) * R(r,c)}{\beta^2 * P(r,c) + R(r,c)}$$

Precision ist der Anteil der Objekte, die sich in Klasse r und Cluster c befinden, durch die Anzahl aller Objekte in Klasse r: $P(r,c) = \frac{n_{rc}}{n_c}$. Recall ist der Anteil der Objekte, die sich in Klasse r und Cluster c befinden, durch die Anzahl aller Objekte in Cluster c: $R(r,c) = \frac{n_{rc}}{n_r}$. Wie bei den vorigen Indizes gilt auch hier, dass ein hohes F-Maß für eine große Übereinstimmung steht.

5.2 Indizes basierend auf der Informationstheorie

Ein neuerer Ansatz zur Bestimmung der Gleichheit von Partitionen sind die externen Indizes, die auf Maßen der Informationstheorie beruhen. Das zentrale Maß der Informationstheorie ist die **Entropie**, die den Informationsgehalt von Daten angibt. In Bezug auf die Clusteranalyse beschreibt die Entropie die Information, die mit der Unsicherheit übermittelt wird, dass ein zufälliges Objekt zu einem bestimmten Cluster gehört (Pfitzner et al., 2008). Die Entropie berechnet sich aus dem aufsummierten Produkt der jeweiligen Clusterwahrscheinlichkeit multipliziert mit der logarithmierten Clusterwahrscheinlichkeit. Die Clusterwahrscheinlichkeit berechnet sich aus dem Verhältnis der Clustergröße und der Größe des Datensatzes: $P(c_k) = \frac{|c_k|}{N}$.

$$H(C) = -\sum_{c_k \in C} P(c_k) \log P(c_k)$$

Die Entropie ist stets positiv. Sie nimmt den Wert 0 nur an, wenn keine Unsicherheit existiert, d.h. es gibt nur einen Cluster. Die Unsicherheit und damit die Entropie werden hauptsächlich von den Clustergrößen beeinflusst.

Ein weiteres Maß ist die **Verbundentropie**, die die Entropie von zwei verbundenen Clusterings misst, d.h. den Informationsgehalt beider Clusterings beschreibt. Die Verbundentropie stellt demnach die Vereinigungsmenge der beiden Entropien dar. Die Definition ist analog zu $H(C)$, wobei sich die verwendete Wahrscheinlichkeit aus dem Anteil der Objekte berechnet, die in beiden Partitionierung zur selben Partition gehören: $P(r_k, c_k) = \frac{|c_k \cap r_k|}{N}$.

$$H(R,C) = -\sum_{r_k \in R} \sum_{c_k \in C} P(r_k, c_k) \log P(r_k, c_k)$$

Die **Transinformation** ist das Gegenstück zur Verbundentropie, da es die Schnittmenge der beiden Entropien darstellt. Es ist also ein Maß der Informationen die beide Clusterings gemeinsam haben. Die Transinformation gibt daher an, wie viel Informationen ein Clustering über ein anderes hat.

$$I(R,C) = \sum_{c_k \in C} \sum_{r_k \in R} P(r_k, c_k) \log \frac{P(r_k, c_k)}{P(r_k)P(c_k)} = H(R) + H(C) - H(R,C)$$

Diese Grundkonzepte der Informationstheorie können verwendet werden, um die Übereinstimmung zwischen den wahren Klassen und den Clustern zu berechnen. Ein verbreiteter Ansatz ist die Berechnung der **normalisierten Transinformation**, bei der sich die Methoden zur Normalisierung unterscheiden. Beispielhaft wird hier der Ansatz von Strehl & Ghosh (2002) dargestellt, die das geometrische Mittel verwenden.

$$NMI(R,C) = \frac{I(R,C)}{\sqrt{H(R)H(C)}}$$

Der Index wird maximal, wenn ähnliche Partitionierungen verglichen werden, da die Schnittmenge der Entropien den einzelnen Entropien entspricht.

Die **Variation of Information** (Meila, 2003) misst den Informationsgehalt, der verloren geht oder gewonnen wird, wenn die Referenzpartitionierung gegen ein Clustering ausgetauscht wird. Berechnet wird der Index, indem der Informationsgehalt der Clusterings addiert wird, über die das jeweilige andere Clustering nicht verfügt.

$$VI(R,C) = H(R) + H(C) - 2 * I(R,C) = \big(H(R) - I(R,C)\big) + \big(H(C) - I(R,C)\big)$$

Wenn zwei Clusterings identisch sind und über dieselben Informationen verfügen, dann wird Variation of Information minimal. Folglich deuten kleine Werte des Index auf ähnliche Partitionierungen hin.

Für eine grafische Übersicht über die Maße der Informationstheorie vergleiche Abbildung 9.

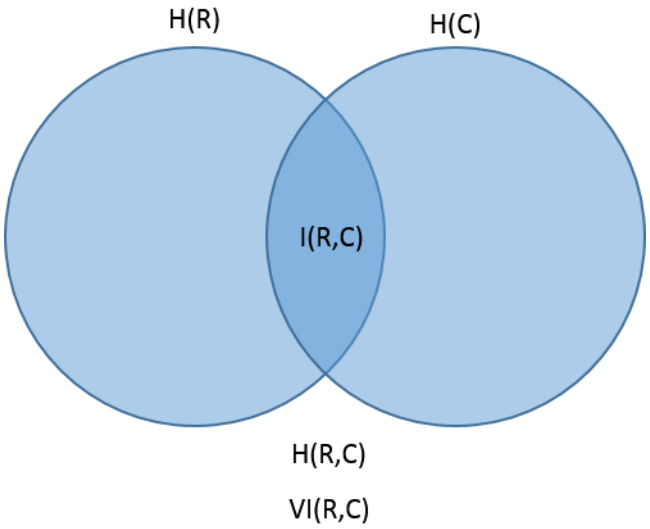

Abbildung 9: Darstellung der Maße der Informationstheorie.

H(R) und H(C) sind die Entropien der Partitionierungen (Kreise), H(R,C) ist die Verbundentropie (blauer Bereich), I(R,C) ist die Transinformation (dunkelblauer Bereich), VI(R,C) ist Variation of Information (hellblauer Bereich).

6 Prototypen der Clustervalidierungsindizes

Bei der Kategorisierung der Clustervalidierungsindizes hat sich gezeigt, dass viele Indizes die gleichen Konzepte, wie Kompaktheit, Separierung und Dichte, verwenden. Teilweise liegt der Unterschied nur in einer anderen Berechnung der Distanzen. Es ist daher möglich, dass bestimmte Indizes die Clusterings identisch beurteilen. Damit ist gemeint, dass die Indizes eine beliebige Auswahl von Clusterings in derselben Reihenfolge bewerten. Die Werte der Indizes müssen dabei nicht zwingend identisch sein. Solche Indizes können zu Prototypen zusammengefasst werden.

Exemplarisch werden die internen Clustervalidierungsindizes für harte Clusterings untersucht, da sie die größte Gruppe von Indizes darstellen. Es werden die Indizes ausgewählt, die die Kompaktheit und die Separierung der Clusterings ermitteln. Diese Konzepte sind oft verwendet worden und die Maße der Distanzen sind systematisch erfasst (vgl. Kapitel 2.2). Verschiedene Kombinationen der Konzepte sind möglich, wobei die Kompaktheit, der Durchmesser der Cluster, minimiert werden muss, während die Separierung, der Abstand zwischen Clustern, maximiert werden muss.

Um die Bewertung der Indizes nachzuvollziehen, werden deren Werte für alle Kombinationen von Kompaktheit und Separierung zwischen 0 und 100 berechnet. Daraus entsteht ein 3D-Plot mit der Kompaktheit als x-Achse, der Separierung als y-Achse und den Werten des Index auf der z-Achse (vgl. Abbildung 10). Auf dieser Grafik ist zu erkennen, für welche Werte der Kompaktheit und Separierung der Index das Optimum annimmt und vor allem welche Kombinationen der Konzepte identisch bewertet werden. Dazu werden die Konturlinien ermittelt, um das Ranking der Indizes nachzuvollziehen. Alle Punkte (Kombinationen von Kompaktheit und Separierung) auf einer Konturlinie werden von dem jeweiligen Index identisch bewertet. Die exakten Werte des Index sind irrelevant, da nur zählt, ob die Werte größer oder kleiner sind. Wichtig ist der Verlauf der Konturlinien, der angibt, wie Clusterings mit verschiedenen Kompaktheits- und Separierungswerten bewertet werden. Durch eine Gegenüberstellung der Konturlinien können die Bewertungen von Indizes verglichen werden. Ein identischer Verlauf der Konturlinien zeigt, dass zwei Indizes alle Clusterings identisch bewerten. Die konkreten Berechnung der Indizes und damit die Werte der Konturlinien können sich aber unterscheiden.

Im Folgenden werden die internen Indizes für harte Clusterings analysiert, wobei eine Trennung nach der Anzahl der Konzepte vorgenommen wurde. Zuerst werden die Indizes verglichen, die Kompaktheit und Separierung der Clusterings erfassen. Anschließend folgen

die Indizes, die nur eines der beiden Konzepte berücksichtigen und daher andere Konturlinien besitzen.

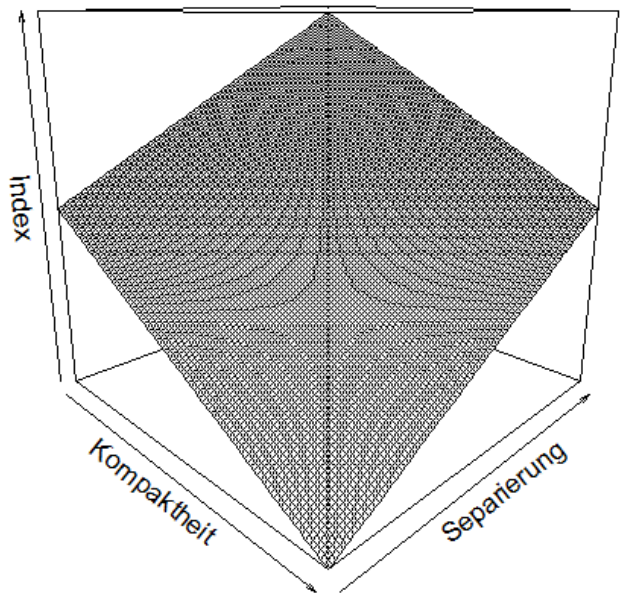

Abbildung 10: Beispielhafte Darstellung der Bewertung von Clusterings durch einen Index.

6.1 Analyse der kombinierenden Indizes

In diesem Abschnitt werden alle Indizes analysiert, die Kompaktheit und Separierung kombinieren und für das gesamte Clustering berechnen. Zur besseren Vergleichbarkeit werden auch nur Indizes betrachtet, die die Kompaktheit mit dem Durchmesser der Cluster bewerten und die Separierung mit dem Abstand zwischen Clustern erfassen. Insgesamt erfüllen acht Indizes diese Bedingungen: Calinski-Harabasz Index, CS Index, Dunn Index, Hartigan Index, PBM Index, Score Function, SD Index und SV Index (vgl. Tabelle 1 und Tabelle 3).

Mögliche Prototypen für die Kombination von Kompaktheit und Separierung sind der Quotient und die Differenz der beiden Konzepte (vgl. Abbildung 11). Wird der Quotient als Index verwendet, rotieren die Konturlinien um den Nullpunkt. Interessant ist dabei, dass Abweichungen von dem Optimum unterschiedlich stark bestraft werden. Abweichungen von der maximalen Separierung werden mit dem Faktor x bestraft, während für Abweichungen von der minimalen Kompaktheit der Faktor $\frac{1}{x}$ angewandt wird. Dies wirkt sich darin aus, dass eine geringere Kompaktheit der Cluster zu einem niedrigeren Wert der Indizes führt als wenn die Separierung der Cluster im gleichen Verhältnis sinkt. Abweichungen von der optimalen

Kompaktheit werden also stärker bestraft als Abweichungen in der Separierung. Bei der Verwendung der Differenz im Index entstehen parallele Konturlinien, die dieselbe Steigung wie die Diagonale haben. Kompaktheit und Separierung werden demnach gleich gewichtet.

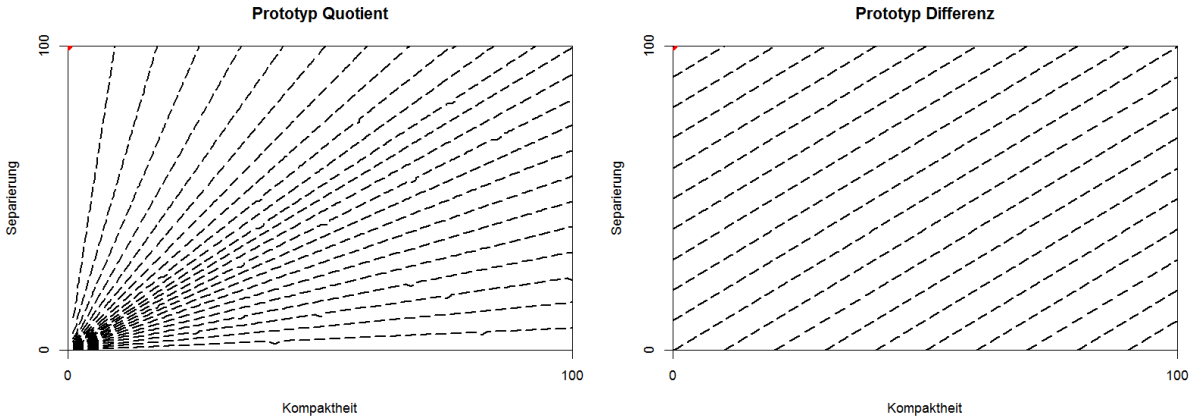

Abbildung 11: Konturlinien der Prototypen „Quotient“ und „Differenz“.

Bei der Analyse zeigt sich, dass alle Indizes das Optimum bei minimaler Kompaktheit und maximaler Separierung finden. Fünf Indizes, die den Quotienten aus Separierung und Kompaktheit berechnen, bilden dieselben Konturlinien wie der Prototyp (vgl. Abbildung 13-17). Auch der invertierte Quotient des CS Index erzeugt die gleichen Konturlinien (vgl. Abbildung 14). Die Gewichtungsfaktoren, wie sie bei dem CH Index verwendet werden, und das Logarithmieren bei dem Hartigan Index hatten ebenfalls keinen Einfluss (vgl. Abbildung 13 und 16). Unterschiede sind hier nur zu erwarten, wenn für das Clustering verschiedene Clusteranzahl oder verschiedene Datensätze verwendet werden. Der PBM Index, der die Kompaktheit des Clustering gewichtet, zeigt einen etwas anderen Verlauf als die Quotienten (vgl. Abbildung 18). Die Konturlinien rotieren ebenfalls um den Nullpunkt, sind aber nicht linear. Je näher die Linie dem Optimum ist, desto runder ist die Kurve. Der SD Index zeigt keinen typischen Verlauf, da die Summe aus der gewichteten Kompaktheit und der invertierten Separierung gebildet wird (vgl. Abbildung 19). Score Function hingegen zeigt den typischen Verlauf, der bei der Differenz von Separierung und Kompaktheit zu erwarten ist (vgl. Abbildung 20). Die Normalisierung mit der Exponentialfunktion hat nur Auswirkungen auf den Wertebereich des Index, nicht aber auf die Konturlinien.

6.2 Analyse der einfachen Indizes

Indizes, die nur ein Konzept, Kompaktheit oder Separierung, berücksichtigen, können nicht optimiert werden, weil beide Konzepte von der Clusteranzahl abhängig sind. Die Kompaktheitsmaße fallen monoton mit der Clusteranzahl, da die Durchmesser der Cluster immer kleiner werden. Die Separierung steigt monoton mit der Clusteranzahl an, aufgrund der größeren Abstände zwischen den Clustern. Zur Ermittlung der optimalen Clusterzahl müssen die Indizes ein „Knie“ oder „Sprung“ in den Werten identifizieren. Die Prototypen für diese Kategorie von Indizes spiegeln das wider (vgl. Abbildung 12). Die Konturlinien sind horizontal für Separierung bzw. vertikal für Kompaktheit und zeigen damit, dass das jeweils andere Konzept komplett ignoriert wird. Die Prototypen werden mit drei vergleichenden Indizes verglichen: Ball-Hall Index, RMSSTD Index und RS Index (vgl. Tabelle 4).

Bei der Analyse der Konturlinien wird offensichtlich, dass alle Indizes das Optimum nicht eindeutig identifizieren können (vgl. Abbildung 21-23). Sie können ein Konzept optimieren, aber keine Aussagen über das andere treffen. Alle Indizes weisen die typischen horizontalen und vertikalen Linien auf. Die Gewichtungen und Normalisierungen, die einige Indizes verwenden, verändern nur die Abstände zwischen den Konturlinien (vgl. Abbildung 22). Möglicherweise vereinfacht das die Ermittlung des „Knie“ zur Bestimmung der optimalen Clusteranzahl. Unterschiede sind erst zu erwarten, wenn verschiedene Clusterzahlen und verschiedene Datensätze verglichen werden.

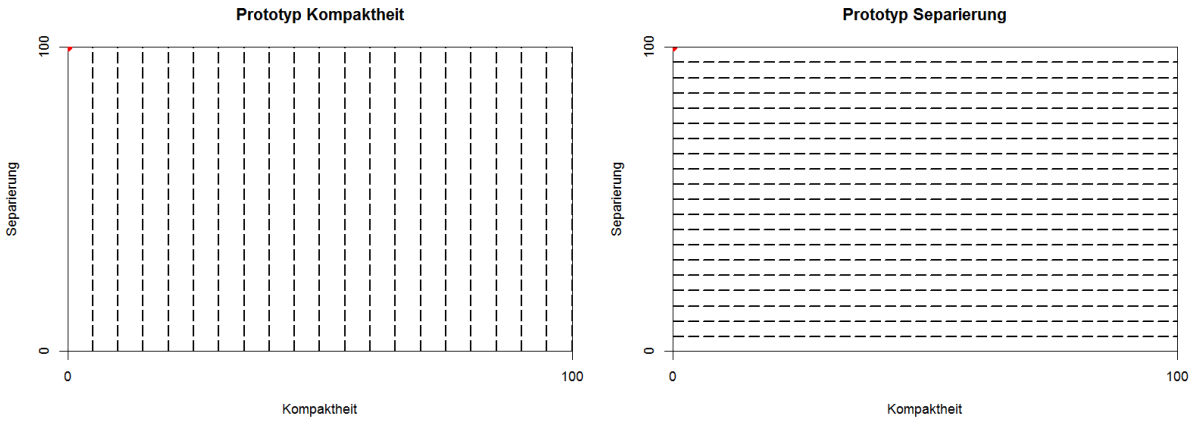

Abbildung 12: Konturlinien der Prototypen „Kompaktheit“ und „Separierung“.

7 Fazit

Das primäre Ziel der Arbeit war es Clustervalidierungsindizes sinnvoll zu kategorisieren. Dies ist mittels vier Kriterien gelungen: den verfügbaren Informationen, dem Ergebnis des Clustering-Algorithmus, der Art der Optimierung und den verwendeten Statistiken. Dadurch konnten die Indizes in elf Gruppen aufgeteilt werden, wobei die Indizes innerhalb der Gruppen jeweils bestimmte Eigenschaften gemeinsam haben. Drei Konzepte, die von Indizes zur Clustervalidierung verwendet werden, konnten identifiziert werden: Kompaktheit, Separierung und Dichte. Indizes, die dieselben Konzepte verwenden, konnten hinsichtlich der Bewertung der Clusterings verglichen werden. Der Vergleich mit vordefinierten Prototypen, die typische Indizes repräsentieren, zeigt, dass viele Indizes die Clusterings ähnlich bewerten oder sogar komplett identisch beurteilen.

Die Literaturrecherche offenbarte, dass es viele Studien gibt, die neue Indizes zur Clustervalidierung entwickeln. Es gibt aber nur wenige Arbeiten, die ausführliche Vergleichstests durchführen und die neusten Erkenntnisse zusammenfassen. Bei der Analyse der vorgeschlagenen Indizes zeigte sich, dass oft dieselben Konzepte und Maße verwendet wurden. Die Gruppierung von gleichartigen Indizes ermöglichte es, eine Struktur in die Clustervalidierung zu bringen und die Kategorien zu benennen. Ähnliche Indizes wurden genauer analysiert, um die Bewertung der Clusterings zu ermitteln und zu vergleichen. Dabei stellte sich heraus, dass nur ein Teil der Indizes in vergleichbare Form gebracht werden kann. Für alle Prototypen konnten trotzdem Indizes gefunden werden, die die Clustering auf die gleiche Weise beurteilen.

In zukünftigen Studien ist es möglich, die Kategorisierung der Clustervalidierungsindizes zu erweitern. Es könnten Indizes von weiteren Clustering-Algorithmen, wie dem modellbasierten Clustering, eingegliedert werden. Dazu müssten dann neue Kategorien und Konzepte definiert werden. Das bestehende Konzept der Dichte kann weiter verfeinert werden, indem versucht wird die typischen Maße zu erfassen. Die Analyse der Bewertungen von Indizes und die Erstellung von Prototypen kann auf weitere Gebiete der Clustervalidierung ausgedehnt werden. Da die Clustervalidierung weiterhin ein aktuelles Thema ist, werden stetig neue Indizes vorgeschlagen, die in diese Systematik einsortiert werden können.

Anhang

Index	Opt.	Referenz
$CH(C) = \frac{N-K}{K-1}\frac{Sep(C)}{Comp(C)}$	Max	(Calinski & Harabasz, 1974)
$Dunn(C) = \frac{Sep(C)}{Comp(C)}$	Max	(Dunn, 1974)
$CI(C) = \frac{Comp(C)-Comp_{min}(C)}{Comp_{max}(C)-Comp_{min}(C)}$	Min	(Hubert & Levin, 1976)
$RL(C) = \frac{1}{\sqrt{K}} * \frac{1}{n}\sum_{f=1}^{F}\sqrt{\frac{Sep_f(C)}{Sep_f(X)}}$	Max	(Ratkowsky & Lance, 1978)
$DB(C) = \frac{1}{K}\sum_{c_k \in C}\max_{c_l \in C \setminus c_k}\left\{\frac{Comp(c_k)+Comp(c_l)}{Sep(c_k,c_l)}\right\}$	Min	(Davies & Bouldin, 1979)
$CJ(C) = \min_{c_k \in C}\left\{\frac{Sep(c_k)}{Comp(c_k)}\right\}$	Max	(Coggins & Jain, 1985)
$Sil(C) = \frac{1}{N}\sum_{c_k \in C}\sum_{x_i \in C}\frac{Sep(x_i,c_k)-Comp(x_i,c_k)}{\max\{Comp(x_i,c_k),Sep(x_i,c_k)\}}$	Max	(Rousseeuw, 1987)
$B_{crit}(C) = \alpha * Comp(C) + Sep(C)$	Min	(Boudraa, 1999)
$SD(C) = \alpha * Comp(C) + Sep(C)$	Min	(Halkidi et al., 2000)
$S_{Dbw}(C) = Comp(C) + Dens(C)$	Min	(Halkidi & Vazirgiannis, 2001)
$PS(C) = \frac{1}{K}\sum_{c_k \in C}\left[\frac{1}{\lvert c_k \rvert}\sum_{x_i \in c_k}\frac{Dens(x_i,c_k)}{Sep(C)}\right]$	Min	(Chou et al., 2002)
$CS(C) = \frac{Comp(C)}{Sep(C)}$	Min	(Chou et al., 2004)
$Cons(C) = Dens(C)$	Max	(Ding & He, 2004)
$PBM(C) = \left(\frac{1}{K}\frac{Sep(C)}{Comp(C)}\right)^2$	Max	(Pakhira et al., 2004)
$Conn(C) = Dens(C)$	Min	(Handl & Knowles, 2005)
$SF(C) = 1 - \frac{1}{e^{e^{Sep(C)-Comp(C)}}}$	Max	(Saitta et al., 2007)
$C_{Dbw}(C) = Coh(C) * Sep_{Dens}(C) * Comp_{Dens}(C)$	Max	(Halkidi & Vazirgiannis, 2008)
$NIVA(C) = \frac{CompC(C)}{Sep(C)}$	Min	(Rendón et al., 2008)
$COP(C) = \frac{1}{N}\sum_{c_k \in C}\lvert c_k \rvert\frac{Comp(c_k)}{Sep(c_k)}$	Min	(Gurrutxaga et al., 2010)
$SV(C) = \frac{Sep(C)}{Comp(C)}$	Max	(Žalik & Žalik, 2011)
$GF(C) = \frac{Sep(C)}{1+Comp(C)}$	Min	(Ghosh & De, 2013)
$HS(C) = \frac{Comp(C)}{Sep(C)}$	Min	(Satapathy et al., 2014)

Tabelle 1: Übersicht über interne, optimierende Clustervalidierungsindizes für hartes Clustering basierend auf dem Clustering C.

Index	Opt.	Referenz
$\tau(P,Q)$: Tau Korrelation	Max	(Rohlf, 1974)
$PB(P,Q)$: Punktbiseriale Korrelation	Max	(Milligan, 1981)
$\Gamma(P,Q)$: Punktseriale Korrelation	Max	(Hubert & Arabie, 1985)
$Coph(P,Q)$: Kophenetische Korrelation	Max	(Halkidi et al., 2001)

Tabelle 2: Übersicht über interne, optimierende Clustervalidierungsindizes für hartes Clustering basierend auf der Distanzmatrix P und der Clusterzugehörigkeitsmatrix Q.

Index	Opt.	Referenz
$BH(C) = SSW$	„Knie“	(Ball & Hall, 1965)
$H(C) = N\ \log_{10}\left(\frac{SSB}{SSW}\right)$	„Knie“	(Hartigan, 1975)
$KL(C) = \frac{\lvert Diff^{(k)}\rvert}{\lvert Diff^{(k+1)}\rvert}$ $Diff^{(k)} = (k-1)^{2/F} * SSW^{(k-1)} - N^{2/F} * SSW^{(k)}$	Max	(Krzanowski & Lai, 1988)
$RMSSTD(C) = \sqrt{\frac{SSW}{F(N-K)}}$	„Knie“	(Sharma, 1996)
$RS(C) = \frac{SSB}{SST}$	„Knie“	(Sharma, 1996)

Tabelle 3: Übersicht über interne, vergleichende Clustervalidierungsindizes für hartes Clustering basierend auf den Quadratsummen.

Index	Opt.	Referenz
$RI(C) = \frac{\lvert T\rvert}{\lvert W\rvert}$	„Knie“	(Friedman & Rubin, 1967)
$TrW(C) = trace(W)$	„Knie“	(Friedman & Rubin, 1967)
$FI(C) = trace(W^{-1}B)$	Sprung	(Friedman & Rubin, 1967)
$MI(C) = K^2\lvert W\rvert$	„Knie“	(Marriott, 1971)
$SS(C) = N\ \log_{10}\left(\frac{\lvert T\rvert}{\lvert W\rvert}\right)$	Sprung	(Scott & Symons, 1971)
$TrCovW(C) = trace\big(cov(W)\big)$	„Knie“	(Milligan & Cooper, 1985)

Tabelle 4: Übersicht über interne, vergleichende Clustervalidierungsindizes für hartes Clustering basierend auf den Streumatrizen.

Index	Opt.	Referenz
$PC(U) = Comp(U)$	Max	(Bezdek, 1974)
$PE(U)$	Min	(Bezdek, 1973)
$WPE(U)$	Max	(Windham, 1981)
$P(U) = Comp(U) - Sep(U)$	Max	(Chen & Linkens, 2001)
$KYI(U) = Sep(U)$	Min	(Kim et al., 2004b)
$OS(U) = \frac{Overlap(U)/Overlap_{max}}{Sep(U)/Sep_{max}}$	Min	(Kim et al., 2004a)

Tabelle 5: Übersicht über interne Clustervalidierungsindizes für weiches Clustering basierend auf der Clusterzugehörigkeitsmatrix U.

Index	Opt.	Referenz
$FS(U,V) = Comp(U,V) - Sep(U,V)$	Min	(Fukuyama & Sugeno, 1989)
$PD(U,V) = \frac{Dens(U)}{Comp(U,V)}$	Max	(Gath & Geva, 1989)
$XB(U,V) = \frac{1}{N}\frac{Comp(U,V)}{Sep(V)}$	Min	(Xie & Beni, 1991)
$CV(U,X) = \frac{Sep(U,X)}{Comp(U,X)}$	Max	(Rhee & Oh, 1996)
$CWB(U,V) = \alpha * Comp(C) + Sep(C)$	Min	(Rezaee et al., 1998)
$SC_{ZLE}(U,V) = \frac{Sep(V)}{Comp(U,V)} - \frac{Sep(U)}{Comp(U)}$	Max	(Zahid et al., 1999)
$Inv(U,V) = \frac{1}{K^2} trace\left(\frac{Sep(U,V)}{Comp(U,V)}\right)$	Max	(Geva et al., 2000)
$SV(V) = Comp_N(V) + \frac{1}{Sep_N(V)}$	Min	(Kim et al., 2001)
$SVI(U,V) = \frac{Comp(U,V)}{Sep(V)}$	Min	(Tsekouras & Sarimveis, 2004)
$GD(U,V) = \frac{(Sep(V))^2}{Comp(U,V,X)}$	Max	(Xie et al., 2002)
$PCAES(U,V) = Comp(U) - Sep(V)$	Max	(Wu & Yang, 2005)
$SC_{BWS}(U,V) = \frac{Sep(U,V)}{Comp(U,V)}$	Max	(Bouguessa et al., 2006)
$W(U,V) = \frac{Comp(U,V)/Comp_{max}}{Sep(U)/Sep_{max}}$	Min	(Zhang et al., 2008)
$SC_R(U,V) = \frac{Sep(U)}{Sep_{max}} + \frac{Comp(U,V)}{Comp_{max}}$	Min	(Rezaee, 2010)

Tabelle 6: Übersicht über interne Clustervalidierungsindizes für weiches Clustering basierend auf der Clusterzugehörigkeitsmatrix U und den Clusterzentren V.

Index	Opt.	Referenz
Indizes basierend auf überlappenden Stichproben		
Figure of Merit	Max	(Levine & Domany, 2001)
Model Explorer	Max	(Ben-Hur et al., 2002)
Consensus	„Knie“	(Monti et al., 2003)
Indizes basierend auf dem Vergleich mit einer Prognose		
Gap Statistik	Max	(Tibshirani et al., 2001)
Clest	Max	(Dudoit & Fridlyand, 2002)
Stabilitätsmaß	Max	(Lange et al., 2004)
Prediction Strength	Max	(Tibshirani & Walther, 2005)
Stabilitätsindex	Min	(Pascual et al., 2010)
Indizes basierend auf Datenmanipulation		
WADP	Min	(Bittner et al., 2000)
Bootstrap Clustering	Max	(Kerr & Churchill, 2001)
Figure of Merit	Min	(Yeung et al., 2001)
Average Proportion of Non-Overlap, Average Distance, Average Distance Between Means	Min	(Datta & Datta, 2003)

Tabelle 7: Übersicht über relative Clustervalidierungsindizes.

Index	Opt.	Referenz
Indizes basierend auf dem Zählen von Paaren		
$RI(R,Q) = \frac{a+d}{a+b+c+d}$	Max	(Rand, 1971)
$JC(R,Q) = \frac{a}{a+b+c}$	Max	(Jaccard, 1901)
$FM(R,Q) = \frac{a}{\sqrt{(a+b)(a+c)}}$	Max	(Fowlkes & Mallows, 1983)
F-Maß	Max	(van Rijsbergen, 1979)
Indizes basierend auf Datenmanipulation		
$NMI(R,C) = \frac{I(R,C)}{\sqrt{H(R)H(C)}}$	Max	(Strehl & Ghosh, 2002)
$VI(R,C) = H(R) + H(C) - 2 * I(R,C)$	Min	(Meila, 2003)

Tabelle 8: Übersicht über externe Validierungsindizes.

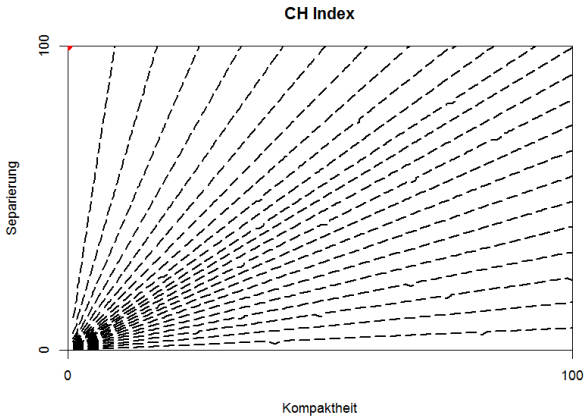

Abbildung 13: Konturlinien des CH Index.

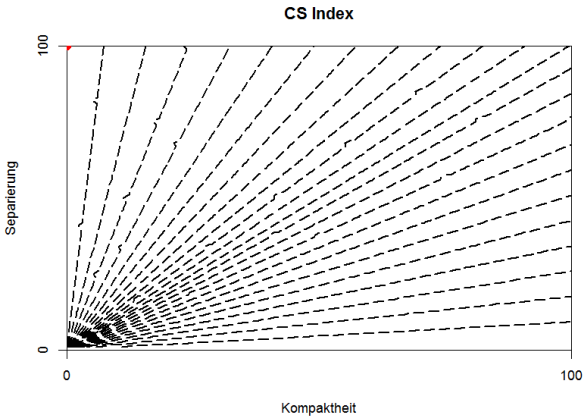

Abbildung 14: Konturlinien des CS Index.

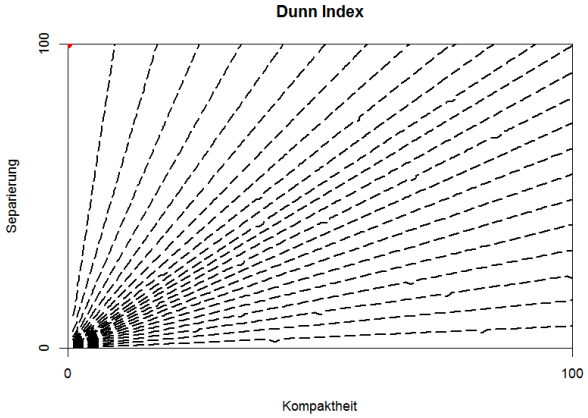

Abbildung 15: Konturlinien des Dunn Index.

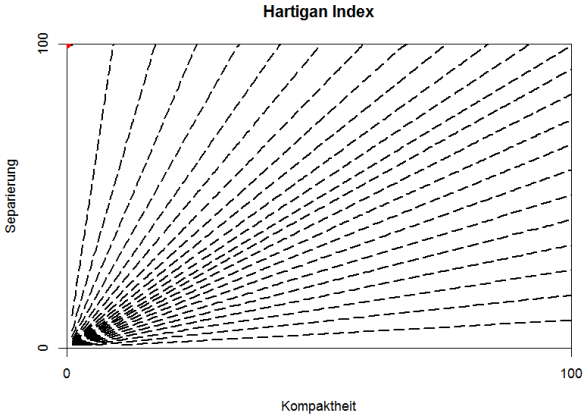

Abbildung 16: Konturlinien des Hartigan Index.

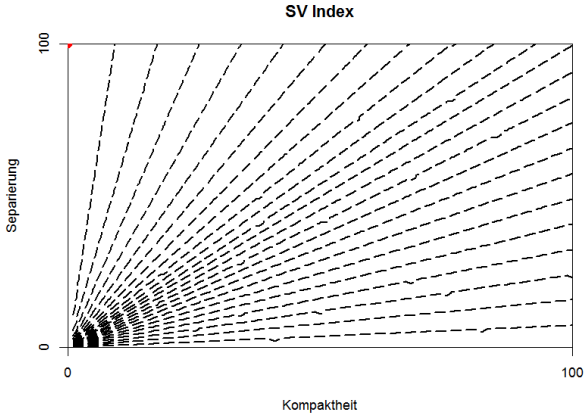

Abbildung 17: Konturlinien des SV Index.

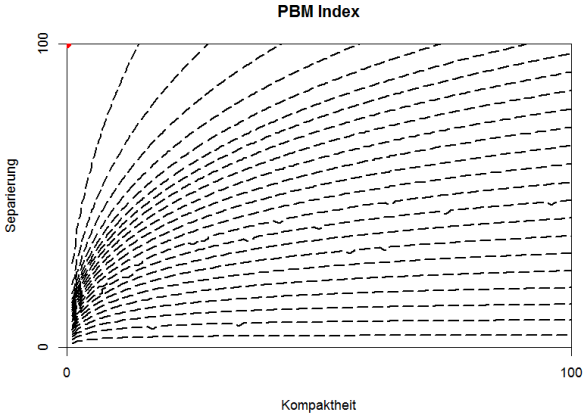

Abbildung 18: Konturlinien des PBM Index.

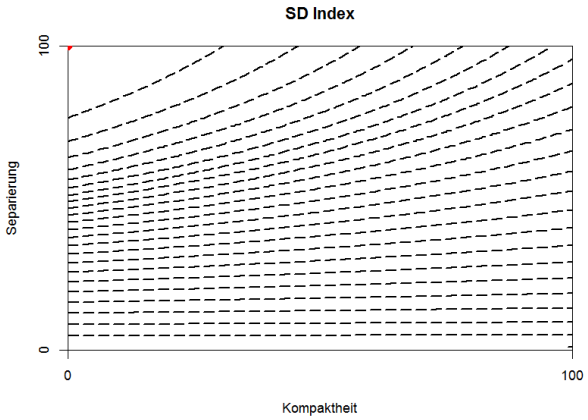

Abbildung 19: Konturlinien des SD Index.

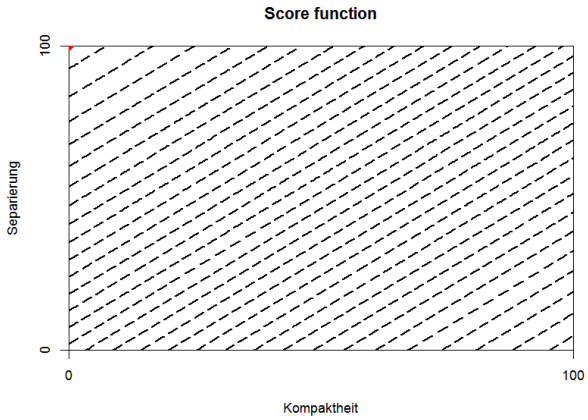

Abbildung 20: Konturlinien des SF Index.

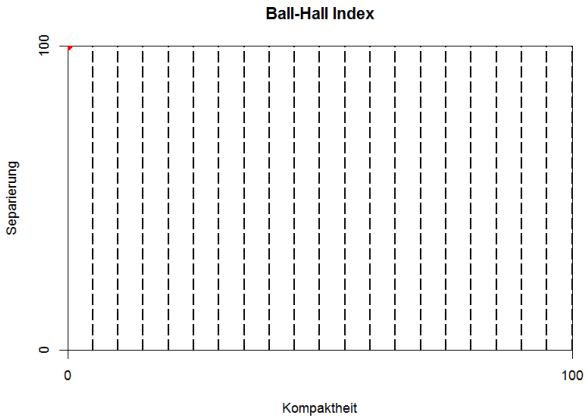

Abbildung 21: Konturlinien des Ball-Hall Index.

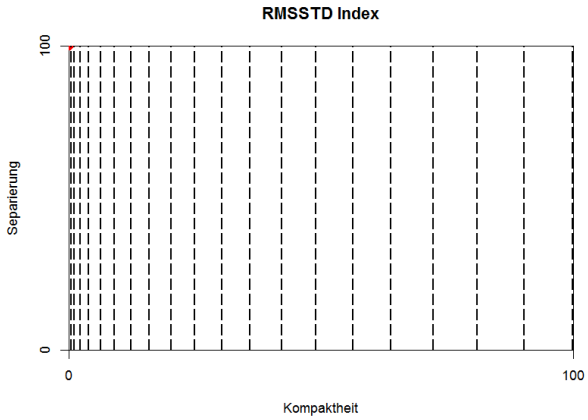

Abbildung 22: Konturlinien des RMSSTD Index.

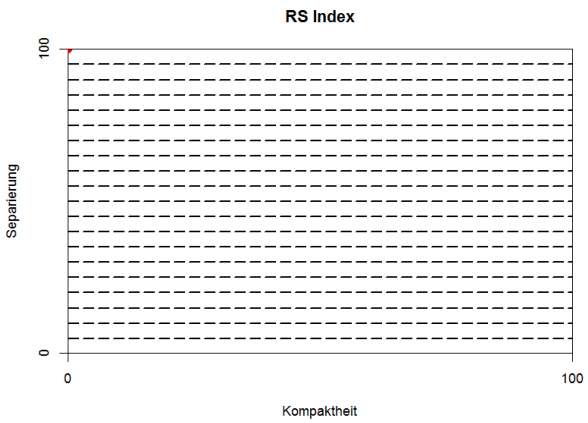

Abbildung 23: Konturlinien des RS Index.

Literaturverzeichnis

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6): 716–723.

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* 46(1): 243–256.

Baker FB, Hubert LJ. 1975. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association* 70(349): 31–38.

Ball GH, Hall DJ. 1965. *ISODATA, a novel method of data analysis and pattern classification*. Stanford Research Institute, Menlo Park, California.

Ben-Hur A, Elisseeff A, Guyon I. 2002. A stability based method for discovering structure in clustered data. In *Prcoceedings of the 2002 Pacific Symposium on Biocomputing*, Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE (eds). World Scientific: Singapore: 6–17.

Bensaid AM, Hall LO, Bezdek JC, Clarke LP, Silbiger ML, et al. 1996. Validity-guided (re)clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems* 4(2): 112–123.

Bezdek JC. 1973. Cluster validity with fuzzy sets. *Journal of Cybernetics* 3(3): 58–73.

Bezdek JC. 1974. Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology* 1(1): 57–71.

Bezdek JC, Pal NR. 1998. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 28(3): 301–315.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, et al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406(6795): 536–540.

Boudraa A-O. 1999. Dynamic estimation of number of clusters in data sets. *Electronics Letters* 35(19): 1606–1608.

Bouguessa M, Wang S, Sun H. 2006. An objective approach to cluster validation. *Pattern Recognition Letters* 27(13): 1419–1430.

Brun M, Sima C, Hua J, Lowey J, Carroll B, et al. 2007. Model-based evaluation of clustering validation measures. *Pattern Recognition* 40(3): 807–824.

Calinski T, Harabasz J. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* 3(1): 1–27.

Chen M-Y, Linkens DA. 2001. Rule-base self-generation and simplification for data-driven fuzzy models. In *10th IEEE International Conference on Fuzzy Systems*. IEEE: 424–427.

Cho S-B, Yoo S-H. 2006. Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data. *Pattern Recognition* 39(12): 2405–2414.

Chong A, Gedeon TG, Koczy LT. 2002. A hybrid approach for solving the cluster validity problem. In *Proceedings of the 2002 14th International Conference on Digital Signal Processing*, Skodras AN, Constantinides AG (eds). IEEE: 1207–1210.

Chou C-H, Su M-C, Lai E. 2002. Symmetry as a new measure for cluster validity. In *Proceedings of the Second WSEAS International Conference on Scientific Computation and Soft Computing*: 209–213.

Chou C-H, Su M-C, Lai E. 2004. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications* 7(2): 205–220.

Coggins JM, Jain AK. 1985. A spatial filtering approach to texture analysis. *Pattern Recognition Letters* 3(3): 195–203.

Datta S, Datta S. 2003. Comparisions and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19(4): 459–466.

Dave RN. 1996. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters* 17(6): 613–623.

Davies DL, Bouldin DW. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2): 224–227.

Dimitriadou E, Dolničar S, Weingessel A. 2002. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika* 67(1): 137–159.

Ding C, He X. 2004. K-nearest-neighbor consistency in data clustering: Incorporating local information into global optimization. In *Proceedings of the 2004 ACM symposium on Applied computing*. ACM: New York, USA: 584–589.

Dudoit S, Fridlyand J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology* 3(7): research0036.1–0036.21.

Dunn JC. 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* 4(1): 95–104.

Fowlkes EB, Mallows CL. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78(383): 553–569.

Friedman HP, Rubin J. 1967. On some invariant criteria for grouping data. *Journal of the American Statistical Association* 62(320): 1159–1178.

Fukuyama Y, Sugeno M. 1989. A new method of choosing the number of clusters for the fuzzy c-means method. In *Proceedings of the 5th Fuzzy Systems Symposium*: 247–250.

Gath I, Geva AB. 1989. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7): 773–780.

Geva AB, Steinberg Y, Bruckmair S, Nahum G. 2000. A comparison of cluster validity criteria for a mixture of normal distributed data. *Pattern Recognition Letters* 21(6-7): 511–529.

Ghosh A, De RK. 2013. Gaussian fuzzy index for cluster validation: Identification of high quality biologically enriched clusters of genes and selection of some possible genes mediating lung cancer. In *Proceedings of the 5th International Conference on Pattern Recognition and Machine Learning*, Maji P, Ghosh A, Murty MN, Ghosh K, Pal SK (eds). Springer: Berlin: 680–687.

Giancarlo R, Scaturro D, Utro F. 2008. Computational cluster validation for microarray data analysis: Experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics and Model Explorer. *BMC Bioinformatics* 9(462): 1–19.

Gurrutxaga I, Albisua I, Arbelaitz O, Martín JI, Muguerza J, et al. 2010. SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition* 43(10): 3364–3373.

Gurrutxaga I, Muguerza J, Arbelaitz O, Pérez JM, Martín JI. 2011. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters* 32(3): 505–515.

Halkidi M, Batistakis Y, Vazirgiannis M. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems* 17(2-3): 107–145.

Halkidi M, Vazirgiannis M. 2001. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, Cercone N, Lin TY, Xindong W (eds). IEEE: Los Alamitos, California, USA: 187–194.

Halkidi M, Vazirgiannis M. 2008. A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters* 29(6): 773–786.

Halkidi M, Vazirgiannis M, Batistakis Y. 2000. Quality scheme assessment in the clustering process. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, Zighed DA, Komorowski J, Zytkow J (eds). Springer: Berlin: 265–276.

Handl J, Knowles J. 2005. Exploiting the trade-off - The benefits of multiple objectives in data clustering. In *Prodceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization*, Coello Coello CA, Hernández Aguirre A, Zitzler E (eds). Springer: Berlin: 547–560.

Handl J, Knowles J, Kell DB. 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15): 3201–3212.

Hartigan JA. 1975. *Clustering algorithms*. Wiley: New York.

Hassar H, Bensaid AM. 1999. Validation of fuzzy and crisp c-partitions. In *18th International Conference of the North American Fuzzy Information Processing Society*, Davé RN, Sudkamp T (eds). IEEE: 342–346.

Hruschka ER, Campello RJGB, de Castro LN. 2006. Evolving clusters in gene-expression data. *Information Sciences* 176(13): 1898–1927.

Hubert LJ, Arabie P. 1985. Comparing partitions. *Journal of Classification* 2(1): 193–218.

Hubert LJ, Levin JR. 1976. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* 83(6): 1072–1080.

Jaccard P. 1901. Distribution de la florine alpine dans la bassin de dranses et dans quelques regiones voisines. *Naturelles Bulletin de la Societe Vaudoise des Science* : 241–272.

Jardine N, Sibson R. 1971. *Mathematical taxonomy*. Wiley: London.

Kerr MK, Churchill GA. 2001. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. In *Proceedings of the National Academy of Sciences of the United States of America*, 98: 8961–8965.

Ketchen Jr. DJ, Shook CL. 1996. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal* 17(6): 441–458.

Kim D-J, Park Y-W, Park D-J. 2001. A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and Systems* 84(2): 281–285.

Kim D-W, Lee KH, Lee D. 2004a. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition* 37(10): 2009–2025.

Kim M, Ramakrishna RS. 2005. New indices for cluster validity assessment. *Pattern Recognition Letters* 26(15): 2353–2363.

Kim Y-I, Kim D-W, Lee D, Lee KH. 2004b. A cluster validation index for GK cluster analysis based on relative degree of sharing. *Information Sciences* 168(1-4): 225–242.

Krzanowski WJ, Lai YT. 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44(1): 23–34.

Kwon SH. 1998. Cluster validity index for fuzzy clustering. *Electronics Letters* 34(22): 2176–2177.

Lago-Fernández LF, Corbacho F. 2010. Normality-based validation for crisp clustering. *Pattern Recognition* 43(3): 782–795.

Lange T, Roth V, Braun ML, Buhmann JM. 2004. Stability-based validation of clustering solutions. *Neural Computation* 16(6): 1299–1323.

Levine E, Domany E. 2001. Resampling method for unsupervised estimation of cluster validity. *Neural computation* 13(11): 2573–2593.

Liu Y, Li Z, Xiong H, Gao X, Wu J. 2010. Understanding of internal clustering validation measures. In *Proceedings of the 10th IEEE International Conference on Data Mining*, Webb GI, Liu B, Zhang C, Gunopulos D, Wu X (eds). IEEE: Los Alamitos, California, USA: 911–916.

Marriott FHC. 1971. Practical problems in a method of cluster analysis. *Biometrics* 27(3): 501–514.

Maulik U, Bandyopadhyay S. 2002. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(12): 1650–1654.

McClain JO, Rao VR. 1975. CLUSTISZ: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research* 12(4): 456–460.

Meila M. 2003. Comparing clusterings by the variation of information. In *Proceedings of the 16th Annual Conference on Learning Theory*, Schölkopf B, Warmuth MK (eds). Springer: Berlin: 173–187.

Milligan GW. 1981. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* 46(2): 187–199.

Milligan GW, Cooper MC. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2): 159–179.

Monti S, Tamayo P, Mesirov J, Golub T. 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 52(1-2): 91–118.

Morey LC, Agresti A. 1984. The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement* 44(1): 33–37.

Pakhira MK, Bandyopadhyay S, Maulik U. 2004. Validity index for crisp and fuzzy clusters. *Pattern Recognition* 37(3): 487–501.

Pal NR, Biswas J. 1997. Cluster validation using graph theoretic concepts. *Pattern Recognition* 30(6): 847–857.

Pascual D, Pla F, Sánchez JS. 2010. Cluster validation using information stability measures. *Pattern Recognition Letters* 31(6): 454–461.

Pfitzner D, Leibbrandt R, Powers D. 2008. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems* 19(3): 361–394.

Rand WM. 1971. Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association* 66(336): 846–850.

Ratkowsky DA, Lance GN. 1978. A criterion for determining the number of groups in a classification. *Australian Computer Journal* 10: 115–117.

Rendón E, Garcia R, Abundez I, Gutierrez C, Gasca E, et al. 2008. NIVA: A robust cluster validity. In *Proceedings of the 12th WSEAS International Conference on Communications*. WSEAS: 209–213.

Rezaee B. 2010. A cluster validity index for fuzzy clustering. *Fuzzy Sets and Systems* 161(23): 3014–3025.

Rezaee MR, Lelieveldt BPF, Reiber JHC. 1998. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters* 19(3-4): 237–246.

Rhee H-S, Oh K-W. 1996. A validity measure for fuzzy clustering and its use in selecting optimal number of clusters. In *Proceedings of the Fifth IEEE International Conference on Fuzzy Systems*. IEEE: 1020–1025.

Van Rijsbergen CJ. 1979. *Information Retrieval*. Butterworths: London.

Rohlf FJ. 1974. Methods of comparing classifications. *Annual Review of Ecology and Systematics* 5(1): 101–113.

Roubens M. 1982. Fuzzy clustering algorithms and their cluster validity. *European Journal of Operational Research* 10(3): 294–301.

Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(1): 53–65.

Saha S, Bandyopadhyay S. 2012. Some connectivity based cluster validity indices. *Applied Soft Computing* 12(5): 1555–1565.

Saitta S, Raphael B, Smith IFC. 2007. A bounded index for cluster validity. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, Perner P (ed). Springer: Berlin: 174–187.

Satapathy SC, Avadhani PS, Udgata SK, Lakshminarayana S. 2014. Homogeneity separateness: A new validity measure for clustering problems. In *Proceedings of the 48th Annual Convention of Computer Society of India*, Murty MR, Murthy JVR, Reddy PVGDP, Naik A, Satapathy SC (eds). Springer: Switzerland: 1–10.

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2): 461–464.

Scott AJ, Symons MJ. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics* 27: 387–397.

Sharma S. 1996. *Applied multivariate techniques*. Wiley: New York, USA.

Sledge IJ, Bezdek JC, Havens TC, Keller JM. 2010. Relational generalizations of cluster validity indices. *IEEE Transactions on Fuzzy Systems* 18(4): 771–786.

Starczewski A. 2012. A cluster validity index for hard clustering. In *Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing*, Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh LA, et al. (eds). Springer: Berlin: 168–174.

Strehl A, Ghosh J. 2002. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3(1): 583–617.

Sun H, Wang S, Jiang Q. 2004. FCM-based model selection algorithms for determining the number of clusters. *Pattern Recognition* 37(10): 2027–2037.

Tang Y, Sun F, Sun Z, Member S. 2005. Improved validation index for fuzzy clustering. In *Proceedings of the 2005 American Control Conference*. IEEE: Los Alamitos, California, USA: 1120–1125.

Tibshirani R, Walther G. 2005. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics* 14(3): 511–528.

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2): 411–423.

Tsekouras GE, Sarimveis H. 2004. A new approach for measuring the validity of the fuzzy c-means algorithm. *Advances in Engineering Software* 35(8-9): 567–575.

Vendramin L, Campello RJGB, Hruschka ER. 2010. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining* 3(4): 209–235.

Wang W, Zhang Y. 2007. On fuzzy cluster validity indices. *Fuzzy Sets and Systems* 158(19): 2095–2117.

Windham MP. 1981. Cluster validity for fuzzy clustering algorithms. *Fuzzy Sets and Systems* 5(2): 177–185.

Wu K-L, Yang M-S. 2005. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters* 26(9): 1275–1291.

Wu K-L, Yang M-S, Hsieh J-N. 2009. Robust cluster validity indexes. *Pattern Recognition* 42(11): 2541–2550.

Xie XL, Beni G. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(8): 841–847.

Xie Y, Raghavan V V., Dhatric P, Zhao X. 2005. A new fuzzy clustering algorithm for optimally finding granular prototypes. *International Journal of Approximate Reasoning* 40(1-2): 109–124.

Xie Y, Raghavan V V., Zhao X. 2002. 3M algorithm: Finding an optimal fuzzy cluster scheme for proximity data. In *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems.* IEEE: 627–632.

Xu Y, Brereton RG. 2005. A comparative study of cluster validation indices applied to genotyping data. *Chemometrics and Intelligent Laboratory Systems* 78(1-2): 30–40.

Yeung KY, Haynor DR, Ruzzo WL. 2001. Validating clustering for gene expression data. *Bioinformatics* 17(4): 309–318.

Yu J, Li C-X. 2006. Novel cluster validity index for FCM algorithm. *Journal of Computer Science and Technology* 21(1): 137–140.

Zahid N, Limouri M, Essaid A. 1999. A new cluster-validity for fuzzy clustering. *Pattern Recognition* 32(7): 1089–1097.

Žalik KR, Žalik B. 2011. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters* 32(2): 221–234.

Zhang Y, Wang W, Zhang X, Li Y. 2008. A cluster validity index for fuzzy clustering. *Information Sciences* 178(4): 1205–1218.