Ludwig-Maximilians-Universität München
Institut für Statistik

Benchmarkanalyse von Clustering-Verfahren mit reellen Datensätzen

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

Autorin:
Myriam Hatz

Betreuerin und Gutachterin:
Prof. Dr. Anne-Laure Boulesteix

München, 18. August 2015

Abstract

Bisher hat sich noch kein Framework etabliert, mit dem der Vergleich zweier Clustering-Verfahren durchgeführt werden kann. Problem hierbei ist unter anderem, dass für diese Verfahren kein natürliches Gütekriterium existiert. Um nun aus zwei Clustermethoden die bessere zu identifizieren, wurde in dieser Arbeit eine Benchmarkanalyse mit 50 reellen Microarray-Datensätzen ausgearbeitet und durchgeführt. Dabei wird die Differenz von externen Validierungsindizes zweier Methoden betrachtet. Diese Indizes ermöglichen es, Clusterlösungen mit einer bekannten wahren Struktur in den Daten zu vergleichen. Aufgrund der Vielzahl externer Indizes wird untersucht, welche sich, angewendet auf die vorliegenden Daten, als nützlich erweisen. Das Ergebnis dieser Untersuchung ist, dass vor allem die beiden Indizes von Baulieu (B1) und Sokal/Sneath (SS3) Unterschiede zwischen den fünf angewendeten Clustering-Verfahren aufzeigen. Der Methodenvergleich mit diesen beiden Indizes ergab daraufhin, dass die Verfahren PAM, K-Means und Ward gegenüber einer hierarchischen Clusterung mit Single-Linkage oder Complete-Linkage vorzuziehen sind. Allerdings bleibt dabei zu berücksichtigen, dass diese Ergebnisse nur für die betrachtete Datensituation gültig sind und für andere Datensätze davon verschieden ausfallen können.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Häufig werden neue Clustermethoden vorgestellt, ohne dass sie mit bereits bekannten Methoden verglichen werden. Das liegt vor allem daran, dass es keine geltenden Normen für das Benchmarking im Bereich des *unsupervised learning* gibt.
In dieser Arbeit soll eine mögliche Herangehensweise vorgestellt werden, wie die Wahl der besten Clustermethode getroffen werden kann. Dabei wird mit 50 Microarray-Genexpressions-Daten aus klinischen Krebsstudien gearbeitet, bei welchen bekannt ist, dass sie in zwei Klassen eingeteilt werden können. Die Klassenzugehörigkeit ist durch eine Zielvariable Y definiert, welche im Bezug zur jeweiligen Krebsdiagnose steht.

Kapitel 2.1 beschreibt folgende fünf Clusteranalyseverfahren, welche auf alle 50 Datensätze angewendet werden: K-Means, Partitioning-Around-Medoids, hierarchische Clusterverfahren mit Single-Linkage bzw. Complete-Linkage und das Ward-Verfahren. Die Übereinstimmung der gebildeten Partitionen mit der wahren Klassenzugehörigkeit wird anhand sogenannter externer Indizes validiert. Die genaue Definition dieser findet sich in Kapitel 2.2. Sie können nur angewendet werden, da die wahre Struktur in den Daten durch die Zielvariable vorgegeben und damit bekannt ist.

Mithilfe des Benchmarking aus Kapitel 3 soll eine Aussage getroffen werden, ob Clustermethode 2 im Vergleich zu Clustermethode 1 eine bessere Einteilung liefert. Die Wahl der beiden Clusterverfahren kann dabei beliebig festgelegt werden. Der Vergleich findet für einen jeweiligen Datensatz durch die Bildung der Differenz der Validierungsindizes zweier verschiedener Clusteranalyseverfahren statt. Um eine allgemeingültige Aussage treffen zu können, welche Clustermethode die bessere ist, werden außerdem Konfidenzintervalle der Differenzen über mehrere Datensätze hinweg betrachtet.

Schlussendlich folgt in Kapitel 4 die Anwendung der vorgestellten Methodik auf die vorliegenden reellen Datensätze und die damit einhergehenden Probleme. Zum Beispiel können viele statistische Verfahren nicht auf Microarray-Daten angewendet werden, da die An-

zahl an Variablen deutlich höher ist, als die Anzahl an Beobachtungen. Auch die Wahl des Validierungsindex ist nicht grundsätzlich festgelegt und wurde für die vorliegenden Daten untersucht, bevor die Benchmarkanalyse mit ausgewählten Indizes durchgeführt werden konnte.
Alle Analysen wurden dabei mit der Statistik-Software R, Versionsnummer 3.1.1, durchgeführt.

Kapitel 2

Methodik

2.1 Clustering-Verfahren

Werden bei einer Studie eine große Anzahl an Beobachtungen mit einer Vielzahl von Merkmalen erhoben, entstehen sehr große Datenmengen. Um Strukturen innerhalb dieser aufzuzeigen, werden häufig Clustering-Verfahren angewendet. Das Ziel einer Clusteranalyse besteht darin, eine Menge von Objekten mit bestimmten Merkmalen in kleinere Teilmengen einzuteilen. Diese Teilmengen werden Klassen bzw. Cluster genannt. Innerhalb eines Clusters sollen die Objekte hinsichtlich ihrer Merkmale möglichst homogen sein. Gleichzeitig sollen allerdings Objekte aus unterschiedlichen Clustern möglichst heterogen sein. (Kaufmann und Pape; 1996, S. 437)

2.1.1 Ähnlichkeits- und Distanzmaße

Um die Ähnlichkeit zwischen Objekten bzw. zwischen Mengen zu messen, wurden Ähnlichkeits- und Distanzmaße definiert. Der Unterschied dieser beiden Maße liegt in der Interpretation: Je ähnlicher sich zwei Objekte oder zwei Mengen sind, desto größer ist der Wert des Ähnlichkeitsmaßes, wohingegen der Wert des Distanzmaßes umso kleiner sein sollte. Kaufmann und Pape (1996, S. 440) definieren sie folgendermaßen:

Sei $I = \{I_1, \ldots, I_N\}$ eine Menge von N Objekten. Die Funktion $s : I \times I \to \mathbb{R}$ heißt *Ähnlichkeitsmaß*, wenn

$$\begin{aligned} s_{nm} &= s_{mn} \\ s_{nm} &\leq s_{nn}, \end{aligned} \tag{2.1}$$

mit $n, m = 1, \ldots, N$. Die symmetrische $N \times N$-Matrix $S = (s_{nm})$ heißt Ähnlichkeitsmatrix.

Sei $I = \{I_1, \ldots, I_N\}$ eine Menge von N Objekten. Die Funktion $d : I \times I \to \mathbb{R}$ heißt *Distanzmaß*, wenn

$$\begin{aligned} d_{nn} = 0 \text{ und } d_{nm} \geq 0 \\ d_{nm} = d_{mn}, \end{aligned} \tag{2.2}$$

mit $n, m = 1, \ldots, N$. Die symmetrische $N \times N$-Matrix $D = (d_{nm})$ heißt Distanzmatrix. Statt d_{nm} kann auch $d(n, m)$ geschrieben werden.

In der Praxis werden häufig metrische Distanzmaße verwendet. Diese erfüllen die Dreiecksungleichung ($d_{nm} \leq d_{nl} + d_{ml}$ mit $n, m, l = 1, \ldots, N$) und entsprechen der räumlichen Vorstellung (Kaufmann und Pape; 1996, S. 441).
Insbesondere für quantitative Merkmale, welche intervall- oder verhältnisskaliert sind, werden metrische Distanzen wie zum Beispiel die L_q-Metrik betrachtet (Kaufmann und Pape; 1996, S. 448):

$$d_q(n, m) = \left(\sum_{i=1}^{p} |x_{ni} - x_{mi}|^q \right)^{\frac{1}{q}} \quad , \; q > 0. \tag{2.3}$$

Dabei gibt p die Anzahl an Variablen an. Diese Distanzen sind translationsinvariant, jedoch nicht skaleninvariant, weswegen Variablen bei ungleichen Einheiten normiert werden müssen.
Gebräuchlich sind vor allem die L_1-Metrik, auch City-Block-Metrik genannt und die L_2-Metrik, die der euklidischen Distanz entspricht:

$$d_2(n, m) = ||x_n - x_m|| = ((x_n - x_m)'(x_n - x_m))^{\frac{1}{2}}. \tag{2.4}$$

2.1.2 Hierarchische Verfahren

Bei den hierarchischen Clustering-Verfahren unterscheidet man zwischen agglomerativen und divisiven Verfahren. Ist ein Verfahren agglomerativ, werden die Daten sukzessive in Teilklassen zusammengefasst, wobei sich die Heterogenität der Klassen schrittweise erhöht. Im Gegensatz dazu stehen die divisiven Verfahren, bei denen bestehende Klassen sukzessive aufgeteilt werden, was die Heterogenität der Klassen schrittweise verringert. In Abbildung 2.1 sind beide Verfahren anschaulich in einem sogenanntem Dendrogramm dargestellt.

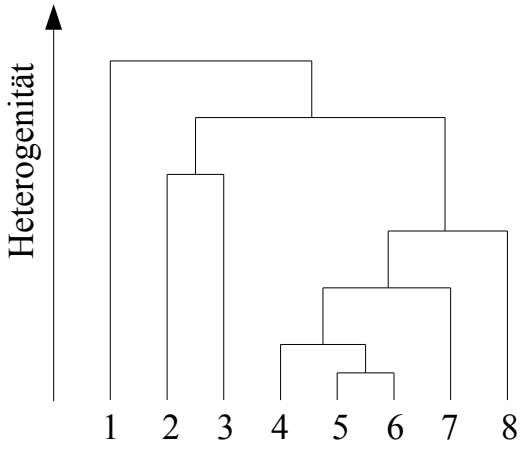

Abbildung 2.1: *Dendrogramm - Darstellung einer hierarchischen Clusterung (Kaufmann und Pape; 1996, S. 453).*

Diese Form von Stammbaum wird bei agglomerativen Verfahren von unten nach oben konstruiert und genau entgegengesetzt bei divisiven Verfahren (Kaufmann und Pape; 1996, S. 453). Da divisive Verfahren für eine große Anzahl an Beobachtungen einen hohen Rechenaufwand mit sich bringen, sind agglomerative Verfahren weiter verbreitet und werden im Folgenden näher betrachtet.

2.1.2.1 Prinzip agglomerativer hierarchischer Verfahren

Das Prinzip der agglomerativen Verfahren lässt sich in drei Schritten darstellen (Kaufmann und Pape; 1996, S. 457-458):

1. Jedes Objekt der Objektmenge $I = \{I_1, \ldots, I_N\}$ entspricht einem Cluster, d.h. es gilt die Anfangspartition $\mathcal{C}^{(0)} = \{\{I_1\}, \ldots, \{I_N\}\}$.

2. Die Partition $\mathcal{C}^{(\nu)}$ ($\nu \geq 1$) wird durch Fusion zweier Klassen aus der Partition $\mathcal{C}^{(\nu-1)}$ gebildet, für die das Distanzmaß D zwischen zwei Klassen minimal wird (ν bezeichnet dabei die Anzahl der Iterationsschritte).

3. Iteration von Schritt 2, bis nur noch ein Cluster besteht, also $\mathcal{C}^{(\nu)} = \{I\}$.

Zu den gebräuchlichsten hierarchischen Verfahren zählen unter anderem der Single-Linkage, Complete-Linkage und das Verfahren von Ward. Durch die Wahl des Verfahrens wird bestimmt, wie die Distanz zwischen den Clustern definiert ist.

2.1.2.2 Single-Linkage-Verfahren

Das Distanzmaß zwischen zwei Klassen C_k und C_j entspricht beim Single-Linkage-Verfahren der kleinsten Distanz zwischen einem Objekt aus C_k und einem Objekt aus C_j, also

$$D(C_k, C_j) = \min_{\substack{n \in C_k \\ m \in C_j}} \{d_{nm}\}.$$

Je größer der Wert des Heterogenitätsindex h, desto unähnlicher sind sich zwei Klassen. Zur Fusion zweier Klassen im ν-ten Iterationsschritt werden die Klassen C_v und C_w der Partition $\mathcal{C}^{(\nu-1)}$ mit dem kleinsten Distanzmaß gewählt, wodurch sich der Index ergibt zu

$$h_\nu = D_\nu = D(C_v, C_w) = \min_{k \neq j} \min_{\substack{n \in C_k \\ m \in C_j}} \{d_{nm}\} \tag{2.5}$$

(Kaufmann und Pape; 1996, S. 461).

2.1.2.3 Complete-Linkage-Verfahren

Analog zum Single-Linkage-Verfahren wird das Complete-Linkage-Verfahren durchgeführt. Mit dem einzigen Unterschied, dass nun das Distanzmaß zwischen zwei Klassen C_k und C_j als größte Distanz zwischen jeweils einem Objekt aus beiden Klassen definiert ist:

$$D(C_k, C_j) = \max_{\substack{n \in C_k \\ m \in C_j}} \{d_{nm}\}.$$

Ebenso ändert sich der Heterogenitätsindex zu

$$h_\nu = D_\nu = D(C_v, C_w) = \min_{k \neq j} \max_{\substack{n \in C_k \\ m \in C_j}} \{d_{nm}\} \tag{2.6}$$

(Kaufmann und Pape; 1996, S. 462).

2.1.2.4 Verfahren von Ward

Das Verfahren von Ward beruht auf der Streuung innerhalb der einzelnen Klassen. Die Heterogenität H der Partition $\mathcal{C}^{(\nu-1)}$ und der Partition $\mathcal{C}^{(\nu)}$ wird durch die Summe der Streuung innerhalb der Klassen ermittelt. Mit diesen Größen kann der Heterogenitätsgewinn bei Fusion der Cluster C_v und C_w ermittelt werden, welcher der Distanz zwischen den beiden Clustern entspricht:

$$H\left(\mathcal{C}^{(\nu)}\right) - H\left(\mathcal{C}^{(\nu-1)}\right) = \frac{n_v n_w}{n_v + n_w} ||\bar{x}_v - \bar{x}_w||^2 =: D(C_v, C_w).$$

Es werden die beiden Klassen aus $\mathcal{C}^{(\nu-1)}$ zur Fusion gewählt, die die Heterogenität nur minimal wachsen lassen, was bedeutet

$$h_\nu = D(C_v, C_w) = \underset{k \neq j}{min} \frac{n_k n_j}{n_k + n_j} ||\bar{x}_k - \bar{x}_j||^2 \tag{2.7}$$

(Kaufmann und Pape; 1996, S. 466).

2.1.3 Nichthierarchische Verfahren: Optimale Partitionen

Bei einer Clusterung mit optimalen Partitionen wird die Qualität der Partition durch ein Gütekriterium gemessen. Es wird die Partition $\mathcal{C}$ gesucht, welche im Hinblick auf das entsprechende Gütekriterium optimal ist (Kaufmann und Pape; 1996, S. 469).

2.1.3.1 Prinzip bei optimalen Partitionen

Ein häufig angewendetes Verfahren für optimalen Partitionen ist das Austauschverfahren, welches von Kaufmann und Pape (1996, S. 472) folgendermaßen beschrieben wird:

1. Sei die zufällige Anfangspartition $\mathcal{C}^{(0)}$ vorgegeben.
2. Nun wird für jedes Objekt geprüft, ob sich das Gütekriterium verbessert, wenn man das Objekt in der Partition $\mathcal{C}^{(\nu)}$ $(\nu \geq 0)$ einem anderen Cluster zuordnet.
3. Das Objekt, welches die größte Verbesserung liefert, wird dem entsprechendem Cluster zugeordnet, wodurch sich die Partition $\mathcal{C}^{(\nu+1)}$ bildet.
4. Die Schritte 2 und 3 werden iteriert, bis keine Verbesserung des Gütekriteriums mehr eintritt.

Da dieses Verfahren auch ein Suboptimum ergeben kann, sollten mehrere Startpartitionen $\mathcal{C}^{(0)}$ gewählt werden. Für jede wird das Verfahren erneut durchgeführt und die optimale Partition ist darunter die, welche das beste Gütekriterium liefert.

Ein großer Unterschied zu den hierarchischen Verfahren ist, dass bei den optimalen Partitionen die Klassenanzahl anfangs bereits festgelegt werden muss.

2.1.3.2 K-Means-Verfahren

Zur Clusterbildung werden beim K-Means-Verfahren Clusterzentren konstruiert. Das Gütekriterium hierbei ist das Varianzkriterium. Dabei wird angenommen, dass ein Cluster mit

ähnlichen Objekten eine kleine Streuung innerhalb des Clusters aufweist. Die Streuungsquadratsumme in den Clustern soll dabei minimiert werden, was sich mithilfe der quadrierten euklidischen Distanz wie folgt darstellen lässt (Kaufmann und Pape (1996, S. 475), Bacher et al. (2010, S. 299)):

$$h(\mathcal{C}) = \sum_{k=1}^{g} \sum_{n \in C_k} ||x_n - \bar{x_k}||^2 \rightarrow \min. \tag{2.8}$$

Daraufhin wird das in Kapitel 2.1.3.1 vorgestellte Prinzip angewendet. In Schritt 2 werden die Clusterzentren der g Cluster als Mittelwertsvektoren der Merkmalsvektoren der Individuen im Cluster berechnet. Anschließend wird für jedes Objekt geprüft, zu welchem Clusterzentrum es die geringste Distanz besitzt und in Schritt 3 entsprechend ausgetauscht. Dadurch minimiert sich $h(\mathcal{C})$ in jedem Iterationszyklus (Bacher et al.; 2010, S. 299).

2.1.3.3 PAM-Verfahren

Das Partitioning-Around-Medoids-Verfahren hat starke Ähnlichkeiten mit dem K-Means-Verfahren. Das PAM-Verfahren bietet allerdings den großen Vorteil, dass es wesentlich robuster gegenüber Ausreißern ist. Zudem können auch Daten verarbeitet werden, die nicht intervall-skaliert sind, da die Distanzmatrix übergeben werden kann. Im Gegensatz zum K-Means-Verfahren werden hier zur Bildung der Cluster nicht Clusterzentren ermittelt, sondern Clustermedoiden. Das sind Objekte innerhalb der Daten, die verschiedene Aspekte der Datenstruktur repräsentieren. Die Anzahl an Repräsentanten entspricht der gewünschten Clusteranzahl. Angelehnt an das Prinzip aus Kapitel 2.1.3.1 wird der erste Schritt beim PAM-Verfahren „Build-Phase“ genannt. In dieser werden mithilfe eines Algorithmus die Repräsentanten gewählt. (Näheres dazu von Kaufman und Rousseeuw (2005, S.102-103).) Letztendlich sollten sie zentral in der Mitte eines Clusters liegen. Anhand der Distanzmatrix wird in dieser Phase auch entschieden, welche Objekte zu den ermittelten Medoiden am nächsten liegen und die Cluster entsprechend gebildet.
Schritt 2 entspricht der „Swap-Phase“, mit dem Unterschied, dass die Objekte nicht in ein anderes Cluster ausgetauscht werden, sondern, dass die Repräsentanten ausgetauscht werden. Es wird geprüft, ob das Gütekriterium optimiert wird, wenn ein Objekt seinen Platz mit einem Repräsentanten wechselt. Dabei soll die durchschnittliche Distanz der n Objekte zu den jeweils am nächsten liegenden Clustermedoiden minimiert werden:

$$h(\mathcal{C}) = \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) z_{ij} \rightarrow \min. \tag{2.9}$$

z_{ij} ist dabei eine Indikatorvariable, die 1 wird, wenn Objekt x_j dem Cluster zugeordnet wird, in dem x_i das repräsentative Objekt ist. (Kaufman und Rousseeuw; 2005)

2.2 Validierung von Clusterlösungen

Soll geprüft werden in welchem Maße sich zwei Clusterlösungen unterscheiden, bieten sich diverse Validierungstechniken an. Grob kann dabei zwischen internen und externen Indizes unterschieden werden. Externe Indizes prüfen dabei die Stabilität einer Partition, d.h. inwieweit die Klassenlabel richtig vergeben wurden. Hierfür muss die wahre Klassenzugehörigkeit bekannt sein, was in der Praxis oft nicht gegeben ist. In diesem Fall können interne Indizes angewendet werden, welche nur aus den Daten, die der Clusteranalyse zur Verfügung stehen, berechnet werden (Scherl; 2010). Im Folgenden werden nur externe Indizes betrachtet, da diese im weiteren Verlauf für die Benchmarkanalyse in Kapitel 3 interessant sind.

2.2.1 Grundprinzip externer Indizes

Mit externen Indizes lässt sich entweder die Ähnlichkeit zweier Clustermethoden quantifizieren oder die ermittelte Clusterlösung mit der wahren Klassenzugehörigkeit vergleichen. All diese Indizes beruhen auf einer Kontingenztabelle, die für alle Objektpaare eines Datensatzes folgende Information enthält (Albatineh et al.; 2006):

		Clustermethode 2	
	Anzahl an Paaren	im selben Cluster	in verschiedenen Cluster
Cluster-methode 1	im selben Cluster	a	b
	in verschiedenen Cluster	c	d

Tabelle 2.1: *Kontingenztabelle von Objektpaaren zweier Clustermethoden.*

- a = Die Anzahl an Objektpaaren, die in beiden Clusterungen demselben Cluster angehören.
- b = Die Anzahl an Objektpaaren, die mit Methode 1 demselben Cluster zugeordnet wurden, jedoch mit Methode 2 nicht.
- c = Die Anzahl an Objektpaaren, die mit Methode 2 demselben Cluster zugeordnet wurden, jedoch mit Methode 1 nicht.
- d = Die Anzahl an Objektpaaren, die in beiden Clusterungen unterschiedlichen Clustern angehören.

2.2.2 Beispiele externer Indizes

Albatineh et al. (2006) liefern eine Übersicht mit 22 externen Indizes, die mithilfe von Tabelle 2.1 und der Anzahl von Beobachtungen m im Datensatz berechnet werden können. Dabei gilt $\binom{m}{2} = a + b + c + d$, was der gesamten Anzahl an Objektpaaren entspricht.

Wie in Tabelle 2.2 zu erkennen ist, unterscheiden sich einige der Indizes nur geringfügig in ihrer Berechnung. Allgemein kann gesagt werden, dass je höher ihr Wert am Maximum des jeweiligen Wertebereichs liegt, desto ähnlicher sind sich die zwei Clustermethoden bzw. desto besser entspricht die ermittelte Clusterung der wahren Klassenzugehörigkeit.

Name	**Symbol**	**Formel**	**Wertebereich**
Rand	R	$\frac{a+d}{a+b+c+d}$	[0,1]
Hubert	H	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[-1,1]
Czekanowski	CZ	$\frac{2a}{2a+b+c}$	[0,1]
Kulczynski	K	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	[0,1]
McConnaughey	MC	$\frac{a^2-bc}{(a+b)(a+c)}$	[-1,1]
Peirce	PE	$\frac{ad-bc}{(a+c)(b+d)}$	[-1,1]
Fowlkes und Mallows	FM	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0,1]
Wallace (1)	W1	$\frac{a}{a+b}$	[0,1]
Wallace (2)	W2	$\frac{a}{a+c}$	[0,1]
Gamma	Γ	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	[-1,1]
Sokal und Sneath (1)	SS1	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$	[0,1]
Baulieu (1)	B1	$\frac{\binom{m}{2}^2-\binom{m}{2}(b+c)+(b-c)^2}{\binom{m}{2}^2}$	[0,1]
Russel und Rao	RR	$\frac{a}{a+b+c+d}$	[0,1]
Fager und McGowan	FMG	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2\sqrt{(a+b)}}$	$[-\frac{1}{2}, 1)$
Pearson	P	$\frac{ad-bc}{(a+b)(a+c)(c+d)(b+d)}$	[-1,1]
Baulieu (2)	B2	$\frac{ad-bc}{\binom{m}{2}^2}$	$[-\frac{1}{4}, \frac{1}{4}]$
Jaccard	J	$\frac{a}{a+b+c}$	[0,1]
Sokal und Sneath (2)	SS2	$\frac{a}{a+2(b+c)}$	[0,1]
Sokal und Sneath (3)	SS3	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	[0,1]
Gower und Legendre	GL	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	[0,1]
Rogers und Tanimoto	RT	$\frac{a+d}{a+2(b+c)+d}$	[0,1]
Goodman und Kruskal	GK	$\frac{ad-bc}{ad+bc}$	[-1,1]

Tabelle 2.2: *Auflistung verschiedener externer Indizes mit Angabe der Entwickler und der entsprechenden Wertebereiche (Albatineh et al.; 2006).*

Kapitel 3

Benchmarking

Im Machine Learning oder der computationalen Statistik werden ständig neue Verfahren vorgestellt, die im Vergleich zur bisher gebräuchlichen eine „bessere“ Performance bieten sollen. Ob das allerdings tatsächlich der Fall ist, wird meist nicht statistisch nachgewiesen. Im Fokus eines Benchmark Experiments liegt daher allgemein nicht nur die Beurteilung der Performance verschiedener Algorithmen, sondern den besten unter ihnen zu identifizieren (Hothorn et al.; 2005).
Ein Benchmarking kann mit simulierten Daten oder mit reellen Datensätzen durchgeführt werden. Simulationsstudien spiegeln dabei allerdings häufig nicht die Komplexität der Verteilung von reellen Daten wieder, weswegen die Betrachtung reeller Datensätze von großer Bedeutung ist. Dabei liegt das Interesse jedoch nicht in der Aussage für nur einen Datensatz, sondern ob ein Verfahren, angewendet auf diverse Datensätze aus einem Themenbereich, im Mittel das bessere Ergebnis liefert. Die beobachtete Performance hängt bei festem Stichprobenumfang und fester Verteilung von der jeweiligen Stichprobe ab. Das „no-free-lunch“-Theorem besagt außerdem, dass nicht davon ausgegangen werden kann, dass Methode 2 für sämtliche Stichprobengrößen und Verteilungen eine bessere Performance liefert wie Methode 1. Daher sollten vor allem bei reellen Daten immer mehrere Datensätze in Betracht gezogen werden (Boulesteix et al.; 2015).

In dieser Arbeit soll nun im Bereich des *unsupervised learning*, zu welchem die Clusteranalyse zählt, solch ein Benchmarking mit reellen Datensätzen durchgeführt werden. Das Ziel ist also, eine Aussage darüber zu treffen, ob Clustermethode 2 im Vergleich zu Clustermethode 1 eine bessere Einteilung liefert.

3.1 Hypothesenformulierung

Mithilfe von Hypothesentests lassen sich zwei verschiedene Methoden vergleichen. Bei solchen Tests werden bestimmte Annahmen über einen Parameter oder eine Verteilung in der Grundgesamtheit getroffen. Diese Annahmen müssen dann als statistisches Testproblem formuliert werden (Fahrmeir et al.; 2011). Boulesteix et al. (2015) stellen eine statistische Testformulierung vor, die im Kontext von Methodenvergleichen angewendet werden kann. Diese Testformulierung beruht auf der Identifizierung der besseren Klassifikationsmethode. Hierfür kann als Parameter die Fehlerrate ε gewählt werden, da diese den Anteil falsch klassifizierter Objekte angibt und somit eine Aussage über die Güte des Verfahrens getroffen werden kann.

Analog dazu gibt es jedoch für Clusterverfahren kein solches natürliches Gütekriterium. Geht man allerdings davon aus, dass die wahre Klassenzugehörigkeit bekannt ist, kann einer der in Kapitel 2.2.2 vorgestellten Validierungsindizes VI als Gütekriterium herangezogen werden. Bei der Untersuchung eines Datensatzes gilt dann, dass Methode 2 besser als Methode 1 ist, wenn $VI_2 > VI_1$.

Wie bereits einleitend erwähnt, liegt das Interesse beim Benchmarking mit reellen Datensätzen jedoch nicht in der Aussage für nur einen Datensatz. Daher werden die Hypothesen mithilfe der Erwartungswerte der Indizes aller berücksichtigten Datensätze folgendermaßen definiert:

$$\begin{aligned} & H_0 : \ \mathbb{E}(VI_1) - \mathbb{E}(VI_2) \geq 0 \\ \text{vs.} \quad & H_1 : \ \mathbb{E}(VI_1) - \mathbb{E}(VI_2) < 0. \end{aligned} \tag{3.1}$$

Wobei gilt, dass $\mathbb{E}(VI_1) - \mathbb{E}(VI_2) = \mathbb{E}(VI_1 - VI_2) = \mathbb{E}(\Delta VI)$, womit die Bedeutung der Differenz der Indizes hervorgehoben wird.

Eine aus theoretischer Sicht vollständige Formulierung der Nullhypothese liegt außerhalb der Zielsetzung dieser Arbeit, wird allerdings von Boulesteix et al. (2015) ausführlich für die Klassifikation diskutiert.

3.2 Bootstrap-Konfidenzintervalle

Um die Präzision einer Schätzung zu quantifizieren, werden häufig Intervallschätzungen für den interessierenden Parameter θ durchgeführt. Dies ist auch für $\theta = \Delta VI$ möglich. Ein übliches 95%-Konfidenzintervall mit Irrtumswahrscheinlichkeit $\alpha = 0.05$ erhält man

zum Beispiel mit

$$\hat{\theta} \pm z_{1-\alpha}\hat{\sigma}. \tag{3.2}$$

Dabei ist $\hat{\theta}$ der Punktschätzer des interessierenden Parameter, $\hat{\sigma}$ die Schätzung der Standardabweichung von $\hat{\theta}$ und $z_{1-\alpha}$ entspricht dem $(1-\alpha)$-Quantil der Standardnormalverteilung. Das Problem bei solchen Standardintervallen ist, dass sie auf einer asymptotischen Approximation beruhen, die in der Praxis nicht immer gegeben ist. Eine Möglichkeit diese Normalverteilungsannahme zu vermeiden, bieten Bootstrap-Konfidenzintervalle (DiCiccio und Efron; 1996).

3.2.1 Bootstrap-Stichprobe

Für die Berechnung eines Bootstrap-Konfidenzintervall ist eine große Anzahl an Bootstrap-Stichproben nötig. Im Folgenden wird kurz die Idee dieser Stichproben aufgezeigt.

Da es nicht immer möglich ist, eine gesamte Population $\mathcal{X} = (X_1, \ldots, X_N)$ zu erheben, wird eine beobachtbare, zufällige Stichprobe $x = (x_1, \ldots, x_n)$ aus dieser Grundgesamtheit gezogen. Dabei ist bekannt, dass x einer bestimmten Verteilung F folgt, wobei die exakte Verteilung unbekannt ist. Meist liegt das Interesse ohnehin keineswegs in der gesamten Verteilung, sondern an einem konkreten Parameter $\theta = T(F)$. Dieser soll auf Basis von x mit $\hat{\theta} = s(x)$ geschätzt werden, dabei gilt oft $s(x) = T(\hat{F})$.
Es wird angenommen, dass $\hat{F}$ die empirische Verteilungsfunktion ist, die jedem Wert x_i, $i = 1, \ldots, n$, die Wahrscheinlichkeit $\frac{1}{n}$ zuweist, womit sich $\hat{F}$ auch schreiben lässt als

$$\hat{F}(x) = \frac{1}{n}\sum_{i}^{n} I(x_i \leq x), \tag{3.3}$$

mit $I(\cdot)$ als Indikatorfunktion.

Somit kann die Bootstrap-Stichprobe $x^* = (x_1^*, \ldots, x_n^*)$ ermittelt werden. Dafür wird n-mal zufällig mit Zurücklegen aus der Stichprobe $x = (x_1, \ldots, x_n)$ gezogen, womit jedes x_i^*, $i = 1, \ldots, n$, unabhängig und identisch $\hat{F}$-verteilt ist.
Beide Stichproben haben also den gleichen Umfang n, wobei die Werte aus x einmal, mehrfach oder gar nicht in x^* vorkommen können.

Einer Bootstrap-Stichprobe x^* kann eine Bootstrap-Replikation von $\hat{\theta}$ zugewiesen werden:

$$\hat{\theta}^* = s(x^*). \tag{3.4}$$

Dabei wird die Schätzfunktion $s(\cdot)$ auf die Bootstrap-Stichprobe angewendet.
Dieses Verfahren kann nun wiederholt durchgeführt werden, sodass B Bootstrap-Replikationen entstehen (Efron und Tibshirani; 1993).

3.2.2 BC$_a$-Bootstrap-Intervall

Es gibt verschiedene Ansätze ein Bootstrap Konfidenzintervall zu schätzen. Eine Möglichkeit bietet die von Efron und Tibshirani (1993) vorgestellte BC_a-Methode (engl. „bias-corrected and accelerated“). Diese liefert mithilfe der Perzentile des Bootstrap-Histogramms approximative Konfidenzintervalle für θ.

Mit den beschriebenen Größen aus Kapitel 3.2.1 lässt sich die kumulierte Verteilungsfunktion $\hat{G}(c)$ von B Bootstrap-Replikationen $\hat{\theta}^*(b)$ aufstellen zu

$$\hat{G}(c) = \frac{\#\left\{\hat{\theta}^*(b) < c\right\}}{B}. \tag{3.5}$$

Nach Definition gilt $\hat{G}^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$, was dem $\alpha\cdot 100$ten Perzentil der Bootstrap-Verteilung entspricht. Liegen B Bootstrap-Replikationen vor, so ist $\hat{\theta}_B^{*(\alpha)}$ das empirische Perzentil. Bei 2000 Replikationen und $\alpha = 0.05$ gleicht $\hat{\theta}_{2000}^{*(0.05)}$ also dem hundertsten ($\hat{=}\, B\cdot\alpha$) Wert aus der geordneten Liste aller Replikationen.
Die Grenzen des BC_a-Intervalls werden nun auch von den Perzentilen der Bootstrap-Verteilung bestimmt, allerdings abhängig von zwei numerischen Parametern $\hat{z}_0$ und $\hat{a}$:

$$\begin{aligned} &\hat{\theta}_{BC_a}(\alpha) = \hat{G}^{-1}\left(\alpha_{adj}\right) \\ &\text{mit } \alpha_{adj} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right). \end{aligned} \tag{3.6}$$

Dabei bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung mit den Quantilen $z^{(\alpha)} = \Phi^{-1}(\alpha)$.
Das 90%-BC_a-Konfidenzintervall ergibt sich dann zum Beispiel zu $\left[\hat{\theta}_{BC_a}(0.05), \hat{\theta}_{BC_a}(0.95)\right]$ sowie jedes andere $(1-2\alpha)$-BC_a-Konfidenzintervall entsprechend zu $\left[\hat{\theta}_{BC_a}(\alpha), \hat{\theta}_{BC_a}(1-\alpha)\right]$.

Die Formel 3.6 für die Intervallgrenzen kann durch folgende Annahmen motiviert werden: Es existiert eine monoton steigende Transformation $\phi = m(\theta)$ mit $\hat{\phi} = m(\hat{\theta})$, so dass für jegliches θ gilt

$$\begin{aligned} &\hat{\phi} \sim N(\phi - z_0\sigma_\phi, \sigma_\phi^2) \\ &\quad \text{mit } \sigma_\phi = 1 + a\phi. \end{aligned} \tag{3.7}$$

Da die Transformation m in Formel 3.6 keine Berücksichtigung findet, können die Intervallgrenzen also auch ohne Wissen über m berechnet werden.
Mithilfe der Wahrscheinlichkeit $P(\hat{\phi} < \phi) = \Phi(z_0)$ lässt sich der Bias-Korrektor z_0 gut interpretieren, womit aufgrund der Monotonie auch gilt $P(\hat{\theta} < \theta) = \Phi(z_0)$.

Die einfachste Form des BC_a-Algorithmus schätzt z_0 daraufhin durch

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\hat{\theta}^*(b) < \hat{\theta}\right\}}{B}\right) = \Phi^{-1}\left(\hat{G}(\hat{\theta})\right). \tag{3.8}$$

Die Beschleunigung a in Formel 3.7 gibt an, wie schnell sich der Standardfehler auf Basis der Standardnormalverteilung ändert. Auch für sie gibt es verschiedene Methoden zur Schätzung. Zum Beispiel können die Jackknife Werte einer Statistik $\hat{\theta} = s(x)$ herangezogen werden. Dabei wird die Stichprobe $x_{(i)}$ verwendet, die aus der ursprünglichen Stichprobe x besteht allerdings ohne den Wert x_i. Es gilt $\hat{\theta}_{(.)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$, wobei $\hat{\theta}_{(i)} = s(x_{(i)})$. Damit ergibt sich für den Schätzer von a:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^2\right\}^{3/2}} \tag{3.9}$$

(DiCiccio und Efron; 1996; Efron und Tibshirani; 1993).
Weitere Schätzer für z_0 und a werden zum Beispiel von DiCiccio und Efron (1996) diskutiert.

Kapitel 4

Anwendung auf 50 Microarray-Datensätze

Ziel der vorgestellten Benchmarkanalyse ist es, beim Vergleich zweier Clusterverfahren das bessere der beiden zu identifizieren. Die in Kapitel 2 und 3 vorgestellten Verfahren sollen nun auf 50 verschiedene Microarray-Datensätze angewendet werden. Hierfür wird zunächst auf die Datenstruktur eingegangen, bevor eine Clusteranalyse mit verschiedenen Verfahren durchgeführt wird. Diese Clusterlösungen werden dann mithilfe der Indizes aus Kapitel 2.2.2 validiert und verglichen. Für alle Analysen wird die Statistik-Software R, Versionsnummer 3.1.1, genutzt.

4.1 Microarray-Daten

Mithilfe der Microarray-Technologie ist es seit Ende der 1990er Jahre möglich, die DNA-Sequenzen eines Organismus zu analysieren. Die DNA transkribiert mRNA. Wird diese sogenannte Messenger-RNA in ein Protein übersetzt, heißt dieser Vorgang Genexpression. Mit einem Microarray kann das Expressionsniveau aller Gene in einem einzigen Experiment festgestellt werden, da die Menge verschiedener mRNA-Moleküle zu einem bestimmten Zeitpunkt in einer Zelle gemessen wird.

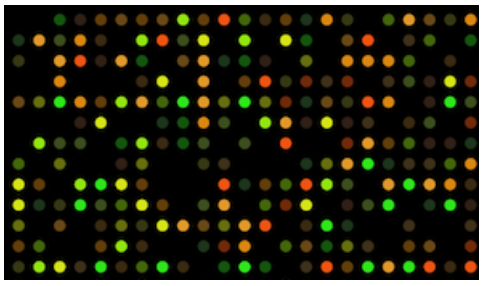

Abbildung 4.1: *Auszug aus einem Microarray.*

Ein Microarray ist ein Objektträger aus Glas, auf dem Millionen von einzelnen DNA-Molekülen auf sogenannten „Spots“ fixiert werden. In Genexpressions-Studien soll mit jedem einzelnen dieser DNA-Moleküle ein mRNA-Molekül im Genom untersucht werden.

Eine weit verbreitete Methode ist das Vergleichen von Expressionsniveaus in zwei verschiedenen Proben (zum Beispiel zwei verschiedene Entwicklungsstadien). Dabei wird die isolierte mRNA in beiden Proben farblich unterschiedlich markiert, z. B. in Probe 1 grün und in Probe 2 rot. Werden die beiden Proben vereinigt und mit einem Laser angeregt, findet eine Hybridisierung statt. Anhand der Fluoreszenzintensität und Farbe der einzelnen Spots (vgl. Abbildung 4.1) kann das relative Expressionsniveau der Gene in beiden Proben geschätzt werden.
Mit Microarrays kann also erforscht werden, welche Gene in welchen Zelltypen aktiv sind bzw. an welchen Zellvorgängen sie teilnehmen und wie sich das Expressionsniveau einzelner Gene z. B. in verschiedenen Krankheitsstadien verhält. Besonders im Bereich der Onkologie erweisen sich Microarrays als informativ, denn ihre Analyse kann dazu beitragen, Tumorarten zu klassifizieren und neue Unterarten zu definieren. Unter anderem können diesen Daten auch genutzt werden, um Vorhersagen zur Prognose und Diagnose für Krebspatienten zu treffen (Causton et al.; 2003).

Die Herausforderung bei der Analyse von Microarray-Daten ist, dass die Anzahl an Variablen deutlich höher ist, als die Anzahl an Beobachtungen, wodurch enorm große Datenmengen entstehen. Werden n Patienten betrachtet, sind diese als Beobachtungen zu verstehen. Von jeder Beobachtung werden bestimmte Gene untersucht, diese können statistisch als p Variablen angesehen werden. Typischerweise werden 20 bis 300 Beobachtungen untersucht, wohingegen p dabei zwischen 5.000 und 50.000 liegen kann. Dadurch sind viele statistische Standardverfahren nicht anwendbar (Boulesteix et al.; 2008).

4.2 Verwendete Datensätze

Im Folgenden sollen Clusteranalyseverfahren auf 50 verschiedene Microarray-Datensätze angewendet werden. Diese Datensätze wurden bereits von Boulesteix et al. (2015) zum Vergleich von Klassifikationsverfahren verwendet. Dabei handelt es sich um 50 reelle Datensätzen aus klinischen Krebsstudien, bei welchen bekannt ist, dass die Beobachtungen in zwei Klassen eingeteilt werden können. Die Klassenzugehörigkeit ist durch eine Zielvariable Y definiert, welche im Bezug zur jeweiligen Krebsdiagnose steht. Dabei kann es sich beispielsweise um den aktuellen Gesundheitszustand handeln (z.B. Tumor ja/nein) oder um eine längerfristige Prognose (z.B. gute/schlechte Prognose).
Für alle Studien steht eine Datenmatrix X zur Verfügung, in der jede Beobachtung v_1 bis v_n eine Zeile $(x_1, \ldots, x_p)$ darstellt. Die Spalten entsprechen den p Variablen, also den gemessenen Genexpressionsniveaus. Je nach Datensatz liegt n zwischen 23 und 286 und p zwischen 1.098 und 54.675.

4.3 Clusteranalyse der Datensätze

Die in Kapitel 2.1 vorgestellten Clustering-Verfahren wurden nun für alle 50 Datensätze durchgeführt. Die gewünschte Clusteranzahl betrug dabei aufgrund der binären Zielvariable jeweils zwei. Weitere Spezifikationen und die jeweiligen R-Funktionen die zum Einsatz kamen, werden im Folgenden aufgeführt:

- Single-Linkage- / Complete-Linkage- / Ward-Verfahren: Funktion *hclust*. Es wurde als Distanzmaß zwischen zwei Objekten die euklidische Distanz gebildet.
- K-Means: Funktion *kmeans*. Um die Rechenzeit zu begrenzen, sollten maximal 20 Iterationen durchgeführt werden. Außerdem wurde das beste Ergebnis aus 10 Startpartitionen gewählt.
- Partitioning-Around-Medoids: Funktion *pam* aus dem Package *„cluster“*. Hier wurde ebenfalls die euklidische Distanz als Distanzmaß verwendet.

Beispielhaft soll anhand eines Datensatzes gezeigt werden, wie die Klasseneinteilungen der unterschiedlichen Verfahren ausgefallen sind. Hierfür wird der Datensatz *leukemia_yagi* verwendet. Bei dieser Studie wurde das Genexpressionsprofil von 53 Patienten mit akuter myeloischer Leukämie in 7241 Variablen gemessen. Die Variable Y beschreibt in diesem Fall das Ansprechen auf die Chemotherapie mit den Ausprägungen „0 = vollständige Remission“ und „1 = Rückfall“.

Verfahren / Cluster	Single-Linkage	Complete-Linkage	Ward	K-Means	PAM
1	52	49	24	17	27
2	1	4	29	36	26

Tabelle 4.1: *Clusterzuordnung verschiedener Clusterverfahren für den Datensatz leukemia_yagi.*

Tabelle 4.1 zeigt die Anzahl der Beobachtungen je Cluster für alle fünf Verfahren. Es ist zu erkennen, dass bei einer hierarchischen Clusterung mit Single-Linkage eine Klasse aus nur einem einzelnen Objekt besteht, was selten bei einer Clusteranalyse gewünscht ist. Der Grund hierfür liegt im Nachteil dieses Verfahrens: Der Single-Linkage hat eine sehr „schwache“ Voraussetzung an die Homogenität in den Clustern. Deutlich getrennte Klassen können daher vermischt werden, wenn sie, wie in Abbildung 4.2, durch eine Brücke verbunden sind. Diese sogenannte Verkettungseigenschaft lässt sich jedoch für die Analyse von Ausreißern nutzen, da diese Objekte weit von den anderen Objekten entfernt liegen und erst in den letzten Iterationsschritten einem Cluster hinzugefügt werden

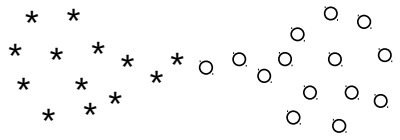

Abbildung 4.2: *Durch eine Brücke verbundene Klassen (Kaufmann und Pape; 1996, S. 462).*

(Kaufmann und Pape; 1996; Bacher et al.; 2010). Es kann also davon ausgegangen werden, dass es sich bei dieser einen Beobachtung in Cluster 2 beim Single-Linkage-Verfahren um einen Ausreißer im Datensatz *leukemia_yagi* handelt. Der Complete-Linkage hingegen weist diese Verkettungseigenschaft nicht auf. Laut Bacher et al. (2010, S. 152) führt er „dagegen oft dazu, dass sehr viele Cluster gebildet werden, da er von einer sehr 'strengen' Vorstellung hinsichtlich der Homogenität in den Clustern ausgeht. Dieser Effekt wird als *Dilatationseffekt* bezeichnet". Die Cluster sind außerdem meist sehr kompakt mit geringen Durchmessern (Kaufman und Rousseeuw; 2005, S. 41). Diese Eigenschaft ist mit diesem Beispieldatensatz, bei nur zwei gebildeten Klassen, nicht zu erkennen.
Auch wenn das Single-Linkage-Verfahren für die vorliegenden Microarray-Daten oftmals keine brauchbaren Clusterungen liefert, wurde es für die nachfolgenden Analysen beibehalten. Grund hierfür ist die Vermutung, dass sich durch die schlechte Anpassung signifikante Unterschiede in den Validierungsindizes der Verfahren ergeben.

4.4 Wahl des Validierungsindex

Zur Validierung einer Clusterlösung kann eine Kontingenztabelle der wahren Klassenzugehörigkeit, die durch Y bestimmt wird, und der Clusterzuordnung einen ersten Überblick liefern.

Ward		Y	
		0	1
Cluster	1	14	10
	2	14	15

K-Means		Y	
		0	1
Cluster	1	9	8
	2	19	17

PAM		Y	
		0	1
Cluster	1	17	10
	2	11	15

Tabelle 4.2: *Kontingenztabellen der Clusterzuordnungen und Y von leukemia_yagi. Die grau hinterlegten Zellen stellen in Summe jeweils die maximale Übereinstimmung der Cluster- mit der wahren Klassenzuordnung dar.*

Die Kontingenztabellen, die sich beispielsweise für den Datensatz *leukemia_yagi* für das Ward-, K-Means- und PAM-Verfahren ergeben, sind in Tabelle 4.2 zusammengefasst. Anhand der Diagonalen dieser Tabellen kann die Übereinstimmung der Clusterzuordnung

und der wahren Klassenzugehörigkeit abgeschätzt werden. Beim Verfahren von Ward wurden maximal 29 Beobachtungen in die wahren Klassen zugeordnet, bei K-Means 27 und bei PAM 32. Um diese Übereinstimmung nun aber in einer konkreten, interpretierbaren Maßzahl auszudrücken, können externe Indizes aus Kapitel 2.2.2 in Betracht gezogen werden.

Wie bereits erwähnt, gibt es eine Vielzahl dieser Indizes. Da einige sich sehr stark ähneln, soll nun beurteilt werden, welcher Index für eine Benchmarkanalyse der vorliegenden Mircoarray-Daten am besten geeignet ist. Hierfür wurden 19 der von Albatineh et al. (2006) zusammengetragenen Indizes (vgl. Tabelle 2.2) für alle 50 Datensätze ermittelt. Ausgeschlossen wurden lediglich Russel und Rao (RR), Fager und McGowan (FMG) und Pearson (P), da eine Benchmarkanalyse der Clusterindizes von Scherl (2010) ergab, dass diese drei einige Schwächen aufweisen und deshalb unzureichende Resultate liefern.
Zur Berechnung der Indizes wird die Kontingenztabelle von Objektpaaren aus Kapitel 2.2.1 benötigt. Für den Datensatz *leukemia_yagi* und die fünf Verfahren resultieren dann folgende Tabellen:

Single		Y	
		s	v
Cluster	s	654	672
	v	24	28

Complete		Y	
		s	v
Cluster	s	580	602
	v	98	98

Ward		Y	
		s	v
Cluster	s	332	350
	v	346	350

K-Means		Y	
		s	v
Cluster	s	371	395
	v	307	305

PAM		Y	
		s	v
Cluster	s	341	335
	v	337	365

Tabelle 4.3: *Kontingenztabellen der Objektpaare von leukemia_yagi. Dabei beschreibt s die Anzahl an Paare, die im selben Cluster zugeordnet wurden und v die Anzahl an Paare, die verschiedenen Cluster zugeordnet wurden.*

Tabelle 4.4 zeigt die Ergebnisse, der aus den Kontingenztabellen resultierenden Indizes, mit welchen dann die Differenzen der Indizes für jede Verfahrenskombination ermittelt werden können (dargestellt in Tabelle 4.5). Dieses Vorgehen wird nicht nur für den Datensatz *leukemia_yagi* durchgeführt, sondern auch für die restlichen 49 Datensätze. Ziel ist es, herauszufinden welcher Index die größten Differenzen liefert.
Dabei fällt auf, dass für einige Datensätze alle Differenzen gleich Null sind. Das liegt daran, dass bei diesen Daten jedes der fünf Clustering-Verfahren die exakt gleiche Clusterzuordnung ergeben hat. In diesem Fall lassen sich folglich keine Unterschiede in den Indizes erkennen, weswegen die sechs Datensätze *breast_veer*, *colon_watanabe*, *leukemia_bullinger_2*,

Index	Single	Complete	Ward	K-Means	PAM
R	0.49	0.49	0.49	0.49	0.51
H	-0.01	-0.02	-0.01	-0.02	0.02
CZ	0.65	0.62	0.49	0.51	0.50
K	0.73	0.67	0.49	0.52	0.50
MC	0.46	0.35	-0.02	0.03	0.01
PE	0.00	-0.00	-0.01	-0.02	0.02
FM	0.69	0.65	0.49	0.51	0.50
W1	0.49	0.49	0.49	0.48	0.50
W2	0.96	0.86	0.49	0.55	0.50
gamma	0.01	-0.01	-0.01	-0.02	0.02
SS1	0.51	0.50	0.49	0.49	0.51
B1	0.72	0.63	0.49	0.49	0.51
B2	0.00	-0.00	-0.00	-0.00	0.01
J	0.48	0.45	0.32	0.35	0.34
SS2	0.32	0.29	0.19	0.21	0.20
SS3	0.10	0.17	0.24	0.24	0.26
GL	0.66	0.66	0.66	0.66	0.68
RT	0.33	0.33	0.33	0.32	0.34
GK	0.06	-0.02	-0.02	-0.03	0.05

Tabelle 4.4: *Externe Indizes der einzelnen Clusteranalysen von leukemia_yagi.*

Index	S-C	S-W	S-K	S-P	C-W	C-K	C-P	W-K	W-P	K-P
R	0.003	0.000	0.004	-0.017	-0.003	0.001	-0.020	0.004	-0.017	-0.022
H	0.006	0.000	0.009	-0.035	-0.006	0.003	-0.041	0.009	-0.035	-0.044
CZ	0.029	0.164	0.139	0.149	0.135	0.110	0.120	-0.026	-0.015	0.010
K	0.056	0.241	0.213	0.225	0.185	0.157	0.169	-0.028	-0.015	0.012
MC	0.112	0.481	0.426	0.450	0.370	0.315	0.339	-0.055	-0.031	0.024
PE	0.009	0.015	0.022	-0.020	0.006	0.013	-0.029	0.007	-0.035	-0.041
FM	0.042	0.202	0.175	0.186	0.160	0.133	0.144	-0.027	-0.015	0.011
W1	0.003	0.006	0.009	-0.011	0.004	0.006	-0.014	0.002	-0.018	-0.020
W2	0.109	0.475	0.417	0.462	0.366	0.308	0.353	-0.058	-0.013	0.044
gamma	0.019	0.022	0.029	-0.012	0.004	0.011	-0.031	0.007	-0.035	-0.042
SS1	0.013	0.014	0.018	-0.003	0.002	0.005	-0.016	0.003	-0.017	-0.021
B1	0.090	0.221	0.221	0.204	0.131	0.131	0.113	0.000	-0.017	-0.018
B2	0.002	0.004	0.005	-0.005	0.001	0.003	-0.007	0.002	-0.009	-0.010
J	0.031	0.161	0.139	0.148	0.130	0.107	0.117	-0.023	-0.014	0.009
SS2	0.027	0.127	0.111	0.117	0.100	0.084	0.091	-0.016	-0.010	0.007
SS3	-0.070	-0.144	-0.139	-0.161	-0.073	-0.068	-0.091	0.005	-0.017	-0.022
GL	0.003	0.000	0.004	-0.015	-0.003	0.001	-0.018	0.004	-0.015	-0.019
RT	0.003	0.000	0.004	-0.016	-0.003	0.001	-0.018	0.004	-0.016	-0.019
GK	0.082	0.084	0.098	0.015	0.002	0.016	-0.067	0.014	-0.069	-0.083

Tabelle 4.5: *Differenzen der externen Indizes der einzelnen Clusteranalysen von leukemia_yagi. In den Spalten werden die Verfahren mit ihren Anfangsbuchstaben abgekürzt, z.B. S-C: Differenz der Indizes aus **S**ingle-Linkage- und **C**omplete-Linkage-Verfahren.*

lung_wigle, *mixed_chowdary* und *ovarian_li_and_campbell* für die Wahl des Index nicht berücksichtigt wurden. Es stehen somit 44 Datensätze zur Verfügung mit jeweils 10 verschiedenen Differenzen je Index.

Für jeden Index wird nun ein Boxplot betrachtet, um im ersten Schritt zu sehen, in welchem Bereich die Differenzen der 44 Datensätze streuen. Abbildung 4.3 zeigt beispielhaft den Boxplot für den Rand-Index (R). Der mögliche Wertebereich der Differenz liegt hier zwischen -1 und 1. Es ist zu erkennen, dass der Median der Differenzen immer nahe Null liegt. Ein ähnliches Bild ergeben auch die Boxplots der folgenden acht Indizes: Hubert (H), Peirce (PE), Wallace (1) (W1), Gamma (Γ), Sokal und Sneath (1) (SS1), Baulieu (2) (B2), Gower und Legendre (GL) und Rogers und Tanimoto (RT)[1]. Nicht nur der Median, sondern einige der Differenzen liegen nahe bei Null. Unterschiede in den Verfahren lassen sich also nicht ausreichend deutlich mit diesen neun Indizes erkennen, weshalb sie für die Benchmarkanalyse der Clustering-Verfahren nur begrenzt geeignet sind.

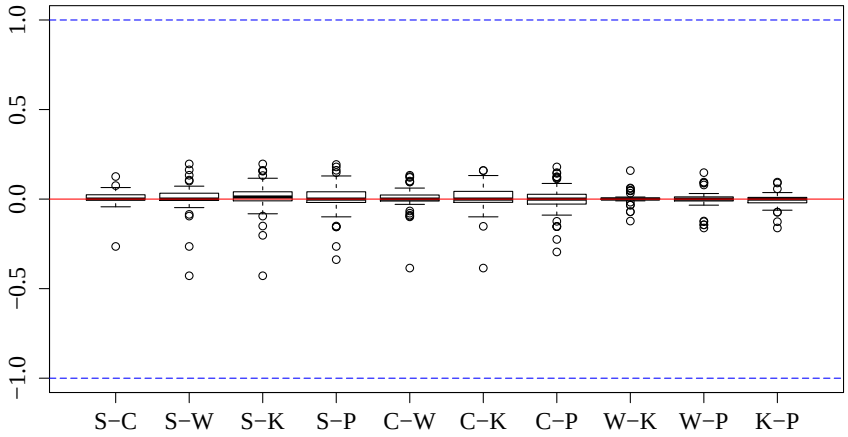

Abbildung 4.3: *Darstellung der Rand-Index-Differenzen zwischen den fünf Clustering-Verfahren. Einbezogen wurden hier 44 der 50 Datensätze, deren Clusterlösungen nicht für alle Verfahren die exakt gleiche Zuordnung ergaben. Die gestrichelte, blaue Linie grenzt den möglichen Wertebereich ab.*

Somit können nun im zweiten Schritt die restlichen 10 Indizes in Abbildung 4.4 genauer betrachtet werden:

Bis auf zwei Indizes von McConnaughey (MC) und Goodman und Kruskal (GK), besitzen alle eine Differenz mit einem möglichem Wertebereich zwischen -1 und 1. Da sich die Tendenz der Boxplots der Indizes Czekanowski (CZ), Kulczynski (K), Fowlkes und Mallows (FM), Baulieu (1) (B1), Jaccard (J) und Sokal und Sneath (2) (SS2) sehr stark ähnelt, wird in Abbildung 4.4 der Boxplot des FM-Index stellvertretend für all diese dargestellt[2].

[1] Die Boxplots hierzu finden sich im Anhang in Abbildung A.1.

[2] Die Boxplots der Indizes, auf deren Darstellung verzichtet wurde, finden sich im Anhang in Abbildung A.2.

Die Differenzen der Indizes MC und GK besitzen einen möglichen Wertebereich zwischen -2 und 2 und werden ebenfalls in Abbildung 4.4 dargestellt.

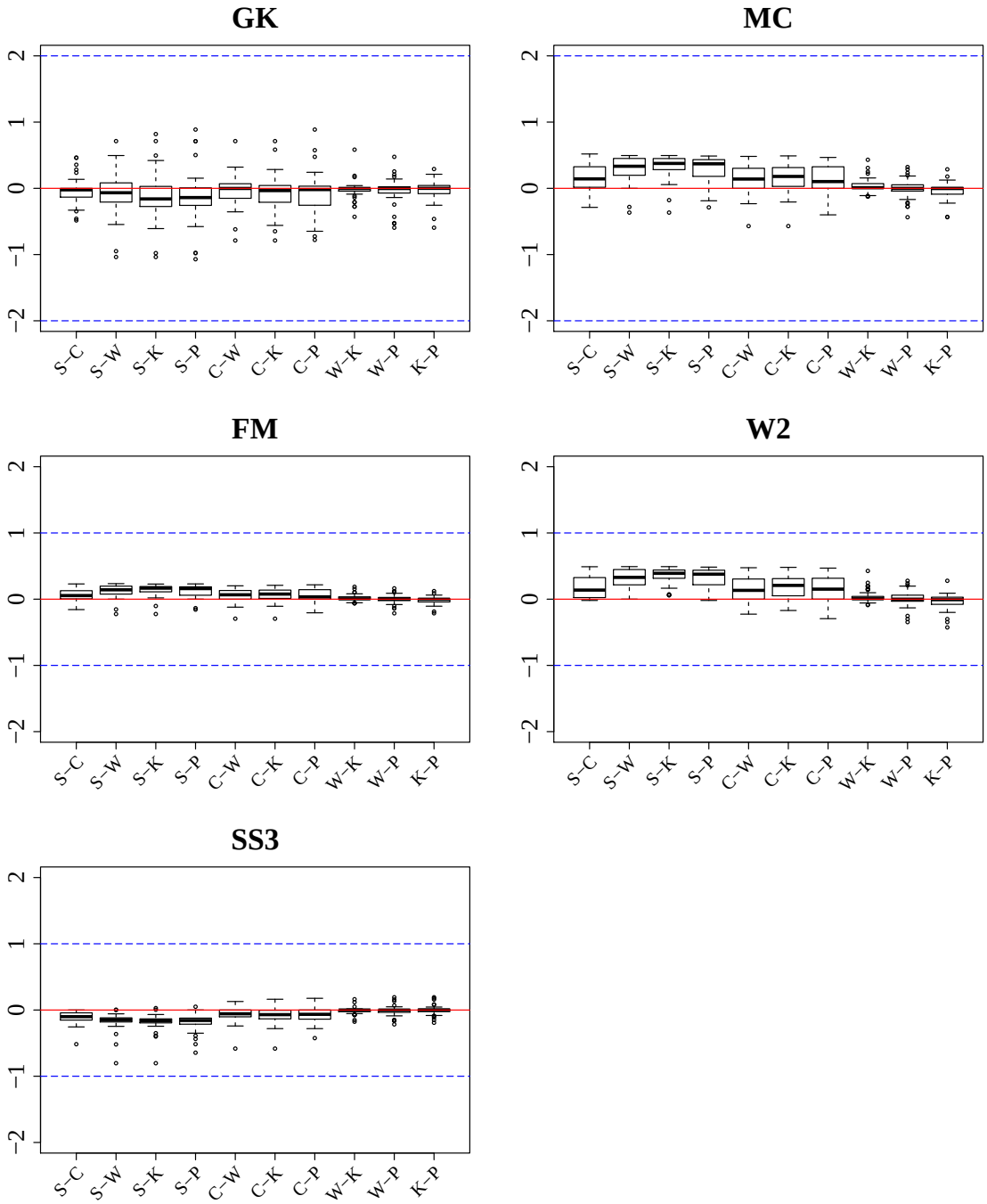

Abbildung 4.4: *Darstellung verschiedener Index-Differenzen zwischen den fünf Clustering-Verfahren. Einbezogen wurden hier 44 der 50 Datensätze, deren Clusterlösungen nicht für alle Verfahren die exakt gleiche Zuordnung ergaben. Die gestrichelte, blaue Linie grenzt den möglichen Wertebereich ab.*

Alle Plots in Abbildung 4.4 haben eine Gemeinsamkeit: die Differenzen zwischen den Verfahren Ward, K-Means und PAM (W-K, W-P, K-P) liegen alle sehr nahe bei Null. Das bedeutet, dass die Validierungsindizes für diese Verfahren jeweils ähnliche Werte ergeben und sich diese drei Verfahren, angewendet auf die 44 Microarry-Datensätze, in ihren Clusterlösungen kaum bis überhaupt nicht unterscheiden. Da die größte Differenz gesucht ist und aus Gründen der Übersichtlichkeit, können also im Weiteren auch nur die restlichen sieben Differenzen betrachtet werden.
Wird der GK-Index mit dem Rand-Index aus Abbildung 4.3 verglichen, treten bei beiden ähnliche Probleme auf. Einige der Mediane liegen sehr nahe bei Null, jedoch nicht alle. Die Streuung der Differenzen wirkt beim GK-Index größer, beachtet man jedoch den möglichen Wertebereich von GK relativiert sich dieser Eindruck. Da alle Boxen der einzelnen Boxplots von GK die Null beinhalten, kann letztendlich dieser Index die möglichen Unterschiede in den Verfahren nicht ausreichend gut verdeutlichen.
Des Weiteren kann allgemein in Abbildung 4.4 erkannt werden, dass die Methodenvergleiche für manche Indizes auch Differenzen liefern, die zwar klein sind, aber nicht ganz Null. Zwei Beispiele hierfür sind die Indizes MC und W2. Vergleicht man diese beiden Indizes, scheinen die Differenzen des W2-Index im Bezug auf den möglichen Wertebereich größer zu sein. Laut Wallace (1983) können allerdings die Werte von W1 und W2 nur gemeinsam zur Interpretation der Ähnlichkeit zweier Clusterpartitionen herangezogen werden, zum Beispiel über den Index von Fowlkes und Mallows (FM), der das geometrische Mittel über W1 und W2 darstellt. W1 wurde bereits im ersten Schritt ausgeschlossen und da der Wert von W2 alleine ebenfalls nicht ausreicht, ist dieser trotz größerer Differenzen nicht für die Benchmarkanalyse der Clustering-Verfahren geeignet.
Zwei Clusterlösungen können theoretisch auch durch Zufall übereinstimmen. Albatineh et al. (2006) stellten fest, dass die Indizes MC und K identisch sind, wenn eine Korrektur dieser zufälligen Übereinstimmung von Partitionen durchgeführt wird (Näheres zur Durchführung und dem theoretischen Hintergrund der Korrektur von Albatineh et al. (2006)). Somit kann einer der beiden im Weiteren vernachlässigt werden. Um die Vergleichbarkeit der Indizes zu verbessern, wird nur noch der K-Index in Betracht gezogen, da er Werte zwischen -1 und 1 annehmen kann. Die Gleichheit nach Korrektur gilt ebenfalls für die Indizes R, H und CZ. Da sich Rand- und Hamann-Index bereits in Schritt 1 als ungeeignet herausgestellt haben, kann auch der Czekanowski-Index (CZ) außer Acht gelassen werden.

Daraufhin verbleiben nun sechs Indizes, die zur Validierung der Clusterlösungen der vorliegenden Microarray-Daten geeignet sind.
Sokal und Sneath (1963) stellen eine mögliche Klassifikation verschiedener Validierungsindizes vor. Danach besteht die fundamentale Formel aller Indizes aus der Anzahl an

Übereinstimmungen dividiert durch einen Term, der die mögliche Anzahl an Vergleichen unterschiedlich einbezieht. Die Einteilung basiert auf der Zusammensetzung von Zähler und Nenner der Indizes. Dabei wird im Zähler nur berücksichtigt, ob die Anzahl an Objektpaaren, die in beiden Clusterlösungen unterschiedlich zugeordnet wurden ($\hat{=}\ d$), einbezogen wurde oder nicht. Fünf der sechs verbleibenden Indizes wurden demnach wie in Tabelle 4.6 zusammengefasst, dabei ist Baulieu (B1) entsprechend ergänzt worden. Dieser Index stellt auch eine Besonderheit dar, da er im Vergleich zu allen anderen, bereits im Zähler eine umfangreichere Berechnung vornimmt. Außerdem misst dieser Index die Unähnlichkeit zweier Clusterlösungen (Baulieu; 1989), weswegen sich die Interpretation von allen anderen unterscheidet: Je niedriger der Wert des Index, desto ähnlicher sind sich zwei Partitionen.

	Zähler: Anzahl an Objektpaaren, die in beiden Clusterlösungen unterschiedlichen Clustern zugeordnet wurden ($\hat{=}\ d$)	
Nenner	ausgeschlossen	einbezogen
gleiche Gewichte auf den Objektpaaren, egal ob sie in gleiche oder verschiedenen Clustern zugeordnet wurden	Jaccard, $J = \frac{a}{a+b+c}$	
Objektpaare, die verschiedenen Clustern zugeordnet wurden, werden doppelt gewichtet	Sokal und Sneath (2), $SS2 = \frac{a}{a+2(b+c)}$	
Randverteilungen (arithmetisches Mittel)	Kulczynski, $K = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$ *als arithmetisches Mittel von W1 und W2*	
Randverteilungen (geometrisches Mittel)	Fowlkes und Mallows, $FM = \frac{a}{\sqrt{(a+b)(a+c)}}$ *als geometrisches Mittel von W1 und W2*	Sokal und Sneath (3), $SS3 = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
Anzahl an Objektpaaren		Baulieu (1), $B1 = \frac{\binom{m}{2}^2 - \binom{m}{2}(b+c) + (b-c)^2}{\binom{m}{2}^2}$

Tabelle 4.6: *Einteilung der verschiedenen Indizes zur Validierung von Clusterlösungen nach Sokal und Sneath (1963), um B1 ergänzt.*

4.5 Bootstrap-Konfidenzintervalle

Nun wurden für alle sechs Index-Differenzen Bootstrap-Konfidenzintervalle mithilfe der R-Funktionen *boot* und *boot.ci* aus dem Package *„boot“* aufgestellt.
Aus der Nullhypothese in Kapitel 3.1 folgt, dass der interessierende Parameter der Benchmarkanalyse von Clustering-Verfahren der Erwartungswert der Differenz zweier Indizes VI_1 und VI_2 ist, somit gilt

$$\theta = \mathbb{E}(VI_1 - VI_2) = \mathbb{E}(\Delta VI). \tag{4.1}$$

Dabei stellen VI_1 und VI_2 Vektoren dar, die jeweils die Validierungsindizes für alle 50 Datensätze beinhalten. Ebenso ist auch ΔVI ein Vektor mit Länge 50, dieser enthält also die Differenzen ΔVI_i für jeden der 50 Datensätze.
Als Schätzfunktion für den Parameter in 4.1 wird das arithmetische Mittel über alle 50 Datensätze gebildet, also

$$\hat{\theta} = s(\Delta VI) = \frac{1}{50}\sum_{i=1}^{50} \Delta VI_i. \tag{4.2}$$

Es wurden, wie von DiCiccio und Efron (1996) empfohlen, 2000 Bootstrap-Replikationen $\hat{\theta}^*$ gebildet und mit diesen ein BC_a-95%-Konfidenzintervall aufgestellt.

Für ein BC_a-Intervall werden Schätzungen der Parameter a und z_0 benötigt. Diese Schätzfunktionen wurden entsprechend Kapitel 3.2.2 aufgestellt und daraufhin α_{adj} berechnet. Damit konnten dann die jeweiligen Perzentile der geordneten Liste aller Bootstrap-Replikationen als Intervallgrenzen bestimmt werden.
Da $B \cdot \alpha_{adj}$ meist keiner ganzen Zahl entspricht, werden die Werte in diesen Fällen durch eine lineare Interpolation auf Basis der Standardnormalverteilungsquantile ermittelt. Davison und Hinkley (1997, S. 195) wählen hierfür $k = \lfloor (B+1)\alpha_{adj} \rfloor$, was der größten ganzen Zahl kleiner als $(B+1)\alpha_{adj}$ entspricht und definieren die Intervallgrenze folgendermaßen:

$$\hat{\theta}_{BC_a}(\alpha) = \hat{\theta}^*_{(B\alpha_{adj})} = \hat{\theta}^*_{(k)} + \frac{\Phi^{-1}(\alpha_{adj}) - \Phi^{-1}(\frac{k}{B+1})}{\Phi^{-1}(\frac{k+1}{B+1}) - \Phi^{-1}(\frac{k}{B+1})} \left(\hat{\theta}^*_{(k+1)} - \hat{\theta}^*_{(k)}\right). \tag{4.3}$$

Dabei sind $\hat{\theta}^*_{(k)}$ und $\hat{\theta}^*_{(k+1)}$ die k- bzw. $(k+1)$-ten Werte aus der geordneten Liste aller Bootstrap-Replikationen.

Abbildung 4.5 zeigt die Konfidenzintervalle, die sich daraufhin für alle sechs Validierungsindizes ergeben. Zur besseren Lesbarkeit wurde auf der x-Achse ein Bereich von -0.5 bis 0.5 gewählt, es sei allerdings darauf hingewiesen, dass alle Differenzen einen möglich Wertebereich von -1 bis 1 besitzen. Auf der y-Achse ist abzulesen, welche Differenz jeweils in

der entsprechenden Zeile aufgetragen ist. Hier sind die Verfahren mit ihren Anfangsbuchstaben abgekürzt, C−P stellt also zum Beispiel die Differenz der Validierungsindizes aus Complete-Linkage- und PAM-Verfahren dar.

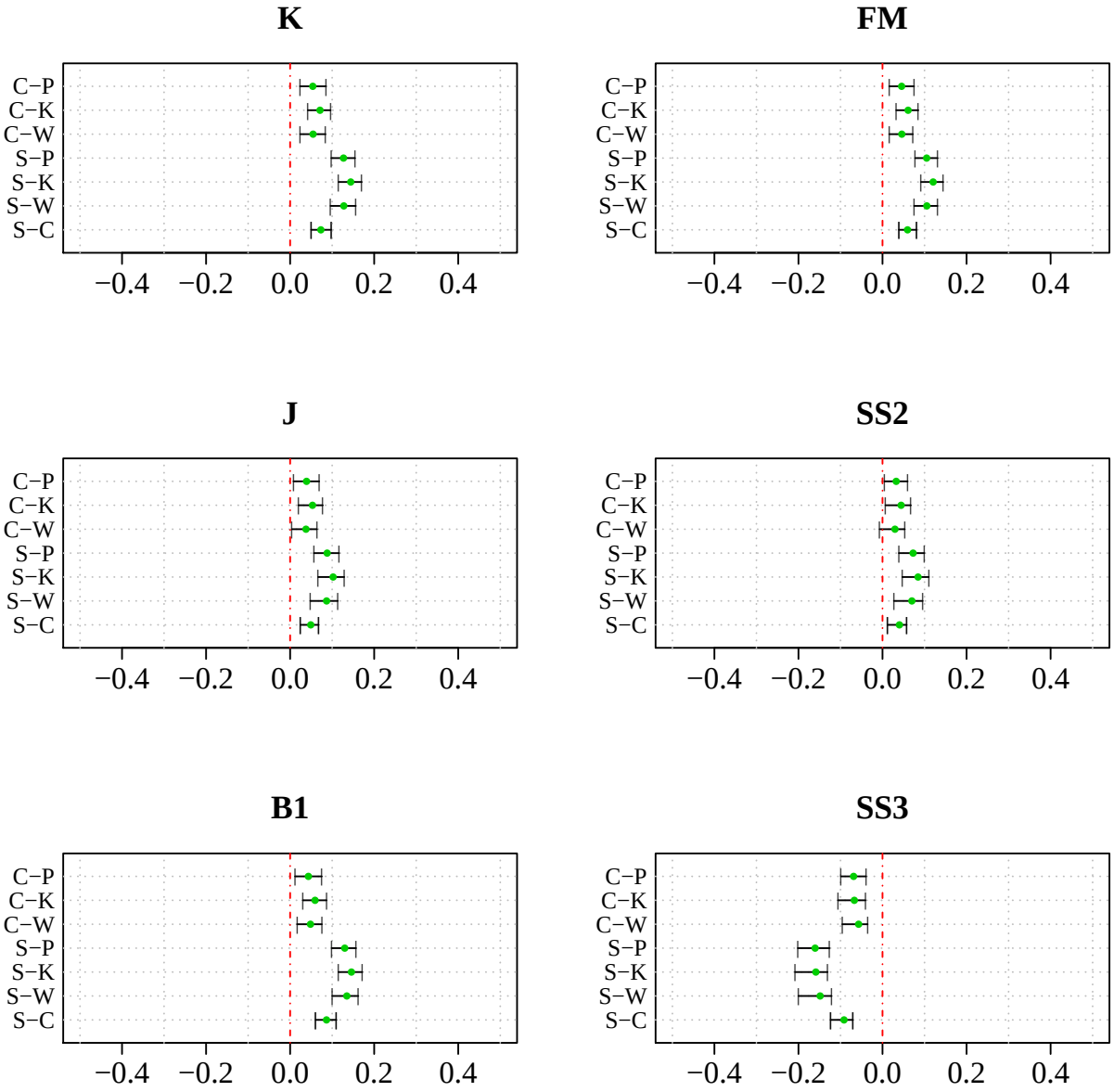

Abbildung 4.5: *Darstellung der BC_a-Konfidenzintervalle der Differenz für sechs Validierungsindizes. Es wurden 2000 Bootstrap-Replikationen gebildet. Der grüne Punkt kennzeichnet jeweils den geschätzten Wert der Differenz aus der ursprünglichen Stichprobe.*

Mit diesen Intervallen soll nun die Nullhypothese aus Kapitel 3.1 überprüft werden. Das heißt, ist die 0 im Konfidenzintervall enthalten, kann kein signifikanter Unterschied zwi-

schen den Erwartungswerten zweier Validierungsindizes festgestellt werden. Beinhaltet es Werte < 0, bedeutet dies einen signifikanten Unterschied zwischen zwei Clustermethoden und Methode 2 liefert bessere Ergebnisse als Methode 1. Im Umkehrschluss bedeuten Werte > 0, dass Clustermethode 1 gegenüber Clustermethode 2 zu bevorzugen ist.

Eine Ausnahme in der Auswertung stellt der Index B1 dar. Da er im Gegensatz zu allen anderen nicht die Ähnlichkeit, sondern die Unähnlichkeit misst, ändert sich die Interpretation des Konfidenzintervalls. Werte > 0 sprechen dabei für Clustermethode 2 und Werte < 0 für Clustermethode 1.

4.6 Interpretation der Ergebnisse

Betrachtet man die Ergebnisse der sechs Indizes in Abbildung 4.5, fallen keine große Unterschiede zwischen den einzelnen Indizes auf. Die einzige Ausnahme bildet hier der Index SS3, welcher für die Differenzen Konfidenzintervalle besitzt, die Werte deutlich kleiner als 0 beinhalten. In diesem Fall werden die Verfahren PAM, K-Means und Ward gegenüber einer hierarchischen Clusterung mit Single- oder Complete-Linkage bevorzugt. Ebenso wird ein Complete-Linkage gegenüber einem Single-Linkage bevorzugt.

Eine zusätzliche, nicht offensichtliche Ausnahme bildet der Index B1. Wie bereits erwähnt, ändert sich hier die Interpretation und Werte > 0 sprechen für Clustermethode 2. Womit sich für jeden Vergleich die gleiche Entscheidung wie bei dem Index SS3 ergibt.

Die Indizes K, FM und J liefern alle Konfidenzintervalle > 0, weshalb man sich entsprechend anders wie bei B1 und SS3 entscheidet: Die hierarchischen Clusterungen mit Single- und Complete-Linkage werden gegenüber den Verfahren PAM, K-Means und Ward bevorzugt. Und der Single-Linkage liefert „bessere“ Partitionen als der Complete-Linkage.

Fast gleiches gilt bei der Analyse mit dem Index SS2, allerdings kann beim Vergleich von Complete-Linkage und Ward-Verfahren kein signifikanter Unterschied festgestellt werden, da dieses Konfidenzintervall die 0 beinhaltet.

Anders als erwartet, wird also eine Clusteranalyse mit dem Single-Linkage-Verfahren mit den Indizes K, FM, J und SS2 als bessere Methode identifiziert, obwohl aufgrund der Verkettungseigenschaft oft keine sinnvollen Cluster gebildet wurden (siehe Kapitel 4.3). Eine Erklärung hierfür liefert beispielsweise Tabelle 4.3: Wenn eins der beiden gebildeten Cluster nur sehr wenige Objekte beinhaltet, ist dementsprechend die Anzahl an Objektpaaren in verschiedenen Clustern für dieses Verfahren sehr gering. Dadurch fallen die

Werte für c und d in den Formeln der Indizes (vgl. Kapitel 2.2) im Vergleich zu a und b ebenfalls klein aus. Da die angesprochenen vier Indizes allerdings nur die Anzahl der mit beiden Methoden gleich zugeordneten Objektpaare ($\hat{=}\ a$) im Zähler berücksichtigen, fallen die Indizes für Partitionen mit sehr ungleichen Clustergrößen auch entsprechend größer aus. Dass diese Indizes für den Single-Linkage die größten Werte liefern, ist auch für den Beispieldatensatz *leukemia_yagi* in Tabelle 4.4 zu erkennen. Ob wirklich eine gute Übereinstimmung mit der wahren Klassenzugehörigkeit Y gegeben ist, ist dabei fraglich. Daher scheint die Anwendung des B1- oder SS3-Index für diese Datensituation angemessener. Beide Indizes berücksichtigen dabei nicht nur a im Zähler, sondern auch die Anzahl an Objektpaaren, die in beiden Clusterlösungen unterschiedlichen Clustern zugeordnet wurden ($\hat{=}\ d$).

Kapitel 5

Zusammenfassung und Ausblick

Das Ziel der Arbeit war es, ein Framework auszuarbeiten, mit dem aus zwei Clustermethoden die bessere identifiziert werden kann. Allgemein ist bei solchen Benchmarkanalysen darauf zu achten, auch reelle Datensätze zu berücksichtigen, da diese meist nicht einer einfachen gemeinsamen Verteilung folgen, wie es bei Simulationsstudien der Fall ist (Boulesteix et al.; 2015). Daher wurde die hier vorliegende Benchmarkanalyse an 50 reellen Datensätzen aus klinischen Krebsstudien durchgeführt.
Der jeweilige Validierungsindex VI einer Clusterlösung bildete dabei das Hauptgütekriterium, anhand dessen eine Entscheidung für oder gegen eine Methode getroffen werden kann. Da das Interesse dabei nicht nur bei einem Datensatz liegt, wurde auf Grundlage von diesen Validierungsindizes zweier Clusteranalyseverfahren ein Hypothesentest für alle betrachteten Datensätze formuliert.
Nachdem fünf verschiedene Clusteranalyseverfahren auf die Datensätze Anwendung fanden, wurden die Validierungsindizes genauer untersucht. Das war nötig, da eine große Anzahl an verschiedenen Validierungsindizes existiert, die alle auf dem selben Grundprinzip beruhen, allerdings unterschiedlich berechnet werden und daraufhin verschiedene Ergebnisse liefern. Es ergab sich, dass von den 22 in Betracht gezogenen Indizes nur sechs überhaupt Unterschiede zwischen zwei Verfahren erkennen ließen. Der Großteil der anderen Indizes ergab Index-Differenzen nahe Null. Das bedeutet für die Benchmarkanalyse, dass mit diesen Indizes für kein Clusterverfahren Vorteile gegenüber einem anderen identifiziert werden konnten. Allerdings besteht auch die Möglichkeit, dass sich die Clusterlösungen der Verfahren zu stark ähnelten und daher tatsächlich keine Unterschiede vorhanden waren. Deshalb wird auch darauf hingewiesen, dass durch die Wahl des größten Validierungsindex der Unterschied zwischen zwei Verfahren möglicherweise überschätzt wird und die Anwendung auf eine andere Datensituation auch andere Ergebnisse liefern kann.
Um den interessierenden Parameter $\mathbb{E}(\Delta VI)$ und dessen Streuung zu schätzen wurden anschließend BC_a-Bootstrap-Intervalle berechnet. Diese kamen zum Einsatz, da die Differenzen der Indizes nicht normalverteilt sind und daher keine Standard-Konfidenzintervalle

verwendet werden konnten.
Für die vorliegenden Datensätze lieferten daraufhin die Indizes von Baulieu (B1) und Sokal und Sneath (SS3) angemessene Validierungen und die Clusterverfahren PAM, K-Means und Ward wurden gegenüber einer hierarchischen Clusterung mit Single- oder Complete-Linkage als „besser“ identifiziert.

Als mögliche Erweiterung der bisherigen Analysen kann eine Korrektur der zufälligen Übereinstimmung von Partitionen in Betracht gezogen werden. Laut Albatineh et al. (2006) erhält dadurch die Wahl des Validierungsindex weniger Gewicht, da sich die Indizes daraufhin einander angleichen oder sogar identische Werte annehmen.
In dieser Arbeit wurden nur fünf verschiedene Clusterverfahren angewendet, weshalb eine Erweiterung der zu betrachtenden Clusteranalyseverfahren dabei natürlich auch denkbar ist.
Da die Struktur der Daten auch oftmals mehr als zwei Klassen beinhaltet, könnten auch Clusteranalysen mit mehr als zwei gesuchten Cluster durchgeführt werden und mit der Klassenzuordnung durch die Zielvariable verglichen werden. Im Gegensatz dazu könnte aber auch der reine Vergleich zweier erhaltenen Clusterlösungen mit Hilfe eines Validierungsindex sehr interessante Ergebnisse liefern. Denn laut Bacher et al. (2010) können Clusterlösungen bei solch einem Vergleich erst als brauchbar erachtet werden, wenn zum Beispiel der Rand-Index einen Wert über 0.7 annimmt. Möglicherweise sollten für andere Indizes ebenfalls bestimmte Grenzwerte Berücksichtigung finden.

Literaturverzeichnis

Albatineh, A. N., Niewiadomska-Bugaj, M. und Mihalko, D. (2006). On similarity indices and correction for chance agreement, *Journal of Classification* **23**(2): 301–313.
URL: *http://dx.doi.org/10.1007/s00357-006-0017-z*

Bacher, J., Pöge, A. und Wenzig, K. (2010). *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*, Oldenbourg, München.

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients, *Journal of Classification* **6**(1): 233–246.
URL: *http://dx.doi.org/10.1007/BF01908601*

Boulesteix, A.-L., Hable, R., Lauer, S. und Eugster, M. (2015). A statistical framework for hypothesis testing in real data comparison studies, *The American Statistician* .
URL: *http://dx.doi.org/10.1080/00031305.2015.1005128*

Boulesteix, A.-L., Strobl, C., Augustin, T. und Daumer, M. (2008). Evaluating microarray-based classifiers: An overview, *Cancer Informatics* **6**: 77–97.
URL: *http://www.la-press.com/evaluating-microarray-based-classifiers-an-overview-article-a577*

Causton, H. C., Quackenbush, J. und Brazma, A. (2003). *Microarray Gene Expression Data Analysis: A Beginner's Guide*, Wiley-Blackwell, Malden.

Davison, A. C. und Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge New York.

DiCiccio, T. J. und Efron, B. (1996). Bootstrap confidence intervals, *Statistical Science* **11**(3): 189–228.
URL: *https://projecteuclid.org/euclid.ss/1032280214*

Efron, B. und Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Fahrmeir, L., Künstler, R., Pigeot, I. und Tutz, G. (2011). *Statistik - Der Weg zur Datenanalyse*, Springer, Berlin.

Hothorn, T., Leisch, F., Zeileis, A. und Hornik, K. (2005). The design and analysis of benchmark experiments, *Journal of Computational and Graphical Statistics* **14**(3): 675–699.
URL: *http://dx.doi.org/10.1198/106186005X59630*

Kaufman, L. und Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis*, Wiley, Hoboken, N.J.

Kaufmann, H. und Pape, H. (1996). Clusteranalyse, *in* L. Fahrmeir, A. Hamerle und G. Tutz (eds), *Multivariate statistische Verfahren*, Walter de Gruyter, Berlin.

Scherl, M. (2010). *Benchmarking of cluster indices*, Diploma thesis, LMU Munich.
URL: *http://epub.ub.uni-muenchen.de/12797/*

Sokal, R. R. und Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*, W.H. Freeman, San Francisco.

Wallace, D. L. (1983). A method for comparing two hierarchical clusterings: Comment, *Journal of the American Statistical Association* **78**(383): 569–576.
URL: *http://dx.doi.org/10.1080/01621459.1983.10478008*

ANHANG

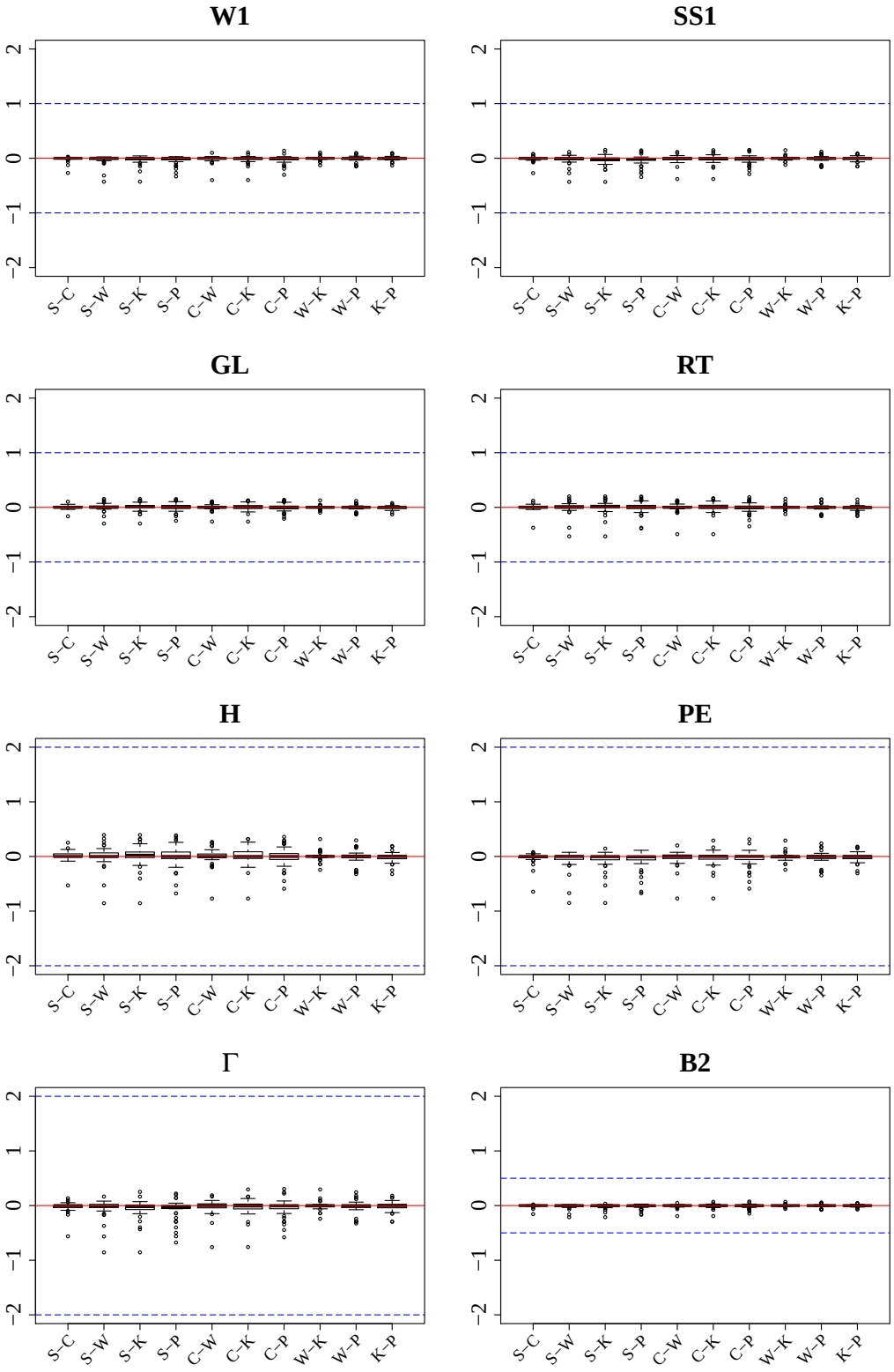

Abbildung A.1: *Darstellung verschiedener Index-Differenzen zwischen den fünf Clustering-Verfahren die stark nahe 0 streuen. Einbezogen wurden hier 44 der 50 Datensätze, deren Clusterlösungen nicht für alle Verfahren die exakt gleiche Zuordnung ergaben. Die gestrichelte, blaue Linie grenzt den möglichen Wertebereich ab.*

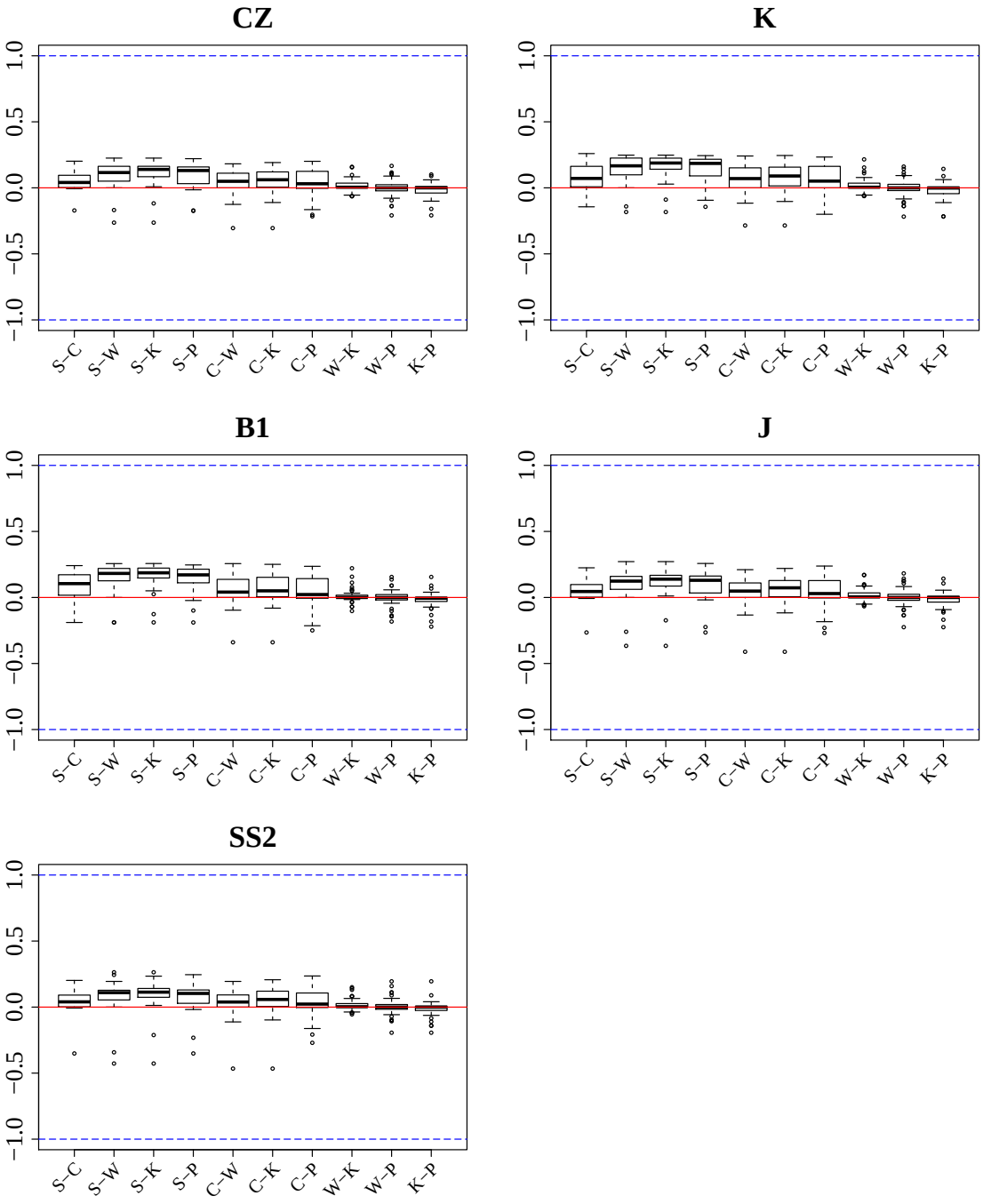

Abbildung A.2: *Darstellung verschiedener Index-Differenzen zwischen den fünf Clustering-Verfahren die tendenziell dem Index FM ähneln. Einbezogen wurden hier 44 der 50 Datensätze, deren Clusterlösungen nicht für alle Verfahren die exakt gleiche Zuordnung ergaben. Die gestrichelte, blaue Linie grenzt den möglichen Wertebereich ab.*

Elektronischer Anhang

Der elektronische Anhang umfasst die folgenden Ordner:

- Daten: Beinhaltet zwei Unterordner
  - datasets: Enthält 65 Microarray-Datensätze im txt-Format.
  - data_R: Enthält noch keine Daten. Er wird beim Importieren der 65 Datensätzen benötigt, um diese im RData-Format dort abzuspeichern.
- Ergebnisse: Beinhaltet zwei Unterordner
  - Clusteranalyse: Enthält für jedes der fünf Clustering-Verfahren eine RData-Datei, die für jeden Datensatz die berechneten 19 Validierungsindizes auflistet.
  - Grafiken: Enthält die Abbildungen, die in dieser Arbeit verwendet wurden als pdf-Datei.
- Programme: Beinhaltet die Syntax-Dateien der Statistiksoftware R, mit denen die Ergebnisse dieser Arbeit reproduziert werden können.

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Arbeit war in dieser oder ähnlicher Form noch nicht Bestandteil einer Prüfungsleistung.

München, 18. August 2015

Myriam Hatz