Technische Universität Darmstadt
Fachbereich Informatik
Fachgebiet Knowledge Engineering
Prof. Dr. Johannes Fürnkranz

Analyse von Heuristiken zur Evaluierung von Assoziationsregeln

Diplomarbeit

Florian Alexander Nattermann

Betreuung

Prof. Dr. Johannes Fürnkranz

Dipl.-Inf. Jan-Nikolas Sulzmann

Januar 2009

Hiermit versichere ich, die vorliegende Diplomarbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus den Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, 23. Januar 2009

Florian Alexander Nattermann

Inhaltsverzeichnis

1. Einleitung

Knowledge Engineering ist die Wissenschaft zur Entwicklung von wissensbasierten Systemen. Ein wissensbasiertes System entsteht durch logische und/oder statistische Analyse (z.B. Data Mining) von Daten mit dem Ziel, eine möglichst genaue Vorhersage für zukünftige Daten zu liefern. Mit dem Begriff des "Data Mining" werden Techniken bezeichnet, um in Daten Muster bzw. Regeln zu erkennen. Das Wort "Mining" (de.: graben) bedeutet, dass nach Informationen gesucht werden soll, die in den Daten vergraben sind. Dazu werden Techniken des "maschinellen Lernens" verwendet bzw. entwickelt.

Im Forschungsbereich "Maschinellen Lernen" werden autonome Programme (Algorithmen) entwickelt um große bestehende Datenmengen auszuwerten [1]. Neben der Entwicklung von Algorithmen beschäftigt man sich beim maschinellen Lernen deshalb auch mit psychologischen und philosophischen Fragen, bzgl. der Wahrnehmung und Denkweise des Menschen in Relation zum Computer.

Folgender Auszug aus [Mit97] vermittelt kurz die vielseitige Forschung des maschinellen Lernens:

> "Machine learning is inherently a multidisciplinary field. It draws on results from artificial intelligence, probability and statistics, computational complexity theory, control theory, information theory, philosophy, psychology, neurobiology, and other fields."
>
> Tom M. Mitchell

1.1. Motivation und Ziel der Diplomarbeit

Beim Auswerten von Daten aus Kaufhäusern, Supermärkten etc. versucht man Relationen zwischen verschiedenen Artikeln (Daten) zu finden. In diesem Fall sind die Relationen gerichtet, dass bedeutet, es gibt Artikel, deren Kauf bedingt den Kauf eines anderen Artikels [RJBA99]. Diese Relationen kann man mit Hilfe von Assoziationsregeln beschreiben.

Zur Vorhersage von Ereignissen auf Grund von vorher gesammelten Daten werden Heuristiken verwendet. Eine Heuristik hat abhängig von den Daten Vor- und Nachteile. Für

[1] Mit dem Bergriff des maschinellen Lernens wird oft auch der Begriff "künstliche Intelligenz" in Verbindung gebracht. Dieser Begriff ist bis heute nicht genau definiert und hat nur eine symbolische Bedeutung für ein Forschungsfeld der Informatik.

bestimmte Ereignisse gibt es darauf abgestimmte Heuristiken - es gibt keine universelle Heuristik. Eine Zusammenstellung verschiedener Heuristiken ist in [Jan06] nachzulesen.

Beim Auswerten von Assoziationsregeln werden werden die Heuristiken Support und Confidence verwendet. Support und Confidence stellen statistische Eigenschaften der Assoziationsregeln dar. In der Auswertung einer Regel gibt es 4 Werte, um zu beschreiben wie viele Datensätze von einer Regel abgedeckt werden. Diese 4 Werte bilden die Konfusionsmatrix [FF05]. Support und Confidence verwenden nur 2 Werte [2] dieser Tabelle. Es gibt andere Algorithmen, welche wesentlich mehr Eigenschaften der Konfusionmatrix betrachten und alle 4 Werte verwenden.

Man kann nicht immer davon ausgehen, dass die zum Teil relativ umfangreichen Daten, aus welchen die Support- und Confidencewerte stammen, für alle Zeit existieren. Deshalb ist es wichtig, Gesetzmäßigkeiten zu finden, um von Support- und Confidnecewerten auf andere Heuristikewerte schließen zu können.

Das Ziel dieser Diplomarbeit ist es, zu prüfen, ob es nur mit den Informationen Support und Confidence möglich ist, den Wert einer anderen Heuristik zu bestimmen. Dazu müssen die verwendeten Heuristiken **analysiert** und danach im **Assoziationsraum** dargestellt (**evaluiert**) werden.

Der Assoziationsraum ist eine graphische Darstellung der Support- und Confidenceheuristiken, welche eine analoge Betrachtungsweise der Heuristiken bietet wie der in [FF05] dargestellte P-N-Raum.

In den folgenden zwei Kapiteln werden dazu die Grundlagen des Knowledge Engineering vorgestellt. Es wird der Assoziationsraum eingeführt und es werden spezielle Eigenheiten dieses Raumes erläutert. In den Kapiteln 4 bis 9 werden einzeln die Heuristiken Accuarcy, Conviction, Lift, Leverage, Phi-Koeffizient und Klösgen untersucht und deren Darstellung im Assoziationsraum analysiert. Im Abschlusskapitel werden die Ergebnisse aus allen Kapiteln zusammengetragen.

[2] bzw. 3 Werte wenn man die Summe aller Einträge hinzuzählt

2. Grundlagen

2.1. Assoziationsregeln für Datenbanken

Die nachfolgende Tabelle 2.1 zeigt ein Beispiel für eine sog. Datenbank. Sie gibt in den Spalten Artikel an, die in einem Supermarkt von verschiedenen Kunden (Zeilen der Tabelle) gekauft wurden. Die Spaltenüberschriften werden als "Attribute" bezeichnet und die Werte in den Spalten als "Attributinformationen". Jede Zeile der Datenbank wird als "Eintrag" bezeichnet und ist von allen anderen Einträgen unabhängig.

	Milch	**Brot**	**Chips**	**Pizza**	**Bier**
1	ja	nein	nein	ja	nein
2	nein	nein	ja	nein	ja
3	nein	ja	ja	nein	ja
4	ja	ja	ja	nein	ja
5	ja	ja	nein	ja	nein
:	:	:	:	:	:
N	nein	ja	ja	ja	nein

Tabelle 2.1.: Beispiel einer Datenbank

Das Interesse des Supermarktes ist es, Regale entsprechend den Kaufgewohnheiten von Kunden anzuordnen. Wenn man z.B. feststellt, dass "Chips-und-Pizza-Kunden" häufig auch Bier kaufen, wird man das Bierregal in die Nähe der Chips- und Pizza-Regale aufstellen. Ziel ist es, Korrelationen zu finden. Eine solche Korrelation kann als eine "Assoziationsregel" ([RJBA99])

$$R: \ A \rightarrow B \tag{2.1}$$

beschrieben werden. A wird als Körper bezeichnet und B als Kopf. Der Kopf B ist in diesem Fall ein Wert des "Klassifikationsattributs" Bier (ja oder nein [1]). Der Körper A ist eine logische Verknüpfung (Konjunktion) einzelner Attributwerte (in der oberen Tabelle: Milch, Brot, Chips und Pizza).

Die Regel R definiert eine Abbildungsvorschrift: Sie macht Vorhersagen über das Kundenverhalten. Die Vorhersage kann zutreffen oder auch nicht. Das Interessante ist nun, wie häufig die Vorhersagen zutreffen: Wieviele Einträge der Datenbank erfüllt die Regel?

[1] Es gibt Klassifikationsattribute die nicht binär sind.

Definition 1. *Ein Eintrag einer Datenbank erfüllt eine Regel $A \rightarrow B$, wenn in den jeweiligen Attributen die selben Werte stehen wie in A und B.*

Beispiel 1. *Der erste Eintrag in der Datenbank 2.1 erfüllt folgende Regel:*

$$R: \ (\textit{Milch: ja}) \ \wedge \ (\textit{Chips: nein}) \ \wedge \ (\textit{Pizza: ja}) \ \rightarrow \ (\textit{Bier: nein}) \quad ,$$

der zweite Eintrag aber nicht.

2.2. Der P - N - Raum

Oben wurde gezeigt, dass jede Regel eine bestimmte Menge an Datensatzeinträgen erfüllt. Danach können verschiedene Regeln unterschiedliche Genauigkeiten haben, man spricht hier von der Güte einer Regel. Zur Definition eines Gütemaßes wird nach [FF05] ein zweidimensionaler P-N-Raum definiert, der durch die Basen P und N aufgespannt wird.

- P ("positive Einträge"): Die Anzahl der Datensätze in der Datenbank, bei welchen der Wert des Klassifikationsattributs mit dem Wert von B übereinstimmt.
- N ("negative Einträge"): Die Anzahl der Datensätze in der Datenbank, bei welchen der Wert des Klassifikationsattributs nicht mit dem Wert von B übereinstimmt.

Hat man eine Regel $A \rightarrow B$ auf einem Datensatz gelernt, dann soll die Regel aus den Eingaben A die Vorhersage B treffen. Eine Regel muss aber nur eine Klassifikation bestimmen, da angenommen werden kann, dass, wenn die Regel nicht zutrifft, die Gegenannahme zutrifft. Aus diesem Grund ist es sinnvoll, dass B konstant wahr(ja) oder falsch(nein) ist, im weiteren Verlauf gilt: $B =$ wahr.

Tabelle 2.2 zeigt die sog. "Konfusionsmatrix". Sie zeigt die Trefferquote einer Regel für einen Datensatz. Die erste Spalte steht für die Anzahl der Datensätze, für die der Körper der Regel eine "positive" Vorhersage macht, die zweite Spalte für die Anzahl der Datensätze, für die der Körper eine "negative" Vorhersage macht.
Die Zeile (P), ist die Anzahl aller Datensatzeinträge in der das Klassifikationsattribut die gesuchte Klassifikation (positiv) hat. Folglich bildet die untere Zeile (N) die Summe der restlichen Datensatzeinträge (negativ).

Daher startet man mit einer leeren Konfusionsmatrix und einer Regel, überprüft jeden Eintrag und erhöht die entsprechende Zelle der Konfusionsmatrix. Am Ende hat man die "Abdeckung" der Regel auf dem Datensatz. Jede Regel $R_i: \ A_i \rightarrow B_i$ stellt dann einen Punkt im P-N-Raum dar:

	positiv vorhergesagt	negativ vorhergesagt	
tatsächlich positiv	true positives (p)	false negatives $(P - p)$	P
tatsächlich negativ	false positives (n)	true negatives $(N - n)$	N
	$p + n$	$p + n - (N + P)$	$P + N$

Tabelle 2.2.: Konfusionsmatrix

- auf der x-Achse die Anzahl an Einträgen in der Datenbank, für die A wahr wird, B aber nicht (in der Tabelle n)
- und auf der y-Achse die Anzahl an Einträgen in der Datenbank, für die A und B wahr werden (in der Tabelle p).

Zur Veranschaulichung der Korrelation zwischen Körper A und Kopf B der Regel wird fortan die Kontingenztabelle [TKS02] verwendet.

	B	$\overline{B}$	
A	C0	C1	
$\overline{A}$	C2	C3	
			N

Tabelle 2.3.: Kontingenztabelle

Es folgt, dass der Wert p gleich dem Wert C0 ist und der Wert n gleich dem Wert C1.

- C0: Anzahl der Datenbankeinträge, bei denen sowohl A als auch B wahr wird.
- C1: Anzahl der Datenbankeinträge, bei denen A wahr wird, B aber nicht.
- C2: Anzahl der Datenbankeinträge, bei denen B wahr wird, A aber nicht.
- C3: Anzahl der Datenbankeinträge, bei denen sowohl A als auch B nicht wahr wird.

2.3. Eigenschaften von Assoziationsregeln

Wie im vorangegangenen Abschnitt gezeigt wurde, liefern Regeln Aussagen über die Einträge einer Datenbank. Soll zwischen zwei Regeln R_1 und R_2 unterschieden werden, sind

diesen Regeln Zahlen (entsprechend ihren Gütewerten [2]) zuzuordnen. Zu diesem Zweck dienen "Heuristiken".

Definition 2. *Eine Heuristik ist eine Abbildung* $h(R) \rightarrow w$ *, mit* $R \in$ *Regelmenge und* $w \in \mathbb{R}$

2.3.1. Assoziationsregel - Support

Die erste Heuristik ist der Support einer Regel. Man betrachte erneut die Darstellung der Regel $A \rightarrow B$ in Tabelle 2.3. Der "Support" einer Regel ist die Wahrscheinlichkeit, mit der ein realer Eintrag in der Datenbank mit der Vorhersage aus der Regel übereinstimmt, also gleich dem Quotienten aus der Anzahl aller Einträge, welche die Regel erfüllen (A und B wahr), zur Anzahl aller Einträge der Datenbank.

$$support(A \rightarrow B) \;=\; P(A,B) \;=\; \frac{C0}{N} \tag{2.2}$$

Der Ausdruck $P(X)$ stellt dabei die Wahrscheinlichkeit des Ereignisses X dar. In dieser Formel ist das Ereignis X das gleichzeitige eintreten der Ereignisse A und B. Der support(A) ist die Wahrscheinlichkeit einer Regel, dessen Kopf eine allgemeingültige Aussage darstellt.

$$support(A) \;=\; \underbrace{P(A,B) + P(A,\bar{B})}_{=\,P(A)} = \frac{C0 + C1}{N} \tag{2.3}$$

Der Support einer Regel ist eine ungenaue Angabe der Güte einer Regel, da die Kontingenztabelle durch 4 Werte genau beschrieben wird aber der Support nur 2 dieser Werte benutzt (C0 und N).

Wenn nur sehr wenige Datensätze das gesuchte Klassifikationsattribut aufweisen, aber alle diese Datensätze von der Regel abgedeckt werden (die Regel ein Klassifikationsattribut sehr genau beschreibt), dann wäre der Supportwert sehr, klein.

Analog kann eine Regel einen kleinen Supportwert haben, wenn die Voraussetzungen im Kopf der Regel nur auf wenige Datensätze zu treffen, was nicht bedeutet das dann die Vorhersagegenauigkeit auf dem Klassifikationsattribut klein sein muss.

2.3.2. Assoziationsregel - Confidence/Precision

Um den Supportwert zu unterstützen verwendet man eine weitere Heuristik "Confidence". Bei der Berechnung des Confidencewert wird der Support einer Regel relativ zum Support des Kopfes dieser Regel dargestellt. Dadurch bekommt man für Regeln, bei welchen der

[2] Ein an dieser Stelle nicht genau definierter Begriff der ausdrücken soll, dass eine Regel R je nach gesuchter Anwendung besser oder schlechter ist als eine andere Regel für die gleichen Daten.

Kopf nur wenige Datensätze abdeckt, aber dafür keine falschen Vorhersagen trifft, einen hohen Wert.

$$confidence(A \rightarrow B) = P(B|A) = \frac{support(A \rightarrow B)}{support(A)} = \frac{C0}{C0 + C1} \tag{2.4}$$

P(X|Y) bezeichnet dabei die "Bedingte Wahrscheinlichkeit" für das Eintreten eines Ereignisses X unter der Bedingung, dass ein Ereignis Y vorher eingetreten ist:

$$P(X|Y) = \frac{P(X \cap Y)}{P(X)} \quad . \tag{2.5}$$

In der Theorie des "Information Retrieval" wird die Confidenceheuristik auch mit dem Wort "Precision" bezeichnet.

2.3.3. Assoziationsregel - Recall

Um die Kontingenztabelle vollständig beschreiben zu können benötigt man noch eine weitere Heuristik, Recall. Im Recallwert ist der Wert C2 ("false negatives") enthalten.

$$recall(A \rightarrow B) = P(A|B) = \frac{support(A \rightarrow B)}{support(B)} = \frac{C0}{C0 + C2} \quad . \tag{2.6}$$

Mit den 3 Heuristiken: Support, Confidence, Recall und der Anzahl aller Datensätze N hat man genügend Informationen, um jeden Wert in der Kontingenztabelle bestimmen zu können.
Es gibt noch andere Heuristiken, welche komplexere Beziehungen der Tabelleneinträge von 2.2 und 2.3 betrachten, jedoch werden diese Heuristiken auch nur durch die gleichen 4 Parameter aus den Tabellen (C0,...,C3) oder (p, n P, N) beschrieben. Oftmals ist es aber so, dass in Zusammenhang mit einer Regel höchstens der Support-, Confidence- und Recallwert sowie N bekannt sind.
Zu Beschreibung von Assoziationsregeln werden im allgemeinen nur die Heuristiken "Support" und "Confidence" verwendet, jedoch benötigt man, wie oben beschrieben, die "Recallheuristik", um die Kontingenztabellen vollständig beschreiben zu können. Deshalb wird im weiteren Verlauf die "Recallheuristik" zur Beschreibung von Assoziationsregeln hinzu gezählt.

3. Der Assoziationsraum

3.1. Der Support-Confidence-Raum

Der Support-Confidence-Raum ist ein von den Basen Support und Confidence aufgespannter zweidimensionaler Raum. Das heißt, die horizontale Achse wird durch den jeweiligen Supportwert einer Regel

$$support(R) = \frac{C0}{N} \tag{3.1}$$

auf einem Datensatz bestimmt. Diese Achse ist für festes N äquidistant geteilt, weil $C0 \in \mathbb{R}$ ist. Die vertikale Achse ist die Confidence-Achse. Aus ihrer Definition

$$confidence(R) = \frac{C0}{C0 + C1} \tag{3.2}$$

erkennt man, dass ihre Achsenunterteilung nicht äquidistant ist, sondern vom jeweiligen C0-Wert (Supportwert) abhängt. Durch wachsenden Umfang N der Datenbank wird die mögliche Unterteilung der Confidenceachse verfeinert (siehe dazu im Anhang die Punktgraphiken mit unterschiedlichen Datensatzgrößen). Man kann dies dadurch erklären, dass bei sehr kleinen C0-Werten mehr Kombinationsmöglichkeiten für den Wert C1 entstehen.

$$max(C1) = N - C0$$

Analog gilt, dass für kleine Support- bzw. C0-Werte der Confidencewert ebenfalls nur klein sein kann. Das bedeutet in der Nähe des Koordinatenpunktes $(0, 0)$ befinden sich viele Punkte, aber in der Nähe des Koordinatenpunktes $(0, 1)$ nur sehr wenige.

Jeder Punkt im Support-Confidence-Raum stellt eine Regel dar, da jede Regel nur einen bestimmten Support- und Confidencewert hat.

Beispiel 2. *Im letzten Kapitel hat man die Regel*

$$R: \ (\textit{Milch: ja}) \ \wedge \ (\textit{Chips: nein}) \ \wedge \ (\textit{Pizza: ja}) \ \rightarrow \ (\textit{Bier: nein})$$

definiert. In der Datenbank 2.1 soll gelten: Die Regel R erfüllt 2 Datensätze. Der Körper alleine erfüllt 4 Datensätze. Der Wert N soll den Wert 10 annehmen. Der Supportwert ist folglich 0.2 und der Confidencewert 0.5 . Also hätte der Punkt der Regel R in dieser Datenbank die Koordinaten $(0.2, 0.5)$ *im Assoziationsraum.*

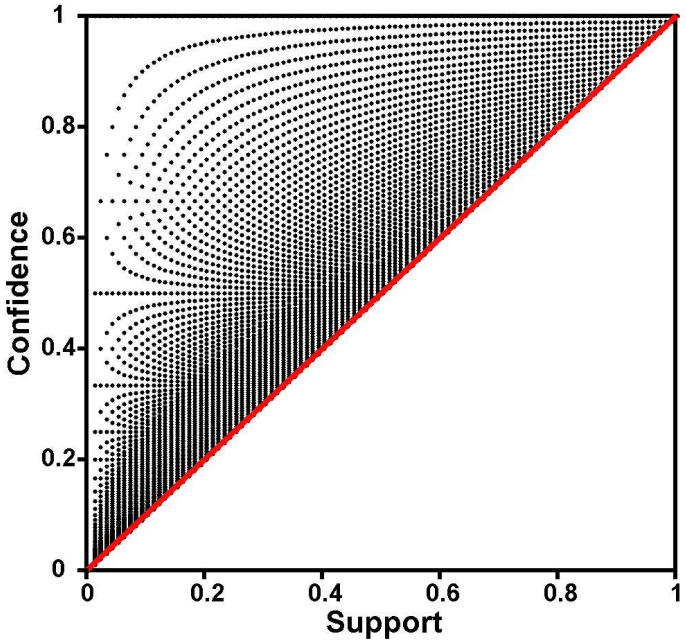

Abbildung 3.1.: Mögliche Punkte im Support-Confidence Raum mit N = 100

In der Abb. 3.1 sind alle Support-Confidencewertkombinationen dargestellt, die sich aus Kombinationen von Kontingenztabellen mit 100 Einträgen darstellen lassen.

Satz 1. *Im Support-Confidence Raum wird jeder Punkt im Dreieck oberhalb der Winkelhalbierenden liegen.*

Beweis:

Liegt ein Wert in der oberen, linken Hälfte gilt: $confidence \geq support$. Da die Anzahl der Einträge der Kontingenztabelle konstant ist und die Gleichung

$$N = C0 + C1 + C2 + C3$$

gilt, lässt sich die Ungleichung umschreiben:

$$\frac{C0}{C0 + C1} \geq \frac{C0}{C0 + C1 + C2 + C3} \tag{3.3}$$

mit $C2 + C3 \geq 0$. Wegen $C2 \geq 0$ und $C3 \geq 0$ wird jeder mögliche Support- und Confidencewert im Assoziationsraum in der oberen linken Hälfte der Abb. 3.1 liegen.

■

Lemma 1. *Jeder Punkt in der linken, oberen Hälfte ist bei genügend großem Datensatz erreichbar.*

Aus den Gl. 3.1 und 3.2 kann für vorgegebene Support- und Confidencewerte die mögliche Kontingenztabelle berechnet werden [1].

$$\begin{aligned} C0 &= supp(R) \cdot N \\ C1 &= C0 \cdot \left(\frac{1}{conf(R)} - 1\right) \\ C2 &= C0 \cdot \left(\frac{1}{recall(R)} - 1\right) \;\leq\; N - C0 - C1 \\ C3 &= N - C0 - C1 - C2 \end{aligned} \tag{3.4}$$

3.1.1. Eigenschaften des Support-Confidence-Raums

(confidence $\to 0$) *und* **(support $\to 0$)**

Ist der Supportwert klein, wird auch der C0-Wert klein. Mit $C0 = 0$ wären sowohl der Support- als auch der Confidencewert immer Null. Allerdings folgt aus $C0 = 1$ nicht automatisch ein kleiner Confidencewert. Ein kleiner Confidencewert hängt dann vom C1-Wert ab, dieser muss groß sein. Der C1-Wert erreicht sein Maximum bei $N - C0$, mit $C0 = 1$ gilt:

> In diesem Datensatz erfüllen alle Einträge die Voraussetzung A, aber nur ein Datensatz hiervon hat die gesuchte Klassifikation B.

(confidence $\to 1$) *und* **(support $\to 0$)**

Wenn der Confidencewert 1 ist gilt: $C0 = C0 + C1$ bzw., $C1 = 0$. Da der Supportwert auch "fast" Null ist folgt:

> Kaum ein Datensatz erfüllt die Regel R. Falls aber ein Datensatz die Voraussetzung A erfüllt, dann erfüllt er immer auch die Klassifikation B.

[1] konstantes N vorausgesetzt

(confidence $\to 1$) *und* (support $\to 1$)

Aus $support = 1$ folgt $C0 = N$. Dann werden A und B für fast alle Datensätze wahr. Aus der vorangegangenen Betrachtung erkennt man, dass erneut $C1 = 0$ gilt, d.h. es gibt keine Datensätze mit

$$A = \text{true} \quad \wedge \quad B = \text{false} \quad .$$

Werte auf der Diagonalen

Liegt ein Punkt auf der Diagonalen, dann gilt:

$$conf(R) = supp(R) \quad .$$

Da der Supportwert die relative Häufigkeit von C0 in der gesamten Datenbank angibt und der Confidencewert die relative Häufigkeit von C0 in allen Datensätzen, welche die Voraussetzung A der Regel R erfüllen, gilt:

$$supp(A) = N \quad .$$

Dies ist gleichbedeutend mit $C0 + C1 = N$ bzw. $C2 = C3 = 0$.

confidence = const

Diese Aussage bedeutet, dass der Confidencewert nicht vom Supportwert abhängt. Anhand der Gl. für den Confidencewert gilt dann, dass für steigendes C0 der Wert von C1 auch proportional ansteigt.

$$\Rightarrow \quad C1 = C0 \cdot \left(\frac{1}{conf(R)} - 1 \right)$$

- Für $conf(R) = 0.5$ gilt, $C0 = C1$.
- Im Bereich $0 \leq conf(R) < 0.5$ ist $C1 > C0$.
- Im Bereich $0.5 < conf(R) \leq 1$ ist $C0 > C1$.

3.2. Der Confidence-Recall-Raum

In Kapitel 1 wurde gezeigt, dass man den Support-Confidence-Raum noch um die Dimension "Recall" erweitern kann. Der Support-Recall-Raum wird hier nicht betrachtet, da mit der Substitution $C1 \rightarrow C2$ in der Confidencefunktion gilt:

$$recall(R) = \frac{C0}{C0 + C2} \quad . \tag{3.5}$$

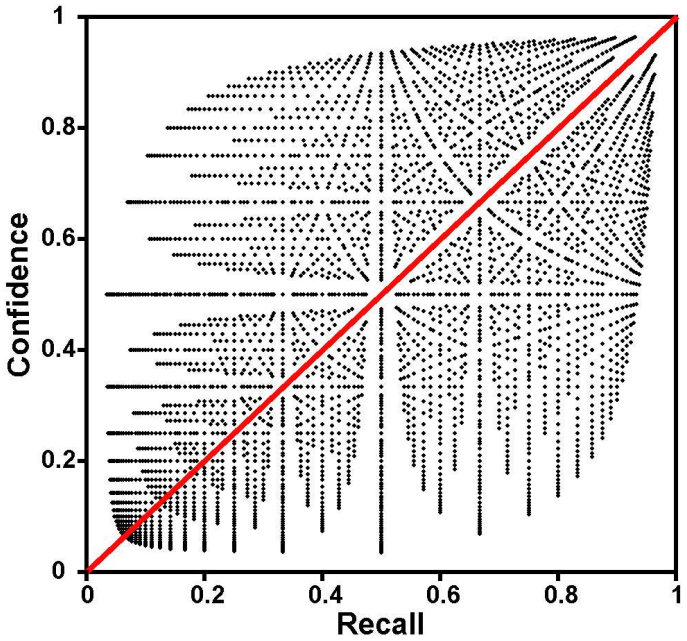

Abbildung 3.2.: Mögliche Punkte im Recall-Confidence-Raum N = 30

Der Confidence-Recall-Raum, siehe Abb. 3.2, zeigt einige Unterschiede zum Support-Confidence-Raum. Die Punktwolke erstreckt sich über den kompletten Raum und nicht nur im oberen linken Teilraum. Zusätzlich entspricht die Punktwolke im Teilbereich oberhalb der Diagonalen einer Spiegelung an der Diagonalen zum Bereich unterhalb der Diagonalen. Analog zum Support-Confidence-Raum gilt:

> Jeder Punkt im Recall-Confidence-Raum stellt eine Regel dar, da jede Regel nur einen bestimmten Recall- und Confidencewert haben kann.

3.2.1. Information Retrieval

Im Forschungsgebiet des "Information Retrieval" [2] (s. [Für09]), wird die Confidenceheuristik mit dem Wort "Precision" beschrieben. Im "Information Retrieval" geht es um die Suche nach Informationen in Dokumenten/Internetseiten in Bezug auf eine Suchanfrage, wie z.B. bei einer Suchmaschine im Internet.

Um Suchmaschinen zu bewerten, wird in einem Testdatensatz gesucht, jeder Datensatzeintrag ist in diesem Fall eine Internetseite. Wenn man eine Sucheingabe, auch "Query" genannt, auf diesen Testdatensatz anwendet, dann wird hinterher bewertet, wie viele relevante und nicht relevante Dokumente die Suchmaschine in dem Datensatz gefunden hat. Dabei weiß man vorher schon, welche Datensätze (Internetseiten) zu einer Query gehören. Die Genauigkeit wird dann in einem 2D-Raum dargestellt. Wobei die Achsenbelegungen identisch mit dem in Abb. 3.2 beschriebenen Recall-Confidence-Raum sind.

- Recall ist der Wert der gefundenen relevanten Dokumente relativ zur Gesamtanzahl aller relervanter Dokumente.
- Precision ist der Wert der gefundenen relevanten Dokumente relativ zur Gesamtanzahl aller gefundener Dokumente.

Dabei geht man davon aus, dass eine Suchmaschine nur die von ihr als relevant klassifizierten Dokumente zurückliefert.

3.2.2. Teilräume des Confidence-Recall-Raumes

Bei einer erneuten Betrachtung der Recall- und Confidencefunktionen stellt man fest, dass man das Aussehen der Kontingenztabelle durch beide Funktionswerte ableiten kann. Dadurch ergeben sich Besonderheiten für die einzelnen Teilräume des Recall-Confidence-Raumes (siehe Abb. 3.3).

Im Quadranten A gilt:

- Der Confidencewert ist größer als 0.5, daraus folgt, $C0 > C1$.
- Der Recallwert ist kleiner als 0.5, daraus folgt, $C2 > C0$.

Beide Schlussfolgerungen ergeben folgende Bedingung für die Kontingenztabelle:

$$C2 > C0 > C1 \tag{3.6}$$

[2] einer weiteren Analysetechnik im Fachgebiet des Web Mining

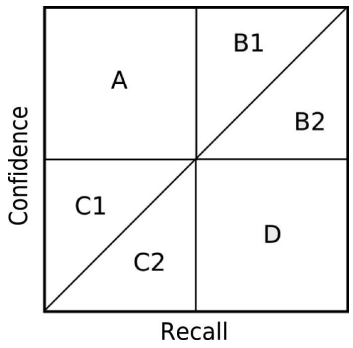

Abbildung 3.3.: Teilräume des Recall-Precision(Confidence)-Raums

Im Teilraum B1 gilt:

- Der Confidencewert ist größer als 0.5, daraus folgt, $C0 > C1$.
- Der Recallwert ist ebenfalls größer als 0.5, daraus folgt, $C0 > C2$.
- Der Teilraum liegt oberhalb der Diagonalen, weshalb der Confidencewert immer größer sein muss als der Recallwert. Daraus resultiert die Aussage $C2 > C1$, sowie $C0 > C2 > C1$

Analog gilt für B2: $C0 > C1 > C2$

Im Teilraum C1 gilt:

- Der Confidencewert ist kleiner als 0.5, daraus folgt $C0 < C1$.
- Der Recallwert ist auch kleiner als 0.5, daraus folgt $C0 < C2$.
- Der Teilraum liegt oberhalb der Diagonalen, also $C2 > C1 > C0$.

Analog gilt für C2: $C1 > C2 > C0$

Im Quadranten D gilt:

- Der Confidencewert ist kleiner als 0.5, daraus folgt, $C0 < C1$.
- Der Recallwert ist größer als 0.5, daraus folgt, $C2 < C0$.

$$\Rightarrow \quad C1 > C0 > C2 \tag{3.7}$$

4. Genauigkeit / Accuracy

4.1. Richtige Positive und richtige Negative

Bisher wurden die Grundlagen erklärt und der Assoziationsraum eingeführt. Damit besitzt man alle Kenntnisse zur Analyse einer weiteren Heuristik, diese Heuristik nennt sich Accuracy (Genauigkeit). Nach [FF05] und [Jan06] ist die Funktion für den Accuracywert im P-N-Raum:

$$acc(R) \;=\; \frac{p + (N - n)}{P + N} \;\cong\; p - n \tag{4.1}$$

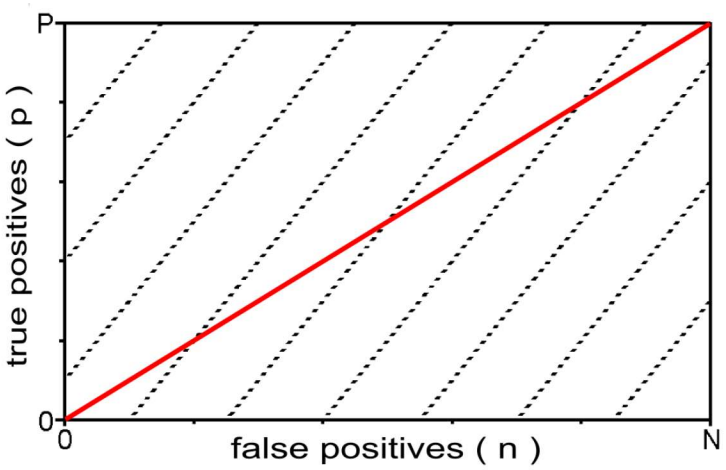

Abbildung 4.1.: Accuracy im P-N-Raum

In Abb. 4.1 ist die Accuracyheuristik im P-N-Raum dargestellt, jede gestrichelte Linie hat den gleichen Accuracywert. Das bedeutet, für jede dieser Linien gibt es eine bestimmte Differenz von "true positives" (p) und "false positives" (n). Es wurde ein Datensatz gewählt, welcher doppelt soviele Datensätze der Klassifikation N wie der Klassifikation P enthält. Wäre die Anzahl beider Klassifikationen gleich, würden die gestrichelten Linien parallel zur Diagonalen verlaufen.

Eine bestimmte Differenz von p und n hat einen bestimmten Accuracywert, deshalb sind die Linien parallel. Jede Regel ist in diesem Raum ein einzelner Punkt, dass bedeutet, zwei

Regeln auf der selben Linie haben die selbe Differenz von p und n (s. Äquivalenzvereinfachung $p - n$).

Eine Linie in der Abb. 4.1 steigt zum Einen, wenn durch den Körper A der Regel R ein Datensatz beschrieben wird, der das gesuchte Klassifikationsattribut in B aufweist. Zum Anderen steigt die Funktion, wenn der Körper A der Regel einen Eintrag, welcher den gesuchten Wert im Klassifikationsattribut nicht hat, nicht erfüllt.

Die erste Menge nennt man "true positives", die zweite "true negatives". Im P-N-Raum werden die "true negatives" durch $N - n$ beschrieben. Da der Wert N immer konstant ist, folgt aus $p + N - n$ die Äquivalenz

$$acc(R) \cong p - n \quad .$$

In der Kontingenztabelle wird dagegen der Wert der "true negatives" durch C3 repräsentiert. Nach Gl. 4.1 ergibt sich für die Accuracyfunktion im Assoziationsraum

$$acc(A \to B) \;=\; acc(R) \;=\; \frac{C0 + C3}{N} \quad . \tag{4.2}$$

$$\widehat{R} \;=\; \overline{A} \to \overline{B} \;=\; P(\overline{A}, \overline{B}) \;=\; \frac{C3}{N} \tag{4.3}$$

$$acc(R) \;=\; supp(R) \;+\; supp(\widehat{R}) \tag{4.4}$$

4.2. Accuracy im Support-Confidence/Recall-Raum

Um die Accuracyheuristik im Support-Confidence-Raum darzustellen, muss man Gl. 4.2 umformen. Mit Hilfe der Gl. 3.4 und der Abgeschlossenheit der Kontingenztabelle,

$$C3 = N - C0 - C1 - C2$$

ergibt sich für die Accuracyfunktion im Assoziationsraum folgende Funktion:

$$acc(R) \;=\; 1 \;-\; supp(R)\left(conf(R)^{-1} \;+\; recall(R)^{-1} \;-\; 2\right) \quad . \tag{4.5}$$

In der Abb. 4.2 ist die Funktion 4.5 für den Accuracywert 0.3 im Support-Confidence-Raum dargestellt. Wenn man diese Abb. mit der Abb. 3.1 vergleicht, erkennt man, dass die Strukturen beider Punktwolken auf den linken Seiten identisch sind. Jedoch ist die Punktwolke in Abb. 4.5 im Gegensatz zu der in Abb. 3.1 nach rechts begrenzt.

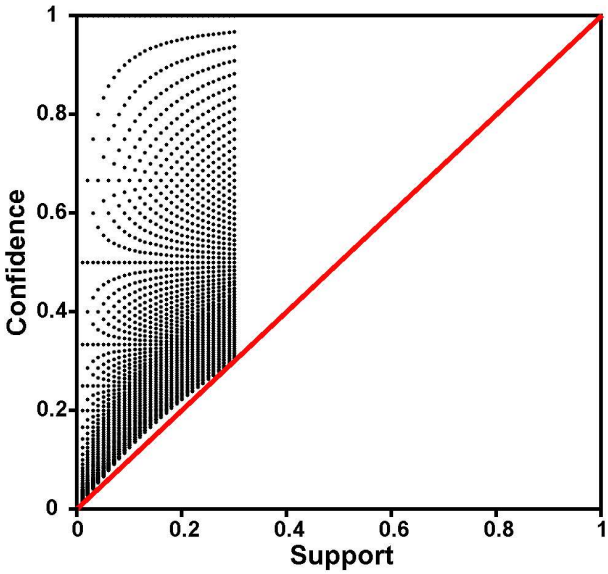

Abbildung 4.2.: Accuracy im Support-Confidence-Raum mit N = 100 und Accuracywert 0.3

Der Anhang A.2.1 zeigt weitere Bsp. von Punktwolken des Support-Confidence-Raums. In allen Abbildungen sind die Punktwolken durch einen Supportwert begrenzt, dieser Supportwert ist immer der jeweilige Accuracywert. Der Accuracywert (siehe Gl. 4.4) bildet sich aus dem Supportwert von R und $\widehat{R}$. Die Werte von $supp(R)$ und von $supp(\widehat{R})$ sind immer positiv, deshalb gilt:

$$supp(R) \quad \leq \quad acc(R) \quad . \tag{4.6}$$

Um die Accuracypunktwolke im Support-Confidence-Raum genauer untersuchen zu können, benötigt man eine Funktion für den Confidencewert, Umformen der Gl. 4.5 liefert:

$$\frac{1}{conf(R)} \quad = \quad \frac{1 - acc(R)}{supp(R)} \quad + \quad 2 \quad - \quad \frac{1}{recall(R)}. \tag{4.7}$$

Wenn man sich in Abb. 4.2 (zusätzlich s. im Anhang A.2.2) die Punktwolke im Support-Confidence-Raum ansieht, dann stellt man fest, dass der Confidencewert auch nach unten begrenzt ist. Diese Begrenzung verläuft nicht parallel zu einer Achse, sondern entlang einer Kurve. Die Kurve startet immer im Punkt $(0, 0)$ und endet immer in $(acc(R), acc(R))$. Unabhängig vom Accuracywert ist der Confidencewert begrenzt durch

$$0 \leq conf(R) \leq 1 \quad .$$

Auf die Funktion 4.7 angewendet, ergibt sich:

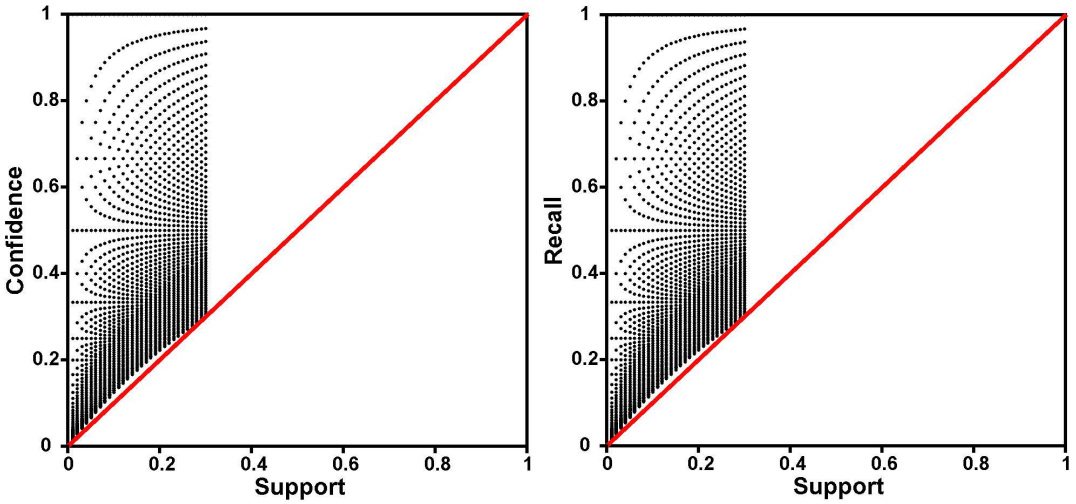

Abbildung 4.3.: Accuracy im Support-Confidence-Raum und Support-Recall-Raum mit N = 100 und Accuracywert 0.3

$$\frac{1 - acc(R)}{supp(R)} + 2 - \frac{1}{recall(R)} \geq 1 \quad . \tag{4.8}$$

Hierbei sollte man berücksichtigen, dass der Confidence- und der Recallwert in der Gl. 4.5 vertauscht werden können. Daher gilt, zusätzlich zu Gl. 4.7 auch:

$$\frac{1 - acc(R)}{supp(R)} + 2 - \frac{1}{conf(R)} \geq 1 \quad . \tag{4.9}$$

Durch Umformen der Gl. 4.8 und 4.9 ergeben sich die Minimumfuntkionen für den Confidencewert und für den Recallwert:

$$min(conf(R)) = min(recall(R)) = \frac{1}{\frac{1-acc(R)}{supp(R)} + 1} \quad . \tag{4.10}$$

4.3. Accuracy im Recall-Confidence-Raum

In der Abb. 4.4 ist die Punktwolke für den Accuracywert 0.3 im Recall-Confidence-Raum dargestellt. Die Punktwolke ist oben rechts beschränkt. Die Form der Beschränkungsfunktion lässt sich durch die Gl. 4.7 erklären:

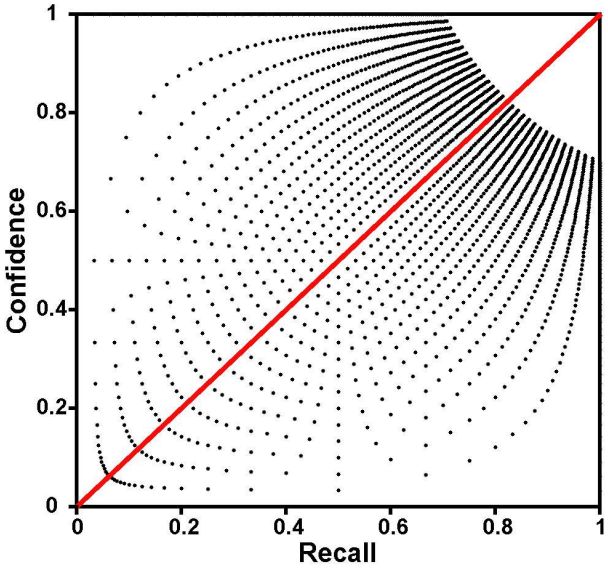

Abbildung 4.4.: Accuracy im Recall-Confidence-Raum mit N = 100 und Accuracywert 0.7

Wächst der Confidencewert, bei konstantem Accuracywert, dann muss der Recallwert schrumpfen und umgekehrt.

Auf der Diagonalen gilt: Der Confidencewert und der Recallwert sind gleich.

$$conf(R) \;=\; recall(R) \;=\; \frac{2}{\frac{1-acc(R)}{supp(R)} \;+\; 2}\;. \tag{4.11}$$

Um den maximalen Wert auf der Diagonalen zu ermitteln, maximiert man

$$max(conf(R))_{ML} \;=\; max(recall(R))_{ML} \;=\; max\left(\frac{2}{\frac{1-acc(R)}{supp(R)} \;+\; 2}\right) \tag{4.12}$$

Das Maximum des Bruchs in Gl. 4.12 ist gleich dem Minimum des Nenners bzw. des maximalen Supportwertes für konstanten Accuracywert. Der maximale Supportwert ist aber gleich dem Accuracywert (s.o.), also ergibt sich für den maximalen Confidence- und Recallwert auf der Diagonalen die Funktion:

$$max(conf(R))_{ML} \;=\; \frac{2}{\frac{1-acc(R)}{acc(R)} \;+\; 2} \;=\; \frac{2}{\frac{1}{acc(R)} \;+\; 1}\quad, \tag{4.13}$$

$$max(recall(R))_{ML} \;=\; \frac{2}{\frac{1}{acc(R)} \;+\; 1}\quad. \tag{4.14}$$

Zusätzlich zur eben betrachteten Beschränkung erkennt man, dass die Punktwolke nicht die gleiche Dichte aufweist wie die Punktwolken aus den Abb. im Anhang A.2.2. Also muss es einen Wert geben, welcher die Punktwolke in Abb. 4.4 "ausdünnt". Im Folgenden wird gezeigt, dass dies der Supportwert ist. Dazu betrachtet man erneut die Formeln für den Confidence- und Recallwert. Wird der Accuracywert erhöht

$$acc(R) \to 1 \quad ,$$

folgt daraus

$$\frac{1 - acc(R)}{supp(R)} + 2 \quad \to \quad 2$$

$$\frac{1}{conf(R)} \quad \to \quad 2 - \frac{1}{recall(R)}$$

$$conf(R) \quad \to \quad \frac{1}{2 - \frac{1}{recall(R)}} \quad .$$

Vor allem anhand der letzten Gl. ist zu erkennen, dass aus kleinen Änderungen im Recallwert größere Änderungen im Confidencewert folgern. Größere Änderungen eines Wertes bedeuten größere Zwischenräume zwischen den Punkten in der Punktwolke. Also bestimmt der Faktor:

$$\frac{1 - acc(R)}{supp(R)} \tag{4.15}$$

die Dichte der Punktwolke im Recall-Confidence-Raum.

4.4. Felder im Support-Confidence-Raum

In den ersten Kapiteln wurde das Szenario des Supermarktes beschrieben, als ein typisches Beispiel für eine Klasse von Einkaufsszenarien. Es ist im Umgang mit Assoziationsregeln üblich, von Produkt(Attribut)-Korrelationen die Support- und Confidencewerte zu bestimmen. Wenn beide Werte von einer Regel R einen "Threshold" (Pegel) erreichen, welcher im jeweiligen Szenario ausreichend ist, dann wird die Aussage der Regel weiter verwendet.

> In unserem Beispiel müsste eine Regel bestimmte Support- und Confidencewerte erreichen, damit man der Regel (Korrelation) genügend Vertrauen schenkt, um die Anordnung der Regale im Supermarkt neu zu strukturieren.

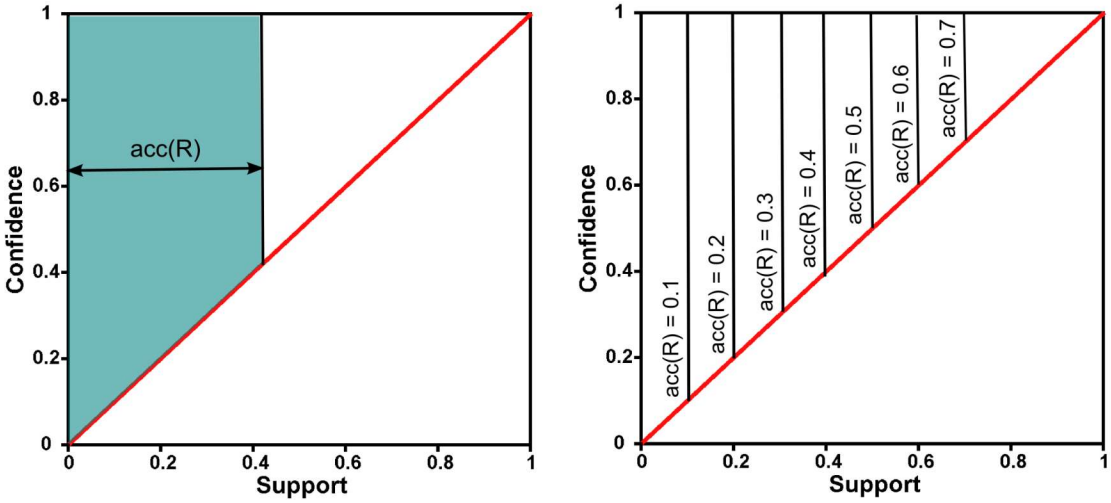

Abbildung 4.5.: Accuracyfeldbegrenzungen

Bild Links: Grünes Feld für alle mögl. Support-Confidence-Wert-Kombinationen mit Accuracywerten kleiner gleich dem eingezeichneten Accuracywert.
Bild Rechts: Die Linien stellen die Grenzen für die einzelnen Accuracywerte dar. Links von einer Linie gilt der Accuracywert noch, rechts nicht mehr.

Dieser Threshold stellt im Support-Confidence-Raum eine Grenze auf den Achsen der Support- und Confidencewerte dar. In einem Szenario geht man folglich davon aus, dass die Werte Support, Confidence und der jeweilige Threshold zu bestimmten Regeln (Produktkorrelationen) vorhanden sind.

Anstatt des Support-Confidence-Thresholdwertes kann durch die Betrachtung der Accuracyheuristik auch ein Accuracy-Thresholdwert angegeben werden. Dazu ist es notwendig, über bestimmte Bereiche des Support-Confidence-Raumes Aussagen über die möglichen Accuracywerte treffen zu können.

In Abb. 4.5 (links) erkennt man das Feld, in dem sich die Accuraycwerte befinden können, die kleiner oder gleich dem maximalen Supportwert sind. Der maximale Accuracywert eines Punktes im Support-Confidence-Raum ist dessen Supportwert. Im rechten Bild erkennt man die maximalen Grenzen für mehrere Accuracywerte.

In Abb. 4.6 visualisiert der helle Bereich die Punkte für die Accuracywerte zwischen 0.3 und 0.7.

- Nach oben wird dieser Bereich durch die Confidencefunktion 4.10 für den Accuracywert 0.7 begrenzt,
- nach unten durch die Diagonale bzw. die Voraussetzung $conf \geq supp$

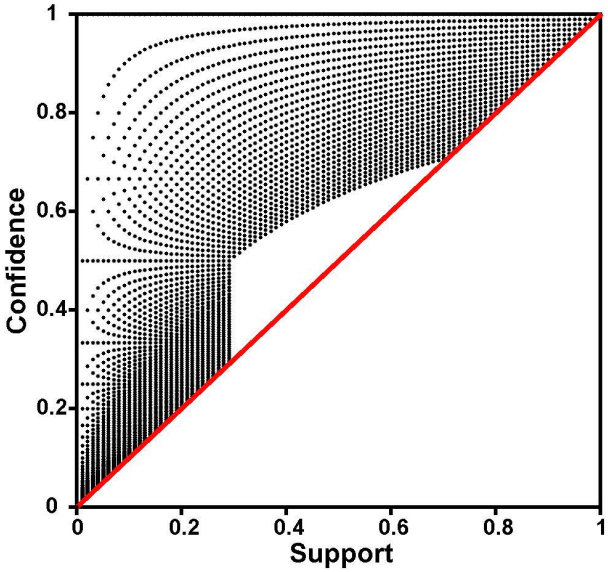

Abbildung 4.6.: Im weißen Feld (eingegrenzt durch die Punktwolke und der Diagonalen) existieren ausschließlich Punkte für Accuracwerte zwischen 0.3 und 0.7.

- und an der linken Seiten durch die Extremalfunktion für den Supportwert.

Aus diesen Ergebnissen kann man also den Accuracywertebereich eines Punktes aus den Support- und Confidencewerten per Ausschlussverfahren ermitteln. Der Accuracywert eines Punktes kann nicht kleiner sein als der Supportwert und nach der Funktion 4.10 kann der Accuracywert nicht größer sein als

$$max(acc(R)) \;=\; supp(R)\left(1 \;-\; \frac{1}{conf(R)}\right) \;+\; 1 \tag{4.16}$$

4.5. Fazit

Der Accuracywert ist die Summe des Supportwertes der Regel R und des Supportwertes der Regel $\widehat{R}$. Die Erweiterung um den Wert von $\widehat{R}$ bedeutet, dass man nicht nur die Information über die Anzahl der "true positives" im Datensatz besitzt, sondern auch der "true negatives".

> Supermarktszenario: Jeder Kunde kauft bestimmte Artikel ein. Jeder Artikel stellt ein Attribut dar. Jedes Attribut/Artikel hat die zwei Klassifikationen "gekauft" (1) und "nicht gekauft" (0).

Ziel ist es, Regeln zu finden, welcher Kauf von Artikeln den Kauf von anderen Artikeln impliziert.

Sowohl der Support- als auch der Confidencewert für eine Regel bestimmen ausschließlich die Wahrscheinlichkeiten:

- einer Produktkombination relativ zu allen Einkäufen des Supermarktes,
- oder den Kauf eines Produktes relativ zu einer Produktkombination.

In der Accuracyheuristik ist die Information des Supportwertes enthalten. Zusätzlich gibt die Accuracyheuristik die Wahrscheinlichkeit an, dass andere Kombinationen von gekauften Produkten auch auf den Klassifikationsartikel schließen lassen.

Die Information dieser "Exklusivität" einer Produktkombination wird durch den Confidence- und Recallwert bestimmt. Da Supportwert und Datensatzgröße bekannt sind, kann man aus C1 und C2 den Wert C3 (bzw. $\widehat{R}$) ermitteln. Da nur die Summe von C1 und C2 zur Bestimmung von C3 notwendig ist, ergibt sich die Vertauschungsunabhängigkeit von Confidence- und Recallwert. Dies ist auch anhand der Achsensymmetrie der Diagonalen in den Abb. vom Recall-Confidence-Raum zu erkennen.

5. Conviction

5.1. Conviction im P-N-Raum

In den bisher beobachteten Heuristiken, Support, Confidence, Recall und Accuracy war der Wertebereich immer zwischen 0 und 1. Eine weitere Klasse von Heuristiken bestimmt das Verhältnis von zwei Heuristiken zueinander. Der Wertebereich ist dabei z.B. zwischen Null und Unendlich. Ein Wert unterhalb von 1 gibt die Dominanz einer Heuristik über der anderen an, genauso wie ein Wert über 1 dem reziproken Fall entspricht. Eine der Heuristiken, die ein solches Verhältnis beschreiben, ist die Convictionheuristik. In [TKS02] wird die Convictionfunktion mit

$$V(A \to B) \;=\; V(R) \;=\; \frac{p(A)p(\overline{B})}{p(A,\overline{B})} \tag{5.1}$$

definiert [1] mit dem Wertebereich $[0.5, \infty]$.

Stellt man die einzelnen Wahrscheinlichkeiten im P-N-Raum dar, dann erhält man

$$p(A) \;=\; \frac{p+n}{P+N} \quad , \quad p(\overline{B}) \;=\; \frac{N}{P+N} \quad \textit{und} \quad p(A,\overline{B}) \;=\; \frac{n}{P+N} \quad .$$

Die Wahrscheinlichkeit von $\overline{B}$ ist eine von R unabhängige konstante Zahl. Sie gibt das Verhältnis von "negativen" Datensätzen in der Datenbank zur Datenbankgröße an. Nach der Definition der Kontingenztabelle hat ein negativer Datensatz im Klassifikationsattribut nicht den Wert, den die Regel R vorhersagen will. Analog hat ein "positiver" Datensatz im Klassifikationsattribut den Wert aus dem Kopf B der Regel R. Demnach ist die Convictionformel für den P-N-Raum

$$V(R) \;=\; \frac{N(p+n)}{n(P+N)} \;=\; \frac{(p+n)}{n}\;\; p(\overline{B}) \quad . \tag{5.2}$$

In der Abb. 5.1 wird die Convictionheuristik im P-N-Raum dargestellt. In dem verwendeten Datensatz gibt es doppelt soviel "negative" (N) wie "positive" Datensätze (P). Aus diesem Grund ist die Steigung der Diagonalen

$$p(\overline{B}) = 2/3 \quad .$$

[1] Hierbei wird auf den komplementären Teil des in der Referenz dargestellten Teils der Formel verzichtet, da dieser Teil den Convictionwert für die Gegenannahme der Regel R bestimmt.

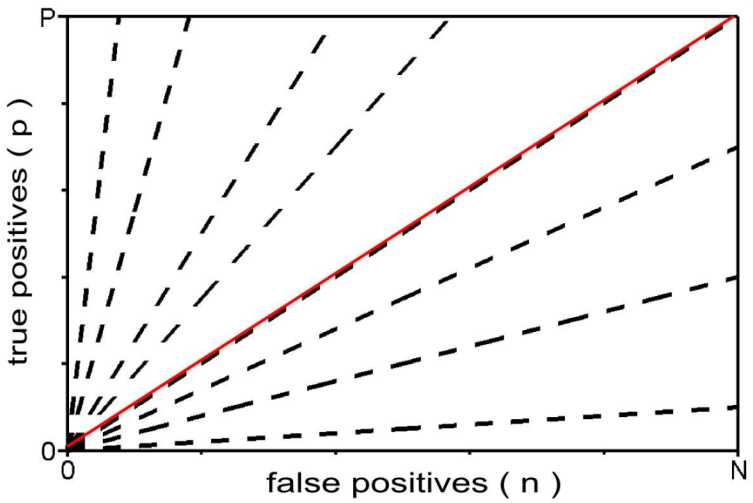

Abbildung 5.1.: Conviction im P-N-Raum

Die Anzahl der Datensätze, die von der Regel als positiv bewertet wurden ist gleich der Summe der "true positives" und "false positives". Für Regeln mit Punkten auf der Diagonalen gilt damit:

> Das Verhältnis von "false positives" (n) relativ zu allen positiv bewerteten Datensätzen ist identisch mit dem Verhältnis von $p(\overline{B})$.

Der Bereich des Convictionwerts (s. [TKS02]) ist $[0.5, \infty]$. Werte unterhalb von 1 haben in der Abb. 5.1 eine kleinere Steigung als die Diagonale, analog haben Werte oberhalb von 1 eine höhere Steigung als die Diagonale. Wenn man den Fall betrachtet, dass alle Datensätze negativ sind, dann gibt es zwei Möglichkeiten für die Regel:

- Entweder die Regel bestimmt für keinen Datensatz, dass dieser positiv ist, und erzeugt damit zu 100% "true negatives"
- oder man erhält für jeden als positiv bewerteten Datensatz einen "false positiv" Eintrag in der Konfusionsmatrix.

$$V(R) = \frac{p(\overline{B})}{\frac{n}{p+n}} \tag{5.3}$$

Im ersten Fall ist der Convictionwert nicht bestimmbar, im zweiten Fall ist der Convictionwert 1. Für den Fall, dass es gleich viele positive und negative Datensätze gibt, gilt für den Zähler $p(\overline{B}) = 1/2$. Die Regel R gibt für jeden Datensatz eine Bewertung ab, dass läßt dann wieder mehrere verschiedene Möglichkeiten zu.

- Bestimmt die Regel für jeden Datensatz die richtige Klassifikation, dann ist der Wert im Nenner 0.
- Bestimmt die Regel jeden Datensatz als positiv, dann ist der Wert im Nenner gleich 1/2.
- Bestimmt die Regel jeden Datensatz als negativ, dann ist der Wert im Nenner 0.
- Bestimmt die Regel mehr "false positives", als "true positives", dann ist der Wert im Nenner größer als 1/2.

Die Regel R ist also um so besser, je kleiner der Wert im Nenner ist [2].

Um im Folgenden die Convictionheuristik genauer untersuchen zu können, muss die Formel für den Convictionwert 5.1 im Assoziationsraum dargestellt werden.

$$V(R) = \frac{1 - \frac{supp(R)}{recall(R)}}{1 - conf(R)} \tag{5.4}$$

$$conf(R) = 1 - \frac{1 - \frac{supp(R)}{recall(R)}}{V(R)} \tag{5.5}$$

$$supp(R) = recall(R)\ (1 - V(R)\ (1 - conf(R))) \tag{5.6}$$

$$recall(R) = \frac{supp(R)}{1 - V(R)\ (1 - conf(R))} \tag{5.7}$$

Anhand von Gl. 5.4 erkennt man, warum der Convictionwert den Wert 0.5 nicht unterschreiten kann. Wäre der Convictionwert kleiner als 0.5, dann gilt nach einigem Umformen der Gl. 5.4:

$$conf(R) < \frac{supp(R)}{recall(R)} - 1 \tag{5.8}$$

Das bedeutet, der Quotient aus Support- und Recallwert müsste größer sein als 1, was unmöglich ist.

[2] Diese Aussage ist relativ zum Zähler zu betrachten.

5.2. Conviction im Support-Confidence-Raum

5.2.1. Convictionwerte kleiner als 1

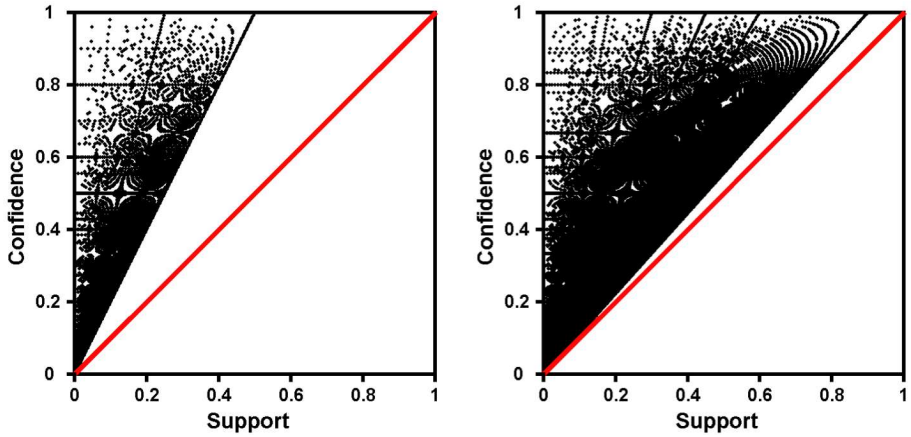

Abbildung 5.2.: Convictionwerte 0.5 (links) und 0.9 (rechts) im Support-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.3.1)

In den verschiedenen Abbildungen im Support-Confidence-Raum ist zu erkennen, dass die Punktgraphiken erneut [3] nach rechts beschränkt sind. Diese Beschränkung ist eine lineare Funktion im Support-Confidence-Raum und verläuft vom Ursprungspunkt $(0, 0)$ zum Punkt $(V(R), 1)$. Daraus folgt:

$$max\{supp(R)\} \;=\; V(R) \;\cdot\; conf(R) \quad .$$

Die Begründung kann man mit Hilfe der Gl. 5.4 herleiten.

$$V(R) \;\cdot\; conf(R) \;=\; \frac{(C0 + C1)\,(C1 + C3)}{N\,C1} \;\cdot\; \frac{C0}{C0 + C1} \tag{5.9}$$

$$V(R) \;\cdot\; conf(R) \;=\; \frac{(C1 + C3)}{C1} \;\cdot\; supp(R) \tag{5.10}$$

Aus einem maximalen Supportwert folgt ein maximaler C0-Wert und mit

$$V(R) \;=\; const. \qquad \wedge \qquad conf(R) \;=\; const.$$

folgt, dass die zwei Parameter C1 und C2 zueinander proportional ansteigen müssen. Der Wert von C3 wird dadurch immer weiter minimiert, da die Anzahl der Datensätze konstant bleibt. Für $C3 \to 0$ gilt:

[3] siehe Accuracyheuristik

$$P(\ A \mid \overline{B}\ ) \ = \ \frac{C1}{C1 + C3} \ \rightarrow \ 1 \quad . \tag{5.11}$$

Damit ist die Behauptung

$$max\,\{supp(R)\} \ = \ V(R) \ \cdot \ conf(R) \tag{5.12}$$

bewiesen. Analog gilt für den minimalen Confidencewert die Gleichung

$$min\,\{conf(R)\} \ = \ \frac{max\,\{supp(R)\}}{V(R)} \quad . \tag{5.13}$$

5.2.2. Convictionwerte größer als 1

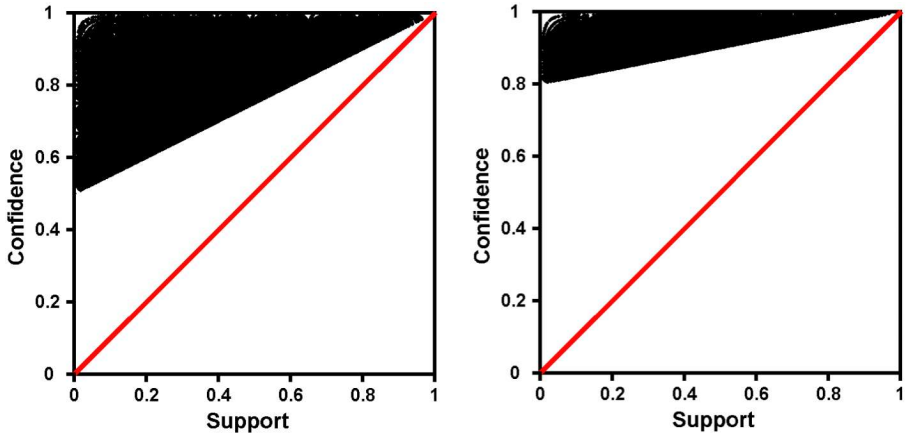

Abbildung 5.3.: Convictionwerte 2 (links) und 5 (rechts) im Support-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.3.2)

Die Berechnung der Beschränkung der Punktwolke erfolgt analog. In den Abb. 5.3 erkennt man, dass sich die Funktion

$$min\,\{conf(R)\} \ = \ 1 \ - \ max\left\{\frac{1}{V(R)}\right\} \ + \ min\left\{\frac{supp(R)}{V(R)\ recall(R)}\right\} \tag{5.14}$$

durch Umformen der Gl. 5.5 ergibt. Da der Convictionwert konstant bleibt, ist der mittlere Term nicht weiter maximierbar. Mit $supp(R) = const$ ergibt sich der Recallwert als einziger

veränderlicher Parameter. Wenn der Confidencewert in Gl. 5.5 sein Minimum annimmt, dann gilt:

$$max\{recall(R)\} \;=\; 1$$

$$min\{conf(R)\} \;=\; 1 \;-\; \frac{1}{V(R)} \;+\; \frac{supp(R)}{V(R)} \quad . \tag{5.15}$$

5.2.3. Cluster im Support-Confidence-Raum

In diesem Abschnitt geht es um "Cluster" im Support-Confidence-Raum (siehe Abb. 5.4). Bei einem Cluster handelt es sich um eine Menge von Punkten, die nur eine minimale Support- und Confidencewertdifferenz zueinander aufweisen. Wenn man zwei Supportwerte betrachtet, dann kann man davon ausgehen, dass mit $N = const$ sich die beiden Supportwerte lediglich im Wert von C0 unterscheiden.

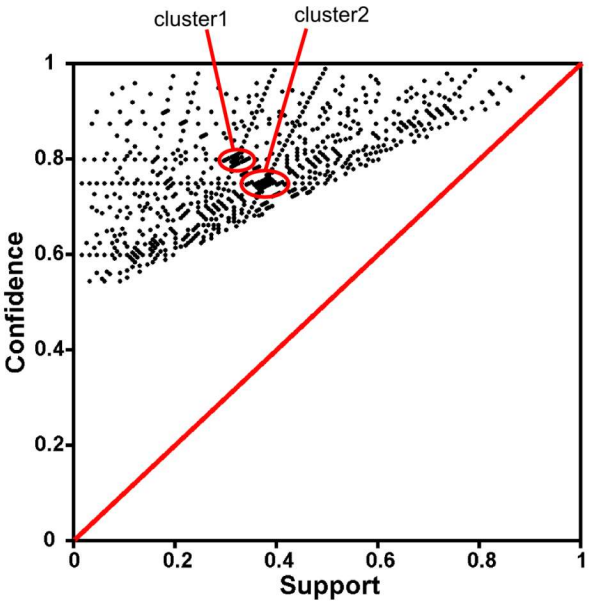

Abbildung 5.4.: Conviction im Support-Recall-Raum mit N = 100

$$min\{\Delta C0\} = 1 \qquad \wedge \qquad \Delta supp(R) \;=\; \frac{1}{N} \tag{5.16}$$

Analog zu dem gerade betrachteten Supportwertverfahren geht man beim Confidencewert vor. Hinzu kommt die Information, dass der neue Confidencewert entscheidend vom gerade ermittelten C0-Differenzwert abhängt (ideal $\Delta C0 = 1$). Da sich beide Confidencewerte kaum unterscheiden, gilt:

$$conf(R_1) = \frac{C0 + \Delta C0}{C0 + \Delta C0 + C1 + \Delta C1} \approx \frac{C0}{C0 + C1} = conf(R_2) \tag{5.17}$$

$$\Delta C1 = \frac{\Delta C0\ C1}{C0} \tag{5.18}$$

Für ganzzahliges $\Delta C1$, muss $\Delta C0 \cdot C1$ ein Vielfaches von C0 sein. Es ergeben sich eine Fülle von konstanten Confidencefunktionen, auf denen die Cluster liegen können.

$$C1 = \frac{C0\ \Delta C1}{\Delta C0} \tag{5.19}$$

$$conf(R) = \frac{\Delta C0}{\Delta C1 + \Delta C0} \quad . \tag{5.20}$$

Diese konstante Funktion beschreibt feste Linien, für kleine $\Delta C0$ und kleine $\Delta C1$, auf denen sich die Cluster befinden können.

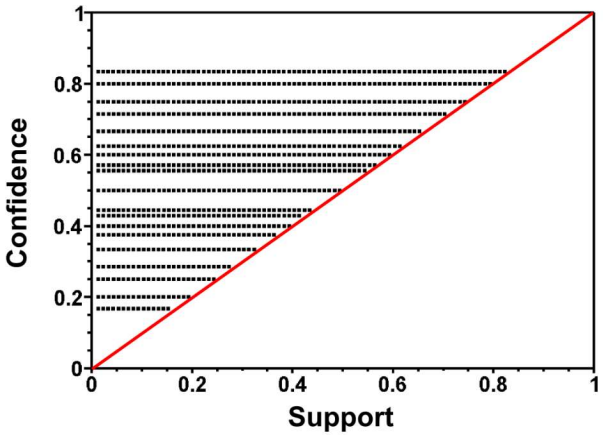

Abbildung 5.5.: Linien mit erhöhter Clusterwahrscheinlichkeit (für $\Delta C0 \leq 5$)

In Abb. 5.4 wird ein Datensatz mit $N = 100$ betrachtet. Die gefundenen Cluster haben eine Breite in Supportrichtung von weniger als 0.05. Eine 0.05 Abweichung in Supportrichtung mit $N = 100$ bedeutet, dass $\Delta C0$ einen Wertebereich von 1 bis 5 umfasst. Die resultierenden Geraden, auf welchen sich mögliche Cluster befinden, sind in Abb. 5.5 dargestellt. Da man nun weiß, an welchen Stellen die Cluster auftreten können, wird im Folgenden die Form der Cluster untersucht. Die Gleichung

$$\frac{\Delta C1}{C1} = \frac{\Delta C0}{C0} \tag{5.21}$$

stellt dazu eine Beziehung zwischen der Position und der Größen eines Clusters her. Wobei $C0$ und $C1$ das Zentrum des Clusters definieren und die Werte $\Delta C0$ und $\Delta C1$ die Abweichung der einzelnen Clusterpunkte widerspiegeln. Mit Hilfe der Gl. 5.17 erkennt man:

> Wenn der Wert $\Delta C0$ steigt, dann muss auch der Wert $\Delta C1$ proportional steigen.

Daraus folgt, dass sich der Confidencewert kaum verändert, da in der Gl. 5.17 das Verhältnis zwischen Nenner und Zähler in etwa gleich bleibt. Analog gilt:

> Wenn sich der Confidencewert stark verändert aber der Supportwert nicht, dann muss der Wert $\Delta C0$ klein sein und der Wert $\Delta C1$ groß.

Zusammengefasst bedeutet das:

- Wenn sich der Confidencewert innerhalb eines Clusters kaum verändert, dann ist ein großer $\Delta C0$-Wert möglich, damit ergibt sich auch eine große Supportwertdifferenz.
- Wenn der $\Delta C0$-Wert innerhalb eines Clusters klein ist, dann ergibt sich schon bei kleinsten Änderungen vom $\Delta C1$-Wert eine große Confidencewertdifferenz.

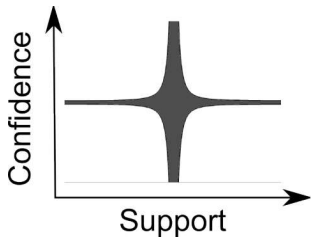

Abbildung 5.6.: Standardform von Cluster

Dieses Verhalten ist unabhängig vom Ort und bewirkt ein einheitliches Aussehen aller Cluster (siehe Abb. 5.6).

5.3. Conviction im Support-Recall-Raum

5.3.1. Support-Recall-Raum mit Convictionwerten kleiner als 1

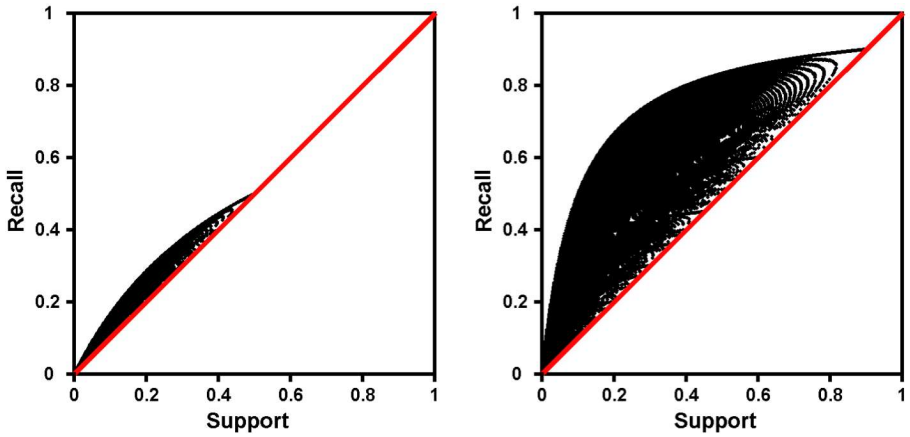

Abbildung 5.7.: Convictionwerte 0.5 (links) und 0.9 (rechts) im Support-Recall-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.3.3)

Wie die Punktwolken im Support-Confidence-Raum sind auch die Punktwolken im Support-Recall-Raum beschränkt, haben jedoch ein anderes Aussehen als im Support-Confidence-Raum, denn in Gl. 5.4 sind der Confidence- und Recallwert nicht vertauschungsunabhängig.

Für den Convictionwertebereich $V \leq 1$ gibt es für die Punktwolken eine Maximalbeschränkung des Recallwerts durch eine Grenzkurve, die in einem Bogen vom Punkt $(0,0)$ bis zum Punkt (V,V) verläuft.

$$max\{recall(R)\} = \frac{supp(R)}{min\{1 - V(R)\ (1 - conf(R))\}} \tag{5.22}$$

Mit konstantem Support- und Convictionwert ist die Maximumsfunktion lediglich vom Confidencewert abhängig, dieser muss dafür minimiert werden.

$$max\{recall(R)\} = \frac{supp(R)}{1 - V(R)\ (1 - min\{conf(R)\})} \tag{5.23}$$

Der minimale Confidencewert wurde im letzten Teilkapitel schon ermittelt (siehe Gl. 5.13).

$$max\{recall(R)\} = \frac{supp(R)}{1 - V(R)\ (1 - \frac{supp(R)}{V(R)})} \tag{5.24}$$

5.3.2. Support-Recall-Raum mit hohen Convictionwerten

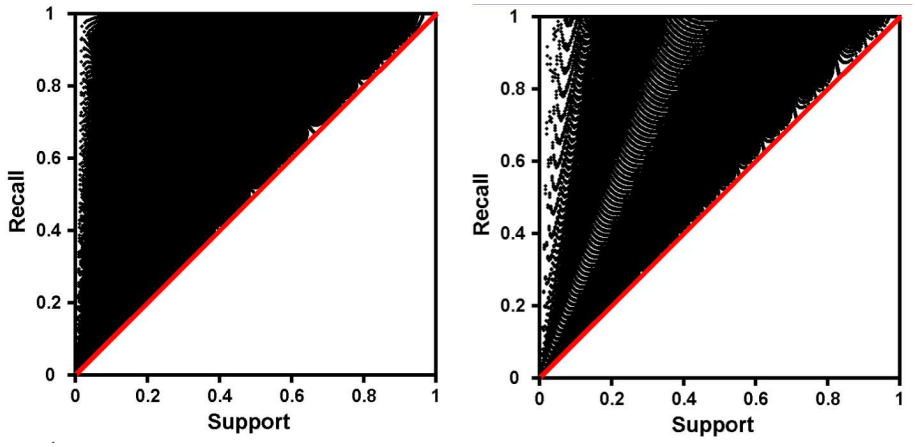

Abbildung 5.8.: Convictionwerte 2 (links) und 5 (rechts) im Support-Recall-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.3.2)

In den Abbildungen in 5.8 kann man erkennen, dass die Punktwolken durch "Bänder/Streifen" dargestellt werden. Der Abstand der "Bänder" wird mit zunehmendem Convictionwert immer größer. Dies liegt daran, dass mit wachsendem Convictionwert sich der Confidencewert immer mehr dem Wert 1 annähert.

Der Nenner der Recallfunktion 5.7 muss positiv bleiben, damit der Recallwert sich noch zwischen 0 und 1 befindet. Wenn man dann für große Convictionwerte den Nenner umformt, entsteht die folgende Minimalbegrenzung für den Confidencewert:

$$1 \; - \; V(R)(1 - conf(R)) \; \geq \; 0 \tag{5.25}$$

$$conf(R) \; \geq \; \frac{V(R) - 1}{V(R)} \; \approx \; 1 \quad . \tag{5.26}$$

Bei großem Convictionwert ergibt sich dadurch ein großes Verhältnis zwischen dem C0 und dem C1 Wert.

$$\frac{C0}{C0 + C1} \; = \; \frac{V(R) - 1}{V(R)} \tag{5.27}$$

$$C1 \; = \; C0 \left(\frac{V(R)}{V(R) - 1} \; - \; 1 \right) \; = \; C0 \; \underbrace{\frac{1}{V(R) - 1}}_{klein} \tag{5.28}$$

Das bedeutet, bei kleinen Supportwerten kann es immer weniger Confidencewerte geben, welche diese Voraussetzung erfüllen. Dies kann man in den Abb. mit sehr hohem Convictionwert erkennen, da sich hier die Linien im niedrigen Supportwertbereich immer mehr "zerstückeln".

5.4. Conviction im Confidence-Recall-Raum

5.4.1. Convictionwerte kleiner als 1

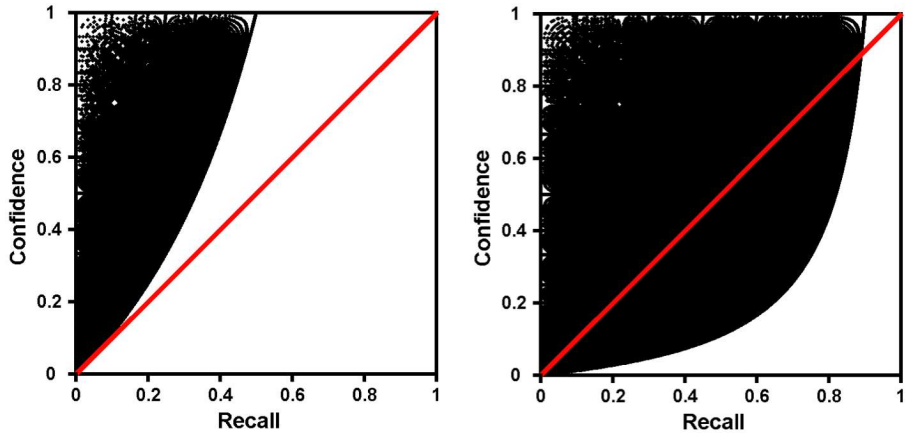

Abbildung 5.9.: Convictionwerte 0.5 (links) und 0.9 (rechts) im Recall-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.3.5)

Im Confidence-Recall-Raum (Abb. 5.9) erkennt man eine klare Abgrenzung des Wertebereichs. Wenn man den maximalen Recallwert ermittelt, dann stellt man fest, dass dies ausschließlich mit dem maximalen Supportwert zusammenhängt. Dieser wurde aber bereits in Gl. 5.12 ermittelt, damit gilt für den maximalen Recallwert

$$max\{recall(R)\} = \frac{V(R) \cdot conf(R)}{1 - V(R)(1 - con(R))} \tag{5.29}$$

5.4.2. Convictionwerte größer als 1

Analog zu den bisherigen Abbildungen sind auch in diesem Fall die Punktwolken beschränkt. Für jede Punktwolke (s. Abb. 5.10) gibt es einen Confidencewert den kein Punkt der

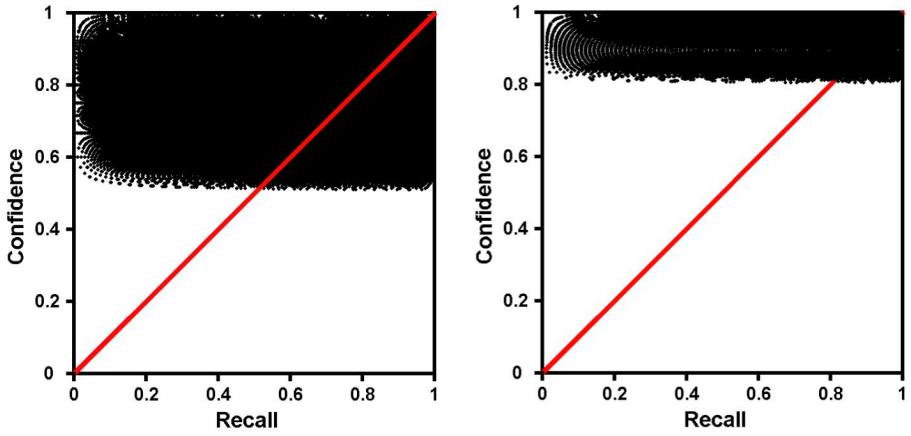

Abbildung 5.10.: Convictionwerte 2 (links) und 5 (rechts) im Recall-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.3.6)

Punktwolke unterschreitet. Dieser Confidencewert hängt vom jeweiligen Convictionwert ab. Für jede einzelne Punktwolke ist der Convictionwert konstant.

$$min\{conf(R)\} \;=\; 1 \;-\; \frac{1}{V(R)} \;+\; \frac{min\{supp(R)\}}{V(R)\; recall(R)} \tag{5.30}$$

Der Recallwert bleibt variabel, da man weiterhin den kompletten Recall-Confidence-Raum betrachten möchte. Damit ist der einzige zu minimierende Faktor der Supportwert. Der minimale Supportwert wird, nach umformen, mit der Funktion

$$min\{supp(R)\} \;=\; recall(R)\;\left(1 \;-\; V(R)\;\left(1 - min\{conf(R)\}\right)\;\right) \tag{5.31}$$

beschrieben. Wenn man den Confidencewert minimiert muss man aufpassen, dass der Supportwert nicht kleiner als 0 wird, dies ist theoretisch möglich (nach dieser Gleichung). Damit hat man auch den minimalen Supportwert gefunden, nämlich 0. Das bedeutet die Gl. 5.30 vereinfacht sich zu

$$min\{conf(R)\} \;=\; 1 \;-\; \frac{1}{V(R)} \quad , \tag{5.32}$$

und man kann die untere Schranke für die einzelnen Punktwolken bestimmen.

5.5. Feldanalyse des Support-Confidence-Raums

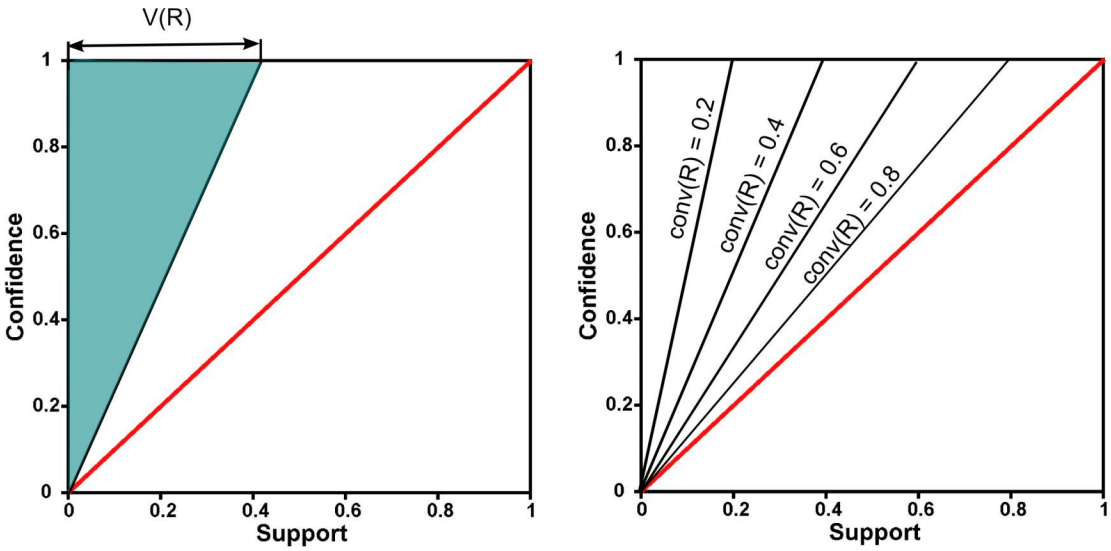

Abbildung 5.11.: Convictionfelder kleiner 1 im Support-Confidence-Raum

Bild Links: Grünes Feld für alle mögl. Support-Confidence-Wert-Kombinationen mit Convictionwerten kleiner gleich dem eingezeichneten Convictionwert.
Bild Rechts: Die Linien stellen die Grenzen für einzelne Convictionwerte dar. Links von einer Linie gilt der Convictionwert noch, rechts nicht mehr.

In der Analyse des Support-Confidence-Raums wurden bisher nur die Punktwolken für jeden Convictionwert einzelnen betrachtet. Um eine fundamentale Aussage für die Convictionheuristik treffen zu können, muss man analog zum letzten Kapitel (Accuracyheuristik) die Convictionwert-Felder des Support-Confidence-Raums analysieren. Dies ist schwieriger als bei der Accuracyheuristik, da man bei Convictionwerten in der Nähe von 1 theoretisch [4] jeden Punkt im Support-Confidence-Raum abdecken kann.

Wenn man annimmt, man hätte durch eine Regel R einen Punkt im Support-Confidence-Raum gegeben

$$P_1 = (supp(R), conf(R)) \quad , \tag{5.33}$$

dann ergibt sich zumindest eine minimale Abschätzung für den Convictionwert.

$$min\{V(R)\} = \frac{supp(R)}{conf(R)} \tag{5.34}$$

[4] Bei einem genügend großem Datensatz ist dies möglich, da es keine Beschränkung der Punktwolke gibt (siehe Anhang A.3.1 und A.3.2)

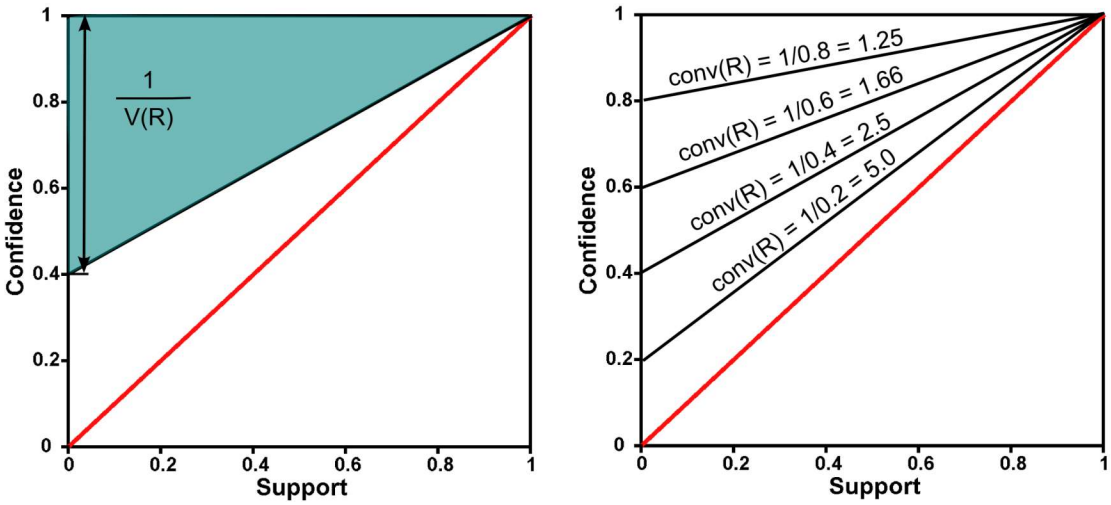

Abbildung 5.12.: Convictionfelder größer 1 im Support-Confidence-Raum

Bild Links: Grünes Feld für alle mögl. Support-Confidence-Wert-Kombinationen mit Convictionwerten kleiner gleich dem eingezeichneten Convictionwert.
Bild Rechts: Die Linien stellen die Grenzen für einzelne Convictionwerte dar. Über einer Linie gilt der Convictionwert noch, unter der Linie nicht.

Dies gilt aber nur für $V(R) < 1$, denn im Fall $V(R) > 1$ ist:

$$min\{V(R)\} = \frac{1}{conf(R)} \quad . \tag{5.35}$$

Anhand dieses Widerspruchs erkennt man, dass man mit dem Support- und Confidencewert alleine keine eindeutige Aussage über den Convictionwert treffen kann. Die benötigten Zusatzinformationen erhält man aus der Lage des Punktes im Support-Recall-Raum.

$$P_2 = (supp(R), recall(R)) \tag{5.36}$$

Der Supportwert dieses Punktes P_2 ist identisch mit dem des Punktes P_1, da es sich um die selbe Regel R handelt. Anhand der Abbildung im Anhang A.3.3 und A.3.2 erkennt man im Support-Recall-Raum, dass sich die Wertebereiche für $V(R)$ kleiner 1 und $V(R)$ größer 1 überschneiden. Also ist anhand der bisherigen Abbildungen der genaue Convictionwert immer noch nicht bestimmbar. Mit $V(R) < 1$ gilt in der Gl. 5.4:

$$\frac{supp(R)}{conf(R)} \geq recall(R) \tag{5.37}$$

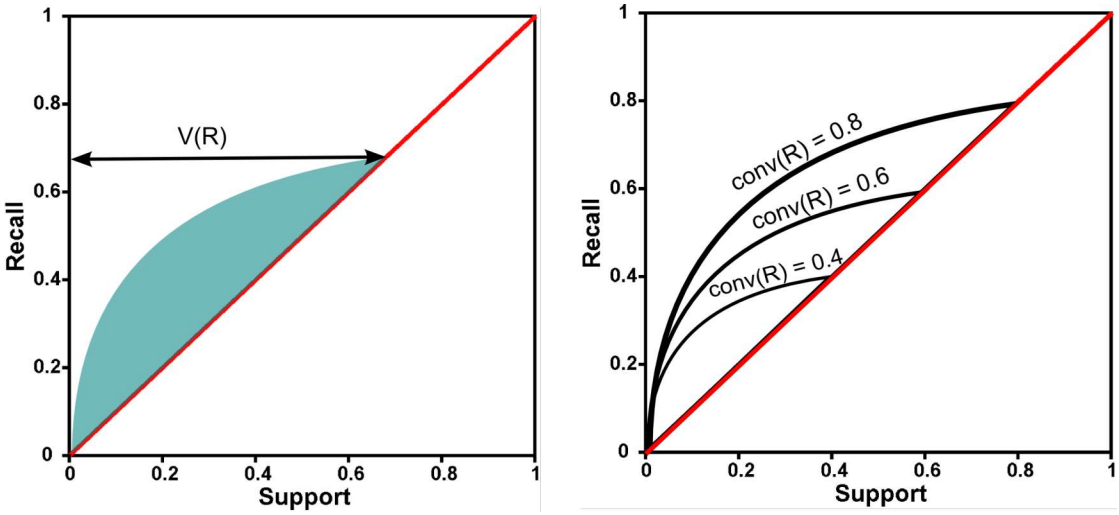

Abbildung 5.13.: Convictionfelder kleiner 1 im Support-Recall-Raum

Bild Links: Grünes Feld für alle mögl. Support-Confidence-Wert-Kombinationen mit Convictionwerten kleiner gleich dem eingezeichneten Convictionwert.
Bild Rechts: Die Bogenlinien stellen die Grenzen für einzelne Convictionwerte dar. Unterhalb einer Linie gilt der Convictionwert noch, oberhalb nicht mehr.

Das bedeutet:

Ist der Quotient aus Support- und Confidencewert größer gleich dem Recallwert, dann hat die Regel einen Convictionwert zwischen

$$[\ \frac{supp(R)}{conf(R)}\ ,\ 1]\quad ,$$

ansonsten liegt der Convictionwert zwischen

$$[\ \frac{1}{conf(R)}\ ,\ \infty]\quad .$$

5.6. Fazit

Es wurde gezeigt, dass die Convictionheuristik das Verhältnis

$$V(R) \;=\; \frac{C0 + C1}{C1} \;\cdot\; p(\overline{B}) \;=\; \frac{1}{p(\overline{B}|A)} \;\cdot\; p(\overline{B})$$

beschreibt, wobei der Wert von $p(\overline{B})$ ein fester Wert des Datensatzes ist, und $p(\overline{B}|A)$ das folgende Verhältnis darstellt:

- Anzahl der "false positives"
- relativ zu der Anzahl aller Datensätze, welche den Körper A erfüllen.

Der Convictionwert wird demnach maximal, wenn möglichst wenige "false positives" durch die Regel erzeugt werden, aber dafür sehr viele "true positives". Dieses Verhalten ist äquivalent zu einem hohen Confidencewert (siehe Gl. 5.4). Ebenso kann der Convictionwert groß werden, wenn der Quotient

$$\frac{supp(R)}{recall(R)} \;=\; \frac{C0 + C2}{N} \;=\; supp(B)$$

klein ist. Die Convictionheuristik bevorzugt Regeln, welche im Kopf nicht die numerisch größte Klassifikationsklasse[5] aufweisen (wenige Datensätze ansprechen) und wenige "false positives" produzieren[6].

5.6.1. Cluster

Zusätzlich zur allgemeinen Analyse wurde in diesem Kapitel eine "Punktwolkenanomalie" (die Cluster) beschrieben. Es wurde gezeigt, dass unabhängig von der jeweiligen Heuristik, Cluster (Punkthaufen) entstehen, wo sie entstehen und welche charakteristische Form sie haben. Cluster sind eine bestimmte Erscheinung des Support-Confidence-Raums und geben eine ungefähre Tendenz an wo sich bei konstantem, kleinem N und konstantem Heuristikwert mit erhöhter Wahrscheinlichkeit Punkte befinden. Im weiteren Verlauf dieser Diplomarbeit wird in den Abbildungen der Support-Confidence-Räume diese Punktwolkenanomalie immer wieder zu sehen sein.

[5] Die a-priori-Wahrscheinlichkeit muss klein sein.
[6] Dann wird die Gleichung 5.4 maximal.

6. Lift

6.1. (Un)abhängigkeiten von Kopf und Körper

Die Convictionheuristik stellt das Verhältnis von

$$V(A \rightarrow B) \;=\; \frac{supp(\overline{B})}{conf(A \rightarrow \overline{B})} \tag{6.1}$$

dar. Im Zähler steht die Quote der negativen Datensätze in der Datenbank und im Nenner die Quote der "false positives" der Regel R relativ zu allen von der Regel als positiv bewerteten Datensätzen. Die Convictionheuristik bevorzugt Regeln, die vorsichtig agieren und lieber einen Datensatz als negativ klassifizieren als möglicherweise einen "false positive" zu produzieren. Dieses Verhalten kann unter Umständen sehr nützlich sein, produziert aber einen großen Teil an "false negatives". Um dies zu verhindern, verwendet man z.B. die Liftheuristik [TKS02], auch Interest genannt.

$$lift(A \rightarrow B) \;=\; \frac{conf(A \rightarrow B)}{supp(B)} \;=\; \frac{supp(R)}{supp(A)supp(B)} \tag{6.2}$$

Bei dieser Heuristik wird der positive Klassifikationswert und nicht der Negative betrachtet. Im P-N-Raum ergibt sich nach Umformen:

$$lift(R) \;=\; \frac{p\ (P+N)}{P\ (p+n)} \quad . \tag{6.3}$$

$$lift(R) \;=\; \frac{p}{(p+n)} \;\cdot\; \frac{1}{supp(B)} \quad . \tag{6.4}$$

In der in Abb. 6.1 verwendeten Datenbank gibt es doppelt soviele negative wie positive Datensätze, dass bedeutet, der Wert von $supp(B)$ ist $1/3$ und stellt exakt die Steigung der Mittellinie dar. Für Punkte unterhalb der Diagonale gilt, dass der Confidencewert kleiner ist als $supp(B)$, der reziproke Fall gilt für Punkte oberhalb der Diagonalen. Im Supportwert des Kopfes

$$supp(B) \;=\; \frac{C0 + C2}{N}$$

steckt sowohl die Information über den Supportwert der Regel (C0-Wert), als auch die Information über den Recallwert (C2-Wert).

$$lift(R) \;=\; \frac{conf(R)\ recall(R)}{supp(R)} \tag{6.5}$$

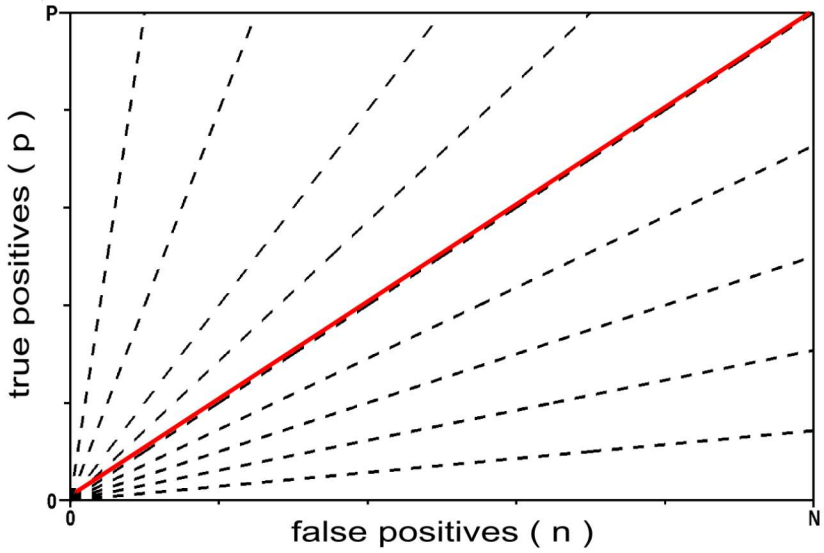

Abbildung 6.1.: Lift im P-N-Raum

$$supp(R) = \frac{conf(R)\ recall(R)}{lift(R)} \tag{6.6}$$

$$conf(R) = \frac{lift(R)\ supp(R)}{recall(R)} \tag{6.7}$$

$$recall(R) = \frac{lift(R)\ supp(R)}{conf(R)} \tag{6.8}$$

In allen Gleichungen erkennt man, dass der Recall- und der Confidencewert erneut [1] vertauschungsunabhängig sind. Der Liftwertebereich ist nach Gl. 6.5 $[0, \infty]$.

[1] siehe Accuracyheuristik

6.2. Lift im Support-Confidence/Recall-Raum

6.2.1. Liftwerte kleiner als 1

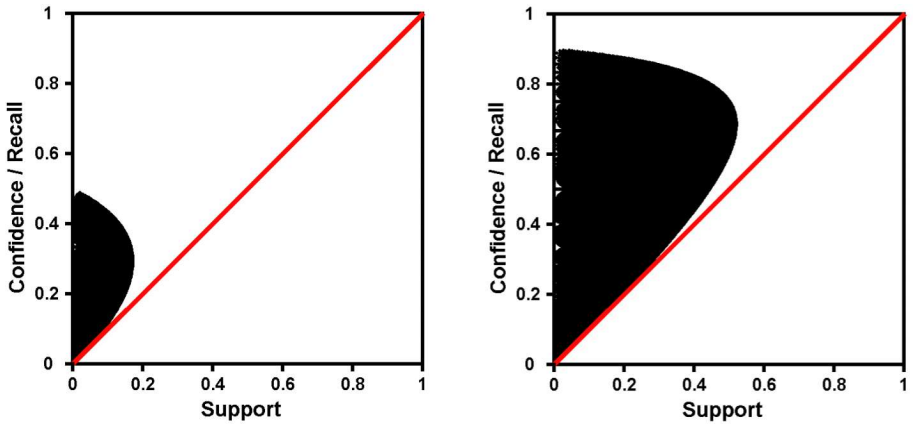

Abbildung 6.2.: Liftwerte -0.05 (links) und -0.2 (rechts) im Support-Recall/Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.4.1)

In den Abbildungen des Support-Confidence- oder des Support-Recall-Raums (Abb. 6.2) erkennt man, dass alle Punktwolken eine einheitliche Struktur haben. Für die Grenzkurven, welche die Punktwolken nach rechts beschränken, gilt:

- Die Kurven fangen alle in dem Punkt $(0, 0)$ an.
- Mit steigendem Confidencewert steigt auch der Supportwert, bis zu einem "Maximalpunkt" P_{max}, welcher sich mit dem Liftwert verändert.
- Nach dem Maximalpunkt fällt der Supportwert, mit steigendem Confidencewert wieder ab und schneidet die Confidence-Achse im Liftwert.
- Für Confidence- und Recallwerte größer als der Liftwert ist die Punktwolke nicht definiert.

Die Punktwolke wird nach links durch die Confidenceachse bzw. den Definitionsbereich des Supportwertes beschränkt. Der Supportwert bildet also sowohl linksseitig als auch rechtsseitig eine Beschränkung für die Punktwolke. Für eine einzelne Punktgraphik ist der Liftwert konstant, deshalb wird der Supportwert innerhalb einer Abbildung ausschließlich vom Produkt des Confidencewerts und des Recallwerts bestimmt.

Der maximal mögliche Supportwert steigt mit dem Confidencewert, wird also von dem Confidencewert nach oben begrenzt. Deshalb verläuft die Grenzkurve auch am Anfang auf der Diagonalen. Weil der Recallwert bei steigendem Confidencewert fallen muss (siehe Gl. 6.8), wächst der Zähler in Gl. 6.6 langsamer als der Confidencewert, damit hat die

Grenzkurve danach eine höhere Steigung als die Diagonale. Die Grenzkurve steigt daraufhin bis zum Punkt P_{max}, in diesem Punkt gilt dann:

$$recall(R) = confidence(R) \quad .$$

Nach P_{max} ist der Confidencewert größer als der Recallwert, deshalb wird der Supportwert für größere Confidencewerte durch den Recallwert nach oben begrenzt. Der Recallwert fällt weiter mit steigendem Confidencewert. Daraus resultiert, dass der Supportwert wieder gegen Null strebt.

Bei kleinen Supportwerten liegt die Grenzkurve sowohl im Support-Confidence-Raum als auch im Support-Recall-Raum auf der Diagonalen, deshalb gilt dann $supp(R) = recall(R)$. Für die Gl. 6.7 folgt

$$conf(R) = lift(R) \quad .$$

Also schneidet die Grenzkurve theoretisch die Confidenceachse im Liftwert, wenn die Support- und Recallwerte klein sind [2]. Die Punktwolke besitzt keine Punkte für Confidencekoordinaten oberhalb des Liftwerts, da der Quotient aus Support- und Recallwert nicht größer als 1 werden kann (siehe Gl. 6.7).

Es wurde noch keine Formel ermittelt, welche die Grenzkurve exakt beschreibt, bzw. man hat keine Formel, um P_{max} bei vorgegebenem Liftwert bestimmen zu können. Bei den bisherigen Heuristiken[3] wurde eine Maxima- oder Minima-Abschätzung benutzt um die Grenzkurven zu beschreiben. Dies ist in diesem Fall nicht möglich, da sich die Grenzkurve nicht durch die Formel für den Liftwert alleine ergibt[4].

Im Punkt P_{max} sind der Confidence- und Recallwert gleich, dies gilt auch noch an anderen Punkten der Punktwolke. In Abb. 6.3 ist die Kurve aller Werte mit

$$recall(R) = confidence(R)$$

weiß dargestellt. Für alle diese Punkte gilt nach Gl. 6.6:

$$supp(R) \;=\; \frac{conf(R)^2}{lift(R)} \;=\; \frac{recall(R)^2}{lift(R)} \quad . \tag{6.9}$$

Man erkennt in Abb. 6.3, dass die weiße Kurve (relativ zum Supportwert) die geringste Steigung im Punkt P_{max} hat. Die Steigung ist im Punkt P_{max} immer kleiner als 1, deshalb kann man eine Minimalabschätzung für den Supportwert treffen.

$$\frac{d}{d\ supp(R)}\ conf(R) \;=\; \frac{d}{d\ supp(R)}\ \sqrt{supp(R)\ lift(R)} \tag{6.10}$$

[2] Wobei zu beachten ist, dass eine wirkliche Schneidung nicht auftritt, da ein Supportwert 0 auch einen Confidencewert 0 zur Folge hätte.

[3] Accuracy und Conviction

[4] Der Supportwert ist durch die jeweiligen Confidence- und Recallwerte beschränkt

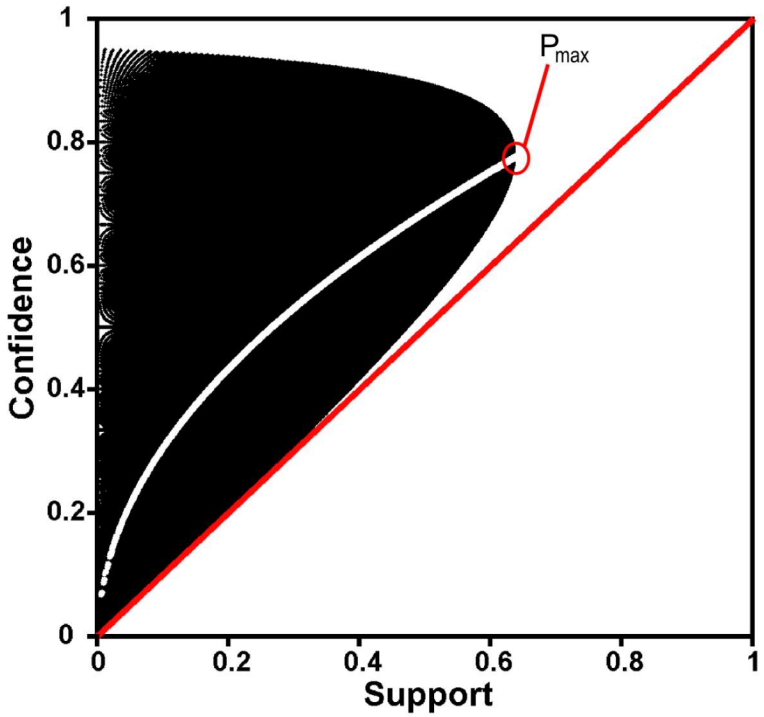

Abbildung 6.3.: Punktwolke für den Liftwert 0.95, die weiße Kurve stellt alle Punkte dar auf denen $conf(R) = recall(R)$ gilt.

$$\frac{d}{d\ supp(R)}\ conf(R) \ = \ \frac{lift(R)}{2\ \sqrt{supp(R)\ lift(R)}} \ < \ 1 \tag{6.11}$$

$$supp(R) \ > \ \frac{lift(R)}{4} \tag{6.12}$$

Man kann den Supportwert also eingrenzen. Eine untere Grenze für den Supportwert folgt aus der Gl. 6.12. Eine Obergrenze des Supportwertes ist der Liftwert selber, da der Confidence- genauso wie der Recallwert nicht den Liftwert überschreiten können.

6.2.2. Liftwerte größer als 1

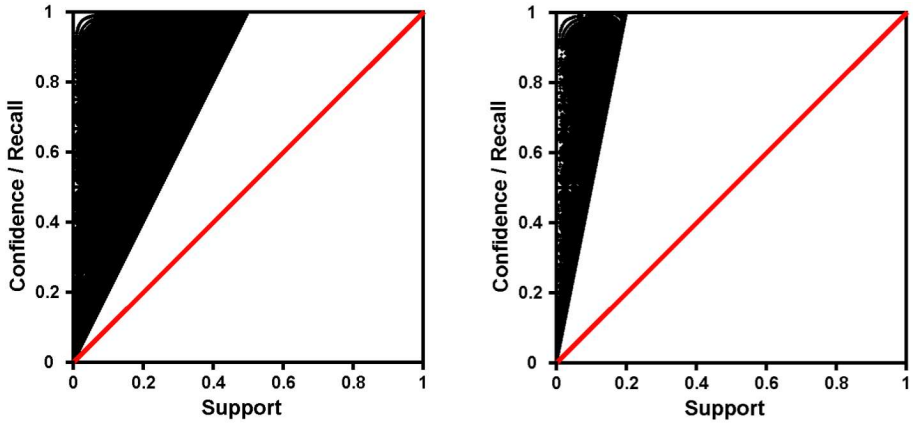

Abbildung 6.4.: Liftwerte 0.05 (links) und 0.2 (rechts) im Support-Recall/Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.4.2)

Für die Grenzkurven der Punktwolken (Abb. 6.4) für Liftwerte größer als 1 gilt:

- Die Kurven starten ebenfalls in dem Punkt $(0, 0)$.
- Der maximale Supportwert (Grenzfunktion) steigt linear mit dem Confidencewert an. Der maximale Confidencewert, jetzt 1, wird bei $supp(R) = 1/lift(R)$ erreicht.
- Die Steigung der Grenzfunktion ist folglich $1/lift(R)$

Anhand der Gl. 6.6 erkennt man, dass für einen Liftwert größer als 1 das Produkt aus Confidence- und Recallwert keiner Beschränkung unterliegt. Es gilt weiterhin, dass nach Gl. 6.7 und 6.8 beide Werte sich antiproportional zu einander verhalten. Es ist möglich, dass beide Werte ihren Maximalwert 1 annehmen. Genau dies geschieht, wenn die Grenzkurve die Confidencewert-1-Linie schneidet, deshalb kann man die Grenzkurve mit Hilfe der Formel

$$max\{supp(R)\}_{conf(R)} = \frac{conf(R)\ max\{recall(R)\}}{lift(R)} = \frac{conf(R)}{lift(R)} \tag{6.13}$$

beschreiben.

6.3. Lift im Recall-Confidence-Raum

6.3.1. Liftwerte kleiner als 1

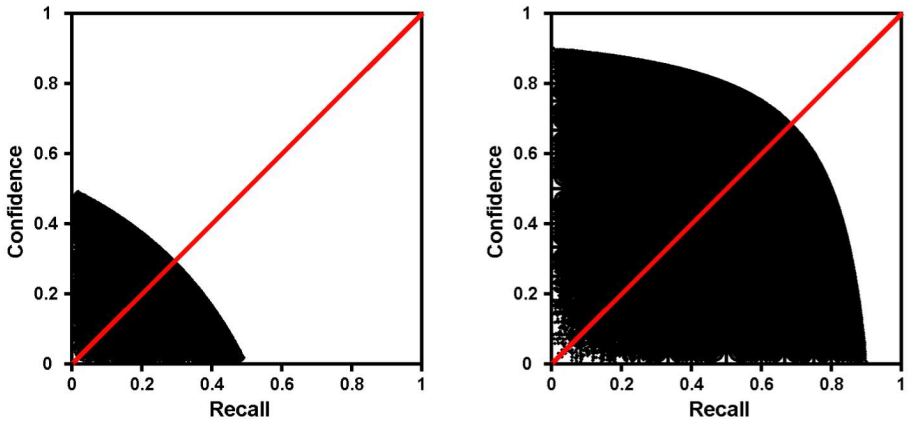

Abbildung 6.5.: Liftwerte -0.05 (links) und -0.2 (rechts) im Recall-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.4.5)

Im Recall-Confidence-Raum (Abb. 6.5) kann man die Vertauschungsunabhängigkeit von Confidence- und Recallwert anhand der Spiegelung der Punktwolke an der Diagonalen erkennen. Für die einzelnen Punktgraphiken gilt:

> Für kleine Liftwerte befindet sich die Punktwolke im Quadraten C (siehe Abb. 3.3), ihre Form ähnelt einem Kreisauschnitt. Die Grenzfunktion der Punktwolke wird durch das Verhältnis $conf(R) \cdot recall(R)$ bestimmt (s.o.).

Der Punkt auf der Grenzfunktion, an welchem der Recall- und Confidencewert gleich sind, ist der äquivalente Punkt zu P_{max} im Support-Confidence/Recall-Raum. P_{max} befindet sich im Recall-Confidence-Raum auf der Diagonalen. Mit steigendem Liftwert steigt auch der Punkt P_{max}, damit verbunden breitet sich die Punktwolke im Recall-Confidence-Raum weiter aus.

6.3.2. Liftwerte größer als 1

Wenn der Liftwert größer als 1 ist (siehe Abb. 6.6), dann erstreckt sich die Punktwolke über den kompletten Confidence-Recall-Raum. Mit steigendem Liftwert wird die Dichte der Punktwolke immer geringer. Das bedeutet, die Flächen zwischen den gefundenen Punkten werden immer größer. Dafür gibt es eine steigende Anzahl an Punkten im rechten oberen Bereich des Quadranten B.

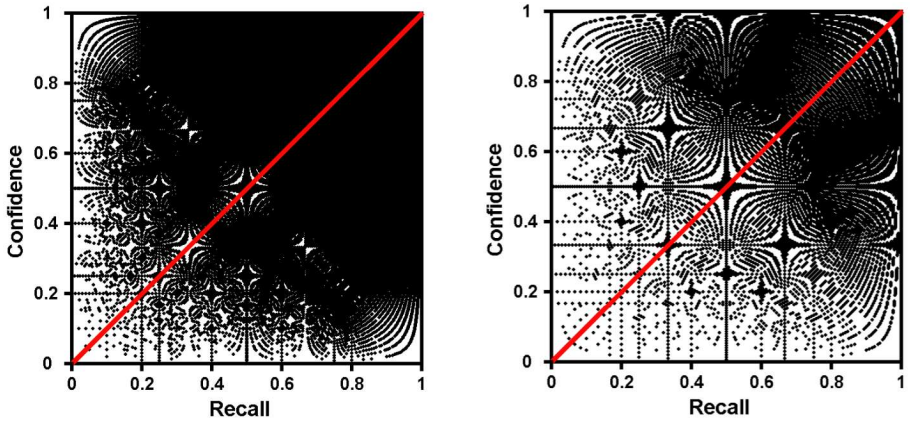

Abbildung 6.6.: Liftwerte 0.05 (links) und 0.2 (rechts) im Recall-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.4.6)

Auch für hohe Liftwerte stellt man fest, dass immer noch Punkte im niedrigen Confidence- und Recallwertebereich existieren. Da N groß ist, wird der Supportwert sehr klein, für kleine C0. Daraus resultiert, dass es für kleine Confidence- oder Recallwerte immer noch viel kleinere Supportwerte gibt. Das bedeutet, der Liftwert ist größer als 1, obwohl Support-, Confidence- und Recallwert klein sind.

Ist der Supportwert viel kleiner als der Confidence- und Recallwert (siehe Gl. 6.5), dann ergeben sich eher große Liftwerte. Darum ist in den Abb. 6.6 auch der obere rechte Quadrant fast komplett "abgedeckt".

6.4. Liftwertfelder im Support-Confidence/Recall-Raum

Durch die Auswertung einer Regel R wird ein Punkt im Assoziationsraum erzeugt. Die Position dieses Punktes wird durch den Support- und Confidencewert bestimmt.

$$P_1 \;=\; (supp(R), conf(R)) \tag{6.14}$$

Für den Punkt P_1 kann man eine Abschätzung über den möglichen Liftwert erstellen. Für Liftwerte unterhalb von 1 (siehe Abb. 6.7) gilt:

$$conf(R) \;\leq\; lift(R) \;\leq\; 1 \quad . \tag{6.15}$$

Mit Hilfe des Supportwertes kann man keine zusätzlichen Aussagen über den Liftwert unterhalb von 1 treffen. Der Supportwert ist zwischen den Grenzen

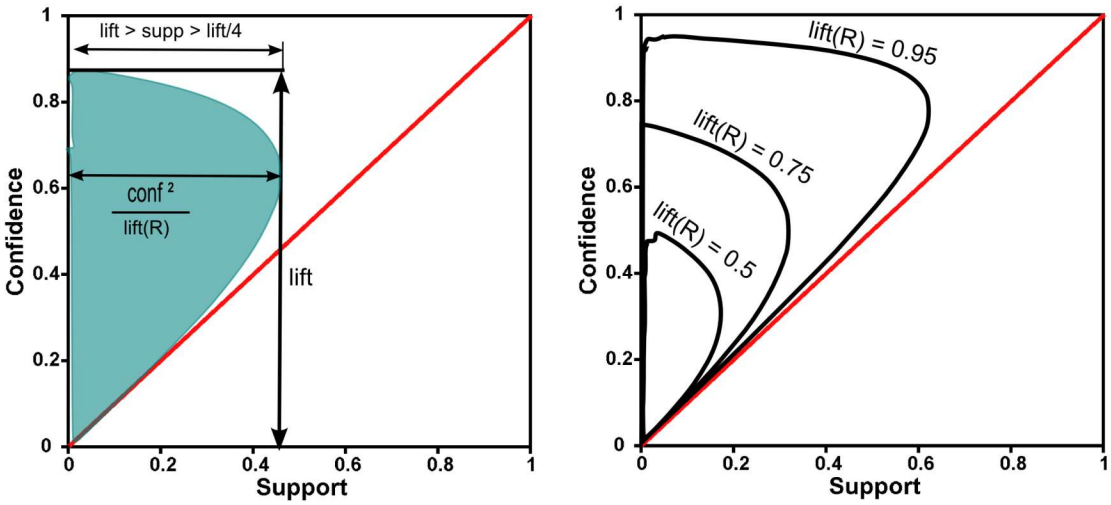

Abbildung 6.7.: Liftwertfelder (Liftwert kleiner 1) im Assoziationsraum

Bild Links: Grünes Feld für alle mögl. Support-Confidence-Wert-Kombinationen mit Liftwerten kleiner gleich dem eingezeichneten Liftwert.
Bild Rechts: Die Bogenlinien stellen die Grenzen für einzelne Liftwerte dar. Links von einer Linie gilt der Liftwert noch, rechts nicht mehr.

$$\frac{lift(R)}{4} \leq supp(R) \leq conf(R) \tag{6.16}$$

definiert worden, dies engt das Liftwertintervall nicht weiter ein. Für Liftwerte über 1 (siehe Abb. 6.8) gilt:

$$supp(R) \leq \frac{conf(R)}{lift(R)} \tag{6.17}$$

$$lift(R) \leq \frac{conf(R)}{supp(R)} \tag{6.18}$$

Damit haben wir eine Obergrenze für den Liftwert definiert. Das mögliche Liftwertintervall beträgt somit

$$lift(R) = \left[conf(R) , \frac{conf(R)}{supp(R)}\right] \quad . \tag{6.19}$$

Anhand des Support- und Confidencewerts kann man nicht entscheiden, ob der zugehörige Liftwert unter- oder oberhalb von 1 liegt, da in beiden Fällen die Punktwolken den

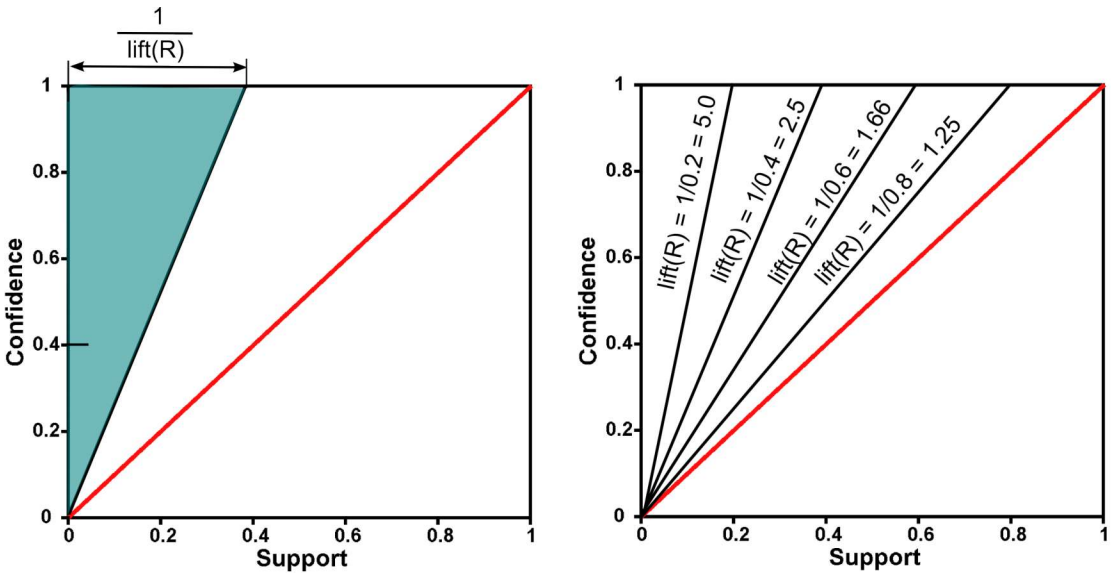

Abbildung 6.8.: Liftwertfelder (Liftwert größer 1) im Assoziationsraum

Bild Links: Grünes Feld für alle mögl. Support-Confidence-Wert-Kombinationen mit Liftwerten größer gleich dem eingezeichneten Liftwert.
Bild Rechts: Die Linien stellen die Grenzen für einzelne Liftwerte dar. Links von einer Linie gilt der Liftwert noch, rechts nicht mehr.

kompletten Support-Confidence-Raum erreichen können [5]. Man weiß allerdings folgendes:

- Liegt P_1 in der Nähe des Punktes $(1, 1)$, dann ist der Liftwert ca. 1.
- Liegt P_1 in der Nähe der Diagonalen, dann ist der Liftwert ebenfalls ca. 1.
  - Es sei denn, P_1 liegt in der Nähe von $(0, 0)$, dann ist fast jeder Liftwert möglich.
- Liegt P_1 in der Nähe von $(0, 1)$, dann ist der Liftwert entweder nahe bei 1 oder sehr groß.

[5] für Liftwerte um den Wert 1

6.5. Fazit

Die Liftheuristik gibt das Verhältnis

$$lift(R) \;=\; \frac{supp(A \rightarrow B)}{supp(A)\ supp(B)} \tag{6.20}$$

an. Wenn die Liftheuristik kleiner als 1 ist (der Nenner dominiert), dann ist es wahrscheinlicher, dass die Produkte aus A und B nicht zusammengekauft werden, oder der Kauf von A oder B den Kauf der anderen Produkte negativ beeinflusst. Ebenso gilt, wenn der Liftwert größer als 1 ist (der Zähler dominiert in Gl. 6.20), dann hat der Kauf der Produkte A einen positiven Einfluss auf den Kauf der Produkte B und umgekehrt.

Wenn man nur den Support und Confidencewert der Regel R kennt, kann man den möglichen Liftwert eingrenzen:

$$lift(R) \;=\; \left[conf(R)\ ,\ \frac{conf(R)}{supp(R)}\right]$$

7. Leverage

7.1. Leveragefunktionen im P-N-Raum und Assoziationsraum

Eine zur Liftheuristik analoge Betrachtung ist die Leverageheuristik (leverage = Hebelwirkung) [TKS02]. Betrachtet wird die Differenz des Supports der Regel R und der separaten Wahrscheinlichkeiten von Körper A und Kopf B [1]. Dies bedingt, dass die Leverageheuristik negative Werte annehmen kann. In [TKS02] ist der Definitionsbereich des Leveragewerts mit $[-0.25, 0.25]$ angegeben [2].

$$leverage(R) = supp(R) - supp(A) \cdot supp(B) \tag{7.1}$$

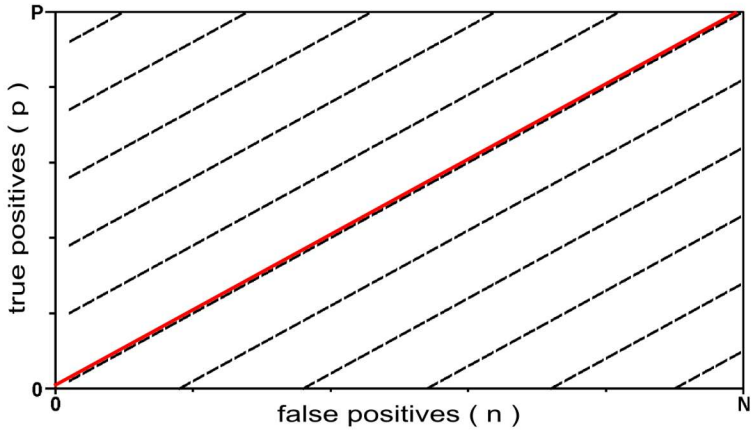

Abbildung 7.1.: Leverage im P-N-Raum

Der Support von B stellt erneut [3] einen Wert dar, welcher sich nicht mit dem Heuristikwert, sondern nur mit dem Klassifikationswert verändert. Der Supportwert von A ist aus dem letzten Kapitel ebenfalls bekannt. Folglich ergibt sich die Darstellung der Leveragefunktion für den P-N-Raum.

[1] Die Liftheuristik dagegen stellt das Verhältnis zwischen diesen Teilheuristiken dar (siehe Gl. 6.20).
[2] Der Grund ist in [TKS02] nicht angegeben, wird aber in den folgenden Teilkapiteln erläutert
[3] siehe vorherige Kapitel

$$leverage(R) \;=\; \frac{p}{P+N} \;-\; \frac{p+n}{P+N} \cdot \frac{P}{P+N} \;=\; \frac{p}{P+N} \;-\; \frac{P\,(p+n)}{(P+N)^2} \tag{7.2}$$

In der Abb. 7.1 sind die "Isometrien" für die Leverageheuristik dargestellt. Die Isometrien sind alle parallel zur Diagonalen. Die Leverageheuristik wird häufig auch "Weighted-Relative-Accuracy" genannt. Das bedeutet bei der Accuracyheuristik werden die Isometrien zum Verhältnis von P und N normiert ($(P+N)$ in Gl. 7.2) , deshalb sind die Isometrien dann parallel zur Diagonalen.

Für Werte oberhalb der Diagonalen gilt, dass der Leveragewert positiv ist und für Werte unterhalb der Diagonalen folglich, dass der Leveragewert negativ ist. Durch die ersten Kapitel weiß man, dass

$$conf(R) \;=\; \frac{supp(R)}{supp(A)} \quad \wedge \quad recall(R) \;=\; \frac{supp(R)}{supp(B)}$$

gilt, deshalb kann man anhand der Gl. 7.1 die Funktionen für den Assoziationsraum herleiten.

$$leverage(R) \;=\; supp(R) \;-\; \frac{supp(R)^2}{recall(R)\;conf(R)} \;=\; supp(R) \;-\; \frac{supp(R)}{lift(R)} \tag{7.3}$$

$$supp(R) \;=\; \frac{recall(R)\;conf(R)}{2}\left(1 \;\pm\; \sqrt{1 \;-\; \frac{4\;leverage(R)}{recall(R)\;conf(R)}}\right) \tag{7.4}$$

$$conf(R) \;=\; \frac{supp(R)^2}{recall(R)\cdot(supp(R) \;-\; leverage(R))} \tag{7.5}$$

$$recall(R) \;=\; \frac{supp(R)^2}{conf(R)\cdot(supp(R) \;-\; leverage(R))} \tag{7.6}$$

Die Recall- und Confidencewerte sind, wie bei der Liftheuristik vertauschungsunabhängig.

$$recall(R)\;conf(R) \;=\; \frac{supp(R)^2}{supp(R) \;-\; leverage(R)} \tag{7.7}$$

7.2. Leverage im Support-Confidence-Raum

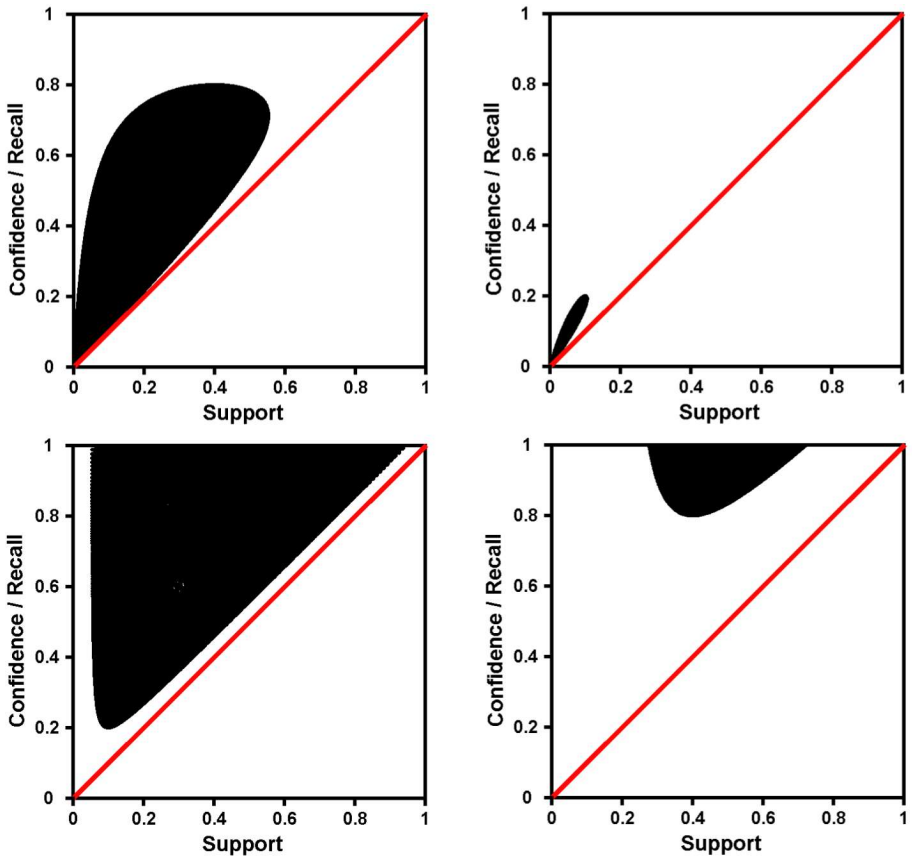

Abbildung 7.2.: Leveragewerte -0.05 (oben links), -0.2 (oben rechts), 0.05 (unten links) und 0.2 (unten rechts) im Support-Recall/Confidence-Raum mit N = 1500 (für weitere Abb. siehe siehe Anhang A.5.1 und A.5.2)

In der weiteren Analyse wird versucht, die Grenzfunktionen der Punktwolken (s. Abb. 7.2) zu beschreiben. Dazu muss man z.B. in Gl. 7.4 Bedingungen finden, die bei bestimmten Leveragewerten auf bestimmte Supportwertebereiche schließen lassen. Da der Supportwertebreich $[0, 1]$ nur für die rationalen Zahlen definiert ist, weiß man, dass der Wert unterhalb der Wurzel positiv sein muss. Für positiven

$$L_+ = |leverage(R)| \tag{7.8}$$

oder negativen Leveragewert

$$L_- = - |leverage(R)| \tag{7.9}$$

ergeben sich damit unterschiedliche Voraussetzungen für den Wert unterhalb der Wurzel in Gl. 7.4.

7.2.1. Supportwertfunktion mit negativem Leveragewert

Falls der Leveragewert negativ ist, gilt nach Gl. 7.3:

$$supp(R) \; < \; \frac{supp(R)^2}{recall(R)\; conf(R)}$$

$$supp(R) \; > \; recall(R) \cdot conf(R)$$

Die Ungleichung bestimmt die Form der Punktwolke, für die Grenzkurve der Punktwolke im Support-Confidence-Raum gilt folglich:

$$supp(R) \; \approx \; recall(R) \cdot conf(R) \quad .$$

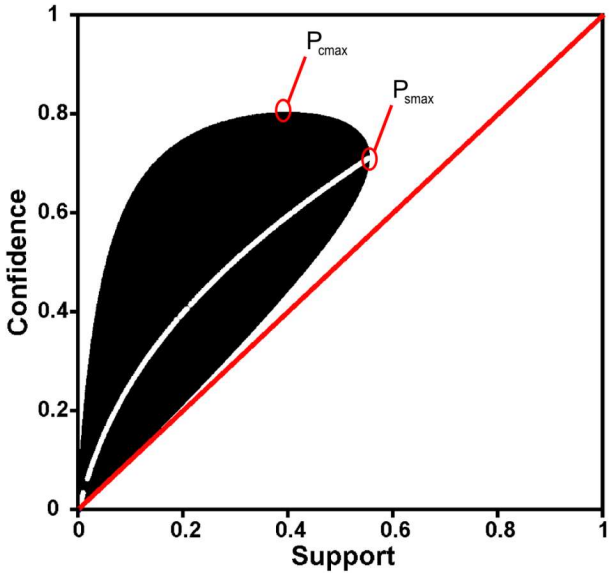

Abbildung 7.3.: Punktwolke für den Leveragewert -0.05, die weiße Kurve stellt alle Punkte dar auf denen $conf(R) = recall(R)$ gilt.

In Abb. 7.3 ist eine Punktwolke für negativen Leveragewert dargestellt, die weiße Kurve bildet alle Punkte innerhalb der Punktwolke ab bei denen $conf(R) = recall(R)$ gilt. Die Kurve schneidet die Grenzkurve der Punktwolke in P_{smax}, dort ist der Supportwert maximal.

Für einen beliebigen Punkt innerhalb der Punktwolke gilt, je geringer die Differenz zwischen der Wurzel des Supportwertes und dem Confidencewert ist, desto näher liegt der Punkt an der weißen Linien in Abb. 7.3.

Analog zu P_{smax} kann man den Punkt P_{cmax} mit dem höchsten Confidencewert innerhalb der Punktwolke bestimmen. Zur Herleitung braucht man die Analyse der positiven Leverагewerte. Deshalb soll an dieser Stelle nur bemerkt werden, dass der Confidencewert von P_{cmax} immer

$$P_{cmax} \;=\; 1 - 4\; leverage(R)$$

beträgt.

Für den negativen Leveragewertebereich kann man das variable Vorzeichen in der Gl. 7.4 eliminieren, dazu dienen die Funktionen L_+ und L_-.

$$supp(R) \;=\; \frac{recall(R)\; conf(R)}{2} \left(1 \;\pm\; \sqrt{1 \;+\; \frac{4\; L_+}{recall(R)\; conf(R)}} \right) \;\geq\; 0 \qquad (7.10)$$

Nach unten ist der Support durch den Wert 0 beschränkt, nach Umformen ergibt sich dann die Bedingung:

$$0 \;\leq\; 1 \;\pm\; \sqrt{1 \;+\; \frac{4\; L_+}{recall(R)\; conf(R)}} \quad . \qquad (7.11)$$

Bei negativem Wurzelvorzeichen muss der Wert innerhalb der Wurzel kleiner als 1 sein, damit die Voraussetzung erfüllt ist. Der Wert unterhalb der Wurzel ist aber immer größer als 1.

> Wenn der Leveragewert negativ wird, dann darf das Vorzeichen vor der Wurzel nur positiv sein!

Mit Hilfe der Funktion L_+ kann man auch die Formeln 7.5 und 7.6 umformen.

$$conf(R) \;=\; \frac{supp(R)^2}{recall(R) \cdot (supp(R) \;+\; L_+)} \qquad (7.12)$$

$$recall(R) \;=\; \frac{supp(R)^2}{conf(R) \cdot (supp(R) \;+\; L_+)} \qquad (7.13)$$

7.2.2. Supportwertfunktion mit positivem Leveragewert

Analog für negative Leveragewerte folgt aus Gl. 7.4:

$$supp(R) \;=\; \frac{recall(R)\;conf(R)}{2}\left(1 \;\pm\; \sqrt{1 \;-\; \frac{4\;L_+}{recall(R)\;conf(R)}}\right) \;\geq\; 0 \quad (7.14)$$

$$0 \;\leq\; 1 \;\pm\; \sqrt{1 \;-\; \frac{4\;L_+}{recall(R)\;conf(R)}} \quad (7.15)$$

Der Wert innerhalb der Wurzel ist immer kleiner gleich 1 und positiv. Damit ist sowohl für ein positives, als auch für ein negatives Vorzeichen der Wurzel die Bedingung aus der Formel 7.15 erfüllt. In den Abb. des Support-Confidence-Raums erkennt man, dass die Punktwolken auf der Supportwertachse den Leveragewert nicht unterschreiten. Für die Grenzkurve der Punktwolke gilt:

- Mit steigendem Supportwert fällt der Confidencewert, bis zu seinem Minimum, danach steigt der Confidencewert wieder an.
- Die Grenzkurve schneidet irgendwann die Linie mit Confidencewert 1. Der Supportwert in diesem Schnittpunkt stellt auch den rechten Grenzwert der Punktwolke dar.

Anhand der Confidence- (7.5) und Recall-Funktionen (7.6) mit positiven Leveragewert kann man diese Beobachtung begründen. Der Supportwert muss immer größer als der Leveragewert sein, damit der Nenner der Formeln nicht negativ wird oder sogar Null, denn damit würde der jeweilige Wert der Gesamtfunktion auch negativ werden oder undefinierbar.

Für den positiven Leveragewertebereich gibt es zu jedem Confidencewert einen Recallwert mit 1 [4]. Die Gl. 7.5 hat bei maximalem Recallwert das Minimum.

$$min\;\{conf(R)\} \;=\; \frac{supp(R)^2}{support(R) \;-\; leverage(R)} \quad . \quad (7.16)$$

Da der Supportwert in dieser Gl. variabel ist, gibt es für jeden Supportwert einen minimalen Confidencewert, aber der absolute minimale Confidencewert ist unabhängig vom Supportwert. Für den absoluten minimalen Confidencewert muss die Supportfunktion die Bedingung

$$supp(R) \;=\; \frac{recall(R)\;conf(R)}{2}\left(1 \;\pm\; \sqrt{1 \;-\; \frac{4\;L_+}{recall(R)\;conf(R)}}\right) \;\geq\; 0 \quad (7.17)$$

erfüllen. Der Wert unterhalb der Wurzel darf nur einen positiven Wert annehmen:

[4] siehe Abb. Recall-Confidence-Raum für positive Leveragewerte

$$recall(R) \cdot conf(R) \geq 4L_+ \quad . \tag{7.18}$$

Das erklärt zum einen den Wertebereich für die Leveragefunktion (s.o.), zum anderem gibt diese Bedingung für einen positiven Leveragewert eine untere Schranke an. Auf der Grenzkurve gilt immer noch $recall(R) = 1$. Deshalb kann der Confidencewert niemals unterhalb des vierfachen Leveragewerts fallen. Analog gilt für das Maximum der Supportwertfunktion im positiven Leveragewertebereich

$$max\{recall(R) \cdot conf(R)\} \quad .$$

Der Recallwert und der Confidencewert sind (s.o.) hier 1. Für die Gleichung der Supportwertfunktion folgt damit

$$max\{supp(R)\} \approx \frac{1}{2}\left(1 + \sqrt{1 - 4\,L_+}\right) \tag{7.19}$$

$$max\{supp(R)\} \approx \frac{conf(R)}{2}\left(1 \pm \sqrt{1 - \frac{4\,leverage(R)}{conf(R)}}\right) \quad . \tag{7.20}$$

7.3. Recall-Confidence-Raum für die Leverageheuristik

In den Abb. des Recall-Confidence-Raums erkennt man sowohl für positiven als auch für negativen Leveragewert, die Vertauschungsunabhängigkeit von Recall und Confidencewert bzgl. der Spiegelung an der Diagonalen. Die Punktwolken bilden für negativen Leveragewert eine Art “Ballon”, welcher mit steigendem Leveragewert immer weiter vom Nullpunkt aus “aufgeblasen” wird.

Dieses Verhalten ist identisch mit dem in den Punktgraphiken der Liftwerte, nur das dort der Wertebereich nicht bei -0.25, sondern bei 0 startet. Beim Liftwertbereich oberhalb von 1 und dem positiven Leveragewertbereich erkennt man ebenfalls ein “gleichartiges” Verhalten. In beiden Fällen befindet sich die Punktwolke, bzw. die Mehrzahl aller Punkte, im oberen rechten Quadranten. Dieses Verhalten entsteht dadurch, dass der Liftwert und der Leveragewert das Verhältnis zwischen denselben Werten darstellen. Trotzdem zeigen sich kleine Unterschiede zwischen den Punktwolken beider Funktionen. Die Leverageheuristikpunktwolken haben exakt definiertere Ränder als die Liftheuristikpunktwolken.

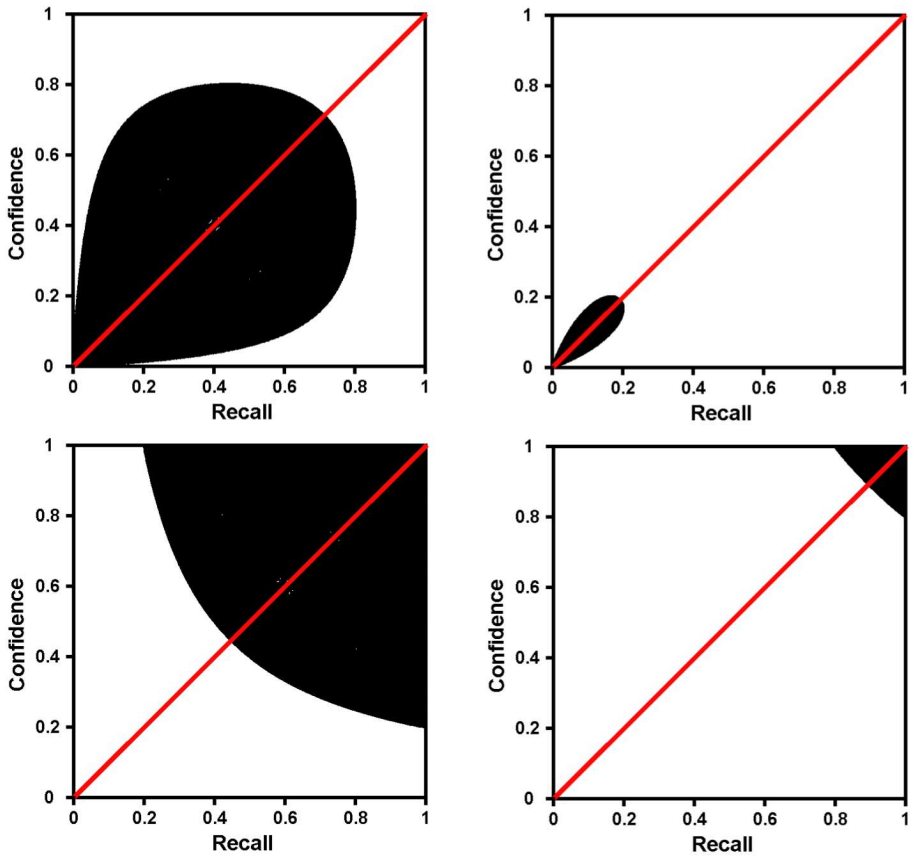

Abbildung 7.4.: Leveragewerte -0.05 (oben links), -0.2 (oben rechts), 0.05 (unten links) und 0.2 (unten rechts) im Recall-Confidence-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.5.3 und A.5.4)

Minimaler Liftwerte- und Leveragewerte-Bereich

Während die Punktwolken für Liftwerte kleiner als 1 eine Art "Viertelkugel" bilden, haben die Punktwolken für negative Leveragewerte die schon angesprochenen "Ballonform". Wie im vorhergehenden Abschnitt gezeigt wurde, gilt für einen negativen Leveragewert die Bedingung

$$supp(R) \geq conf(R)\, recall(R) \quad . \tag{7.21}$$

Die "Äußere-Hülle" der Punktwolke in Anhang A.5.3 und A.5.4 stellt das jeweilige Maximum dar an dem sich die Voraussetzung gerade noch erfüllt. Das bedeutet, außer im Ursprung des Confidence-Recall-Raums, gibt es keinen Punkt mehr an dem einer der Werte Null wird. Durch diese Erkenntnisse ist es möglich die potentiellen Felder des Support-Confidence- (bzw. Support-Recall-) Raums zu beschreiben.

7.4. Leveragewertfelder

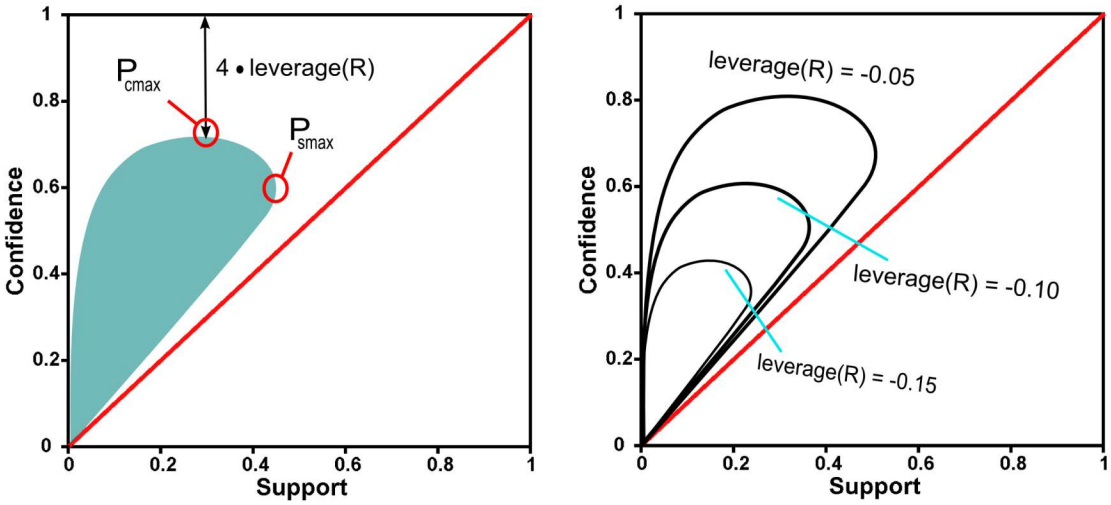

Abbildung 7.5.: Leveragewertfelder (negativer Leveragewert) im Assoziationsraum

Bild Links: Grünes Feld für alle Leveragewerte kleiner gleich dem eingezeichneten Leveragewert. Bild Rechts: Die Bogenlinien stellen die Grenzen für einzelne Leveragewerte dar. Innerhalb einer Linie gilt der Leveragewert noch, ausserhalb nicht mehr.

Man habe einen Punkt P_1 im Support-Confidence-Raum gegeben, dieser Punkt wird durch den Support- und Confidencewert repräsentiert. Der mögliche Leveragewert ist gesucht. Zu jedem möglichen Leveragewert gibt es eine Punktwolke. Der Punkt P_{smax} (Abb. 7.5) einer jeden Punktwolke zeichnet sich durch 2 Eigenschaften aus: Zum Einen müssen der Confidence- und Recallwert gleich sein, zum Anderen die Voraussetzung für die Grenzkurve (s.o.) erfüllt sein. Insgesamt muss also für den Punkt P_{smax} gelten:

$$supp(R) \geq recall(R)^2 = conf(R)^2 \tag{7.22}$$

Den möglichen Leveragewert für den Punkt P_{smax} kann man dann durch

$$min\{L_+\} = supp(R) - \left(\frac{supp(R)}{conf(R)}\right)^2 \tag{7.23}$$

$$leverage \geq - min\{L_+\} \tag{7.24}$$

ermitteln. Mit steigendem Leveragewert vergrößert sich die Punktwolke. Ein Punkt P_1 liegt immer innerhalb irgendeiner Punktwolke. Der minimale Leveragewert von P_1 gilt für eine

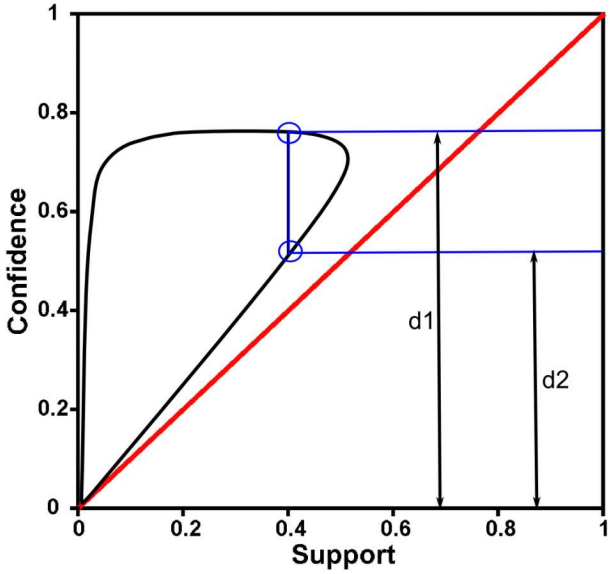

Abbildung 7.6.: Wenn der Confidencewert gleich d1 ist, dann ist der Recallwert gleich d2 und umgekehrt.

Punktwolke, bei welcher P_1 und der Punkt P_{smax} gleich sind. Ist der Punkt P_1 nicht P_{smax} in der Punktwolke, welche P_1 einschließt, dann ist die Bedingung $conf(R) = recall(R)$ nicht zwingend erfüllt. Dann kann es sein, dass der Wert der Gl. 7.23 größer oder kleiner ist als der mögl. Leveragewert (Gl. 7.26). Deshalb ist die Gleichung 7.23 nicht für eine Minimal-Leveragewert-Abschätzung geeignet.

Da der Recall- und der Confidencewert antiproportional zueinander sind (siehe Gl. 7.5 und 7.6), kann man beide Werte anhand der Punktwolke ablesen [5] (siehe Abb.7.6). In der Abb. 7.6 erkennt man, dass auf der Grenzkurve zu jedem Supportwert 2 Confidencewerte existieren. Diese beiden Werte verhalten sich zueinander wie der Confidence- zum Recallwert. Das bedeutet, wenn der Confidencewert einen von beiden Punkten als Wert annimmt, dann hat der Recallwert genau den Confidencewert des anderen Punktes, und umgekehrt. Für die Multiplikation beider Werte gilt weiterhin die Bedingung

$$supp(R) \geq conf(R)\, recall(R) \quad .$$

Man gehe erneut vom Anfangsproblem aus, man habe Punkt P_1 und möchte genau wissen, welchem Leveragewert dies entspricht. Der Recallwert zum Punkt P_1 muss also kleiner sein als $supp(R)/conf(R)$ bzw. der Confidencewert kleiner als $supp(R)/recall(R)$. Ohne den jeweiligen Recallwert zu den Koordinatenpunkten von P_1 kann man die genaue Form der möglichen Punktwolke nicht bestimmen.

[5] Das gilt nur für den Rand/Grenzkurve

7.4.1. Der Punkt Pcmax

Es ist jedoch möglich, den größten Confidencewert P_{cmax} einer Punktwolke[6] zu bestimmen. Bei genauer Betrachtung der Punktwolken im Anhang A.5.1 stellt man fest, dass der Confidencewert des Punktes P_{cmax} immer den vierfachen L_+-Wert als Abstand zur Achse mit Confidencewert 1 hat. Also ergibt sich augenscheinlich die Formel

$$conf_{cmax}(R) = 1 - 4L_+ \tag{7.25}$$

für den Confidencewert im Punkt P_{cmax}. Der Supportwert ist immer der halbe Confidencewert.

$$supp_{cmax}(R) = \frac{1 - 4L_+}{2} \tag{7.26}$$

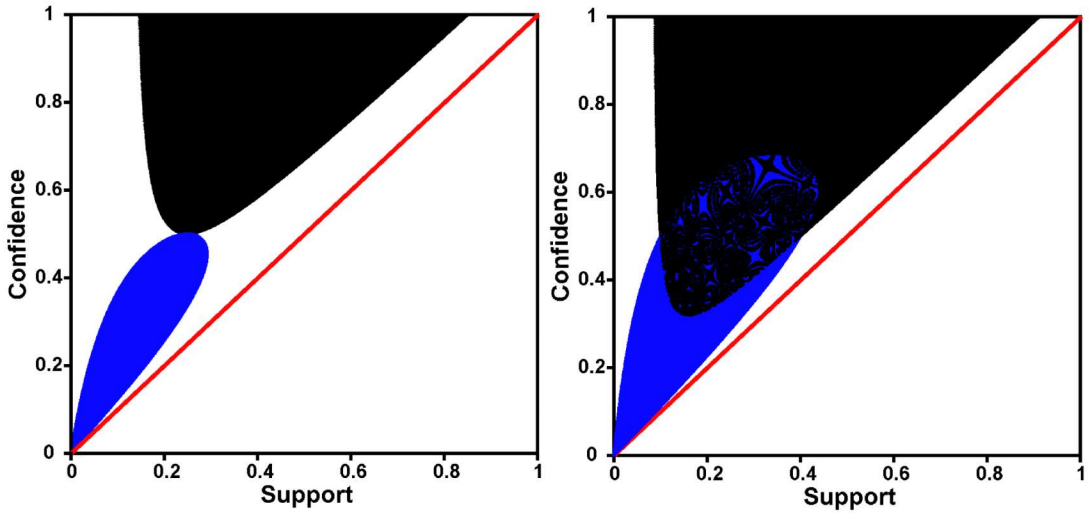

Abbildung 7.7.: Positive(schwarz) und Negative(blau) Leveragewertpunktwolken 0.125(-0.125) (links) und 0.08(-0.08) (rechts)

In der Abb. 7.7 (links) erkennt man, dass der maximale Confidencewert bei negativem Leveragewerte (-0.125) und der minimale Confidencewert für positivem Leveragewert (0.125) gleich sind. Im rechten Bild in Abb. 7.7 erkennt man dann für den Leveragewert -0.08 (0.08), dass die beiden Confidence-Extrempunkte sich gleich weit vom Confidencewert 0.5 (Leveragewert 0.125 bzw. -0.125) entfernt haben.

Das bedeutet (siehe alle Abb. im Anhang A.5.1 und A.5.2), der Extrempunkt einer Punktwolke wächst und fällt linear mit dem Leveragewert. Der Abstand ist sowohl für pos.

[6] für negative Leveragewerte

als auch für neg. Leveragewert immer der Gleiche. Da für pos. Leveragewerte der Wert des minimalen Confidencewerts der vierfache Leveragewert ist, kann man folgern, dass sich der maximale Confidencewert für den komplementären negativen Leveragewert durch die Subtraktion

$$conf(R)_{cmax} \leq 1 - 4L_{+} \quad . \tag{7.27}$$

ergibt. Wenn sowohl der negative als auch der positive Leveragewert fast 0 sind, dann ist der neg. Leveragewert vom Ursprung zur Confidencewert-1-Linie gewachsen und der pos. Leveragewert umgekehrt von der Confidencewert-1-Linie zum Ursprung (siehe Abb. im Anhang A.5.1 und A.5.2). Da der Supportwert im Punkt P_{cmax} einmal, z.B. beim Leveragewert -0.125, halb so groß war wie der Confidencewert und sich die Punktwolke wie gerade geschildert konstant ausdehnt, gilt die Aussage

$$conf(R)_{cmax}(R) \leq 2 \cdot supp(R)_{cmax}(R) \tag{7.28}$$

für alle mögliche Punkte P_{cmax} mit verschiedenen Leveragewerten.

7.4.2. Maximaler Lift- und Leveragewerte-Bereich

Der maximale Liftwert ist nicht exakt definiert und kann beliebig groß werden, wogegen der Leveragewert maximal den Wert 0.25 erreichen kann. Dies wurde bewiesen, indem man für positive Leveragewerte eine Abschätzung des Wertes unterhalb der Wurzel in Gl. 7.4 getroffen hat. Es folgt die Erkenntnis, dass das Produkt aus Confidence- und Recallwert niemals den vierfachen Leveragewert übersteigen kann.

$$recall(R) \cdot conf(R) \geq 4L_{+} \tag{7.29}$$

Da beide Werte nicht größer als 1 werden, wird das Produkt auch nicht größer als 1, da der Leveragewert 0.25 als obere Grenze hat.

Bei der Liftwertfunktion hat man erkannt, dass selbst bei hohen Liftwerten noch Kombinationen von Confidence- und Recallwerten entstehen können welche sich im Bereich um den Nullpunkt befinden. Dies gilt für die Leveragepunktwolken in Abb. Anhang A.5.4 nicht.

Es gilt analog zu den negativen Leveragewerten die umgekehrte Voraussetzung für positiven Leveragewerte:

$$supp(R) \leq conf(R)\, recall(R) \tag{7.30}$$

Durch erneutes Betrachten der Formel 7.29 erkennt man, weshalb die Punktwolken in den Abb. Anhang A.5.4 nach unten exakt beschränkt sind.

In Abb. 7.5 sind die Felder der positiven Leveragewerte dargestellt. Man kann die Lage der Punktwolke alleine aus dem Leveragewert rekonstruieren. Die Felder überschneiden sich mit

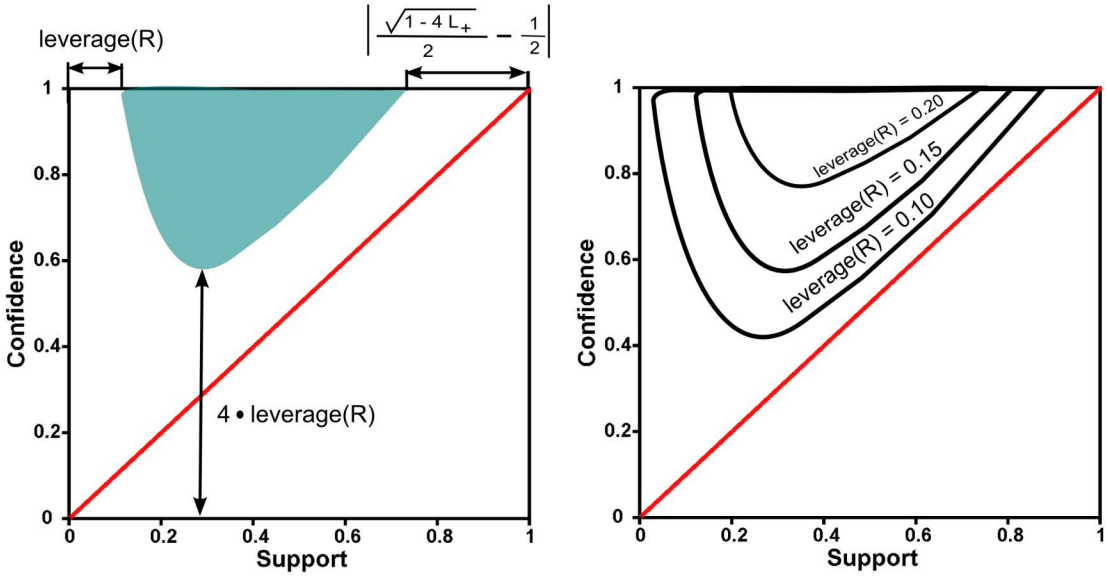

Abbildung 7.8.: Leveragewertfelder (pos. Leveragewert) im Assoziationsraum

Bild Links: Grünes Feld für alle Leveragewerte größer gleich dem eingezeichneten Leveragewert. Bild Rechts: Die Bogenlinien stellen die Grenzen für einzelne Leveragewerte dar. Innerhalb von einer Linie gilt der Leveragewert noch, ausserhalb nicht mehr.

den Feldern aus den negativen Leveragewertfeldern. Eine exakte Aussage aus dem Punkt P_1 über den Leveragewert ist nicht möglich, wenn man den Recallwert nicht kennt.

7.5. Der Coverage-Confidence-Raum

Besser sichtbar wird das Verhalten der Support- und Confidencewerte des Punktes P_{cmax} im Coverage-Confidence-Raum (s. Abb.7.9). Der Coverage-Confidence-Raum entsteht indem man den jeweiligen Supportwert eines Punktes im Support-Confidence-Raums durch seinen Confidencewert teilt (normiert).

$$coverage = \frac{supp(R)}{conf(R)} = supp(A) \tag{7.31}$$

Die entsprechenden Kurven in diesem Raum sind in Abb. 7.9 dargestellt. Man erkennt an diesen Punktwolken, dass der negative Leveragewertebereich das Komplement des positiven

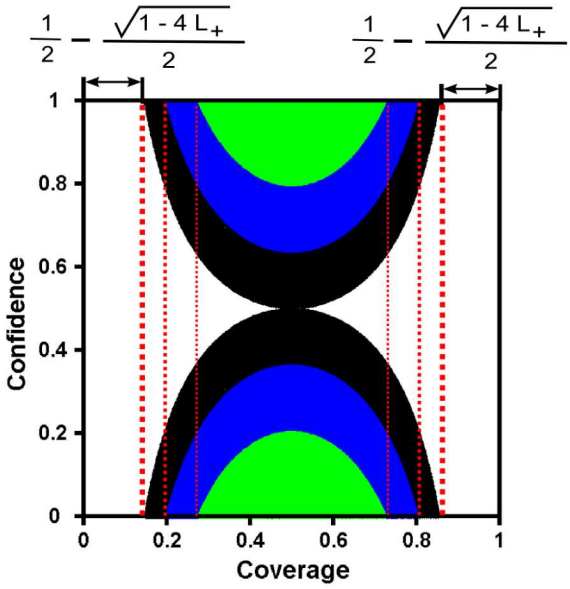

Abbildung 7.9.: Leveragepunktwolken im Coverage-Confidence-Raum

Positive(obere) und negative(untere) Coverage-Confidencewert-Punktwolken mit den Leveragewerten ± 0.125 *(schwarz),* ± 0.16 *(blau) und* ± 0.20 *(grün). Die "roten Linien" stellen die maximalen und minimalen Coveragewerte der einzelnen Leveragewerte dar.*

Wertebereichs ist. Desgleichen gilt, dass für alle Punktwolken der Mittelpunkt (der Coveragewert mit maximalem Confidencewert) immer bei $coverage(R) = 0.5$ liegt, unabhängig des Leveragewerts. Alle Punktwolken sind symmetrisch um die Achse mit $coverage(R) = 0.5$. Deshalb kann man den Abstand der Punktwolken von der Supportwert-0-Linie genauer betrachten.

Der Abstand wurde mit dem Leveragewert abgeschätzt, weil in den Confidence- und Recall-Funktionen der Nenner positiv sein muss. Dieses Minimalmaß muss auch weiterhin erfüllt sein, jedoch ist der Abstand eher mit dem größeren Wert

$$\frac{1}{2} - \frac{\sqrt{1 - 4\,L_+}}{2} \tag{7.32}$$

beschreibbar. Dies gilt an der linken sowie an der rechten Seite der Punktwolke.

Anmerkung:
Der Coverage-Confidence-Raum kann als Ergänzung für alle Heuristiken verwendet werden. Es zeigt sich, dass in den meisten Fällen, wenn man die Darstellungen komplementärer

Heuristikwerte [7] vergleicht, symmetrische Punktwolken entstehen. Jedoch lassen sich nicht immer ergänzende Informationen aus diesen Darstellungen gewinnen. Bei der Leverageheuristik und einer folgenden Heuristik (Klösgen) dient dieser Raum jedoch genau dazu.

7.6. Das Leveragewertintervall

Aus der Gl. 7.27 folgt die linke Grenze für den Leveragewertebereich des Punktes P_1

$$leverage(R) \; > \; - \; \frac{1 - conf(R)}{4} \; = \; \frac{conf(R) - 1}{4} \quad . \tag{7.33}$$

Aus der Gl. 7.29 folgt für den Recallwert 1 die rechte Seite des Leveragewertebereiches und damit das komplette Intervall:

$$leverage(R) \; = \; \left[\frac{conf(R) - 1}{4} \; , \; \frac{conf}{4} \right] \quad . \tag{7.34}$$

Man kann den Leveragewertebreich nur mit Hilfe des Confidencewertes einschränken. Dieses Intervall hat immer die Breite 0.25. Damit wird der Leveragewertebereich lediglich halbiert. Zusätzlich kann man sofort feststellen ob der Punkt P_1 ein Punkt P_{cmax} ist.

> Nur wenn der Coveragewert in diesem Punkt den Wert 0.5 hat, kann es sein, dass der Punkt P_1 ein Punkt P_{cmax} einer Punktwolke ist. Jedoch nicht jeder Punkt mit $coverage = 0.5$ ist ein Punkt P_{cmax}, da die Punktwolken mit kleinerem Wert L_+ die Punktwolken mit größerem Wert L_+ einschließen. Auf der linearen Funktion mit Steigung 2 im Support-Confidence-Raum hat jeder Punkt den Coveragewert 1/2.

Darüber hinaus kann die Abschätzung des Leveragewerts noch eingeschränkt werden durch die Verwendung des Supportwertes. Man kann für den Punkt P_1 den Coveragewert

$$coverage(R)_{P1}$$

berechnen. In der Abb. 7.9 erkennt man das jeder Leveragewert durch einen minimalen und maximalen Coveragewert beschränkt ist.

$$max\{coverage(R)\} \; = \; 1 \; - \; min\{coverage(R)\}$$

Also kann man durch Umformen der Formel 7.37 den minimalen bzw. maximalen Leveragewert bestimmen. Falls $coverage(R)_{P1}$ größer als 0.5 ist, dann liegt der Punkt in Abb.

[7] Im Fall der Leverageheuristik ist die ein Wert im pos. und neg. bereich mit dem glechen betrag

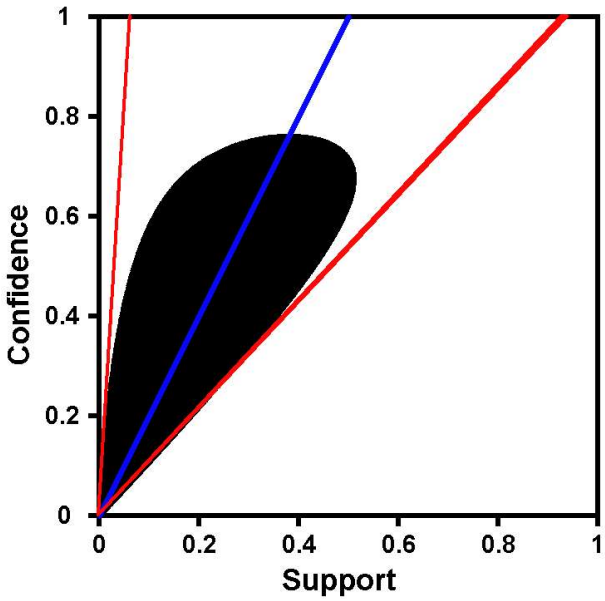

Abbildung 7.10.: Leveragepunktwolke mit Coveragelinien

7.10 rechts der blauen Linie andernfalls links dieser Linie mit Coveragewert 0.5. Je höher der L_+-Wert ist, desto kleiner ist die Coveragewertspanne (siehe Abb. 7.9). Wenn der Coveragewert größer als 0.5 ist, dann muss man den Wert erst relativieren, indem man den Wert von 1 subtrahiert[8]. Danach kann man dann mit diesem Wert den maximalen L_+-Wert bestimmen.

$$cov = 1 - coverage(R) \quad , falls\ coverage(R) > 0.5 \tag{7.35}$$

$$cov = coverage(R) \quad , falls\ coverage(R) < 0.5 \tag{7.36}$$

$$cov = \frac{1}{2} - \frac{\sqrt{1 - 4\,L_+}}{2} \tag{7.37}$$

$$L_+ = \frac{1 - (1 - 2cov)^2}{4} \tag{7.38}$$

Es gilt:

- Ist $L_- = -L_+$ größer als die linke Grenze des Intervalls aus der Gl. 7.34, dann kann diese Grenze korrigiert werden.
- Ist der Wert von L_+ kleiner als die rechte Grenze des Intervalls aus der Gl. 7.34, dann kann diese Grenze ebenfalls korrigiert werden.

[8] Ist der Coveragewert kleiner als 0.5 erübrigt sich dieser Schritt

Allgemein gilt dann für den möglichen Leveragewert:

$$min\{leverage(R)\} \;=\; max\left\{\frac{conf(R)-1}{4}\;,\;\frac{(1\;-\;2cov)^2-1}{4}\right\} \tag{7.39}$$

$$max\{leverage(R)\} \;=\; min\left\{\frac{conf}{4}\;,\;\frac{1-(1\;-\;2cov)^2}{4}\right\} \tag{7.40}$$

7.7. Fazit

Die Heuristiken Lift und Leverage geben Werte wieder, mit welchen man für eine Regel R vorhersagen kann, ob das Auftreten von Kopf und Körper sich gegenseitig positiv oder negativ beeinflusst. Wenn man für eine Regel R nicht den Liftwert oder Leveragewert generiert hat, sondern den Support-, Confidence- oder Recallwert, dann kann man anhand der Koordinatenpunkte bestimmen, für welche Lift- und Leveragewerte dieser Punkt gilt. Dank der genaueren Punktwolken im Recall-Confidence-Raum ist dies für den Leveragewert wesentlich einfacher als für den Liftwert. Man kann für Koordinatenpunkte im Recall-Confidence-Raum folgende Aussagen treffen:

> Die Punktwolken werden durch "harte" Kanten abgegrenzt. D.h. auf der einen Seite dieser Kanten/Grenzen liegt die Punktwolke und kann jeden Punkt des Raums theoretisch treffen. Auf der anderen Seite der Kanten/Grenzen liegt kein Punkt der Punktwolke.

8. Der Phi-Koeffizient

8.1. Gewichtete Leverageheuristik

Analog zur Leverageheuristik ermittelt auch der Phi-Koeffizient [1] die Differenz zwischen

> der Wahrscheinlichkeit, dass der Körper und der Kopf der Regel R erfüllt werden,
> und
> der Wahrscheinlichkeit, dass Kopf und Körper unabhängig von einander sind.

In [TKS02] wird diese Differenz durch die Anzahl aller auftretenden Mengen gewichtet.

$$phi(R: A \to B) = \frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)\ P(B)\ (1-P(A))\ (1-P(B))}} \tag{8.1}$$

Für den P-N-Raum folgt:

$$phi(R: A \to B) = \frac{p(P+N) - P(p+n)}{\sqrt{(p+n)\ (P+N-n-p)\ P\ N}} \quad . \tag{8.2}$$

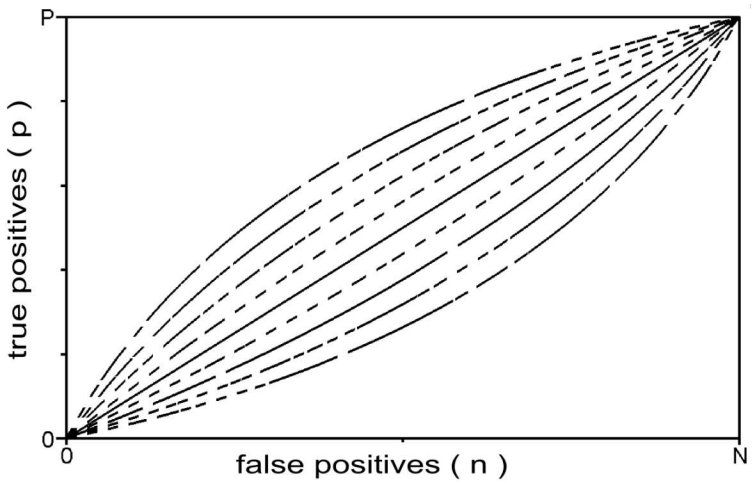

Abbildung 8.1.: Phi im P-N-Raum

[1] Der Phi-Koeffizient ist ein Spezialfall des Chi-Quadrat-Unabhängigkeitstests

In der Abb. 8.1 verlaufen die Isometrien jeweils vom Ursprungspunkt zum Punkt (N, P). Auf jeder Isometrie gilt weiterhin der gleiche Phi-Koeffizient. Isometrien unterhalb der Diagonalen haben einen negativen Phi-Koeffizient, Isometrien oberhalb der Diagonalen einen positiven, Werte auf der Diagonalen haben einen Phi-Koeffizient gleich 0. Auf der Diagonalen ist das Verhältnis zwischen "true positives" und $p(A)$ (dies entspricht dem Confidencewert) gleich der Menge von positiven Datensätzen in der Datenbank.

$$(P+N)\ p \ = \ P\ (p+n) \qquad \Rightarrow \qquad \frac{p}{p+n} \ = \ \frac{P}{P+N} \ = \ conf(R)$$

Analog gilt für Isometrien unter/über der Diagonalen:

$$\underbrace{\frac{p}{p+n} \ < \ \frac{P}{P+N} \ = \ conf(R)}_{\text{unter der Diagonalen}} \quad \vee \quad \underbrace{\frac{p}{p+n} \ > \ \frac{P}{P+N} \ = \ conf(R)}_{\text{über der Diagonalen}}$$

Der Zähler der Gl. 8.2 stellt die Leverageheuristik dar, diese hat Isometrien parallel zur Diagonalen. Der Nenner ist für die Beugung der Isometrien zum Ursprung und zum Punkt (N, P) verantwortlich. Im Assoziationsraum muss man diese Normierungswerte umformen. Die Werte $P-n$ und $N-n$ können durch die Confidence- und Recall-Heuristik beschrieben werden.

$$P(A) \ = \ \frac{C0+C1}{N} \tag{8.3}$$

$$P(B) \ = \ \frac{C0+C2}{N} \tag{8.4}$$

$$1-P(A) \ = \ P(\overline{A}) \ = \ \frac{C2+C3}{N} \ = \ \frac{N-C0-C1}{N} \tag{8.5}$$

$$1-P(B) \ = \ P(\overline{B}) \ = \ \frac{C1+C3}{N} \ = \ \frac{N-C0-C2}{N} \tag{8.6}$$

Durch Umformen ergibt sich:

$$\phi \ = \ phi(R) \quad = \quad \frac{leverage(R)\ lift(R)}{\sqrt{(conf(R)-supp(R))\ (recall(R)-supp(R))}} \tag{8.7}$$

$$= \quad \frac{recall(R)conf(R)-supp(R)}{\sqrt{(conf(R)-supp(R))\ (recall(R)-supp(R))}} \tag{8.8}$$

Wie in den zuletzt analysierten Heuristiken Lift und Leverage gilt auch hier die Vertauschungsunabhängigkeit vom Recall- und Confidence-Wert. Mit Hilfe der quadratischen Ergänzung bekommt man die Lösungen für die drei Basen des Assoziationsraums.

Supportwertfunktion

$$supp(R) \;=\; -\frac{1}{2s_1}\left(s_2 \pm \sqrt{s_2^2 + 4s_1 s_x}\right) \tag{8.9}$$

$$\begin{aligned} s_1 &= \phi^2 - 1 && (8.10)\\ s_2 &= recall(R)\;\phi^2 \;+\; conf(R)\;(\phi^2 - 2recall(R)) && (8.11)\\ s_x &= (recall(R)\;conf(R))^2 \;-\; \phi^2\;recall(R)\;conf(R) && (8.12) \end{aligned}$$

Confidencewertfunktion

$$conf(R) \;=\; -\frac{1}{2recall(R)^2}\left(c_1 \pm \sqrt{c_1^2 - 4recall(R)^2 c_x}\right) \tag{8.13}$$

$$\begin{aligned} c_1 &= \phi^2\;(supp(R) \;-\; recall(R)) - 2\;supp(R)\;recall(R) && (8.14)\\ c_x &= supp(R)\;\phi^2(recall(R) - supp(R)) \;-\; supp(R)^2 && (8.15) \end{aligned}$$

Recallwertfunktion

$$recall(R) \;=\; -\frac{1}{2conf(R)^2}\left(r_1 \pm \sqrt{r_1^2 + 4conf(R)^2 r_x}\right) \tag{8.16}$$

$$\begin{aligned} r_1 &= \phi^2\;(supp(R) \;-\; conf(R)) - 2\;supp(R)\;conf(R) && (8.17)\\ r_x &= supp(R)\;\phi^2(conf(R) - supp(R)) \;-\; supp(R)^2 && (8.18) \end{aligned}$$

In allen Heuristikfunktionen kommt der Phi-Koeffizient immer nur in der Verbindung ϕ^2 vor. Das bedeutet, bei der weiteren Analyse ist es unerheblich, ob der Phi-Koeffizient positiv oder negativ ist.

Die Abbildungen der Phi-Koeffizienten-Heuristik (siehe Anhang A.6.1 und A.6.2) sind denen der Lift und Levergeheuristiken ähnlich. Die Phi-Koeffizienten-Heuristik nimmt negative Werte zwischen $[-1, 0]$ an, wenn der Körper A und der Kopf B der Regel R unabhängig sind. Analog gilt, die Phi-Koeffizienten-Heuristik nimmt positive Werte zwischen $[0, 1]$ an, wenn A und B abhängig von einander sind.

8.2. Der Phi-Koeffizient im Support-Confidence-Raum

Die Supportfunktion besteht aus 2 Teilen: Den Wert innerhalb der Klammer und den Vorfaktor der Klammer.

$$-\frac{1}{2s_1} > 0$$

Weil der Wert vor der Klammer immer positiv ist, muss der komplette Ausdruck in der Klammer auch positiv sein. Das bedeutet, wenn das Vorzeichen vor der Wurzel negativ ist, dann muss s_2 größer oder gleich dem Wurzelwert sein.

$$s_2 \geq \sqrt{s_2^2 + 4s_1 s_x} \tag{8.19}$$

Wenn man diese Gleichung weiter auflöst und die Hilfsfunktionen einsetzt, dann ergibt sich die Bedingung:

$$\phi^2 \geq conf(R)\ recall(R) \quad . \tag{8.20}$$

Für die Supportwertfunktion folgt:

> Immer wenn die Voraussetzung 8.20 gilt ist das Vorzeichen vor der Wurzel egal, ansonsten muss das Vorzeichen vor der Wurzel positiv sein.

Die Beschränkungen, welche sich z.B. für positive Phi-Koeffizienten ergeben, müssen auch für negative Phi-Koeffizienten gelten, da sich der Supportwert nicht ändert, wenn man das Vorzeichen beim Phi-Koeffizient verändert.

8.2.1. Positive Phi-Koeffizienten-Werte

Anhand der Punktwolken im Support-Confidence-Raum (s. Abb. 8.2) kann man erkennen, dass der unterer Rand dieser Punktwolken vom Punkt $(0, \phi^2)$ zum Punkt $(1, 1)$ verläuft. Diese Werte ergeben sich durch Gl. 8.20. Wenn der Confidencewert sein Minimum ϕ^2 annimmt, muss der Recallwert maximal werden (s. Recall-Confidence-Raum). Mit variablem Support- und Phi-Wert folgt für den minimalen Confidencewert:

$$min\{conf(R)\} = -\frac{1}{2}\left((supp(R)-1)\phi^2 - 2\,supp(R) \pm (supp(R)-1)\phi^2\right) \quad .$$

Da die Funktionen c_1 und r_1 immer negativ sind,

$$supp(R) \leq conf(R) \quad \wedge \quad supp(R) \leq recall(R) \quad ,$$

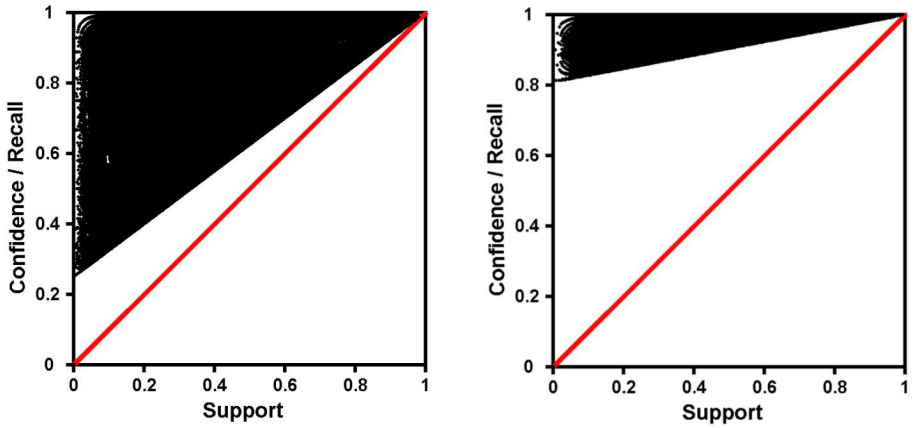

Abbildung 8.2.: Liftwerte 0.5 (links) und 0.9 (rechts) im Support-Confidence/Recall-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.6.2)

muss das Vorzeichen vor der Wurzel der Confidence- und Recallfunktionen immer positiv sein.

$$min\{conf(R)\} \;=\; (1 - supp(R))\phi^2 \;+\; supp(R) \tag{8.21}$$

Dies stellt die Grenzfunktion für pos. Phi-Koeffizienten dar. Die Funktion hat ihr Minimum auf der Confidenceachse und steigt von da aus monoton an (s. Abb. Anhang A.6.2). Für den Supportwert 0 ergibt sich aus Gl. 8.21:

$$min\{conf(R)\} \;=\; \phi^2 \quad . \tag{8.22}$$

Da Confidence- und Recallwert vertauschungsunabhängig sind, folgt der minimale Recallwert analog. Für die Steigung und die Grenzfunktion (analog zu Gl. 8.21) gilt:

$$m \;=\; \frac{\Delta conf(R)}{\Delta supp(R)} \;=\; 1 - \phi^2 \quad . \tag{8.23}$$

$$conf(R) \;=\; (1 - \phi^2)\;\; min\{supp(R)\} \;+\; \phi^2 \tag{8.24}$$

Nun kann man die minimale Supportfunktion an der Grenzfunktion ermitteln.

$$min\{supp(R)\} \;=\; \frac{conf(R) \;-\; \phi^2}{1 - \phi^2} \tag{8.25}$$

8.2.2. Negative Phi-Koeffizienten-Werte

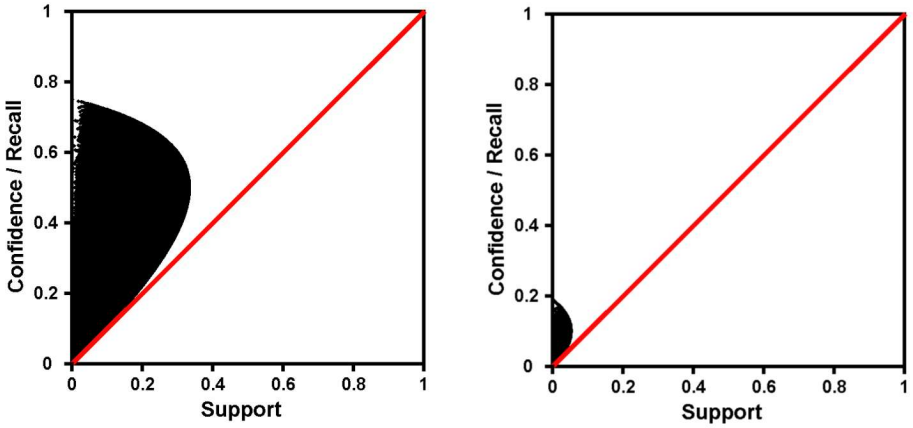

Abbildung 8.3.: Phiwerte -0.5 (links) und -0.9 (rechts) im Support-Confidence/Recall-Raum mit N = 1500 (für weitere Abb. siehe Anhang A.6.1)

Bei positivem Phi-Koeffizient ist der Supportwert kleiner als das Produkt aus Recall- und Confidencewert (siehe Gl. 8.8). Analog gilt: Wenn der Phi-Koeffizient negativ ist, dann ist das Produkt aus Recall- und Confidencewert kleiner als der Supportwert. Daraus folgt, dass die Punktwolken mit positivem und negativem Phi-Koeffizient ein anderes aussehen haben. Die Art der Punktwolken für negative Phi-Koeffizienten erinnert an die Punktwolken für die Liftwerte kleiner als 1 (s. Abb. 8.3).

In den einzelnen Abb. in 8.4 erkennt man jeweils 2 Punktwolken mit den gleichen ϕ^2-Werten. Wenn der ϕ^2-Wert sich dem Wert 0 nähert, dann werden sich sowohl die blauen als auch die schwarzen Punktwolke immer weiter vom Confidencewert 0.5 entfernen. Der min. Confidencewert der Punktwolke für pos. Phi-Koeffizienten-Werte nähert sich dem Punkt $conf(R) = 1$ und der max. Confidencewert für den jeweils komplementären negativen Phi-Koeffizienten-Wert dem Ursprungspunkt. Bei komplementären Phi-Koeffizienten-Werten ist der Abstand zum Punkt $conf(R) = 0.5$ gleich. Da man die Formel für den min. Confidencewert bei pos. Phi-Koeffizienten-Werten kennt (siehe Gl. 8.22) ergibt sich analog für die neg. Phi-Koeffizienten-Werte die Gleichung

$$max\{conf(R)\} \;=\; 1 - \phi^2 \quad . \tag{8.26}$$

Der nächste markante Punkt in den Abb. im Anhang A.6.1 ist der maximale Supportwert. Dieser Punkt wird analog zu den letzten Kapiteln P_{smax} genannt. Durch Ablesen der Werte in den Abbildungen stellt man fest, dass der Confidencewert des Punktes P_{smax} sich für jede Abbildung mit $1 - \phi$ ergibt (s. Abb. 8.4). Im Punkt P_{smax} gilt stets $conf(R) = recall(R)$. Für pos. Phi-Koeffizient hat man die Voraussetzung 8.20 definiert.

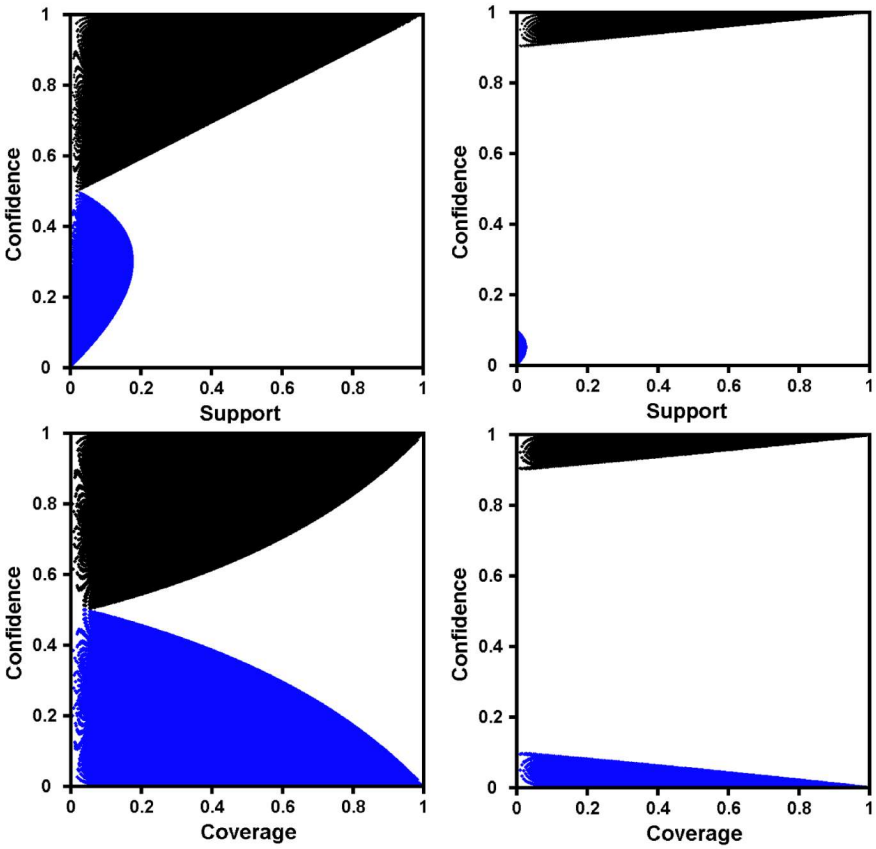

Abbildung 8.4.: Positive(schwarz) und Negative(blau) Leveragewertpunktwolken 0.7(-0.7) (links) und 0.95(-0.95) (rechts). Support-Confidence-Räume oben und Coverage-Confidence-Räume unten.

$$\phi^2 \;=\; conf(R)\; recall(R) \;=\; conf(R)^2$$

Analog zur Erklärung von Gl. 8.26 kann man auch hierbei verfahren. Der mögliche Confidencewert von P_{smax} verändert sich mit der Höhe der Punktwolke $max\{conf(R)\}$. Also ergibt sich auch der Confidencewert von P_{smax} durch Subtraktion von 1.

$$conf(R)_{P_{smax}} \;=\; 1 - |\phi| \tag{8.27}$$

Der Supportwert ergibt sich nachdem man $conf(R) = recall(R) = 1 - \phi$ in die Gl. 8.9 einsetzt.

$$supp(R)_{P_{smax}} \;=\; -\frac{1}{1-\phi^2}\left(-\phi^3 \;+\; 2\phi \;-\; 1 \;+\; \sqrt{(\phi-1)^2\phi^4}\right) \tag{8.28}$$

Da die Voraussetzung 8.20 in diesem Fall nicht gilt, muss das Vorzeichen vor der Wurzel positiv sein.

$$supp(R)_{smax} = \frac{(1 - \phi)^2}{1 - \phi^2} \tag{8.29}$$

8.3. Felder und Fazit

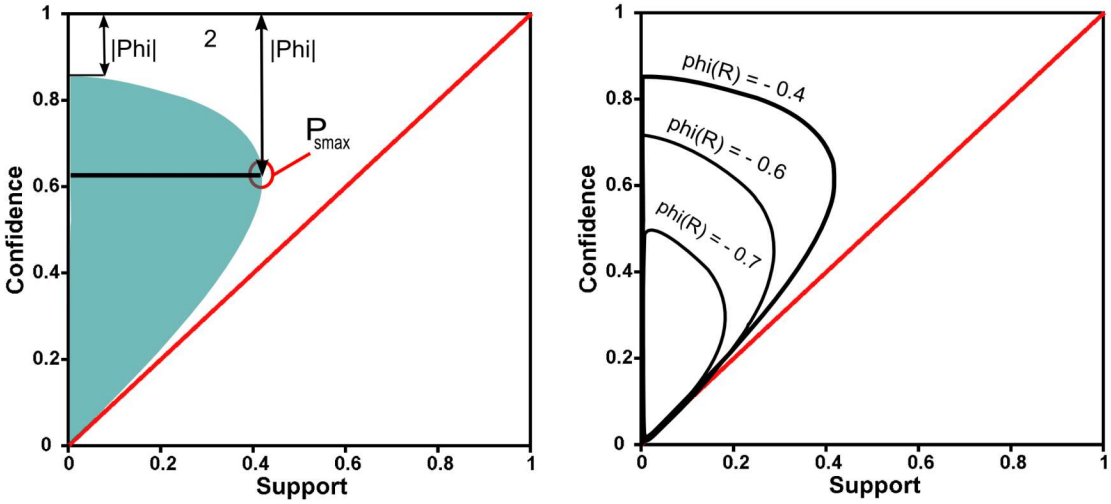

Abbildung 8.5.: Phiwertfelder (negativer Phiwert) im Assoziationsraum

Bild Links: Grünes Feld für alle Phiwerte kleiner gleich dem eingezeichneten Phiwert.
Bild Rechts: Die Bogenlinien stellen die Grenzen für einzelne Phiwerte dar. Innerhalb einer Linie gilt der Phiwert noch, ausserhalb nicht mehr.

Da der Recall- und Confidencewert antiproportional zu den Supportwerten des Körpers A und des Kopfes B der Regel R sind,

$$conf(R) = \frac{supp(R)}{supp(A)} \quad \wedge \quad recall(R) = \frac{supp(R)}{supp(B)}$$

gilt:

- Wird der Recall- und/oder Confidencewert groß (≈ 1), dann sind die Werte von $supp(A)$ und/oder $supp(B)$ klein.

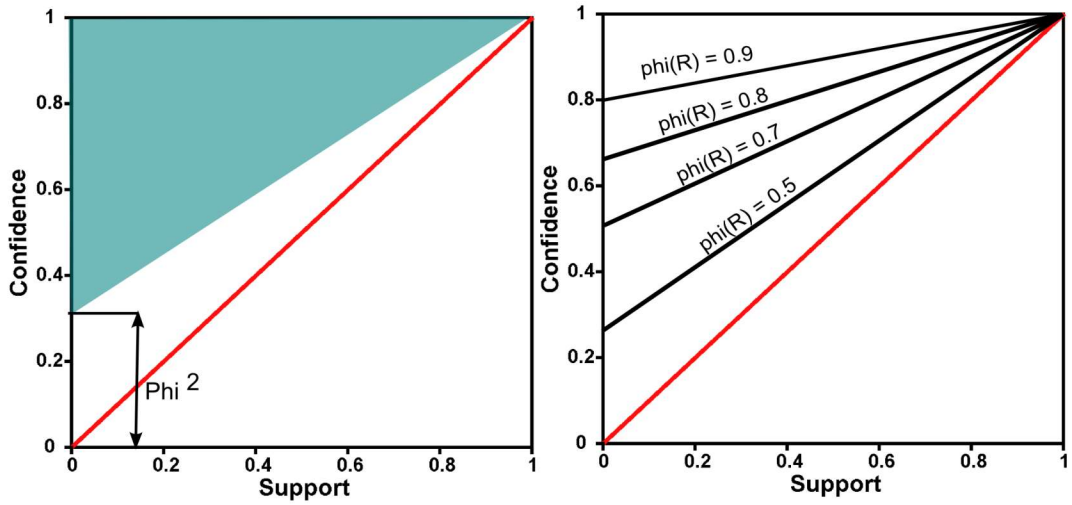

Abbildung 8.6.: Phiwertfelder (positivem Phiwert) im Assoziationsraum

Bild Links: Grünes Feld für alle Phiwerte größer gleich dem eingezeichneten Phiwert.
Bild Rechts: Die Linien stellen die Grenzen für einzelne Phiwerte dar. Über einer Linie gilt der Phiwert noch, unter der Linie nicht.

- Wird der Recall- und/oder Confidencewert klein (≈ 0), dann sind die Werte von $supp(A)$ und/oder $supp(B)$ groß.

Anhand der Abbildungen des Recall-Confidence-Raumes gilt für die Phi-Koeffizienten-Heuristik:

- Falls $supp(R)$ dominiert ($\phi > 0$), dann kann man für jeden Confidencewert den Recallwert beliebig erhöhen und umgekehrt. Es folgt, dass C1 und C2 in der Kontingenztabelle beliebig klein werden können. Erkennbar ist dies an der Punktwolke im Recall-Confidence-Raum als "Viertelkugel" um $(1, 1)$.
- Falls $supp(A) \cdot supp(B)$ den Phi-Koeffizienten-Wert dominiert, dann können der Recall- und Confidencewert beliebig klein werden ("Viertelkugel" um den Ursprungspunkt).

8.3.1. Felder im Phi-Koeffizienten-Wertebereich

Wie in den letzten Kapiteln geht man davon aus, dass man nur den Punkt

$$P_1 = (supp(R), conf(R))$$

gegeben hat. Anhand des Support- und Confidencewerts, kann man auf den jeweiligen Phi-Koeffizient schließen. Der Punkt P_1 ist im Extremfall ein Punkt auf der Grenzkurve der Punktwolken. Mit Hilfe der Informationen über die Grenzkurven kann man das Phi-Koeffizienten-Intervall bestimmen.

$$|\phi| \;\leq\; \sqrt{1 - conf(R)} \tag{8.30}$$

d.h.:

$$\phi \;\in\; [-\sqrt{1 - conf(R)}, \sqrt{conf(R)}] \tag{8.31}$$

Zusätzlich zum Confidencewerten kann man die Information über den Supportwert mit einbeziehen. Der Supportwert einer Punktwolke ($\phi < 0$) wird am Punkt P_{smax} maximal.

$$supp(R)_{smax} \;=\; \frac{(1 \;-\; |\phi|)^2}{1 \;-\; \phi^2} \quad . \tag{8.32}$$

$$\phi \;=\; \frac{1 \;-\; supp_{smax}(R)}{1 \;+\; supp_{smax}(R)} \tag{8.33}$$

Bei pos. Phi-Koeffizienten kann man mit Hilfe der Gl. 8.25

$$min\{supp(R)\} \;=\; \frac{conf(R) \;-\; \phi^2}{1 - \phi^2}$$

die rechte Grenze des Intervalls 8.31 korrigieren, falls der Phi-Koeffizient

$$\phi \;=\; \sqrt{\frac{supp(R) \;-\; conf(R)}{supp(R) \;-\; 1}}$$

kleiner ist als der Wert $\sqrt{conf(R)}$. Das Gesamtintervall ergibt sich durch zusammenführen aller Ergebnisse.

$$\phi \;=\; \left[max\left\{ -\sqrt{1 - conf(R)} \;,\; \frac{supp(R) \;-\; 1}{1 \;+\; supp(R)} \right\} \;,\; min\left\{ \sqrt{conf(R)} \;,\; \sqrt{\frac{supp(R) \;-\; conf(R)}{supp(R) \;-\; 1}} \right\} \right]$$

Mit Hilfe des zusätzlichen Recallwerts ist es aber erst möglich den genauen Phi-Koeffizienten zu bestimmen.

9. Klösgen

9.1. Trade-Off

Es gibt Heuristiken, welche einen Parameter beinhalten mit dem man einen Wert innerhalb der Heuristik gewichten kann. Üblicherweise wird dies benutzt um unterschiedliches Verhalten von zwei Heuristiken sichtbar zu machen. Je nach Parameterwert ist die eine Heuristik stärker gewichtet als die andere Heuristik, man spricht hierbei von einem Trade-Off der Basisheuristiken [Jan06].

Die Klösgenheuristik [Jan06] ist ein Trade-Off zwischen den Heuristiken Coverage und Precision-Gain [1]; es gibt einen Parameter w mit dem man den Anteil des Coveragewerts im Klösgenwert steuern kann. Der Parameter bewirkt, dass eine genaue Untersuchung oder Umrechnung der Klösgenheuristik in Support-, Confidence- und Recallheuristik unmöglich ist. Im Folgenden wird gezeigt, dass man die Klösgenheuristik auf die schon bekannte Leverageheuristik zurückführen kann und es sich bei den Klösgenpunktwolken um "gemorphte" Leveragepunktwolken handelt.

9.2. Coverage und Precision-Gain

Die Coverageheuristik ist identisch mit dem Support des Kopfes A,

$$coverage(R) = \frac{supp(R)}{conf(R)} = supp(A) \quad . \tag{9.1}$$

Da das Verhältnis von Support- und Confidencewert für einen konstanten Coveragewert auch konstant bleiben muss, sind die Punktwolken im Support-Confidence-Raum Punktlinien (s. Abb. 9.1).

Bei der Precision-Gain-Heuristik wird von der Precisionheuristik [2] der Wert $supp(B)$ subtrahiert.

$$precgain(R) = conf(R) - supp(B) = \frac{supp(R)}{supp(A)} - supp(B) \tag{9.2}$$

[1] Erläuterung folgt in diesem Kapitel

[2] In der Einleitung wurde bereits erwähnt das dies der Confidenceheuristik entspricht.

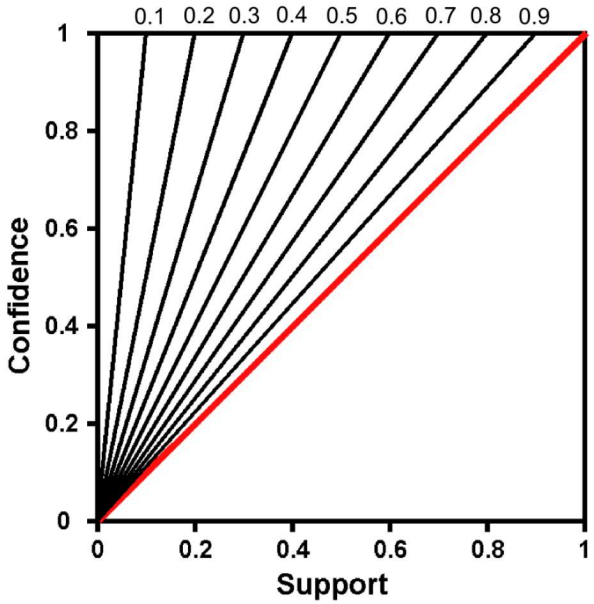

Abbildung 9.1.: Coveragewerte 0.1 bis 0.9 im Support-Confidence-Raum

Wenn der Confidencewert groß ist, dann verhindert die a-priori-Verteilung [3] $supp(B)$ einen zu hohen Precision-Gain-Wert. Genauso wie bei den Heuristiken Lift und Leverage wird auch hierbei ein Verhalten zwischen $supp(R)$, $supp(A)$ und $supp(B)$ generiert. Somit ist die Darstellung der Precision-Gain-Heuristik mithilfe der Leverageheuristik naheliegend.

$$precgain(R) = \frac{leverage(R)}{supp(A)} \tag{9.3}$$

[3] Die a-priori-Verteilung einer Datenbank entspricht $supp(B)$, dem Anteil an positiven Datensätzen in der Datenbank. Einen höheren Wert kann z.B. der Supportwert einer Regel, mit der Klassifikation B, niemals erreichen.

9.3. Trade-Off von Coverage und Leverage

Die Klösgenheuristik ist das Produkt der Heuristiken Coverage und Precision-Gain. Mit $x = w - 1$ ergeben sich die folgenden Formeln für den Klösgenwert:

$$klosgen(R) = coverage(R)^w \; precgain(R) \tag{9.4}$$
$$klosgen(R) = supp(A)^w \; (conf(R) - supp(B)) \tag{9.5}$$
$$klosgen(R) = supp(A)^x \; leverage(R) \quad . \tag{9.6}$$

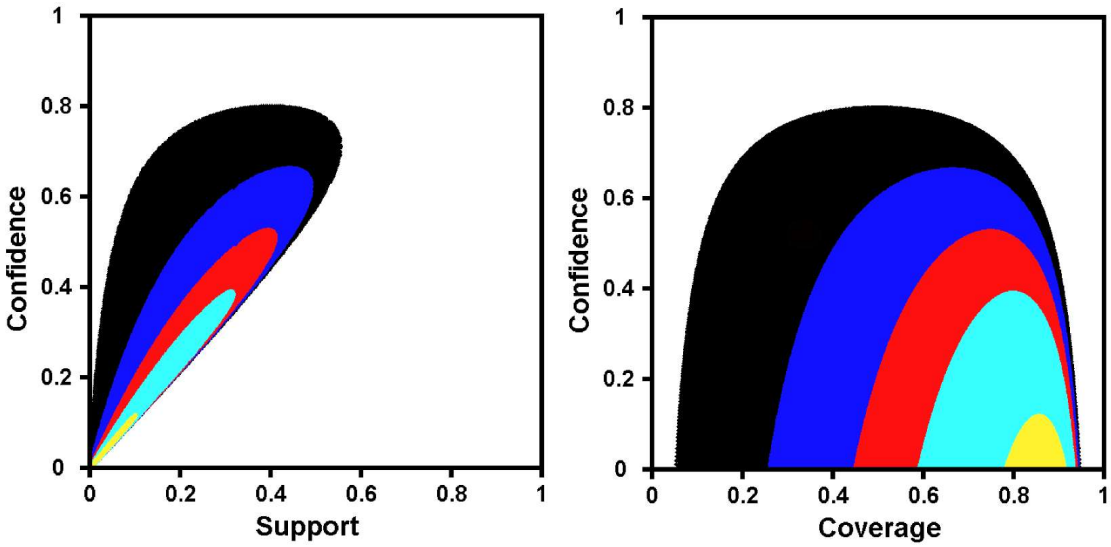

Abbildung 9.2.: Klösgenwert -0,05 mit verschiedenen Parametern w im Support-Confidence-Raum (links) und Coverage-Confidence-Raum (rechts): ($w = 1$(schwarz), $w = 2$(blau), $w = 3$(rot), $w = 4$(türkis)

Da der Coveragewert $supp(A)$ eine lineare Funktion im Support-Confidence-Raum ist, müssen die Punktwolken der Klösgenheuristik "gemorphten" Leverageheuristikpunktwolken entsprechen. Der "gemorphte" Faktor ist die mit w gewichtete Coveragefunktion. In Abb 9.2 erkennt man verschiedene Punktwolken des gleiche Klösgenwerts mit unterschiedlichen Parametern w.

9.3.1. Negative Klösgenwerte

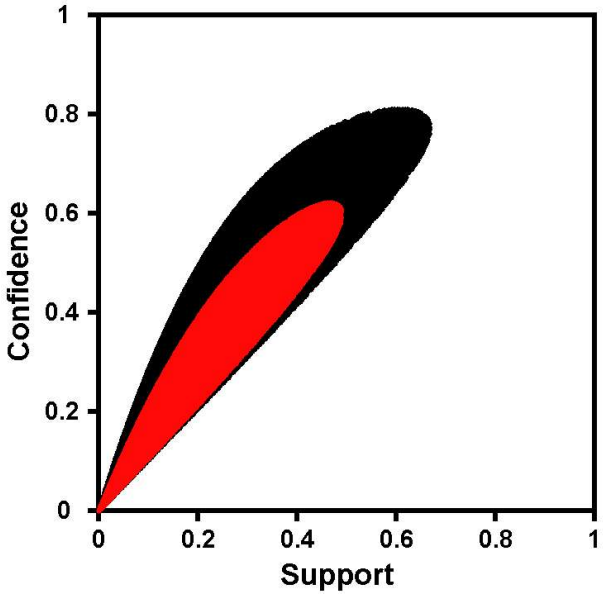

Abbildung 9.3.: Klösgenwert 0,02 (schwarz) und Klösgenwert 0,04 (rot) mit Parameter $w = 3$

Der Coveragewert ist immer ≤ 1 und wird für höhere w-Werte (x-Werte) immer kleiner. Das bedeutet, der Leveragewert muss immer größer werden um den gleichen Klösgenwert darzustellen (wie in Abb 9.2). Die verschiedenen Punktwolken in der Abb. 9.2 zeigen für steigende w-Werte (x-Werte), eine Annäherung an die Diagonale.

Auf der Diagonalen ist der Coveragewert 1 und eine Klösgenpunktwolke die exakt auf der Diagonalen liegt müsste gleich dieser Coverageheuristik sein, dass bedeutet der Leverageanteil fällt weg (sehr hoher Parameter w) und der Supportwert wäre gleich dem Confidencewert [4].

In der Abb. 9.2 sind 2 Eigenschaften der Punktwolken erkennbar:

- Jede Klösgenpunktwolke mit $w = i$ liegt innerhalb einer Klösgenpunktwolke mit dem gleichen Klösgenwert aber $w = j$ und $j < i$.

  Verändert man den Klösgenwert, dann verändert sich z.B. für $w = 1$ auch die Form der Leveragepunktwolke. Weil alle höheren Parameter innerhalb dieser Leveragepunktwolke liegen und alle Leveragepunktwolken[5] mit niedrigerem

[4] ist auf der Diagonalen immer gegeben
[5] bei negativem Klösgenwert bzw. negativem Leveragewert

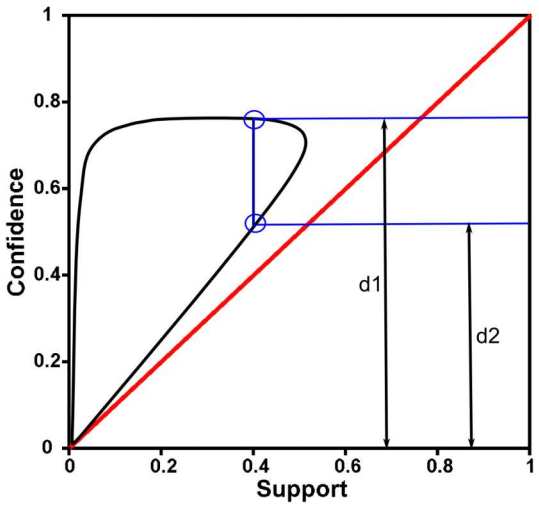

Abbildung 9.4.: Ist der Confidencewert gleich d1, dann ist der Recallwert gleich d2 und umgekehrt.

Leveragewert innerhalb Leveragepunktwolken mit höherem Leveragewert liegen, liegt auch jede Klösgenpunktwolke mit Klösgenwert $-k_1$ und Parameter w innerhalb einer Punktwolke mit Klösgenwert $-k_2$

$$-k_2 > -k_1 \qquad \vee \qquad k_1 > k_2$$

und Parameter w (siehe Abb. 9.3).

- Die zweite Eigenschaft der Punktwolken in Abb. 9.2 ist, dass alle unteren Grenzkurven fast identisch sind, also auch identisch mit dem unteren Teil der Grenzkurve der Leveragepunktwolke. In der Analyse der Leverageheuristik ist die Grenzkurve dadurch gegeben, dass für einen negativen Leveragewert

$$supp(R) < \frac{supp(R)^2}{conf(R)\ recall(R)}$$

$$supp(R) > conf(R)\ recall(R)$$

gilt. Das bedeutet diese Aussage bestätigt sich erneut, da selbst für gemorphte Leveragepunktwolken diese Bedingung so stark ist, dass sie immer noch die Grenzkurve am unteren Rand vorgibt. Die Leveragepunktwolke ist zumindest für den unteren Rand eine sehr gute Abschätzung der Klösgenpunktwolke.

Der obere Rand der Grenzkurve hingegen zieht sich zusammen, dies ist dadurch zu erklären, dass, wie in Abb. 9.4 aus der Leverageheuristik bekannt, der Recallwert aus dem Confidencewert auf einem Punkt der Grenzlinie folgt. Da nun dieses Verhältnis zwischen Confidence- und Recallwert durch den Coveragewert gestaucht wird, ergibt sich eine "dünnere" Punktwolke.

9.3.2. Positive Klösgenwerte

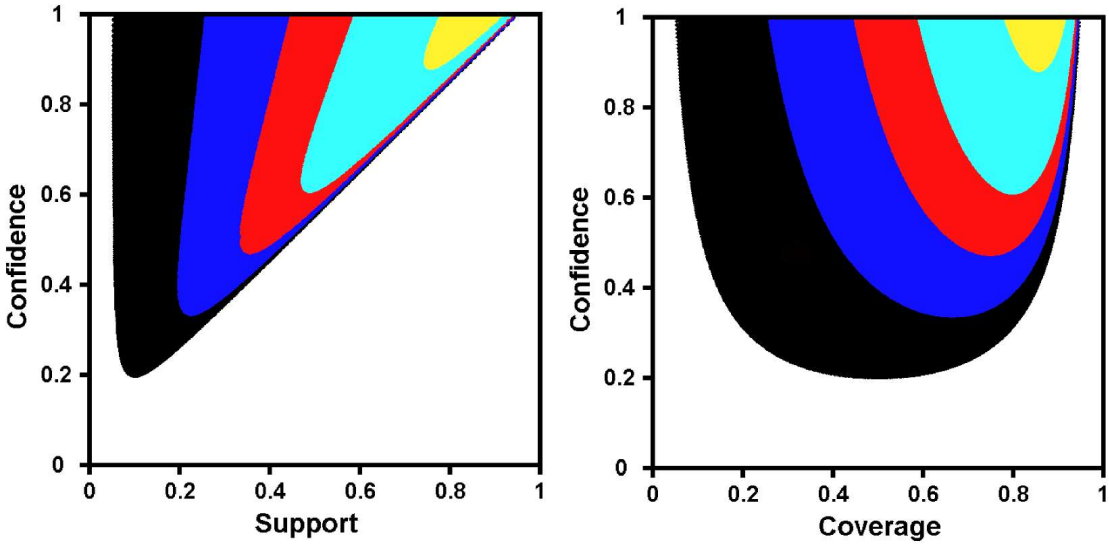

Abbildung 9.5.: Klösgenwert 0,05 mit verschiedenen Parametern w im Support-Confidence-Raum (links) und Coverage-Confidence-Raum (rechts): ($w = 1$(schwarz), $w = 2$(blau), $w = 3$(rot), $w = 4$(türkis), $w = 6$(gelb))

Bei den positiven Klösgenwerten hat man eine Punktwolke, die mit dem Faktor 4 des Leveragewerts sich immer weiter der Confidencewert-1-Linie annähert. Dieser Faktor ergibt sich in der Leverageheuristik, weil der Wert unterhalb der Wurzel im Supportwert nur für den positiven reellen Bereich definiert ist. Wenn die Potenz der Coverageheuristik steigt, sinkt die Potenz der Confidenceheuristik (mit konstantem Supportwert):

$$supp(A)^w \propto conf(R)^{-w}$$

Da $supp(A)$ kleiner gleich 1 und $conf(R)^{-1}$ größer gleich 1 ist folgt:

$$supp(A)^w \leq conf(R)^{-w} \quad .$$

Mit wachsendem Parameter wächst der Confidencewert. Dadurch erhöht sich der Abstand des niedrigsten Confidencewerts der Punktwolke von der Supportwert-0-Linie. Bei $w = 1$ ist dieser Abstand noch der gleiche wie bei der Levergeheuristik

$$min\{conf(R)\} \; = \; 4 \cdot klosgen(R)$$

jedoch schon eine Potenz weiter ($w = 2$) ist der Confidencewert bei

$$min\{conf(R)\} \; > \; 4 \cdot klosgen(R) \quad .$$

Analog zur Levergeheuristik ergibt sich aus dem Abstand von $min\{conf(R)\}$ im positiven Klösgenwertebereich der Wert $max\{conf(R)\}$ im negativen Klösgenwertebereich durch

$$max\{conf(R)\}_{klosgen(R)<0} \quad = \quad 1 - min\{conf(R)_{klosgen(R)>0}\} \quad .$$

Die Erklärung ergibt sich analog zur Levergeheuristik.

9.4. Verschiedene Parameter

In den folgenden Teilkapiteln werden verschiedenen Parameter genauer analysiert. Für die Darstellung im P-N-Raum der einzelnen Klösgenheuristiken wird auf [Jan06] und [FJ06] verwiesen. Die Abbildungen für die Support-Confidence-Recall-Teilräume sind im Anhang zu sehen.

9.4.1. Parameter w = 0, die Precision-Gain-Heuristik

Da der Parameter Null ist fällt der Coverageanteil im Klösgenwert weg und es bleibt lediglich der Precision-Gain-Anteil erhalten, also:

$$klosgen(R)_{w=0} \; = \; precgain(R) \; = \; conf(R) - supp(B) \quad . \tag{9.7}$$

In der Analyse der Levergeheuristik wurde das Verhältnis zwischen dem Support der Regel und der Wahrscheinlichkeit der Unabhängigkeit von Kopf und Körper der Regel untersucht. Die Precision-Gain-Heuristik beschränkt sich auf das Verhältnis zwischen dem Confidencewert und dem Kopf B der Regel.

In dem Beispiel der Supermarktdatenbank gilt für die Precision-Gain-Heuristik folgendes:

- Keine Regel die eine Vorhersage auf diese Klassifikation treffen will kann einen höheren Support haben, als der Wert $supp(B)$ [6].

[6] a-priori-Abschätzung

- Für Precision-Gain-Wert 0 gilt: Die Wahrscheinlichkeit, dass B gekauft wird, wenn vorher schon Artikel aus A gekauft wurden, ist ebenso groß, wie die Wahrscheinlichkeits dass überhaupt B gekauft wird.

- Falls der Precision-Gain-Wert kleiner als 0 ist, dann hat der Kauf der Artikel A einen negativen Einfluss auf den Kauf des Artikels B.

$$supp(B) > conf(R)$$

- Im letzten Fall ist der Precision-Gain-Wert größer als 1. Es ist wahrscheinlicher die Ware B zu kaufen, wenn A gekauft wurde, als unabhängig die Ware B zu kaufen.

Die Precision-Gain-Heuristik ist der Levergeheuristik ähnlich. Bei der Prec.-Gain-Heuristik sind allerdings die Punktwolken im Recall-Confidence-Raum nicht symmetrisch an der Diagonalen.

Im Recall-Confidence-Raum kann man erkennen, dass je kleiner der Precision-Gain-Wert wird desto mehr liegen die Punktwolken oberhalb der Diagonalen. Die Punktwolke nähert sich immer weiter der Confidencewert-0-Linie an. Da der Recallwert antiproportional zu $supp(B)$ ist, gilt: Mit schrumpfendem Precision-Gain-Wert wird $supp(B)$ größer als der Confidencewert. Der mögliche Recallwert schrumpft durch wachsenden $supp(B)$-Wert ebenfalls und ist schnell viel kleiner als der Confidencewert.

$$klosgen(R)_{w=0} = precgain(R) = conf(R) - \frac{supp(R)}{recall(R)} \tag{9.8}$$

$$precgain(R) > 0 \quad \Rightarrow \quad conf(R) > \frac{supp(R)}{recall(R)} \tag{9.9}$$

Für kleines $supp(B)$ klein ist muss der Recallwert groß sein. Für $supp(B) = 0$, nimmt der Confidencewert sein Minimum an, dies ist dann beim Precision-Gain-Wert. Die Punktkonzentration in den Punktwolken im Recall-Confidence-Raum mit positiven Precision-Gain-Werten ist haupsächlich um den Punkt $(1, 1)$. Bei hohen Recall- und Confidencewerten gibt es viel Spielraum für den Supportwert und damit die Möglichkeit, auch den gewünschten Klösgenwert der Punktwolke zu erreichen. Für kleinen Recallwert oder sogar kleinen $supp(B)$-Wert (Randgebiete der Punktwolke) gibt es nur wenige Supportwertkombinationen, da der Supportwert mit den zwei eben genannten Werten fallen muss.

9.4.2. Parameter w = 1, die Leverageheuristik

Man kann anhand der Formel

$$klosgen(R) \;=\; supp(A)^x \; leverage(R)$$

erkennen, dass für $x = w - 1$ sich der Wert $x = 0$ ergibt. Also ist die Klösgenheuristik hierfür gleichwertig mit der Leverageheuristik.

$$klosgen(R)_{w=1} \;=\; leverage(R) \tag{9.10}$$

In [Jan06], [Klö92], [Klö96] und [FJ06] wird gezeigt, dass die Klösgenheuristik mit diesem Parameter gleich der "Weighted Relative Accuracy"-Heuristik ist. Diese Heuristik ist die genormte Accuracyheuristik und hat im P-N-Raum (s. [Jan06] und [FF05]) die Darstellung

$$weighted - relative - acc(R) \;=\; \frac{p+n}{P+N}\left(\frac{p}{p+n} \;-\; \frac{P}{P+N}\right) \quad . \tag{9.11}$$

9.4.3. Parameter w = 0.5

$$klosgen(R)_{w=0.5} \;=\; \frac{leverage(R)}{\sqrt{supp(A)}} \tag{9.12}$$

Die Punktwolken im Support-Confidence-Raum für diesen Parameter sind größer als die Punktwolken im Leverageraum mit dem Klösgenwert als Leveragewert

$$klosgen(R)_{w=0.5} \;=\; leverage(R) \quad .$$

Dies ist verständlich, da der $supp(A)$-Wert kleiner als 1 ist, kann der Klösgenwert hier größer werden als der Leveragewert.

Bei der Leverageheuristik sind die Punktwolken im Recall-Confidence-Raum jeweils symmetrisch an der Diagonalen. Diese Symmetrie geht durch den Faktor $supp(A)^{-1}$ verloren, deshalb erkennt man, für steigenden Klösgenwerte, in den Abb. des Recall-Confidence-Raums für die Klösgenheuristik (siehe Anhang A.7.19 und A.7.20) eine Verschiebung der Punktwolken in positiver Confidence-Richtung.

Durch Division mit $supp(A)$ wird der $supp(A)$-Anteil in der Leverageheuristik reduziert. Dadurch steigt die Möglichkeit für Confidencewertkombinationen[7], die Punktwolke verschiebt sich in positiver Confidencewertrichtung. Beim Parameter $w = 0.5$ hat der Körper einen geringeren Anteil als der Kopf der Regel. Dies ist analog zum Fall $w = 0$ (s.o.), wo der Anteil des Kopfes in der Klösgenheuristik Null ist.

[7] Durch reziprokes Verhalten von $supp(A)$ und Confidencewert

9.4.4. Parameter w = 2

$$klosgen(R)_{w=2} \;=\; supp(A)^1 \; leverage(R) \tag{9.13}$$

Der Anteil von $supp(A)$ ist hier höher gewichtet als der Supportanteil des Kopfes. Dadurch ergeben sich die umgekehrten Aussagen wie zum Fall $w = 0.5$, die Mehrzahl der Punkte in den Punktwolken im Confidence-Recall-Raum liegen unterhalb der Diagonalen.

9.4.5. Parameter w > 2

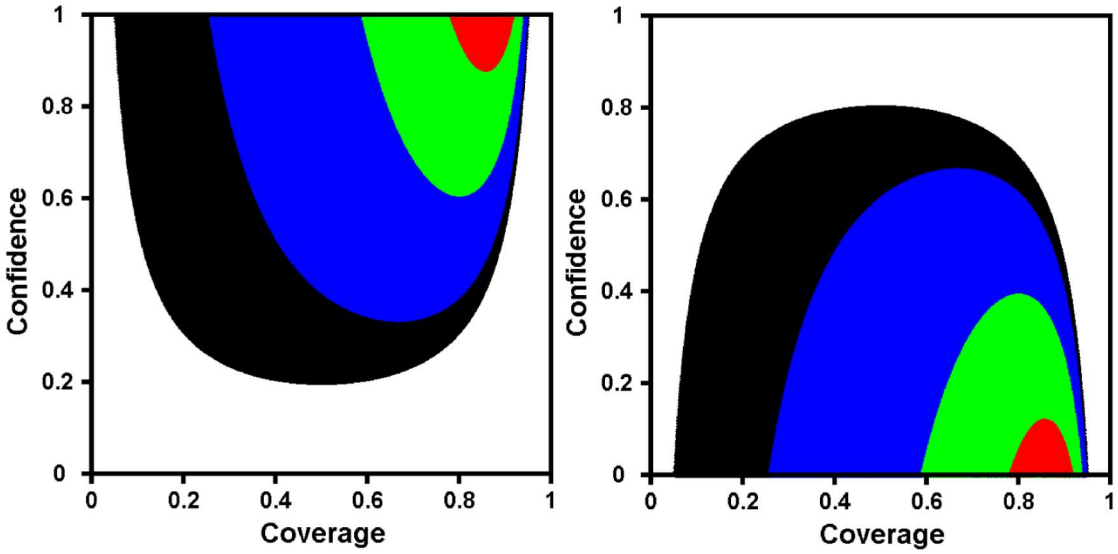

Abbildung 9.6.: Klösgenpunktwolken im Coverage-Confidence-Raum

Bild links: Klösgenwert 0,05 mit verschiedenen Parametern w ($w = 1$(schwarz), $w = 2$(blau), $w = 4$(grün), $w = 6$(rot))
Bild rechts: Klösgenwert -0,05 mit verschiedenen Parametern w ($w = 1$(schwarz), $w = 2$(blau), $w = 4$(grün), $w = 6$(rot))

Analog zu Erklärung $w = 2$ gilt : Der Anteil von $supp(A)$ ist höher gewichtet als der $supp(B)$ Anteil der Funktion. Der Wert von $supp(A)^w$ (bzw. $supp(A)^x$) wird bei jeder Potenzierung kleiner.

$$klosgen(R)_{w>2} \;=\; \underbrace{supp(A)^{(w-1)}}_{\text{sehr klein}} \; leverage(R) \tag{9.14}$$

Zwei Punktwolken mit selbem Klösgenwert aber verschiedenen Parametern werden verglichen: Der höhere Parameter w_2 erzeugt einen kleineren Wert $supp(A)^x$ als der Parameter w_1.

$$\begin{aligned} klosgen(R)_{w_1} &= klosgen(R)_{w_2} \\ klosgen(R)_{w_2} &= supp(A)^{(w_2-1)}\ leverage(R)_{w_2} \\ klosgen(R)_{w_1} &= coverage(A)^{(w_1-1)}\ leverage(R)_{w_1} \end{aligned}$$

$$\frac{supp(A)^{(w_2-1)}}{supp(A)^{(w_1-1)}}\ leverage(R)_{w_2} = leverage(R)_{w_1}$$

Da der Wert von $w2$ größer ist als der Wert $w1$ folgt:

$$leverage(R)_{w_2} > leverage(R)_{w_1}$$

Bei wachsendem Parameter und gleich bleibendem Heuristikwert muss der Coveragewert wachsen. In Abb. 9.6 ist im linken, sowie im rechten Bild der Heuristikwert konstant und nur die Parameter w (bzw. x) verändern sich. Man erkennt, dass der minimale Coveragewert jeder Punktwolke mit wachsendem Parameter w steigt. Mit steigendem w werden die Punktwolken immer kleiner.

9.5. Fazit und Klösgenfelder

In [Jan06], [Klö92] und [Klö96] wird die Klösgenheuristik als ein Trade-Off zwischen den Heuristiken Coverage ($supp(A)$) und Precision-Gain angegeben. Durch den Parameter w wird der Anteil des Coveragewerts gewichtet. Für den Assoziationsraum stellt die Klösgenheuristik den Trade-Off zwischen Coverage und Leverage dar.

Mit steigendem w werden die Punktwolken immer kleiner und liegen innerhalb der Punktwolken kleinerer Parameter w. Die Leveragepunktwolken kann man hinreichend genug beschreiben (siehe Kapitel: Levergeheuristik), dass bedeutet für jedes $w \geq 1$ kann man die Grenzen des Leveragewertintervalls auch für das Klösgenwertintervall voraussetzen. Mit steigendem w verjüngt sich das Klösgenwertintervall immer weiter. Die genauen Werte sind jeweils vom Parameter w abhängig.

Anhand des Coverage-Confidence-Raums kann man eine Abschätzung treffen in wie weit sich die Punktwolken mit steigendem Parameter jeweils verjüngen. Während bei der Levergeheuristik der minimale (maximale) Confidencewert mit der Formel $4 \cdot L_+\ (1 - 4 \cdot L_+)$ berechnet werden kann, ist eine derartige Berechnung bei den Klösgenpunktwolken mit $w > 1$ nicht erkennbar. Nur der maximale Coveragewert gilt selbst bei steigenden Parametern als "gute" Grenze für den unteren Rand der Punktwolken im Support-Confidence-Raum (siehe Abb. 9.5)

Folgende Änderungen ergeben sich bei Klösgenpunktwolken mit $w > 1$ zu den Klösgenpunktwolken mit $w = 1$ (Leverageheuristik)

- Der Coveragewert im Punkt P_{cmax} ist nicht mehr $1/2$.
- Der minimale und maximale Coveragewert ergeben sind nicht mehr symmetrisch um die Achse mit Coveragewert $1/2$ im Coverage-Confidence-Raum.
- Die Punktwolken sind nicht mehr symmetrisch um die Achse mit Coveragewert $1/2$ im Coverage-Confidence-Raum.
- Mit steigendem Heuristikwert verjüngen sich die Leveragepunktwolken zum Punkt $(1/2, 0)$ bzw. $(1/2, 1)$ im Coverage-Confidence-Raum. Bei den Klösgenpunktwolken ist der Coveragewert dieser beiden Punkte leicht verschoben (s. Anhang A.8.9).

10. Fazit

In der Einleitung dieser Diplomarbeit wurde folgendes Ziel formuliert:

> "Das Ziel dieser Diplomarbeit ist es, zu prüfen, ob es nur mit den Informationen Support und Confidence möglich ist, den Wert einer anderen Heuristik zu bestimmen. Dazu müssen die verwendeten Heuristiken **analysiert** und danach im **Assoziationsraum** dargestellt (**evaluiert**) werden."

Es wurde gezeigt, dass eine Darstellung der Heuristiken Accuarcy, Conviction, Lift, Leverage, Phi-Koeffizient und Klösgen im Assoziationsraum möglich ist. Im Gegensatz zum P-N-Raum ist die Darstellung der Heuristiken im Assoziationsraum durch "Punktwolken" gegeben, deren Lage und Form analysiert wurde. Bei allen Heuristiken, mit Ausnahme der Klösgenheuristik, war es möglich, Felder zu definieren, welche die Wertebreiche der verwendeten Heuristiken bei vorgegebenem Support- und Confidencewert eingrenzen.

Bei jeder analysierten Heuristik konnte man die Heuristikfunktion aus dem P-N-Raum in den Assoziationsraum überführen. Dazu wurden die Werte der Konfusionmatrix mit Hilfe der Gl. 3.4 als Support-, Confidence- und Recallwerte dargestellt. Dieses Verfahren war bei allen analysierten Heuristiken anwendbar (siehe Tabelle 10.1), nur bei der Klösgenheuristik gelang dies nicht.

Die Klösgenheuristik gehört zur Klasse von Heuristiken, bei denen es nicht möglich ist, rein mathematisch die bestehenden Heuristikfunktionen aus dem P-N-Raum in den Assoziationsraum zu überführen. Bei ihr konnte man stattdessen eine Analogie zur vorher betrachteten Leverageheuristik herstellen, wodurch dann eine eingeschränkte Analyse der Klösgenheuristik möglich war.

Aus den umgeformten Funktionen erfolgte anschließend die Analyse der Punktwolken. Bei allen Punktwolken im Support-Confidence-Raum waren die Kanten, welche die Form der Punktwolken bestimmten, "hart". Deshalb lässt sich die Form einer Punktwolke durch eine Funktion beschreiben. Diese Funktion legt exakt fest, dass auf der einen Seite der "Grenze" die Punktwolke existiert und auf der anderen Seite nicht, unabhängig von der Größe der Datenbank.

Bei den meisten Funktionen konnte man die Kanten der Punktwolken durch mathematische Eigenschaften oder die Grunddefinitionen des Assoziationsraums ermitteln. So musste z.B. der Wert unterhalb einer Wurzel größer gleich Null sein, damit der jeweilige Heuristikwert innerhalb der reellen oder rationalen Zahlen liegt.

Für einen vorgegebenen Punkt im Assoziationsraum war man dann in der Lage, den möglichen Wert der analysiert Heuristik einzugrenzen (Feldanalysen). Weil bei den Feldanalysen

	Support	**Confidence**	**Recall**
Accuracy	$\frac{1 - acc(R)}{c^{-1} + r^{-1} - 2}$	$\frac{1}{\frac{1 - acc(R)}{s^{-1}} - r^{-1} + 2}$	$\frac{1}{\frac{1 - acc(R)}{s^{-1}} - c^{-1} + 2}$
Conviction	$\frac{1 - V(R)\ (1-conf(R))}{recall(R)^{-1}}$	$1 - \frac{1 - \frac{suppR)}{recall(R)}}{V(R)}$	$\frac{supp(R)}{1 - V(R)\ (1-conf(R))}$
Lift	$\frac{conf(R)\ recall(R)}{lift(R)}$	$\frac{lift(R)\ supp(R)}{recall(R)}$	$\frac{lift(R)\ supp(R)}{conf(R)}$
Leverage	$\frac{r\ c}{2}\left(1 \pm \sqrt{1 - \frac{4\ l}{r\ c}}\right)$	$\frac{supp(R)^2}{recall(R)\cdot(supp(R) - leverage(R))}$	$\frac{supp(R)^2}{conf(R)\cdot(supp(R) - leverage(R))}$
Phi-Koeffizient	$-\frac{1}{2s_1}\left(s_2 \pm \sqrt{s_2^2 + 4s_1 s_x}\right)$ $s_1 = \phi^2 - 1$ $s_2 = r\ \phi^2 + c)\ (\phi^2 - 2r)$ $s_x = (r\ c)^2 - \phi^2\ r\ c$	$-\frac{1}{2r^2}\left(c_1 \pm \sqrt{c_1^2 - 4r^2 c_x}\right)$ $c_1 = \phi^2\ (s - r) - 2\ s\ r$ $c_x = s\ \phi^2(r - s) - s^2$	$-\frac{1}{2c^2}\left(r_1 \pm \sqrt{r_1^2 + 4c^2 r_x}\right)$ $r_1 = \phi^2\ (s - c) - 2\ s\ c$ $r_x = s\ \phi^2(c - s) - s^2$

Tabelle 10.1.: Auflistung aller Umformungen der analysierten Heuristiken in den Assoziationsraum ($s = supp(R)$, $c = conf(R)$, $r = recall(R)$, $l = leverage(R)$).

auf den Recallwert verzichtet und damit ein Freiheitsgrad variabel gelassen wurde, konnte man den Heuristikwert nicht exakt bestimmen.

Für die Klösgenheuristik funktionierte diese Vorgehensweise nicht, stattdessen wurde ein Ausblick darauf gegeben, wie man die Ergebnisse aus vorherigen auf zukünftige Heuristiken wiederverwenden kann. Wie oben schon beschrieben wurde, nimmt die Klösgenheuristik eine Sonderstellung ein, da man nicht davon ausgehen kann, dass die Mehrzahl der sonstigen Heuristiken in den Assoziationsraum umgeformt werden kann. Aus diesem Grund liegt in der weiteren Analyse der Klösgenheuristik vielleicht der Schlüssel zu einer fundamentaleren Theorie des Assoziationsraums.

Accuracy	$acc(R) = 1 - supp(R)\left(conf(R)^{-1} + recall(R)^{-1} - 2\right)$
Conviction	$V(R) = \frac{1-\frac{supp(R)}{recall(R)}}{1-conf(R)}$
Lift	$lift(R) = \frac{conf(R)\ recall(R)}{supp(R)}$
Leverage	$leverage(R) = supp(R) - \frac{supp(R)^2}{recall(R)\ conf(R)} = supp(R) - \frac{supp(R)}{lift(R)}$
Phi-Koeffizient	$\phi = \frac{recall(R)conf(R)-supp(R)}{\sqrt{(conf(R)-supp(R))\ (recall(R)-supp(R))}}$
Klösgen	$klosgen(R) = \left(\frac{supp(R)}{conf(R)}\right)^w (conf(R) - \frac{supp(R)}{recall(R)})$

Tabelle 10.2.: Auflistung aller Umformungen der analysierten Heuristiken in den Assoziationsraum ($s = supp(R)$, $c = conf(R)$, $r = recall(R)$, $l = leverage(R)$).

Abschließend kann man folgende Aussage treffen:

> Der Assoziationsraum ist eine Analyseraum analog zum P-N-Raum. Es kann jede Heuristik anhand der Formeln 3.4 dargestellt werden. Die Analyse und Bewertung einer Regel kann dann grapisch erfolgen (s. Kapitel 3) wie beim P-N-Raum in [FF05]. Die Möglichkeit einer mathematische Analyse/Umformung ist von der jeweiligen Heuristik abhängig.

10.1. Ausblick

In dieser Diplomarbeit wurde an einigen Stellen der Coverage-Confidence-Raum verwendet. Dieser Raum stellt eine Normierung des Support-Confidence-Raums relativ zur Supportachse dar. Anhand der Punktwolken im Coverage-Confidence-Raum kann man die Steigungen der Punktwolken im Support-Confidence-Raum untersuchen. Dies wurde bereits bei den

	Untere Grenze	**Obere Grenze**
Accuracy	$supp(R)$	$supp(R)\ \left(1\ -\ \frac{1}{conf(R)}\right)\ +\ 1$
Conviction	$\frac{supp(R)}{conf(R)}$	∞
Lift	$conf(R)$	$\frac{conf(R)}{supp(R)}$
Leverage	$max\left\{\frac{conf(R)-1}{4}\ ,\ \frac{(1\ -\ 2cov)^2-1}{4}\right\}$	$min\left\{\frac{conf}{4}\ ,\ \frac{1-(1\ -\ 2cov)^2}{4}\right\}$
Phi-Koeffizient	$max\left\{-\sqrt{1-conf(R)}\ ,\ \frac{supp(R)\ -\ 1}{1\ +\ supp(R)}\right\}$	$min\left\{\sqrt{conf(R)}\ ,\ \sqrt{\frac{supp(R)\ -\ conf(R)}{supp(R)\ -\ 1}}\right\}$

Tabelle 10.3.: Auflistung aller Intervalle der analysierten Heuristiken in den Assoziationsraum (Wichtig: $coverage(R) = supp(R)/conf(R)$ ist nicht cov siehe dazu Gl. 7.35 und 7.36) im Kapitel der Leverageheuristik

Heuristiken Leverage und Klösgen angewendet und könnte bei der Klösgenheuristik fortgeführt werden, um detailliertere Aussagen über die Lage der Klösgenpunktwolken zutreffen. Beim Phi-Koeffizieten wurde der Coverage-Confidence-Raum verwendet um die Symmetrieeigenschaften des pos. und neg. Wertebereiches in Abb. 8.4 zu verdeutlichen.

Analog kann man die Analyse der Punktwolken im Coverage-Confidence-Raum bei allen Heuristiken verwenden, damit könnte die Diplomarbeit fortgesetzt werden. Eine Darstellung der einzelnen Heuristiken im Coverage-Confidence-Raum wird in Tabelle 10.4 gezeigt.

Durch eine genauere Analyse des Coverage-Confidence-Raums gibt es möglicherweise eine Erklärung für die monotone Veränderung der Punktwolken (im Coverage-Confidence-Raum) bei gleich bleibendem Heuristikwert und steigendem Parameter w der Klösgenheuristik. Falls das so ist, dann kann man hieraus ein Klösgenwertintervall entwickeln, wie bei den restlichen anaylsierten Heuristiken dieser Diplomarbeit.

Accuracy	$acc(R) = 1 - coverage(R)\left(1 + \frac{conf(R)}{recall(R)} - 2conf(R)\right)$
Conviction	$V(R) = \frac{1 - \frac{conf(R)\ coverage(R)}{recall(R)}}{1 - conf(R)}$
Lift	$lift(R) = \frac{recall(R)}{coverage(R)}$
Leverage	$leverage(R) = conf(R)coverage(R) - \frac{conf(R)coverage(R)^2}{recall(R)}$
Phi-Koeffizient	$\phi = \frac{conf(R)(recall(R) - coverage(R))}{\sqrt{conf(R)(1-coverage(R))\ (recall(R)-conf(R)coverage(R))}}$
Klösgen	$klosgen(R) = coverage(R)^w\ (conf(R) - \frac{conf(R)coverage(R)}{recall(R)})$

Tabelle 10.4.: Auflistung aller Umformungen der analysierten Heuristiken in den Coverage-Confidence-Raum mit $coverage(R) = supp(R)/conf(R)$

Literaturverzeichnis

[AMJ05] Paulo J. Azevedo Alípio M. Jorge. An Experiment with Association Rules and Classification: Post-Bagging and Conviction. In Hoffmann A Carbonell JG, Motoda H, editor, *Proceedings of the 8th International Conference on Discovery Science (DS 2005)*, volume 3735 of *Lecture Notes in Computer Science*, pages 137–149, Singapore, October 2005. Springer. ISI, ISIProc.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (ISBN:0-387-31073-8)*. Springer, 2006.

[FF05] Johannes Fürnkranz and Peter A. Flach. ROC 'n' Rule Learning - Towards a Better Understanding of Covering Algorithms. *Machine Learning*, 58(1):39–77, 2005.

[FJ06] Johannes Fürnkranz and Frederik Janssen. On Trading Off Consistency and Coverage in Inductive Rule Learning. *KDML 2006: Knowledge Discovery, Data Mining, and Machine Learning*, 2006.

[Fla03] P.A. Flach. The geometry of ROC space: Understanding Machine Learning Metrics through ROC Isometrics. In *Proc. 20th International Conference on Machine Learning (ICML'03)*, pages 194–201. AAAI Press, January 2003.

[Für09] Johannes Fürnkranz. Folien der Vorlesungen: Maschinelles Lernen - Symbolische Ansätze, Web Mining und Einführung in die Künstliche Intelligenz. TU Darmstadt, Fachbereich: Knowledge Engineering, 2009.

[Jan06] Frederik Janssen. Eine Untersuchung des Trade-Offs zwischen Precision und Coverage bei Regel-Lern-Heuristiken. Master's thesis, TU Darmstadt, 2006.

[JBB05] Regis Gras Julien Blanchard, Fabrice Guillet and Henri Briand. Using Information-Theoretic Measures to Assess Association Rule Interestingness. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 66–73, Washington, DC, USA, 2005. IEEE Computer Society.

[JL06] H. Wegmann J. Lehn. *Einführung in die Statistik (ISBN:978-3835100046)*. Vieweg+Teubner, 5 edition, Juni 2006.

[Klö92] Willi Klösgen. Problems for Knowledge Discovery in Databases and their Treatment in the Statistics Interpreter Explora. *International J. Of Intelligent Systems*, pages 7:649–673, 1992.

[Klö96] Willi Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. *In Advances in Knowledge Discovery and Data Mining*, pages 249–271, 1996.

[LHCM00] Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma. Analyzing the Subjective Interestingness of Association Rules. *IEEE Intelligent Systems (issn:1541-1672)*, 15(5):47–55, 2000.

[Mit97] Thomas M. Mitchell. *Machine Learning (ISBN:0070428077)*. McGraw-Hill, 1997.

[RAS93] Tomasz Imielinski Rakesh Agrawal and Arun N. Swami. Mining Association Rules between Sets of Items in Large Databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C.*, pages 207–216. ACM Press, 1993.

[RJBA99] Jr. Roberto J. Bayardo and Rakesh Agrawal. Mining the Most Interesting Rules. In *KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ISBN: 1581131437)*, pages 145–154, New York, NY, USA, 1999.

[SB95] Hans-Joachim Novak Stephan Busemann. Generierung Natürlicher Sprache. In Günther Görz, editor, *Einführung in die Künstliche Intelligenz*, pages 492–540. Addison-Wesley, 1995. 2. Auflage.

[TKS02] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the Right Interestingness Measure for Association Patterns. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 32–41, New York, NY, USA, 2002. ACM.

[WCH07] Tianyi Wu, Yuguo Chen, and Jiawei Han. Association Mining in Large Databases: A Re-examination of Its Measures. In *PKDD 2007: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 621–628, Berlin, Heidelberg, 2007. Springer-Verlag.