Analyse von Heuristiken zur Evaluierung von Assoziationsregeln

Diplomvortrag
im Fachbereich „Knowledge Engineering“

Prof. Dr. Johannes Fürnkranz
TU-Darmstadt

Betreuer: Jan-Nikolas Sulzmann

Diplomand: Florian Nattermann

Übersicht

- Einleitung

  Regeln, Kontingenztabelle

- Assoziationsregeln

  Support, Confidence, Recall und der Assoziationsraum

- Accuracy
- Lift
- Leverage
- Phi-Koeffizient

- Fazit

Ein Datensatz und eine Regel

Beispieldatenbank eines Supermarktes:

Attribut

Eintrag/
Datensatz

	Pizza	Brot	Chips	Milch	Bier
1	ja	nein	nein	ja	ja
2	nein	nein	ja	nein	nein
3	nein	ja	ja	nein	nein
4	ja	ja	ja	nein	nein
5	ja	ja	nein	ja	ja
6	nein	nein	nein	nein	nein
7	nein	ja	ja	nein	ja
8	nein	nein	nein	nein	ja
9	nein	ja	nein	nein	nein
10	ja	ja	nein	nein	nein

Regel R: $\underbrace{\text{Pizza:ja} \wedge \text{Chips:nein}}_{\text{Körper}} \rightarrow \underbrace{\text{Bier:ja}}_{\text{Kopf}}$

Ein Datensatz und eine Regel

Nattermann

Regel R:	A → B	: 1, 5
Körper A:	Pizza:ja ∧ Chips:nein	: 1, 5, 10
Kopf B:	Bier:ja	: 1, 5, 7, 8

	Pizza	Brot	Chips	Milch	Bier
1	ja	nein	nein	ja	ja
2	nein	nein	ja	nein	nein
3	nein	ja	ja	nein	nein
4	ja	ja	ja	nein	nein
5	ja	ja	nein	ja	ja
6	nein	nein	nein	nein	nein
7	nein	ja	ja	nein	ja
8	nein	nein	nein	nein	ja
9	nein	ja	nein	nein	nein
10	ja	ja	nein	nein	nein

True positives (C0=2): 1, 5

False positives (C1=1): 10

False negatives (C2=2): 7, 8

True negatives (C3=5): 2, 3, 4, 6, 9

Kontingenztabelle:

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

	B	$\overline{B}$
A	2	1
$\overline{A}$	2	5

Anzahl aller Einträge in der Datenbank: N = C0 + C1 + C2 + C3

Support, Confidence und Recall

Nattermann

R: Pizza:ja ∧ Chips:nein → Bier:ja

A B

	B	$\overline{B}$
A	C0=2	C1=1
$\overline{A}$	C2=2	C3=5

	Pizza	Brot	Chips	Milch	Bier
1	ja	nein	nein	ja	ja
2	nein	nein	ja	nein	nein
3	nein	ja	ja	nein	nein
4	ja	ja	ja	nein	nein
5	ja	ja	nein	ja	ja
6	nein	nein	nein	nein	nein
7	nein	ja	ja	nein	ja
8	nein	nein	nein	nein	ja
9	nein	ja	nein	nein	nein
10	ja	ja	nein	nein	nein

$$support(R) = supp(R) = \frac{C0}{N} = \frac{2}{10} = 0.2$$

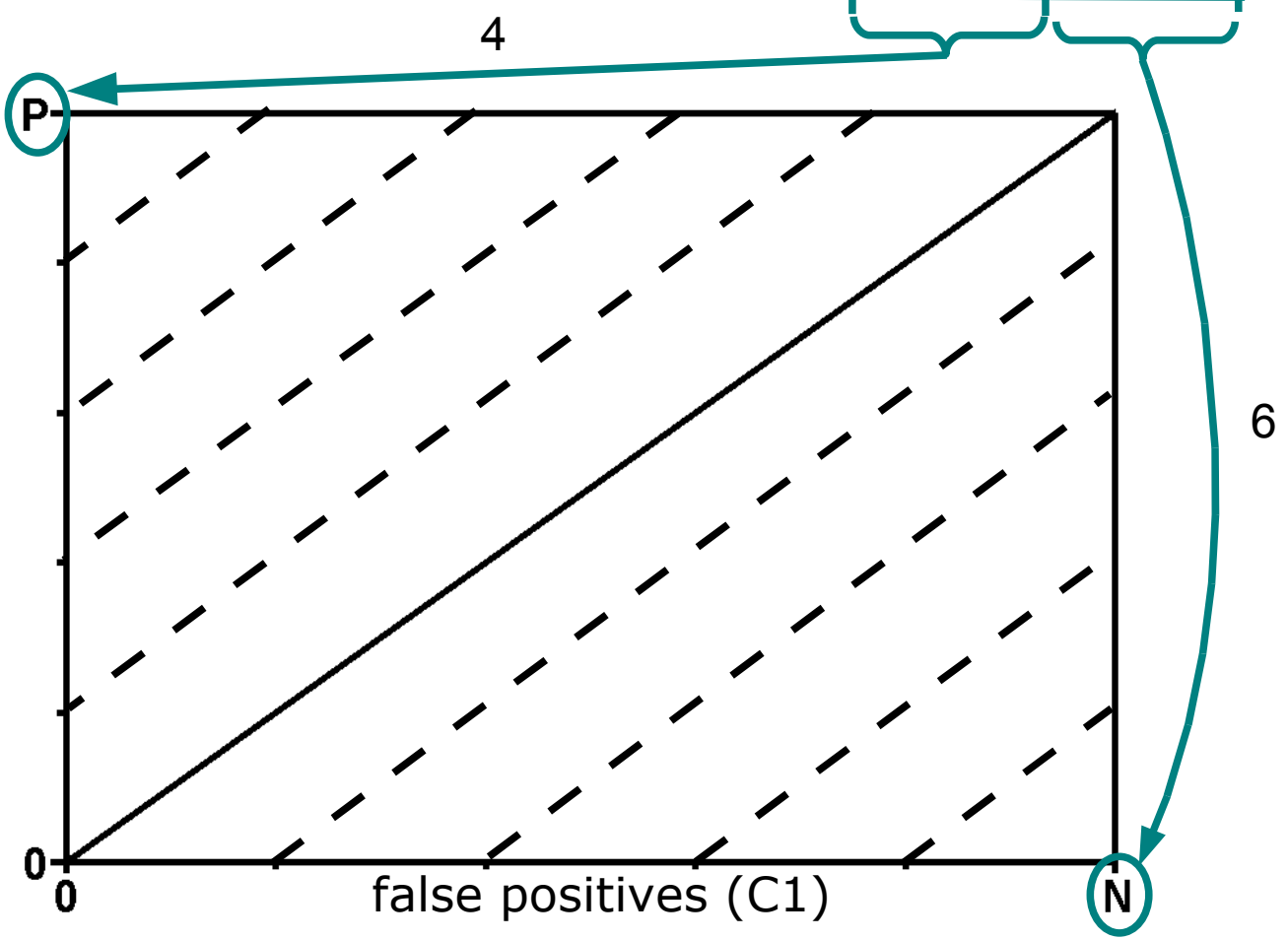
$$confidence(R) = conf(R) = \frac{supp(R)}{supp(A)} = \frac{C0}{C0+C1} = \frac{2}{3}$$

$$recall(R) = \frac{C0}{C0+C2} = \frac{supp(R)}{supp(B)} = \frac{2}{4} = 0.5$$

Der P-N-Raum

R: A → B

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



Assoziationsraum

Nattermann

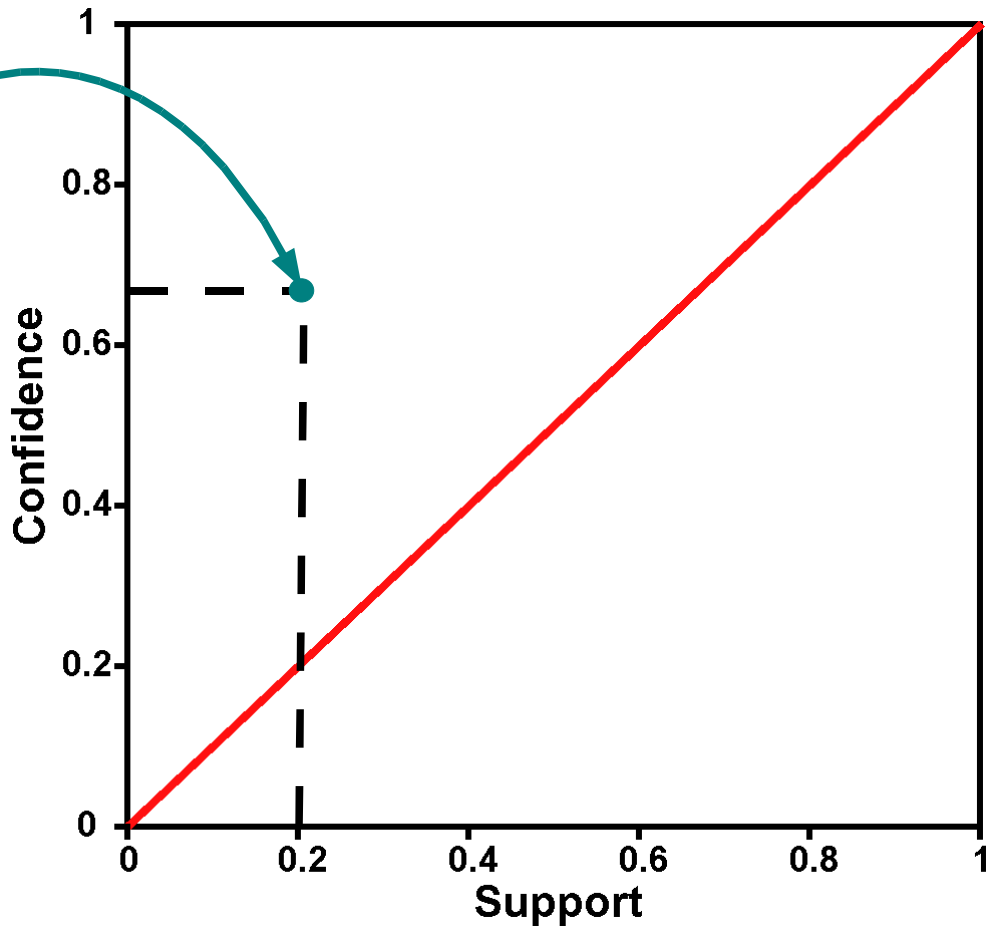
R: Pizza:ja $\wedge$ Chips:nein $\rightarrow$ Bier:ja

A B

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

$$supp(R) = 0.2$$

$$conf(R) = \frac{2}{3}$$



Assoziationsraum

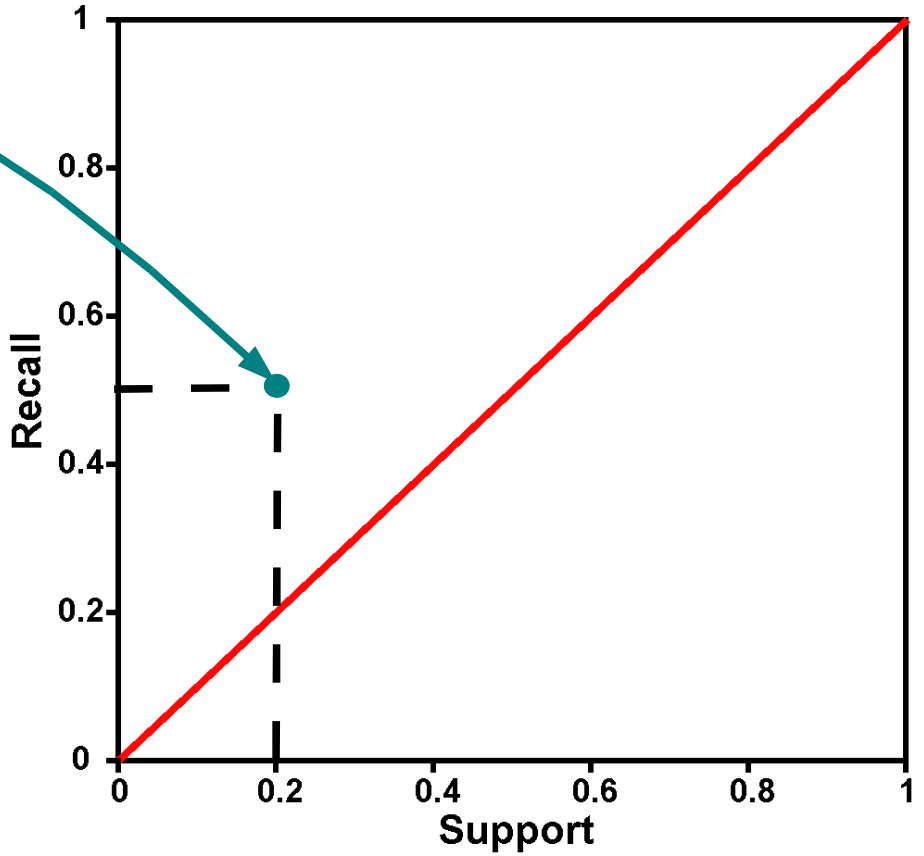
R: Pizza:ja ∧ Chips:nein → Bier:ja

A B

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

$supp(R) = 0.2$

$recall(R) = 0.5$



Assoziationsraum

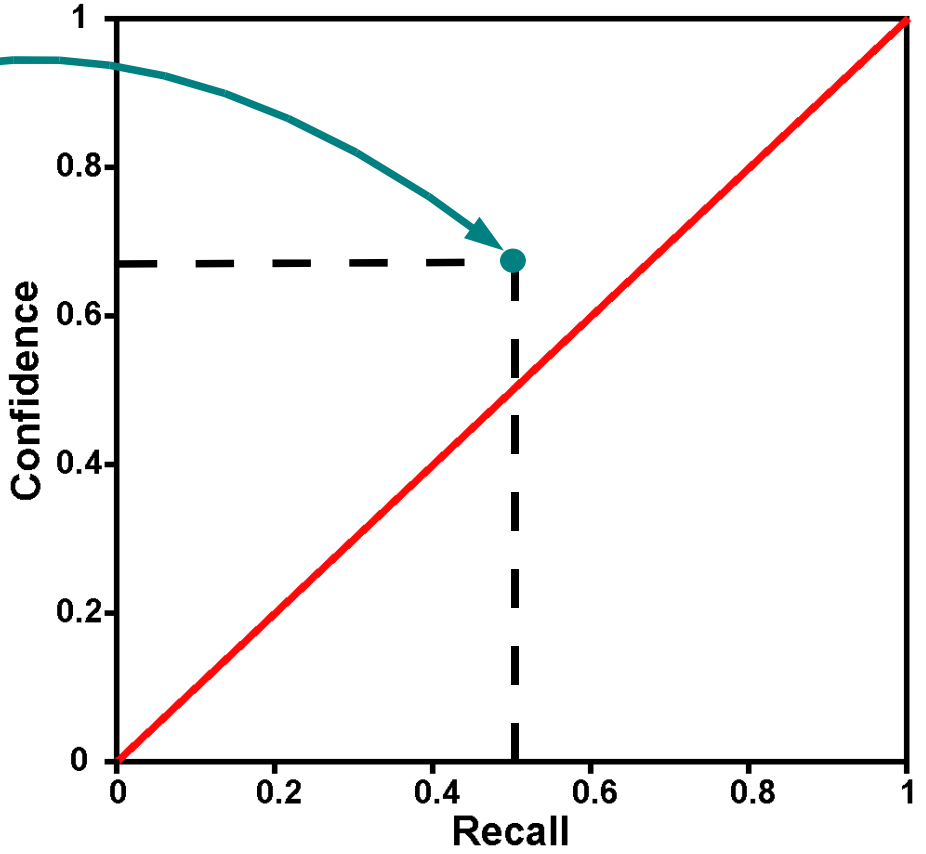
R: Pizza:ja ∧ Chips:nein → Bier:ja

A B

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

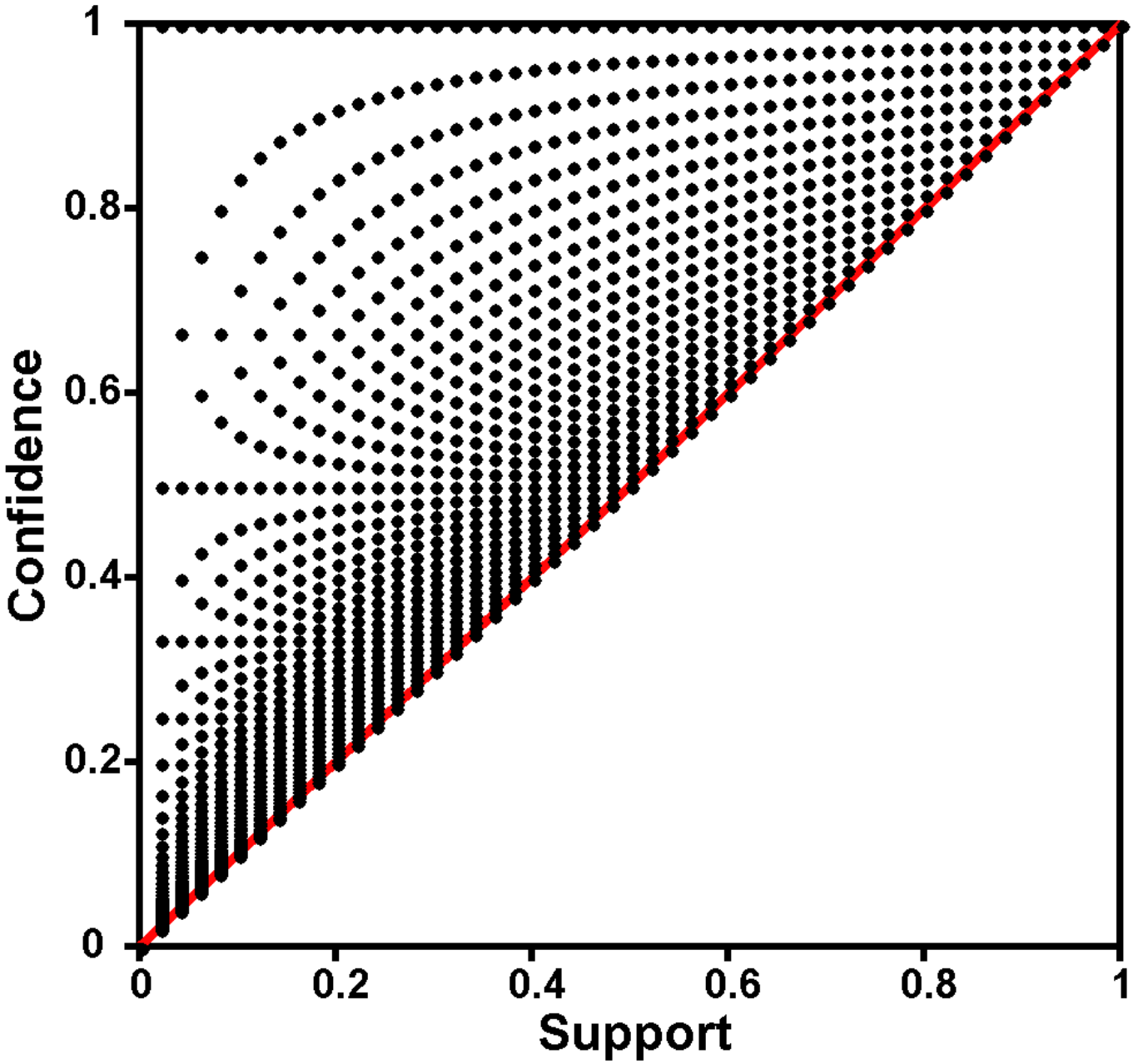
$$recall(R) = 0.5$$

$$conf(R) = \frac{2}{3}$$



Support-Confidence-Raum

Nattermann



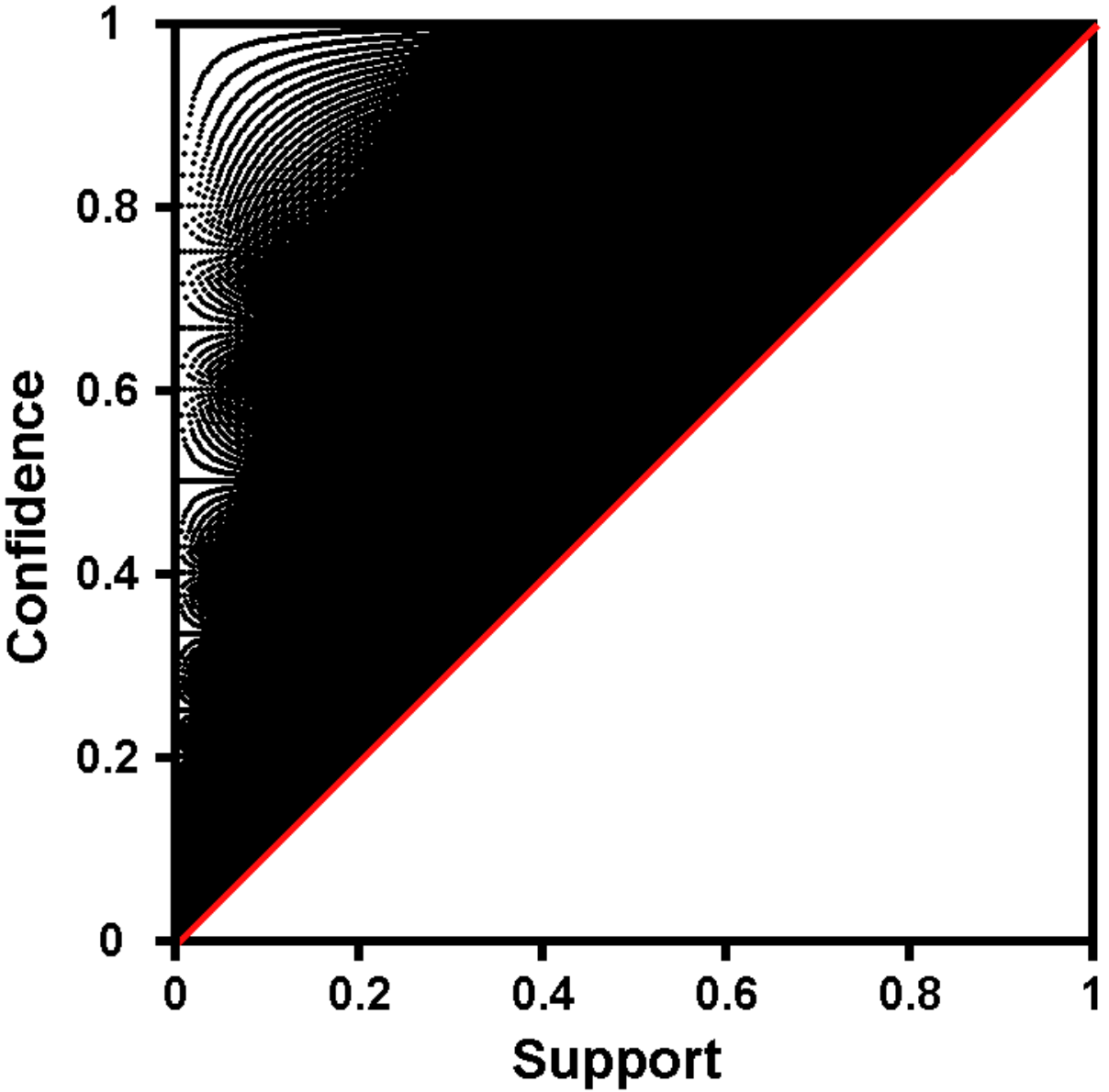
N = 50

Support-Confidence-Raum


```
For (C0=0 ; C0<=N ; C0++)
{
    For (C1=0 ; C1<=N-C0 ; C1++)
    {
        For (C2=0 ; C1<=N-C0-C1 ; C2++)
        {
            .........
        }
    }
}
```

Support-Confidence-Raum

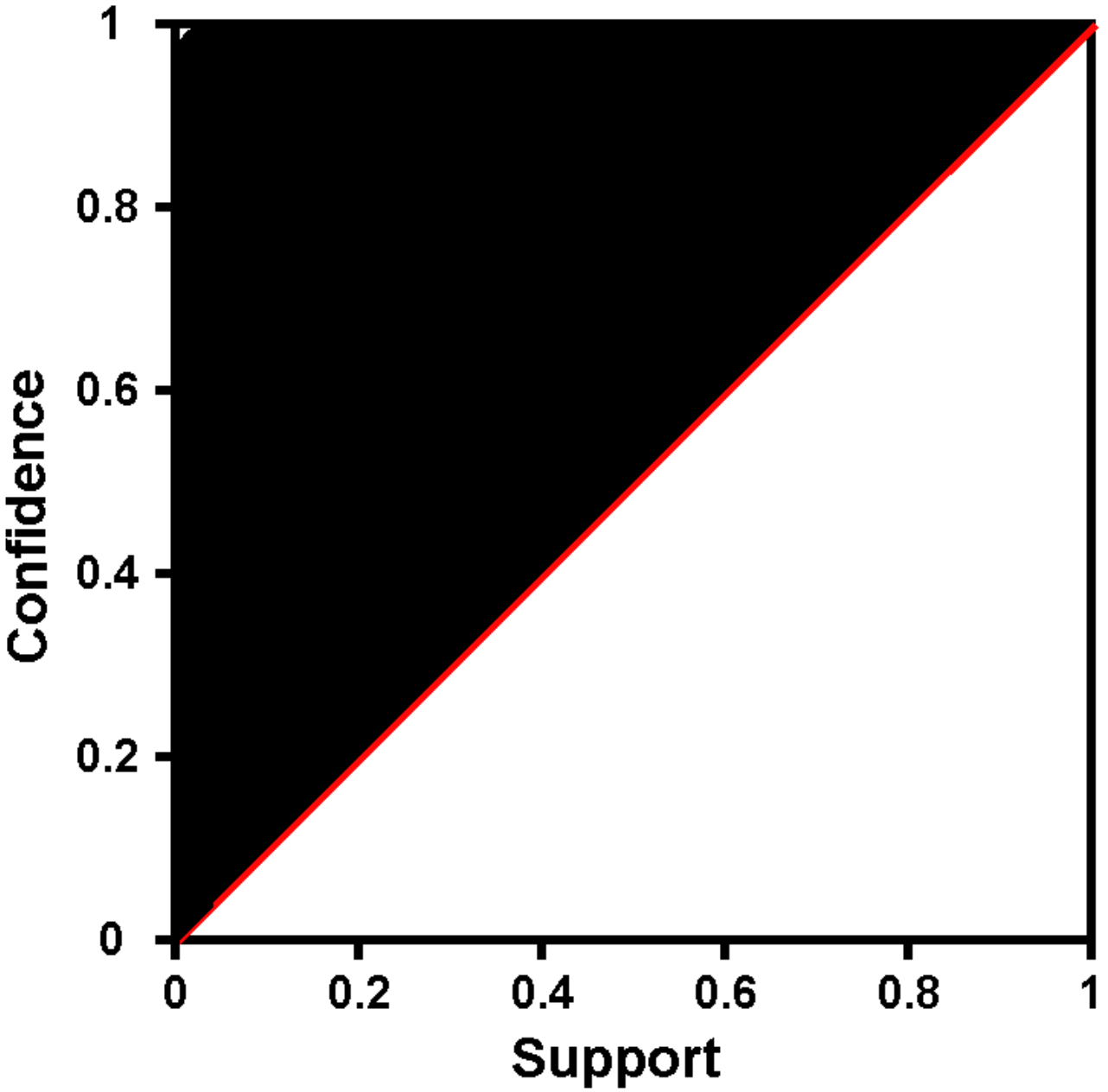
Nattermann



N = 500

Support-Confidence-Raum

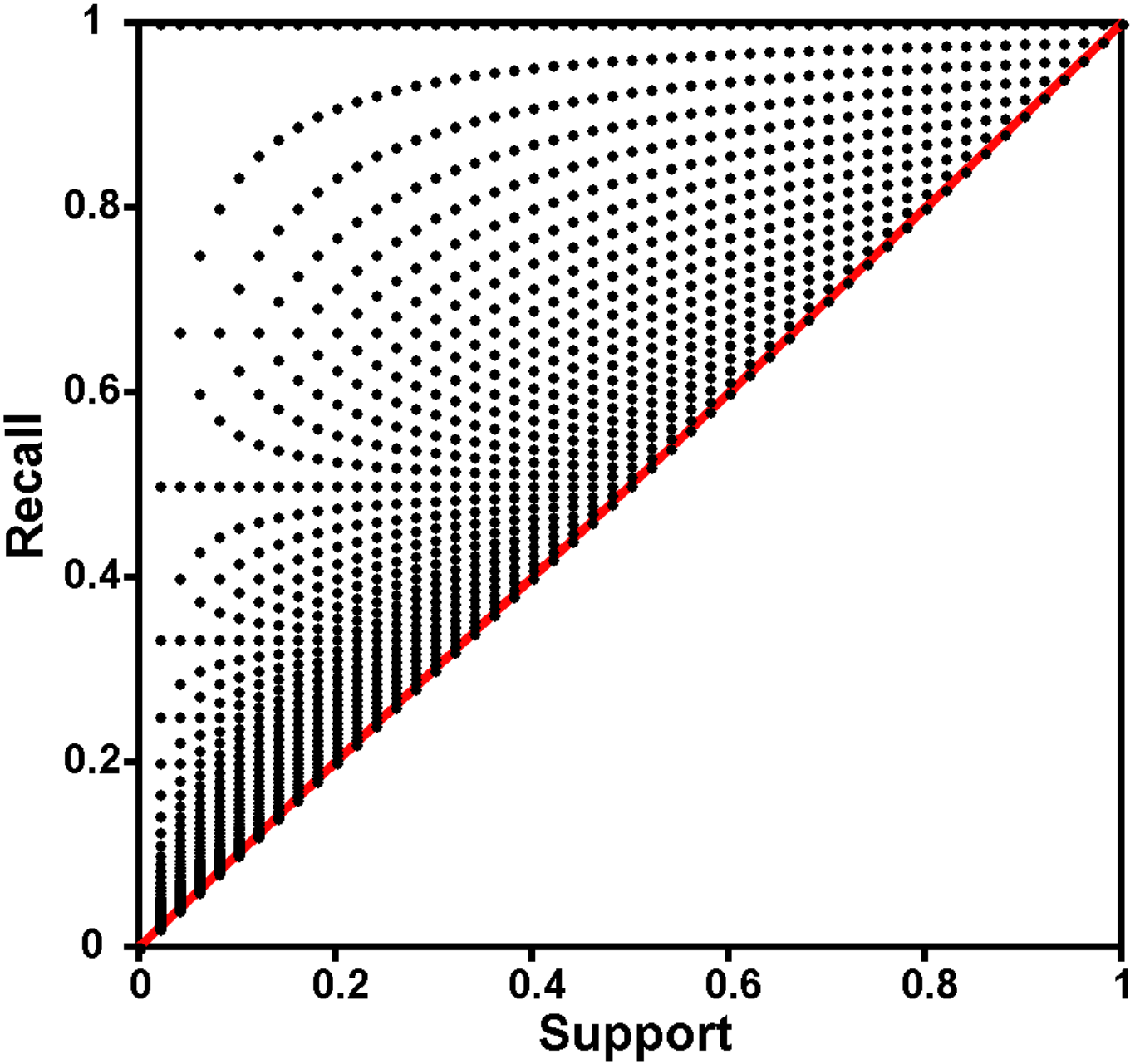
Nattermann



N = 5000

Support-Recall-Raum

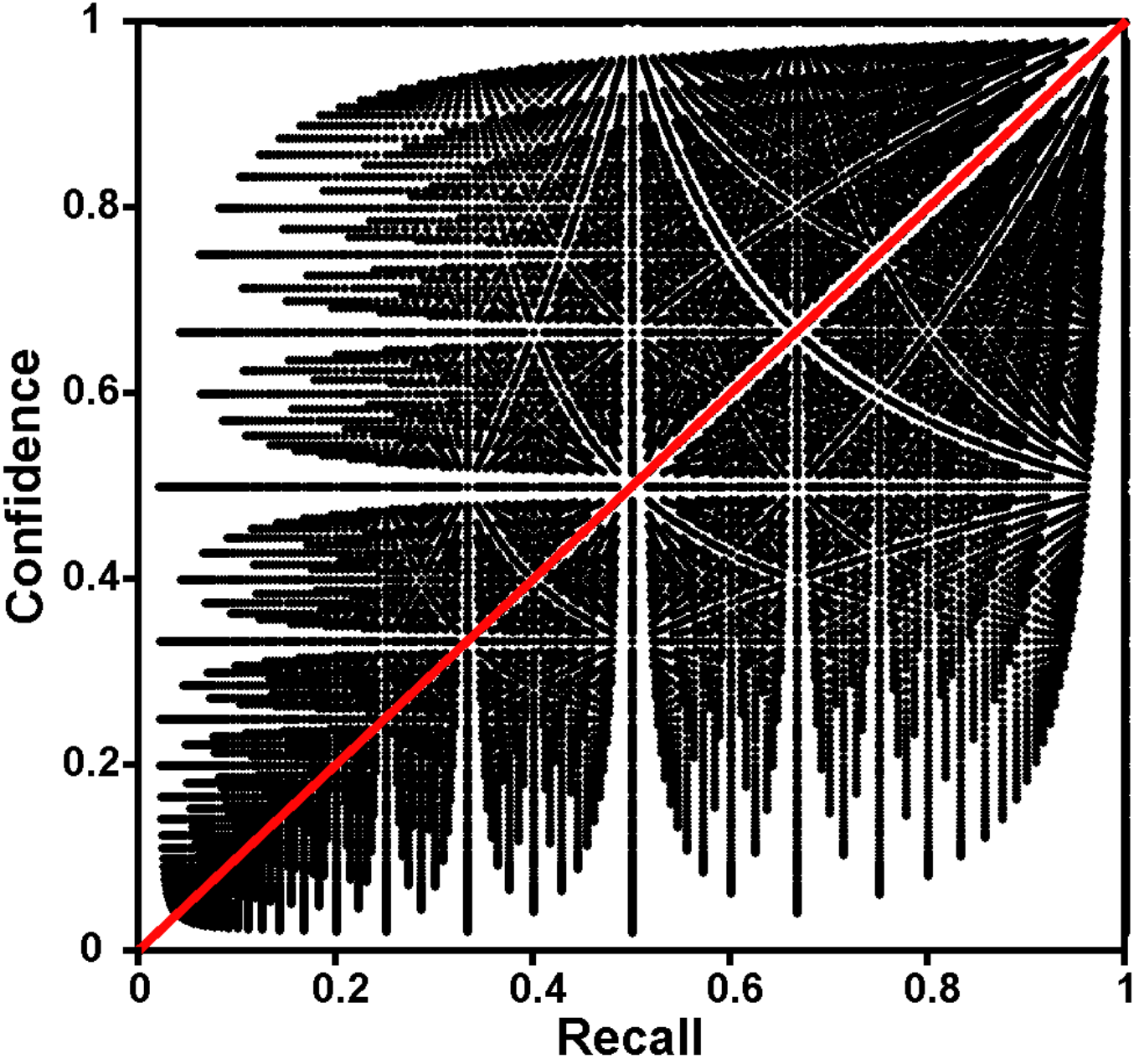
Nattermann



N = 50

Confidence-Recall-Raum

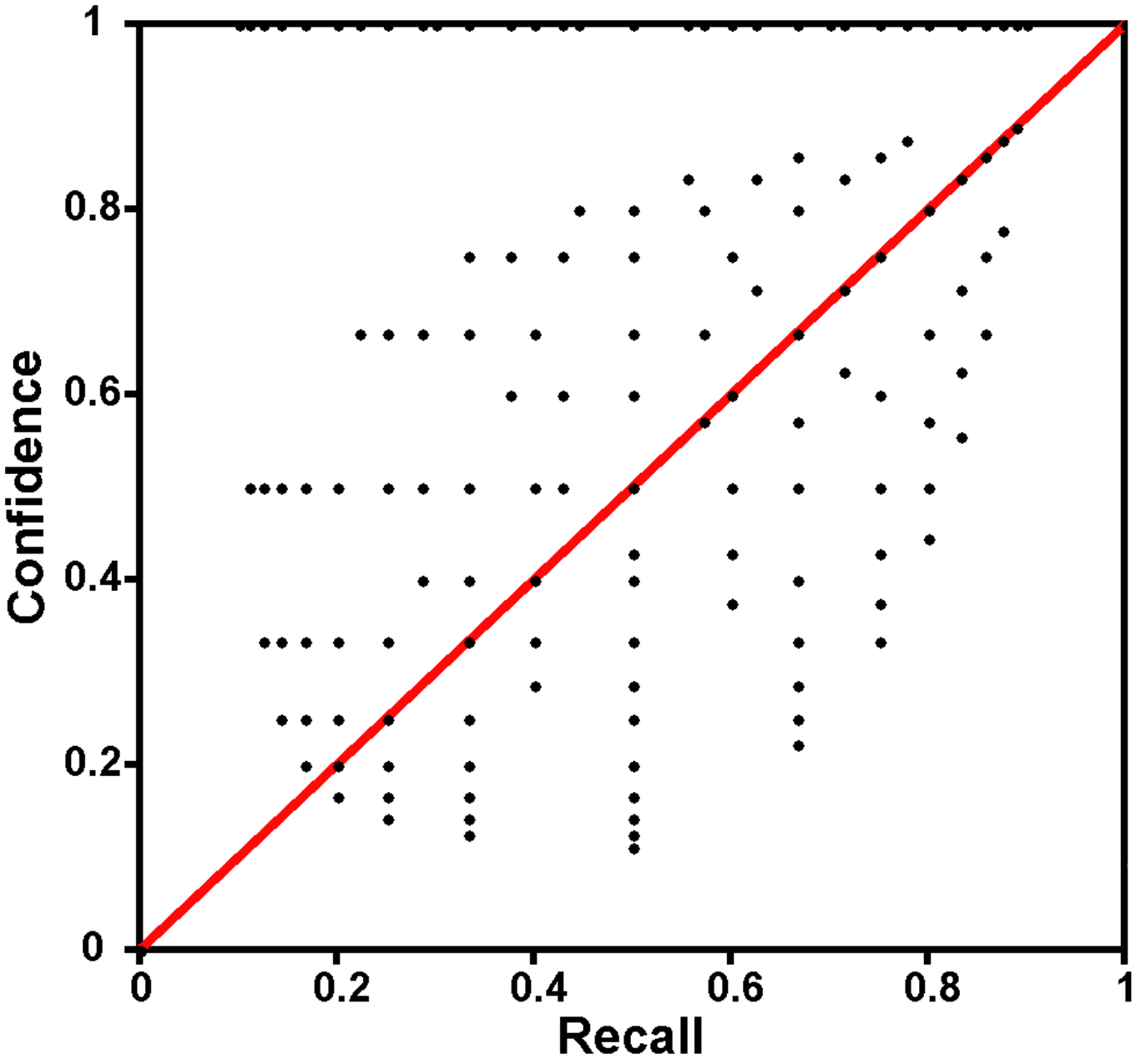
Nattermann



N = 50

Confidence-Recall-Raum

Nattermann



N = 10

Kontingenztabelleneinträge

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

$$supp(R) = \frac{C0}{N} \Rightarrow C0 = supp(R)\ N$$

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

$$recall(R) = \frac{C0}{C0+C2} \Rightarrow C2 = C0\left(\frac{1}{recall(R)} - 1\right)$$

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

Kontingenztabelleneinträge

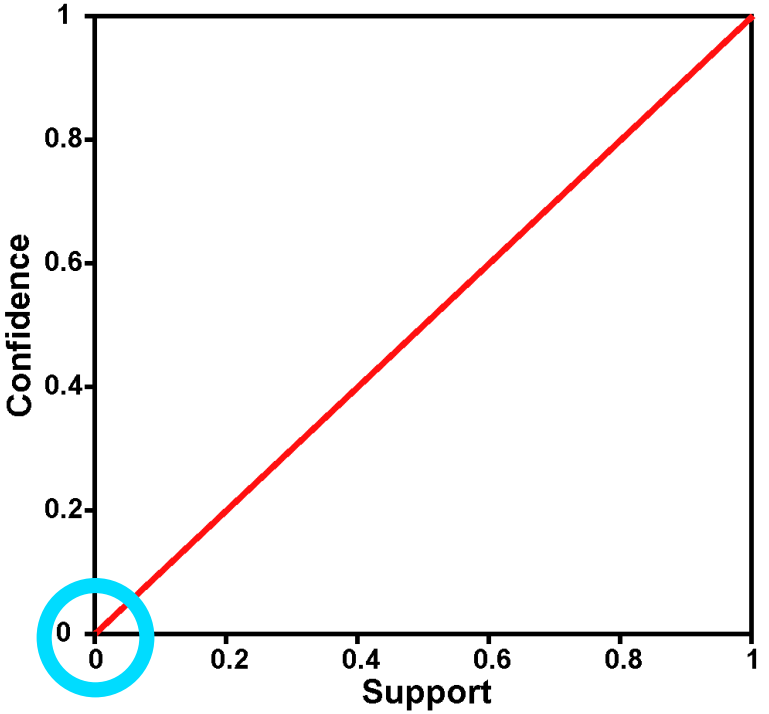
- Supportwert klein, also auch C0-Wert klein

Falls C0 = 0 ➡ supp(R) = 0

conf(R) = 0

Falls C0 = 1 ➡ supp(R) = sehr klein

conf(R) = [0 , 1]



C1 groß (= N-C0) und C0 klein ➡ supp(R) = sehr klein

conf(R) = sehr klein

Für Regel R: $A \rightarrow B$ gilt:

In diesem Datensatz erfüllen alle Einträge die Vorrausetzung A, aber nur wenige Datensätze hiervon haben die gesuchte Klassifikation B.

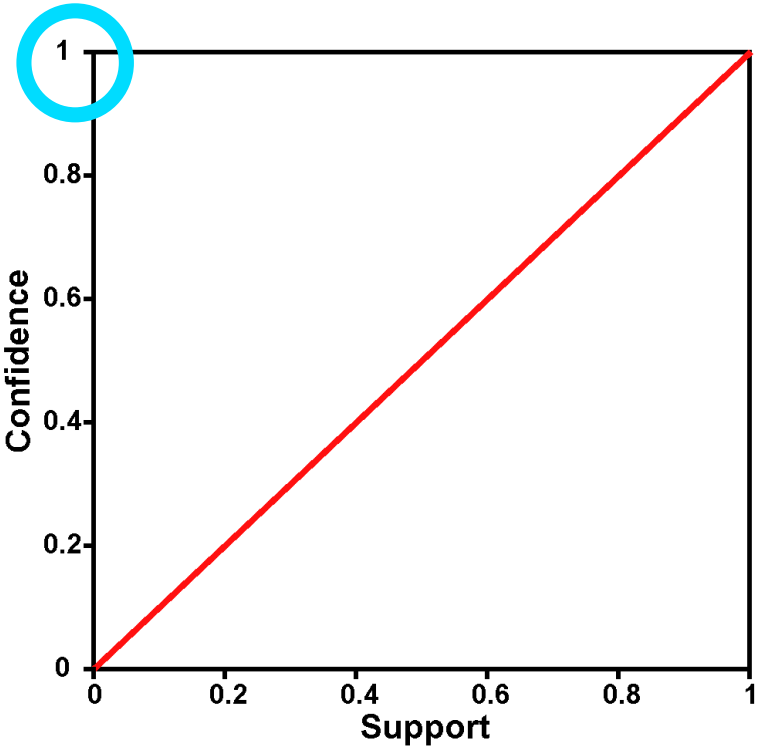
Kontingenztabelleneinträge

- Supportwert klein, also auch C0-Wert klein

Falls C0 = 0 ➡ supp(R) = 0

Falls C0 = 1 ➡ supp(R) ≈ 0

Falls C1 = 0 ➡ conf(R) = 1



Für Regel R: $A \rightarrow B$ gilt:

Es erfüllen gar keine oder kaum Datensatzeinträge die Bedingung A.
Wenn doch dann sehr wenige und dann ist B auch immer erfüllt.

Kontingenztabelleneinträge

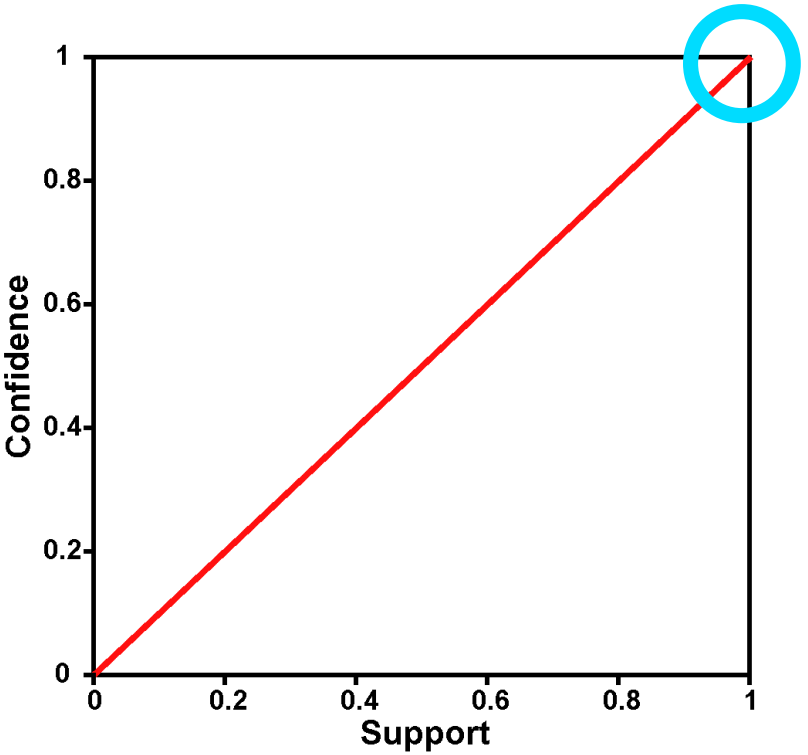
- Supportwert **groß**, also auch C0-Wert **groß**

Falls C0 ≈ N ➡ supp(R) = 1

➡ $$conf(R) = \left[\frac{C0}{C0+N-C0} \; , \; \frac{C0}{C0+0} \right]$$

$$conf(R) = [supp(R) \; , \; 1]$$

$$conf(R) = 1$$



Für Regel R: A → B gilt:

Alle Datensätze erfüllen die Regel R. Fast jeder Datensatz der die Voraussetzung A erfüllt, erfüllt auch die Klassifikation B.

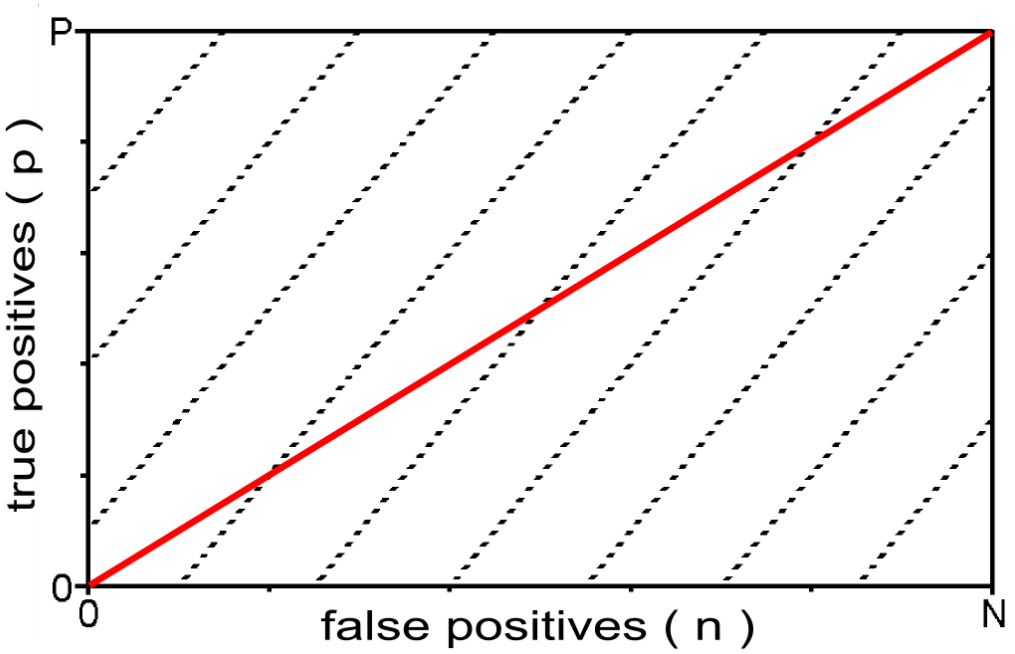
Die Accuracyheuristik

Accuracy

- Im P-N-Raum:

$$acc(R) = \frac{p+(N-n)}{P+N} \approx p-n$$

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



- Im Assoziationsraum: $acc(R) = \frac{C0+C3}{N} = supp(R)+\frac{C3}{N}$

Accuracy

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

- Im P-N-Raum:

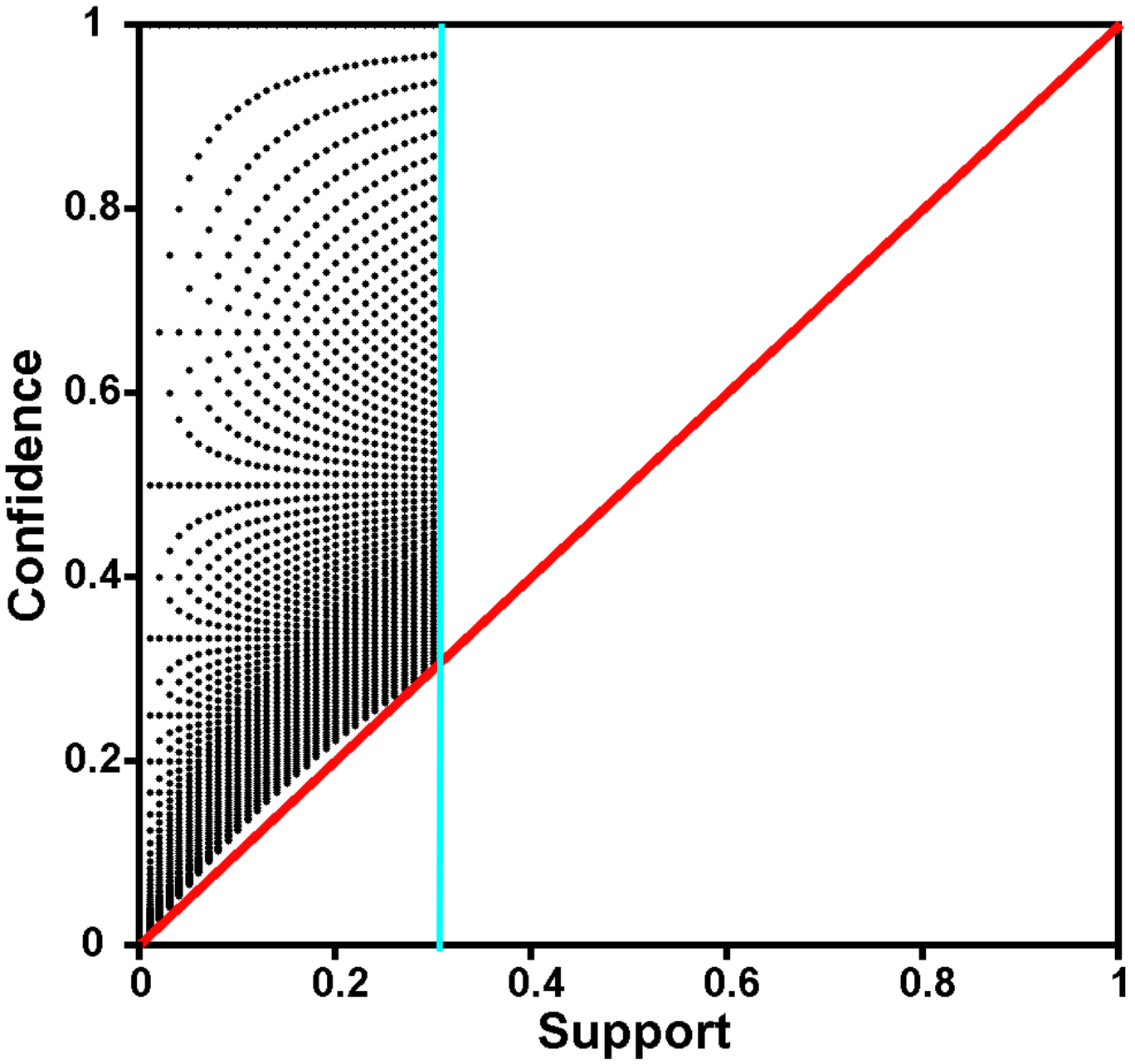
$$acc(R) = \frac{p+(N-n)}{P+N} \approx p-n$$

- Im Assoziationsraum: $acc(R) = \frac{C0+C3}{N} = supp(R)+\frac{C3}{N}$

  - $acc(R) = 1 - supp(R)\left(conf(R)^{-1} + recall(R)^{-1} - 2\right)$

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

Accuracy im Support-Confidence-Raum

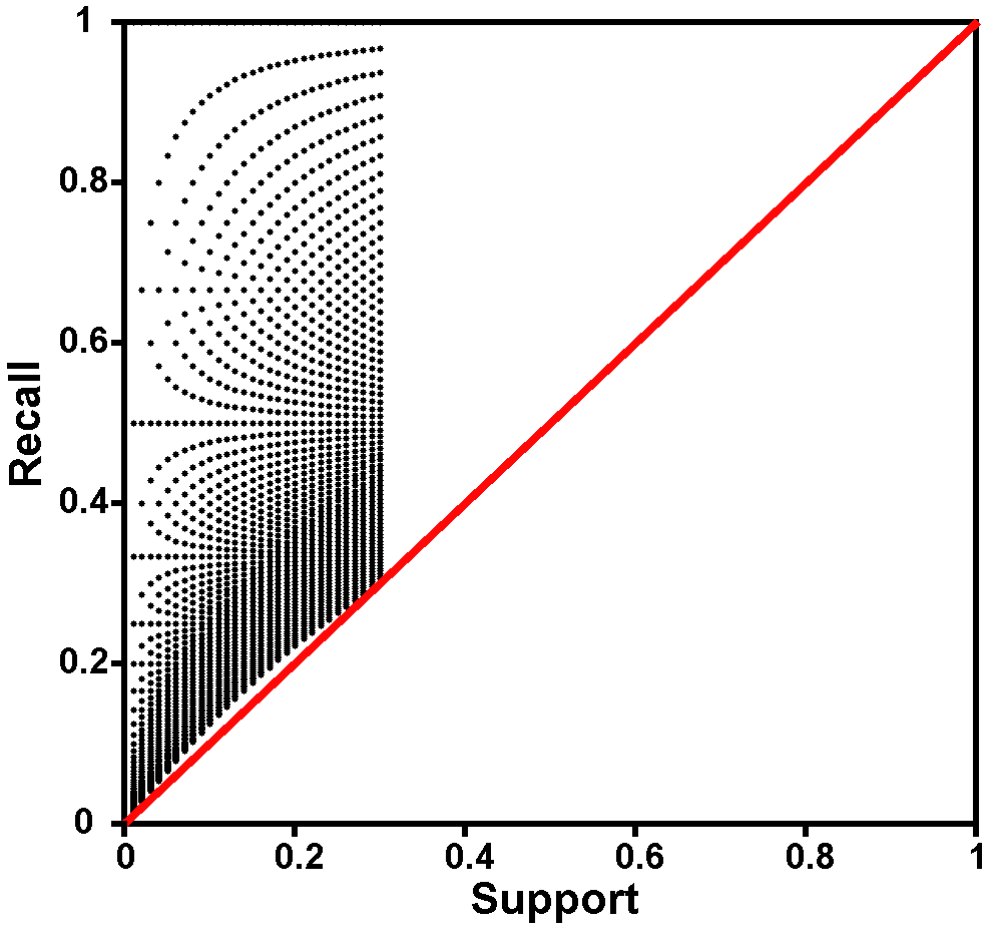


N = 100

acc(R) = 0.3

Accuracy im Support-Confidence-Raum

Nattermann



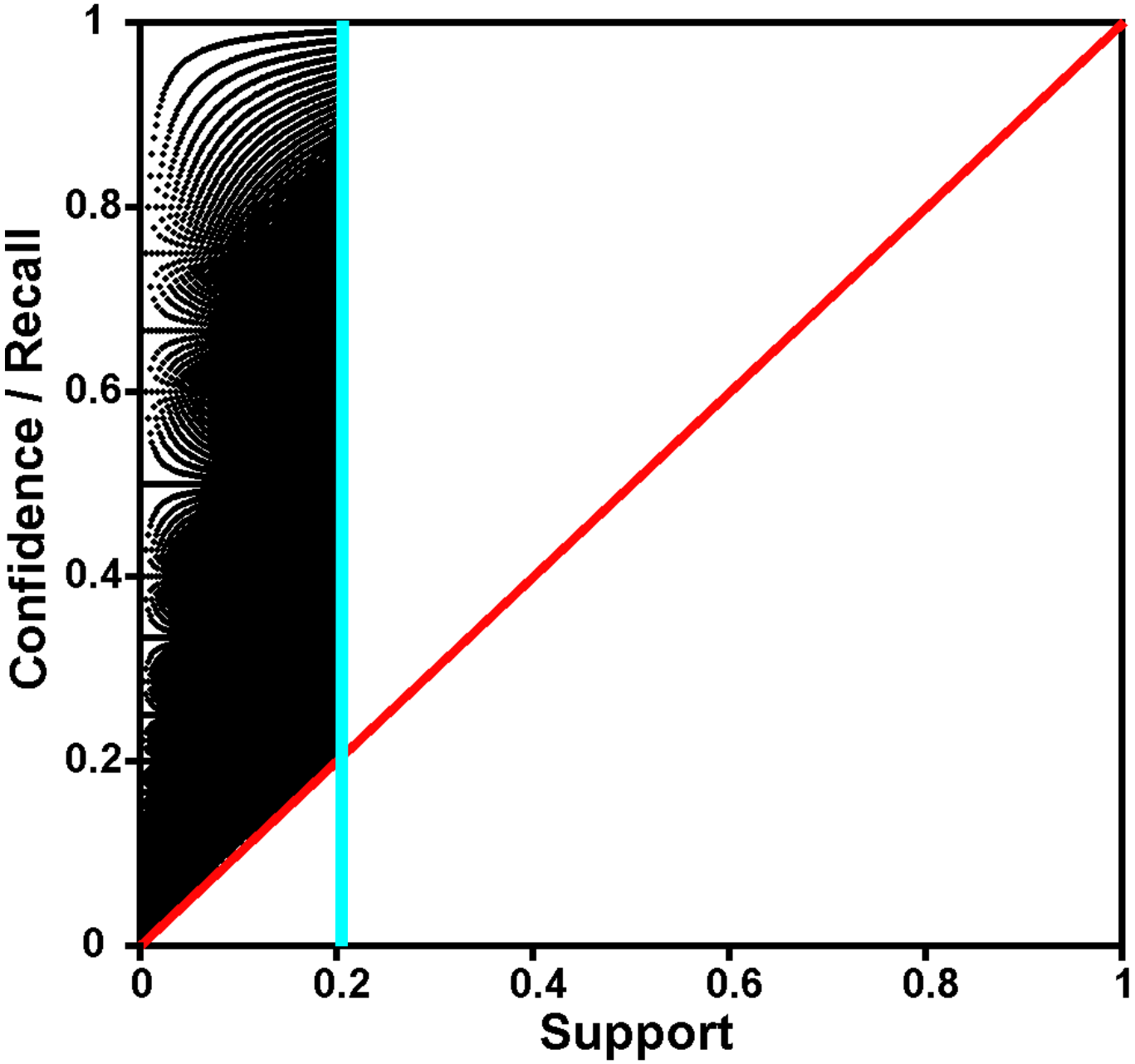
N = 100

acc(R) = 0.3

$$acc(R) = 1 - supp(R)\left(conf(R)^{-1} + recall(R)^{-1} - 2\right)$$

Accuracy im Support-Confidence-Raum

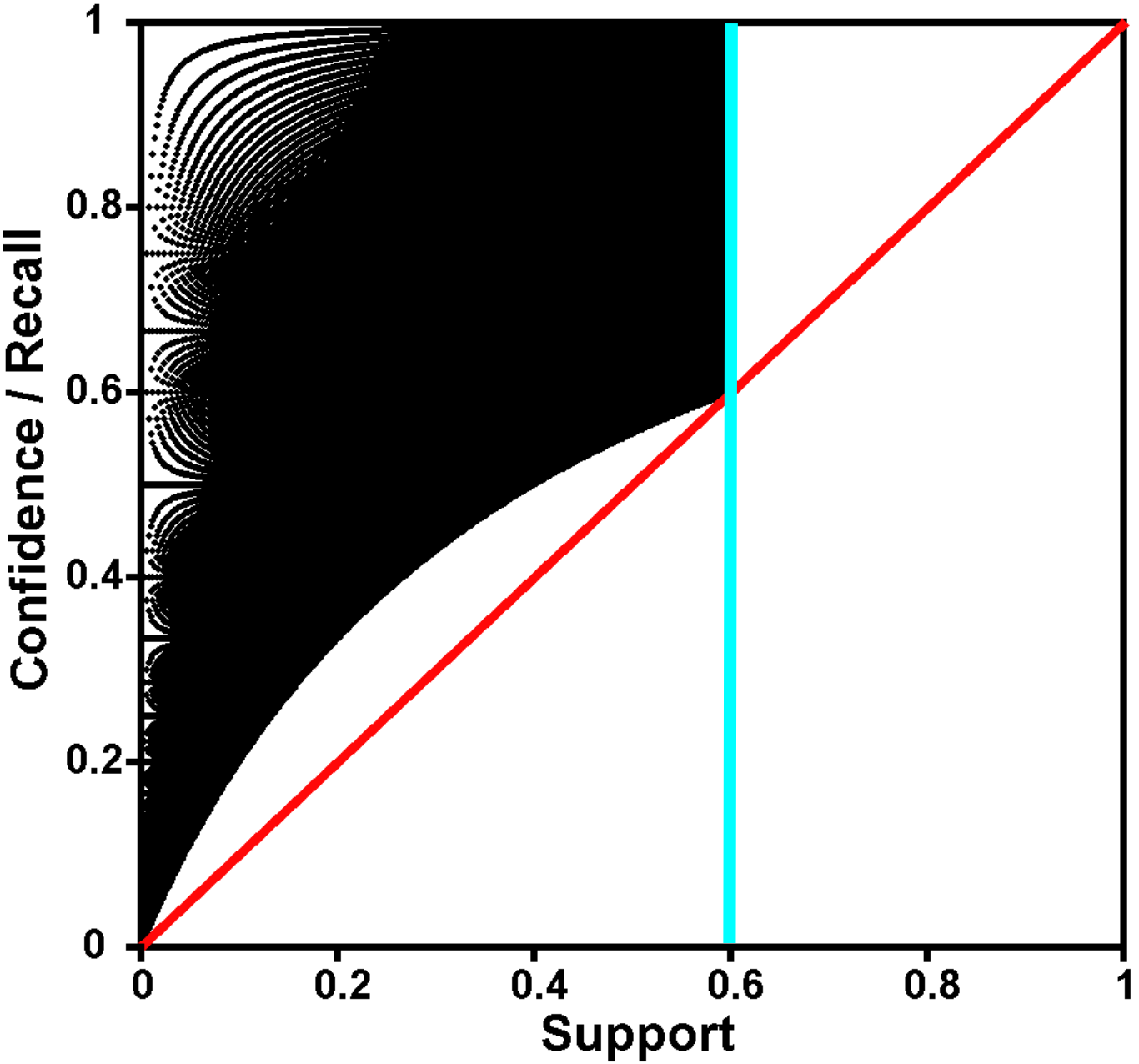
Nattermann



N = 100

acc(R) = 0.2

Accuracy im Support-Confidence-Raum

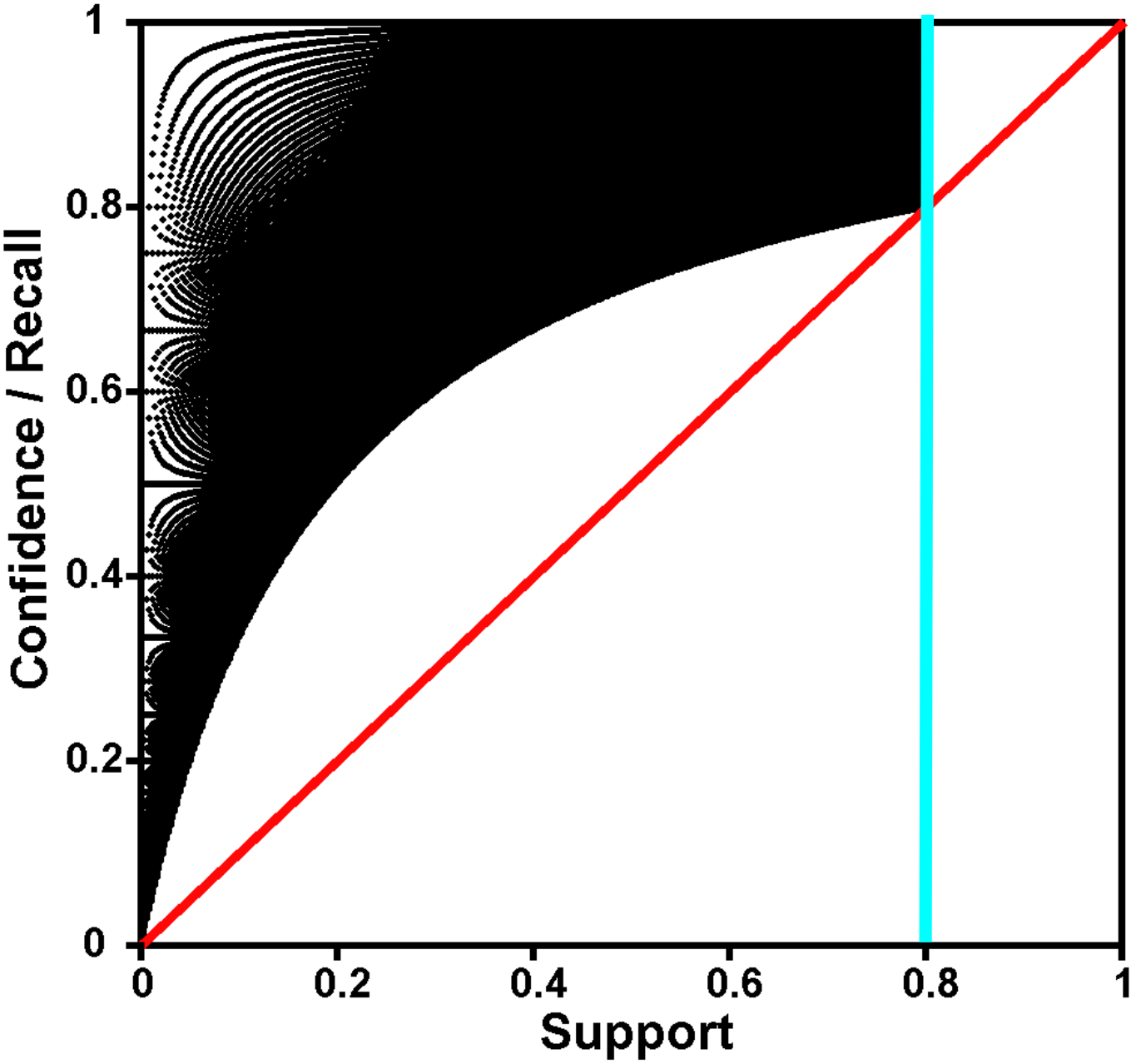


N = 100

acc(R) = 0.6

Accuracy im Support-Confidence-Raum

Nattermann

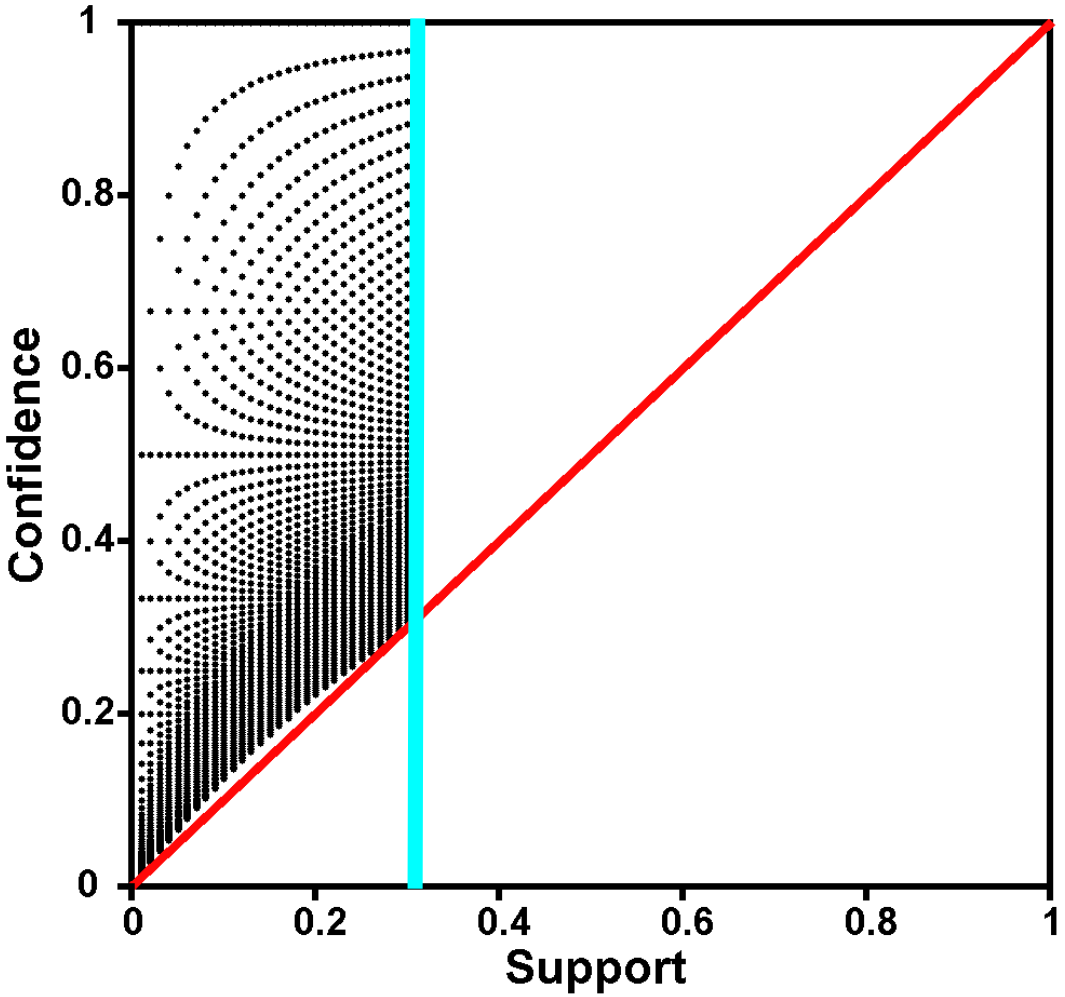


N = 100

acc(R) = 0.8

Accuracy im Support-Confidence-Raum

Nattermann



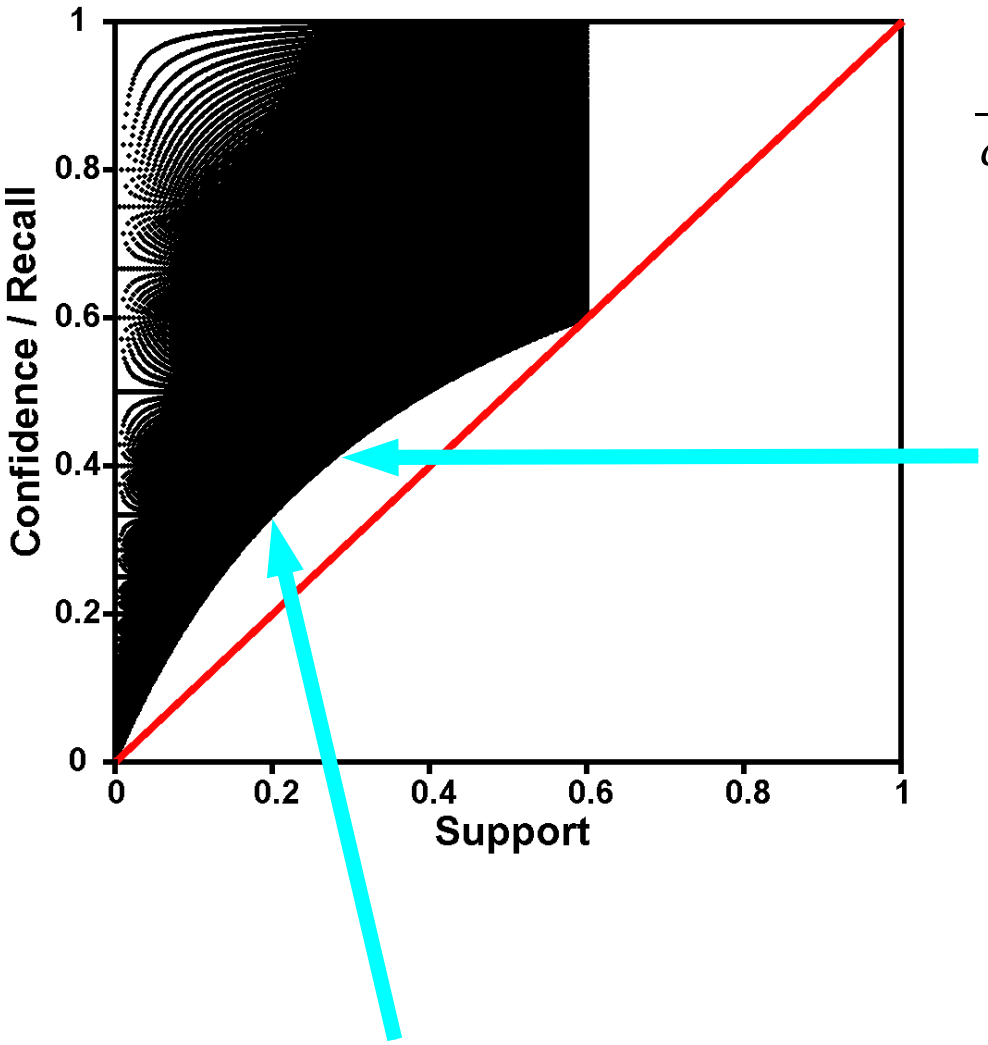
N = 100

acc(R) = 0.3

$$acc(R) = \frac{C0 + C3}{N} = supp(R) + \frac{C3}{N} \quad \Rightarrow \quad acc(R) \geq supp(R)$$

Accuracy im Support-Confidence-Raum

Nattermann



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

Kurve verläuft immer
vom Punkt (0,0)
zum Punkt (acc(R) , acc(R))

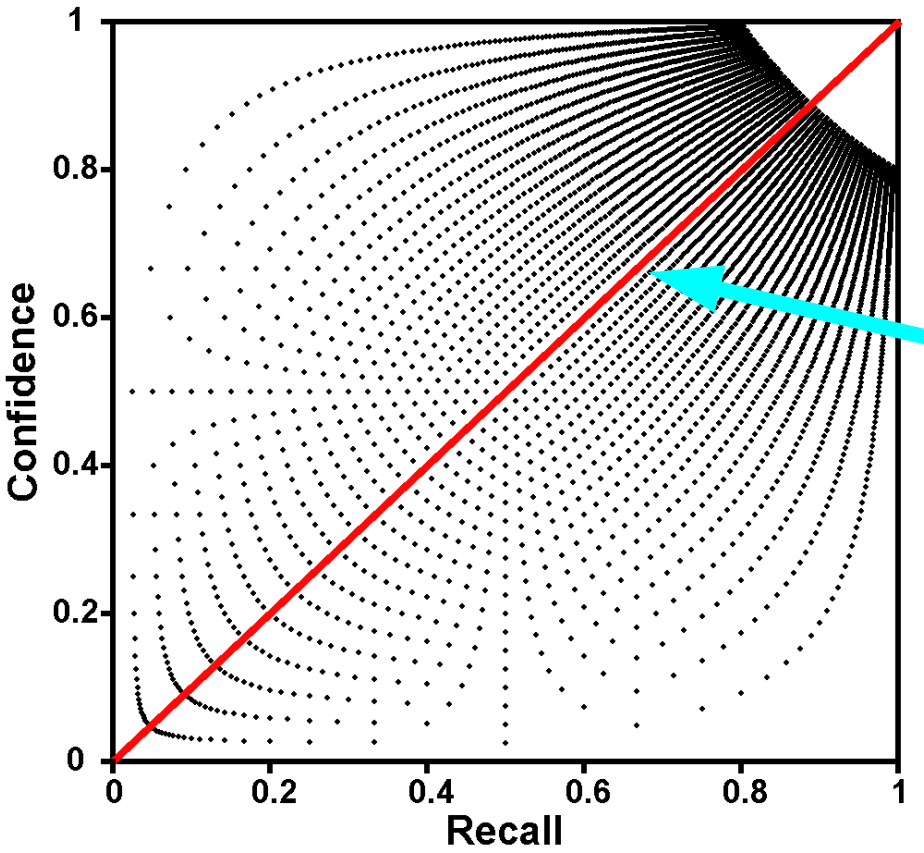
$$0 \leq conf(R) \leq 1$$

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

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

Accuracy im Recall-Confidence-Raum

Nattermann



Auf der Diagonalen gilt:

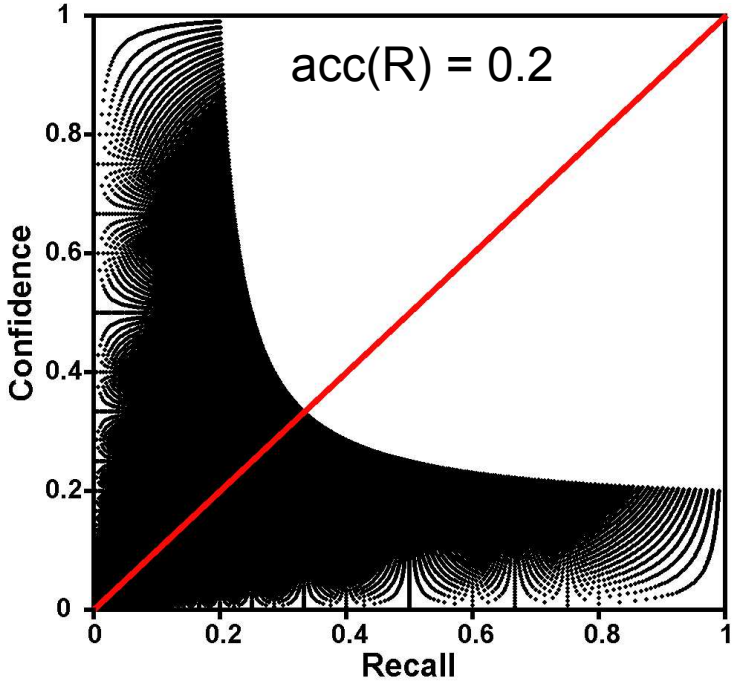
$$conf(R) = recall(R) = \frac{2}{\frac{1-acc(R)}{supp(R)} + 2}$$

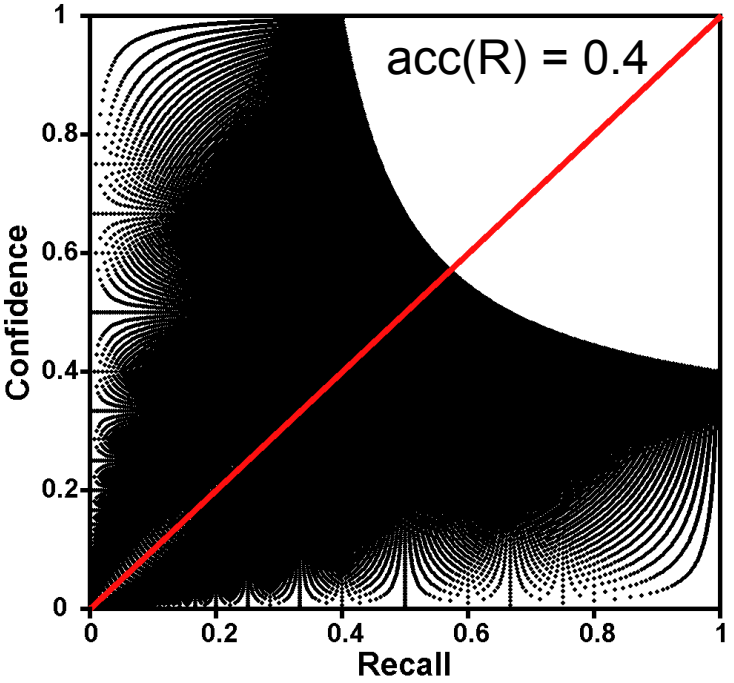
$$max\{supp(R)\} = acc(R)$$

$$max\{conf(R)\} = max\{recall(R)\} = \frac{2}{\frac{1}{acc(R)} + 1}$$

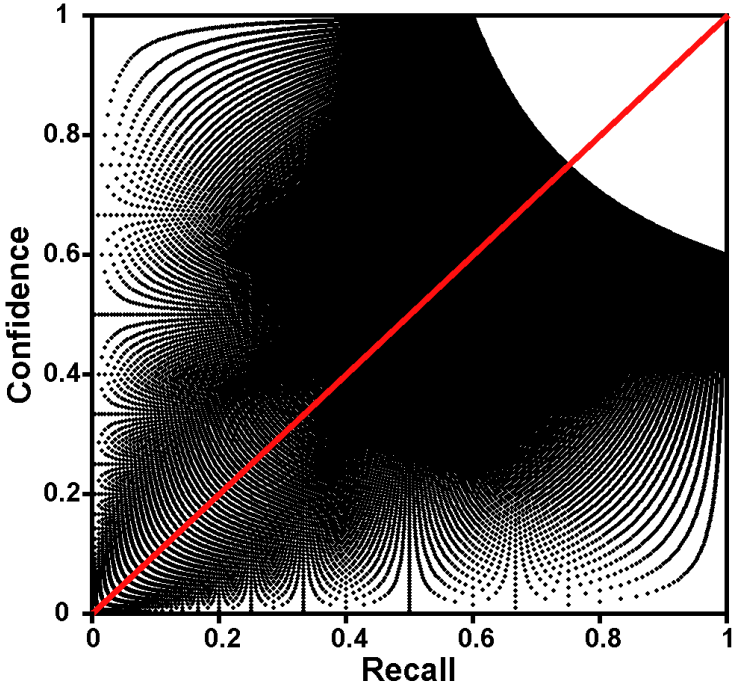
Accuracy im Recall-Confidence-Raum

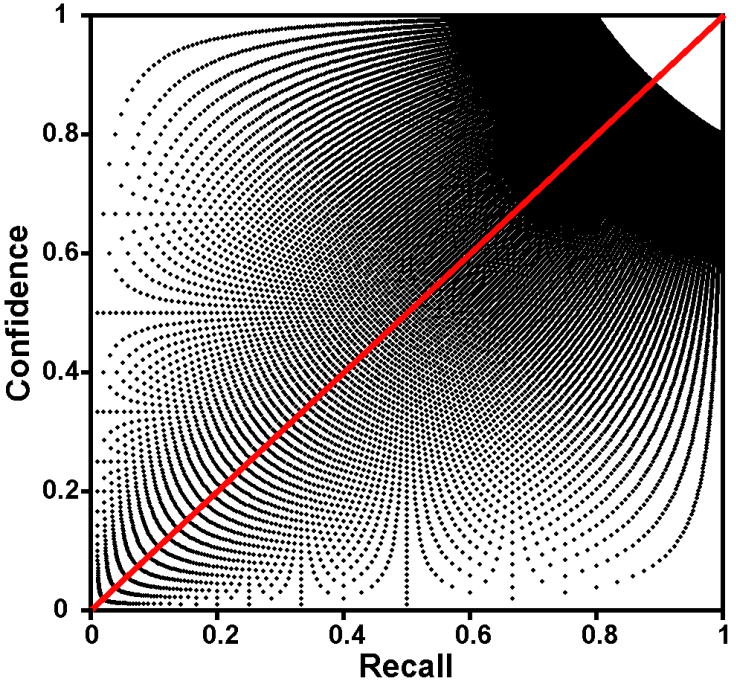
Nattermann





acc(R) = 0.6





acc(R) = 0.8

Accuracy im Recall-Confidence-Raum

Nattermann

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

Wenn der Accuracywert sehr groß ist: $acc(R) \to 1$

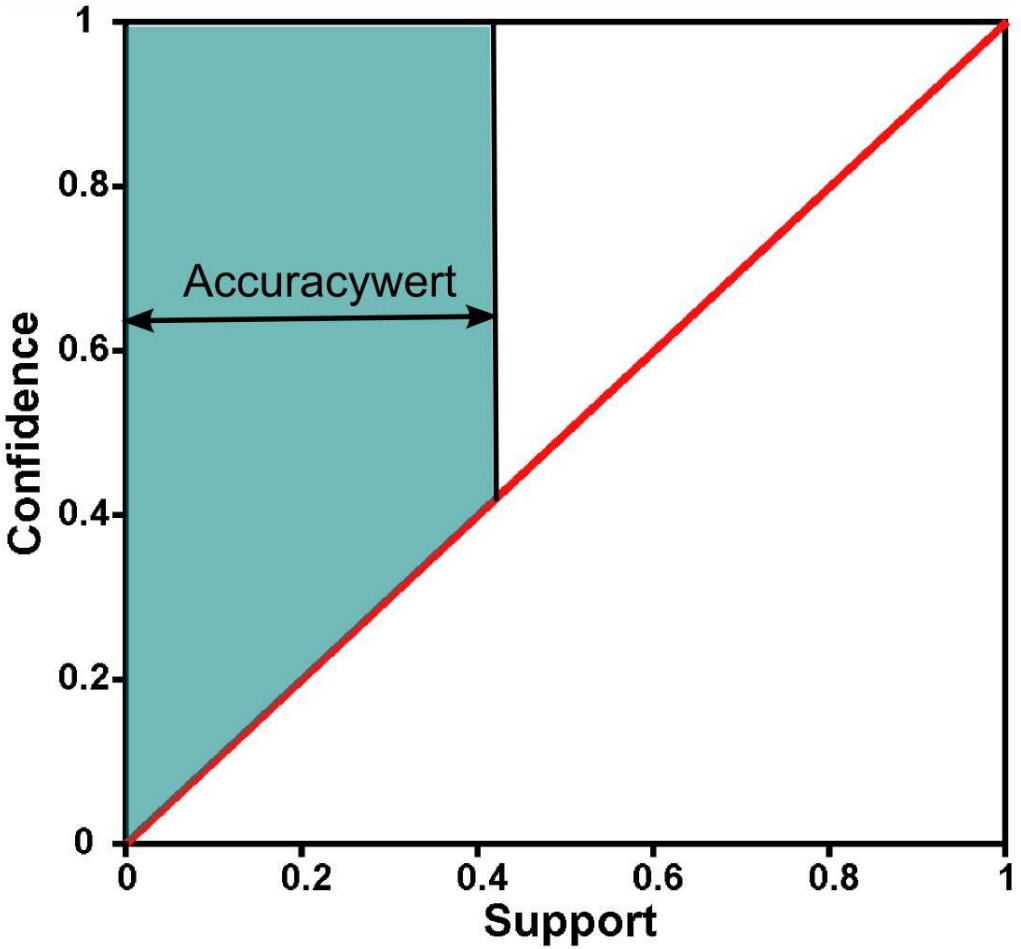
$$\frac{1-acc(R)}{supp(R)} + 2 \quad \to \quad 2$$

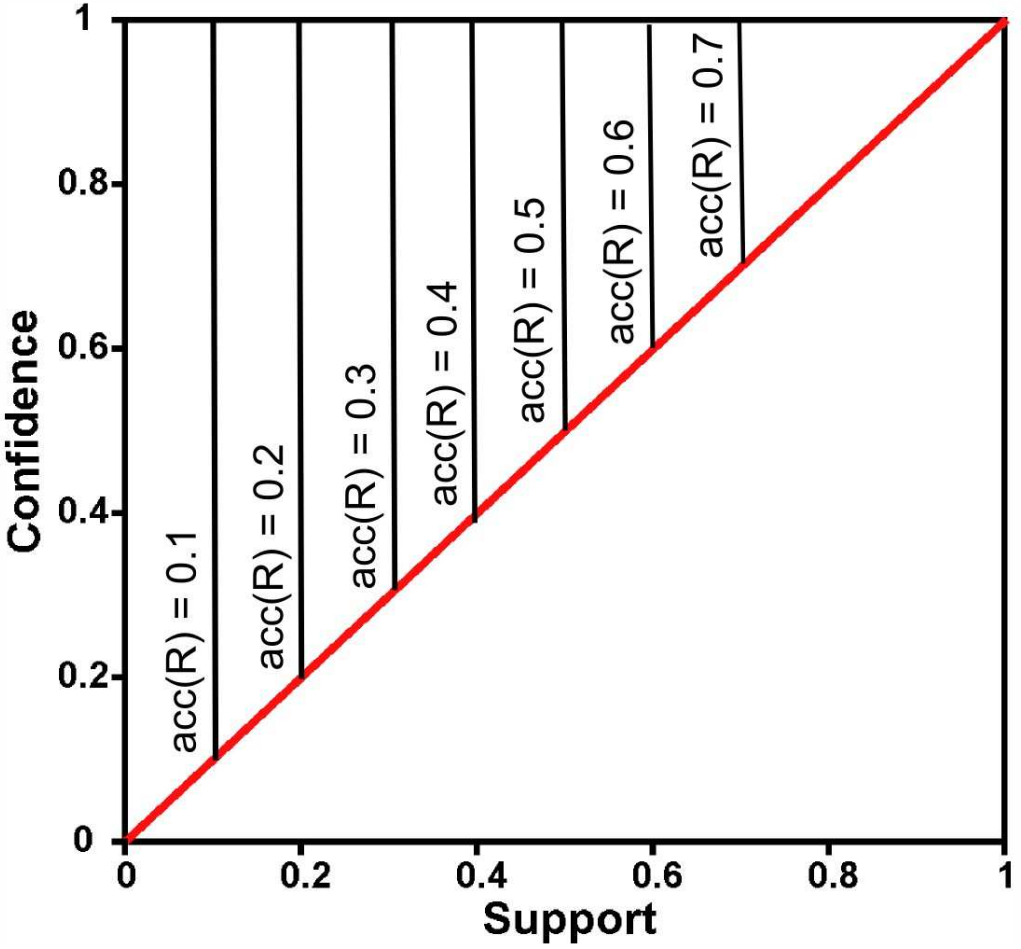
$$conf(R) \quad \to \quad \frac{1}{2 - \frac{1}{recall(R)}}$$

Wenn der Recallwert sich nur minimal verändert,
dann kommt es zu einer großen Veränderung des Confidencewertes

Accuracy - Felder

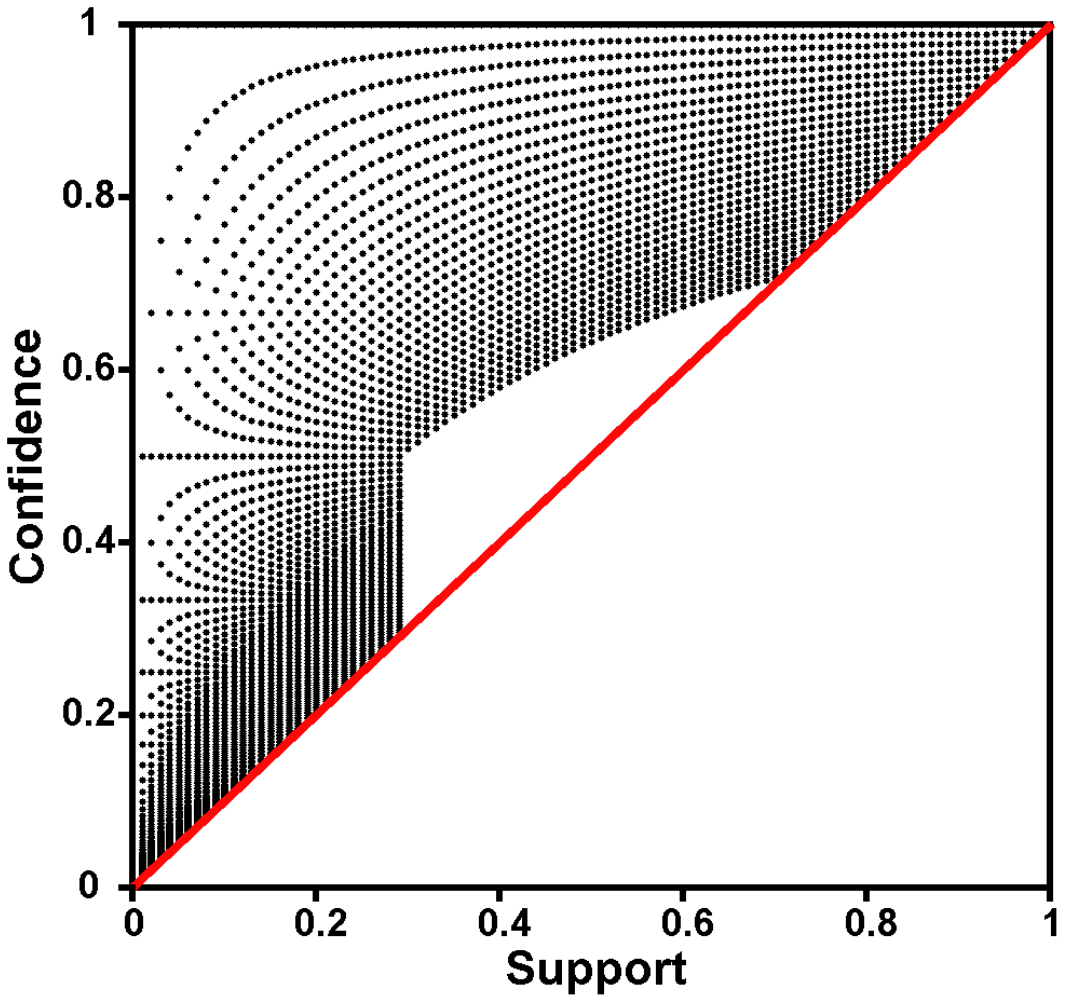
Nattermann





Accuracy - Felder

Nattermann



Minimum Confidencewert:

$$min(conf(R)) = \frac{1}{\frac{1-acc(R)}{supp(R)} + 1}$$

$$supp(R) \leq acc(R) \leq supp(R)\left(1-\frac{1}{conf(R)}\right) + 1$$

Fazit - Accuracy

Die Accuracyheuristik gibt die Wahrscheinlichkeit an, dass andere Kombinationen von gekauften Produkten auch auf den Klassifikationsartikel schließen lassen:

- Die Information der Exklusivität einer Produktkombination wird durch den Confidence- und Recall-Wert bestimmt.
- Da Supportwert und Datensatzgröße bekannt sind, kann man aus C1 und C2 den Wert C3 ermitteln.
- Da nur die Summe von C1 und C2 zur Bestimmung von C3 notwendig ist, ergibt sich die Vertauschungsunabhängigkeit von Confidence- und Recall-Wert.
- Dies ist auch anhand der Achsensymmetrie der Mittellinie in den Abb. vom Recall-Confidence-Raum zu erkennen.

Die Liftheuristik

Lift

- $lift(R) = \frac{supp(R)}{supp(A)\,supp(B)} = \frac{conf(R)}{supp(B)}$

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

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

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

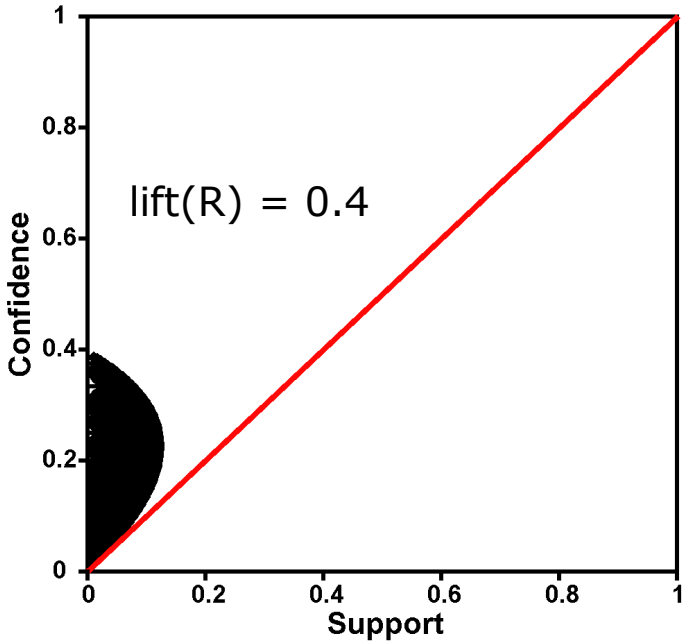
$$supp(R) = \frac{conf(R)\,recall(R)}{lift(R)}$$

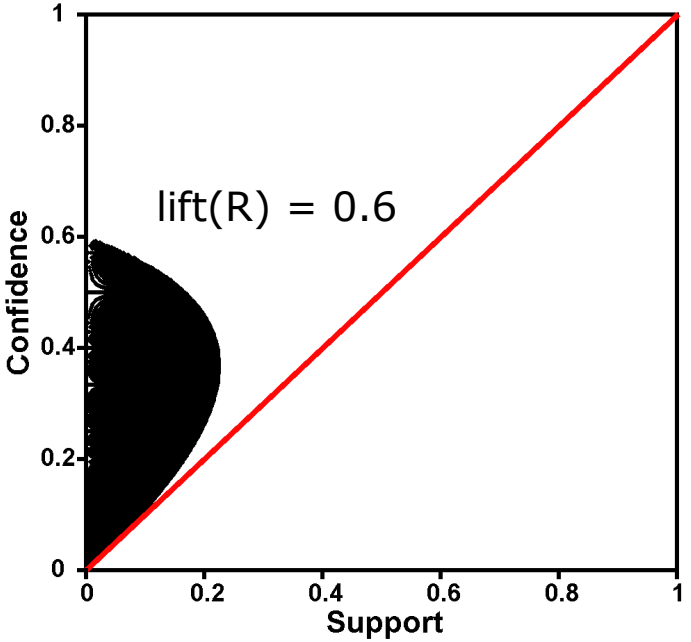
$$conf(R) = \frac{lift(R)\,supp(R)}{recall(R)}$$

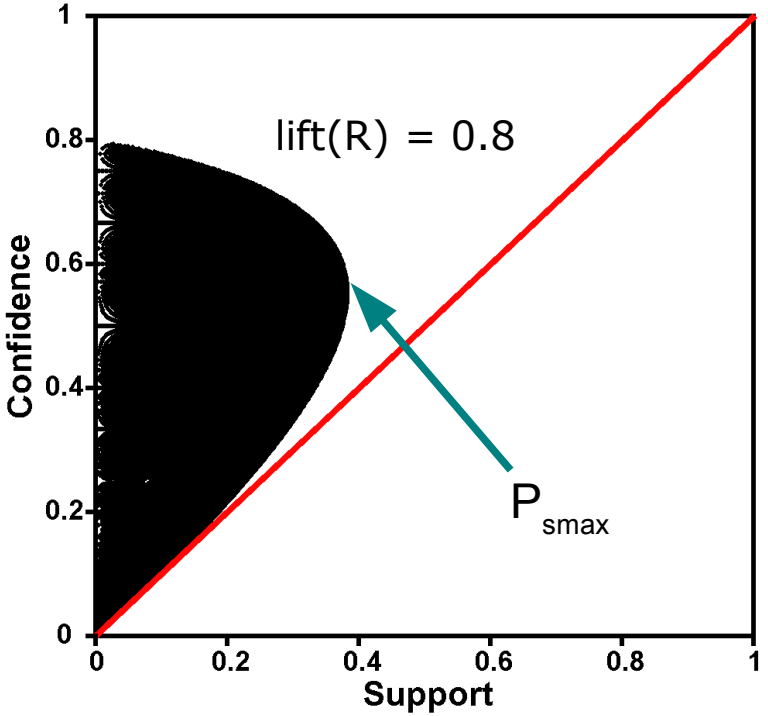
$$recall(R) = \frac{lift(R)\,supp(R)}{conf(R)}$$

Lift im Support-Confidence-Raum

Nattermann



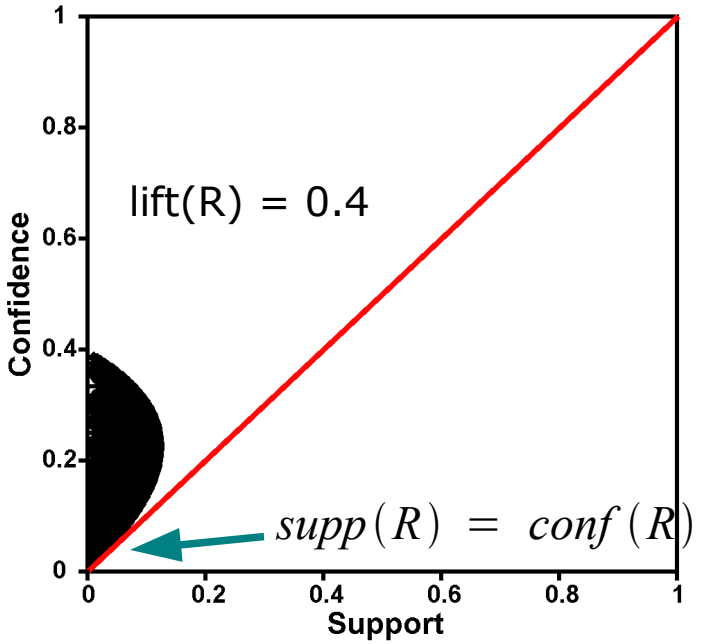


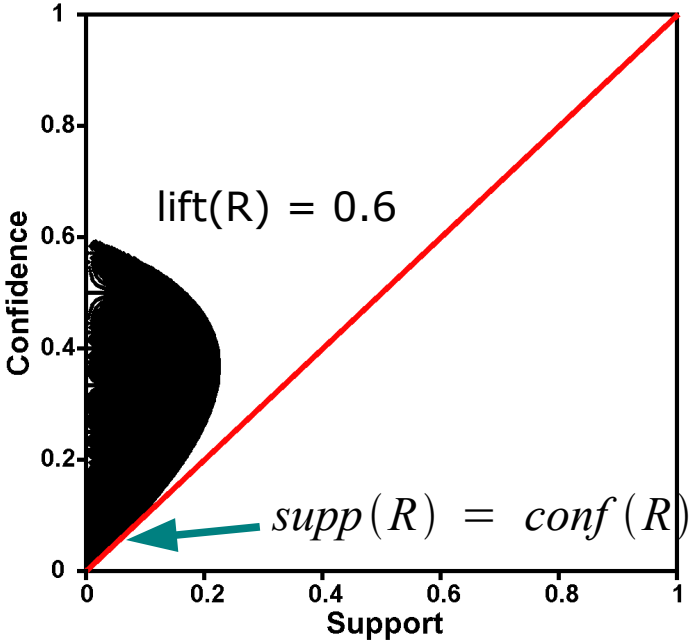


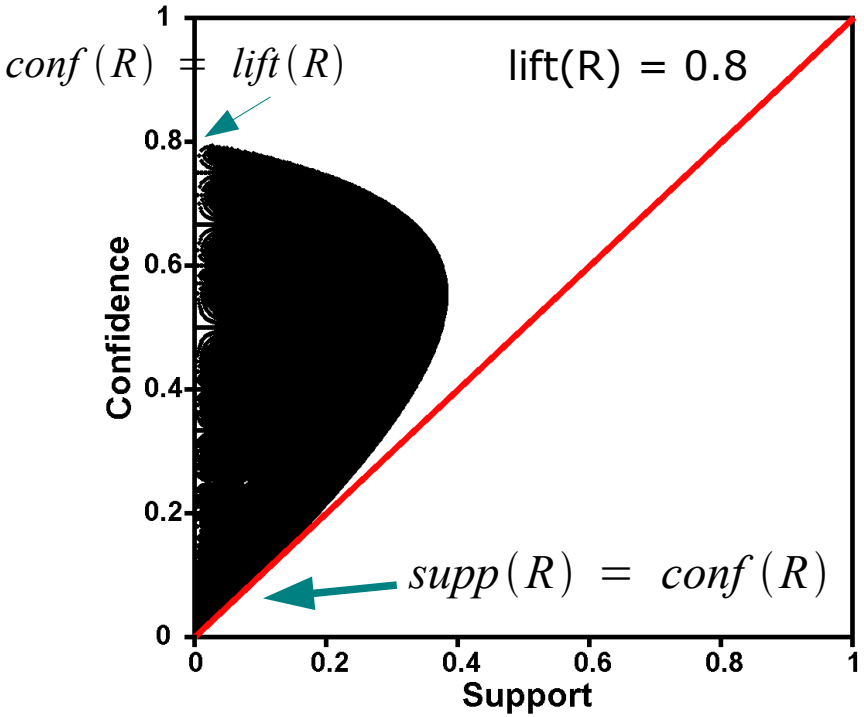
- Die Kurven fangen alle in dem Punkt (0,0) an.
- Mit steigendem Confidencewert steigt auch der Supportwert immer weiter an, bis zu einem Maximalpunkt P_{smax}, welcher sich mit dem Liftwert verändert.
- Nach dem Maximalpunkt fällt der Supportwert, mit steigendem Confidencewert, wieder ab und schneidet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

Lift im Support-Confidence-Raum

Nattermann







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

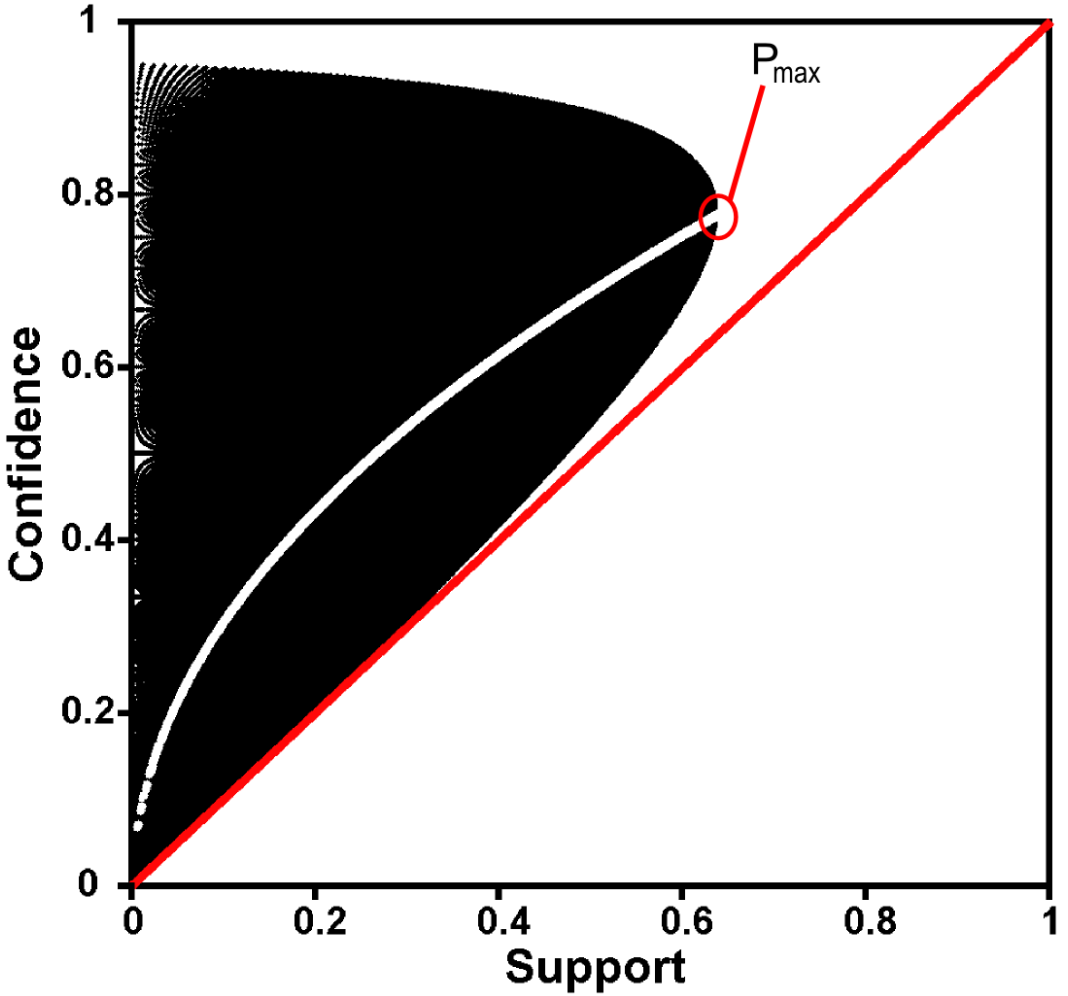
$$supp(R) = \frac{conf(R)\,recall(R)}{lift(R)}$$

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

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

Lift im Support-Confidence-Raum

Nattermann



Am Punkt P_{max}:

$$supp(R) = \frac{conf(R)^2}{lift(R)}$$

$$\frac{d}{d\,supp(R)} conf(R)$$

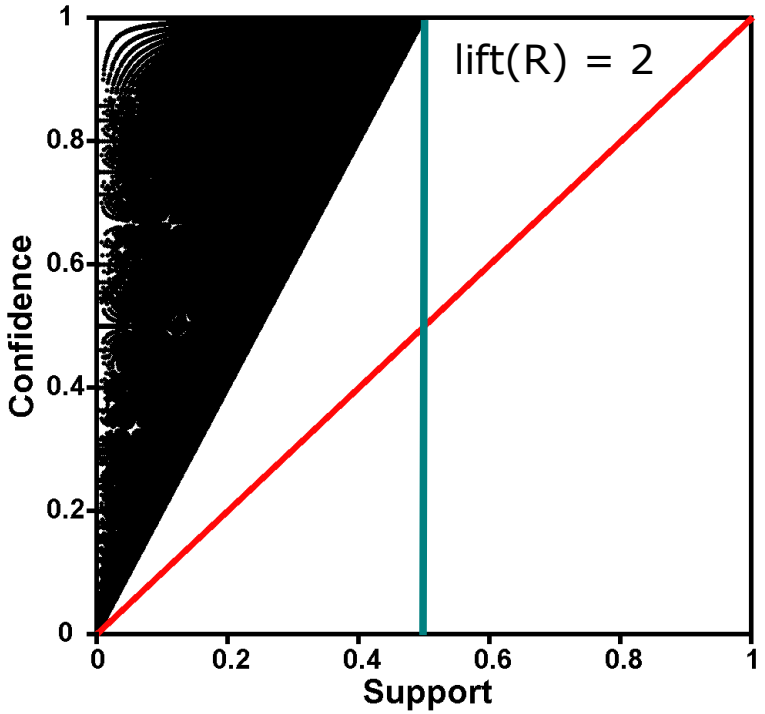
$$= \frac{d}{d\,supp(R)} \sqrt{(supp(R)\,lift(R))}$$

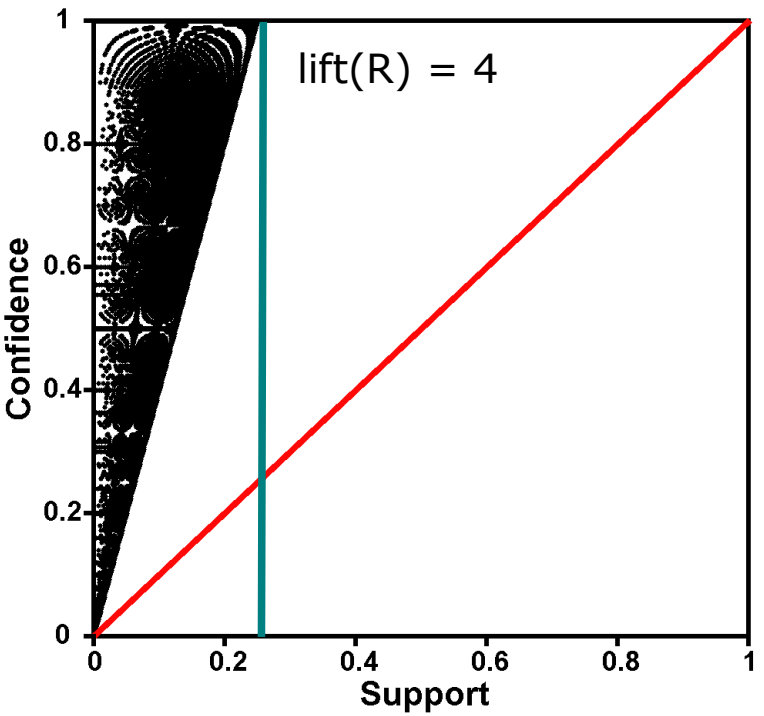
$$= \frac{lift(R)}{2\ \sqrt{(supp(R)\,lift(R))}} < 1$$

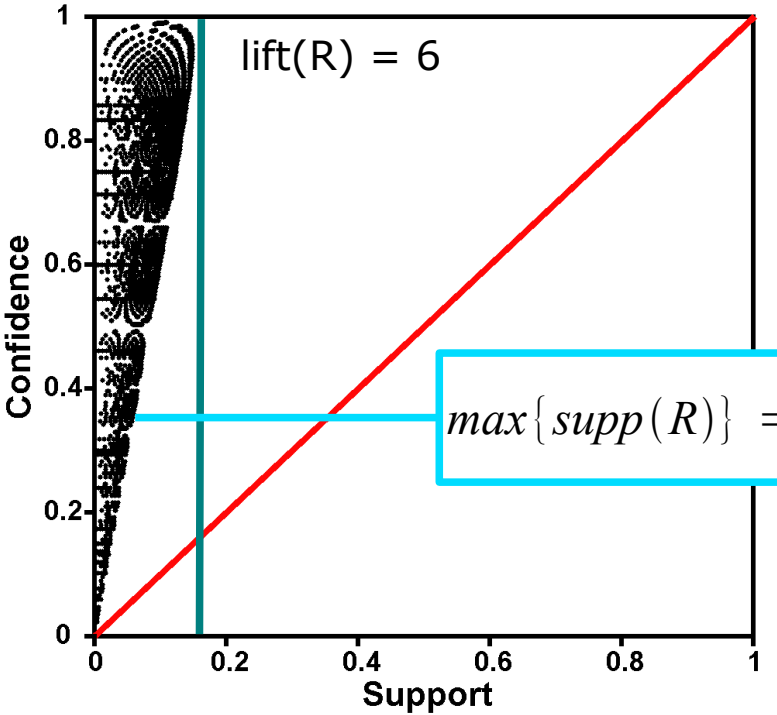
$$lift(R) > supp(R) > \frac{lift(R)}{4}$$

Lift im Support-Confidence-Raum

Nattermann







$$max\{supp(R)\} = \frac{conf(R)\overbrace{max\{recall(R)\}}^{1}}{lift(R)}$$

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

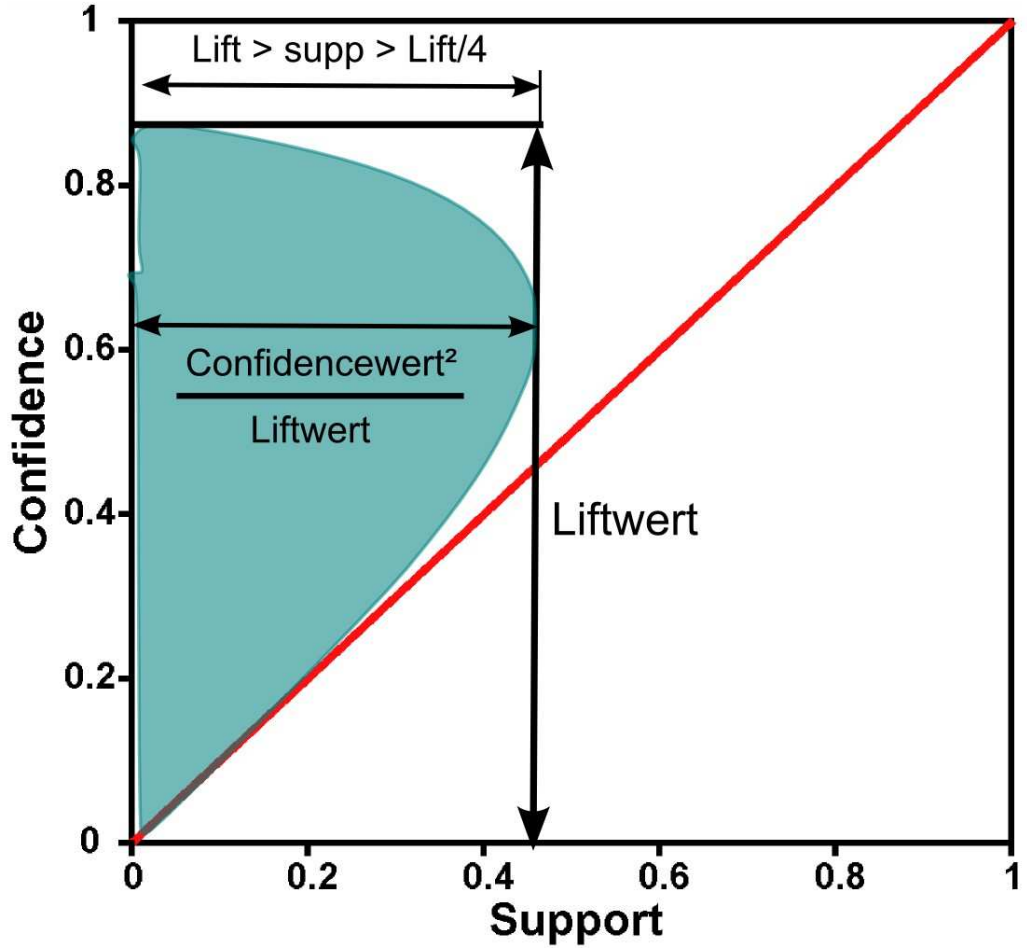
$$supp(R) = \frac{conf(R)recall(R)}{lift(R)}$$

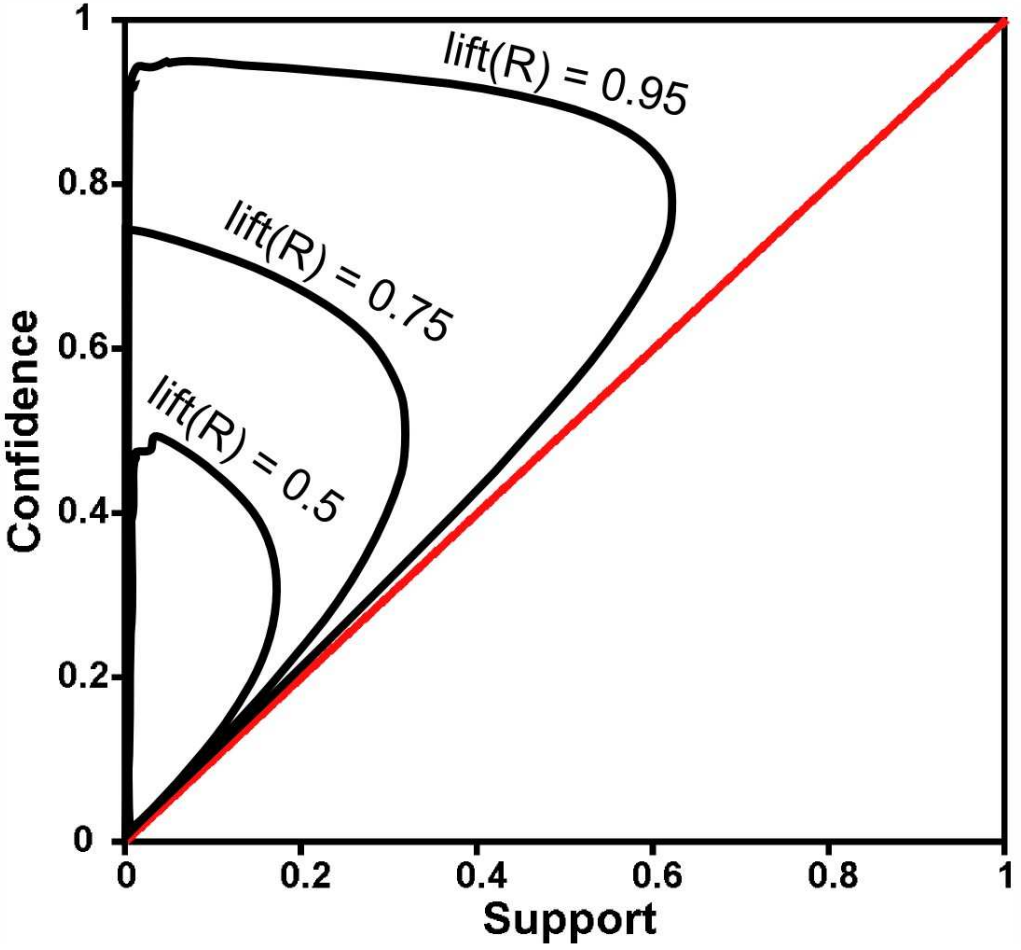
$$conf(R) = \frac{lift(R)supp(R)}{recall(R)}$$

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

Liftwertfelder

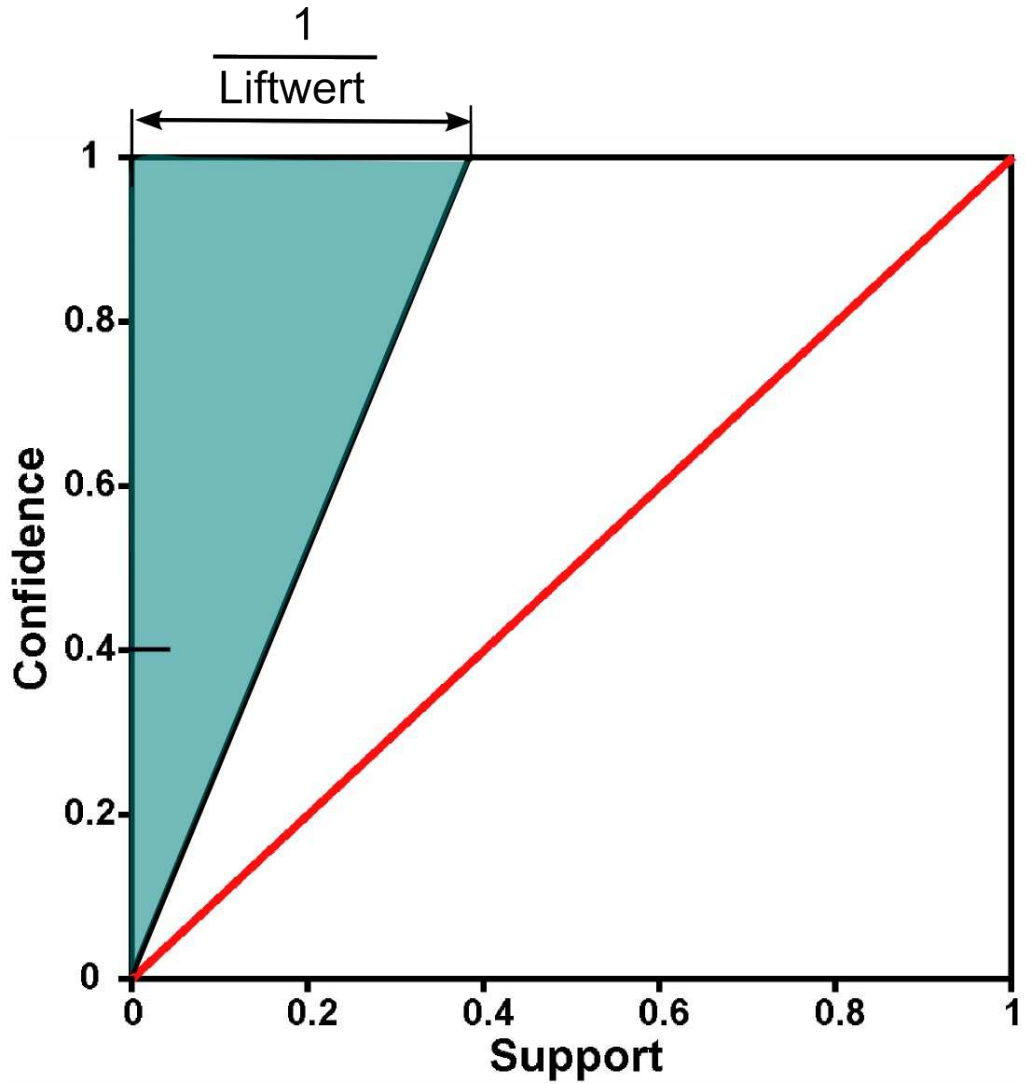
Nattermann

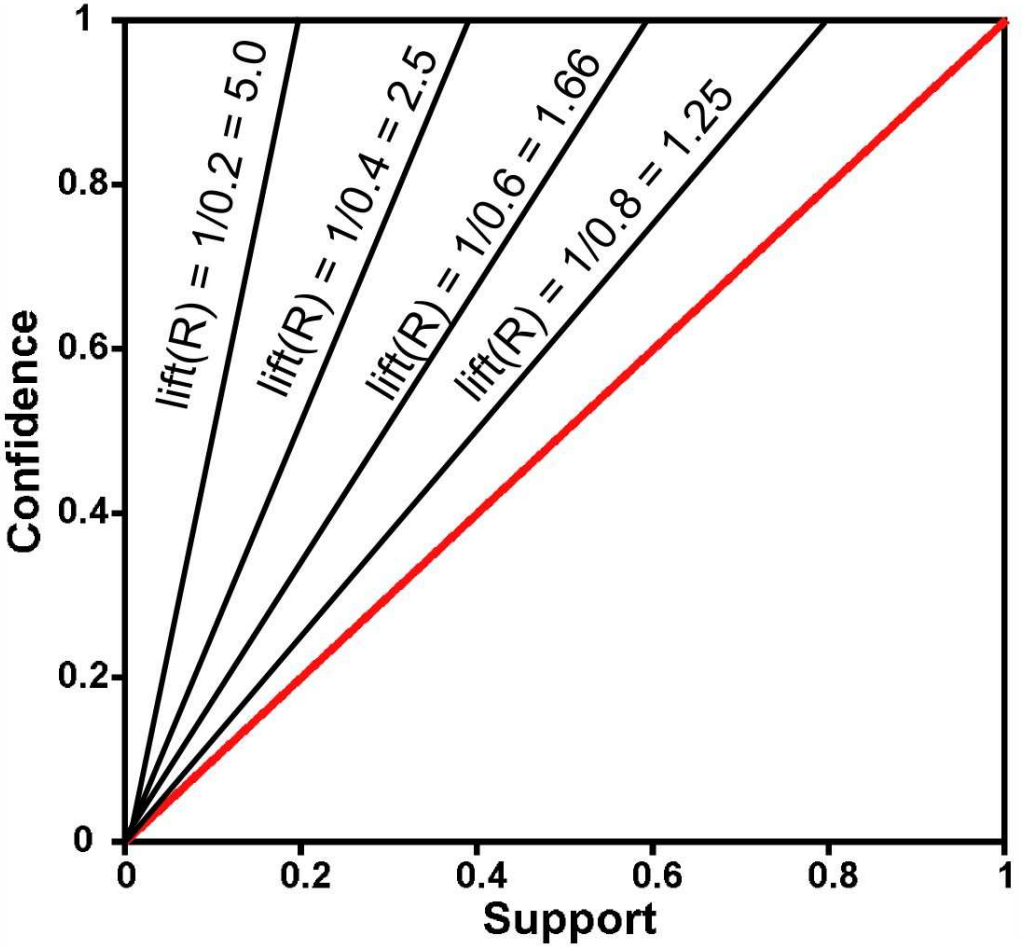




$$1 \quad \geq \quad lift(R) \quad \geq \quad conf(R)$$

Liftwertfelder





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

Lift

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

- Wenn ein Punkt in der Nähe des Punktes (1,1) liegt, dann ist der Liftwert ca. 1.
- Wenn ein Punkt in der Nähe der Diagonalen liegt, dann ist der Liftwert ebenfalls ca. 1.
  - Es sein denn der Punkt liegt in der Nähe des Punktes (0,0), dann ist jeder Liftwert möglich.
- Wenn der Punkt in der Nähe des Punktes (0,1) liegt, dann ist der Liftwert entweder ca. 1 oder sehr groß.

Die Leverageheuristik

Leverage

Nattermann

- $leverage(R) = supp(R) - supp(A)supp(B)$

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

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

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

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

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

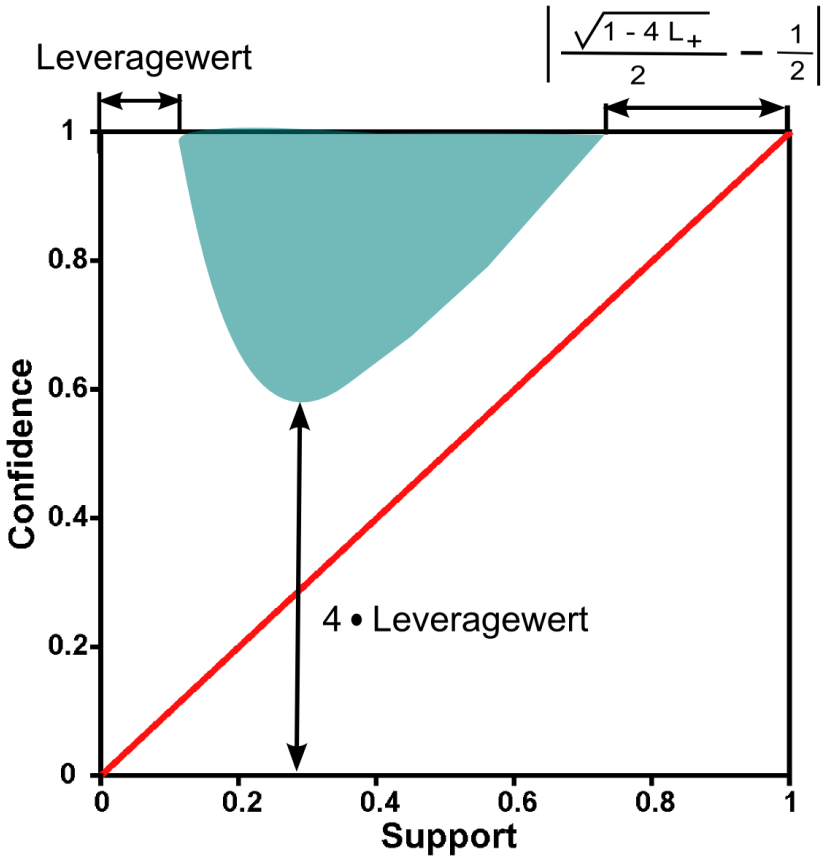
$$recall(R)conf(R) = \frac{supp(R)^2}{supp(R) - leverage(R)}$$

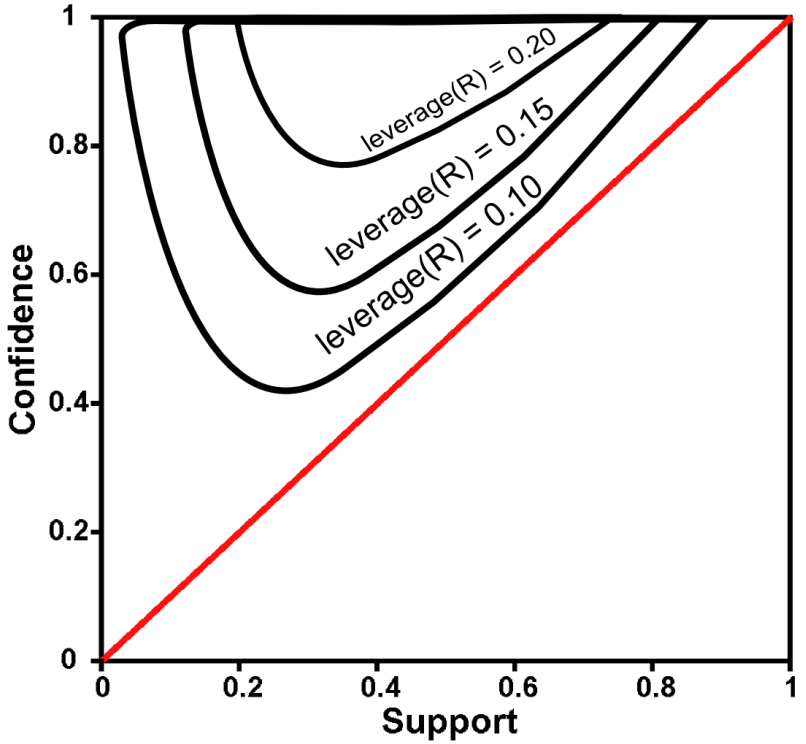
$$L_+ = |leverage(R)|$$

$$L_- = -|leverage(R)|$$

Positiver Leveragewert

Nattermann





$$supp(R) = \frac{conf(R)\,recall(R)}{2}\left(1 \pm \sqrt{1-\frac{4L_+}{conf(R)\,recall(R)}}\right) > 0$$

$$4L_+ \leq conf(R)\underbrace{recall(R)}_{1}$$

Positiver Leveragewert

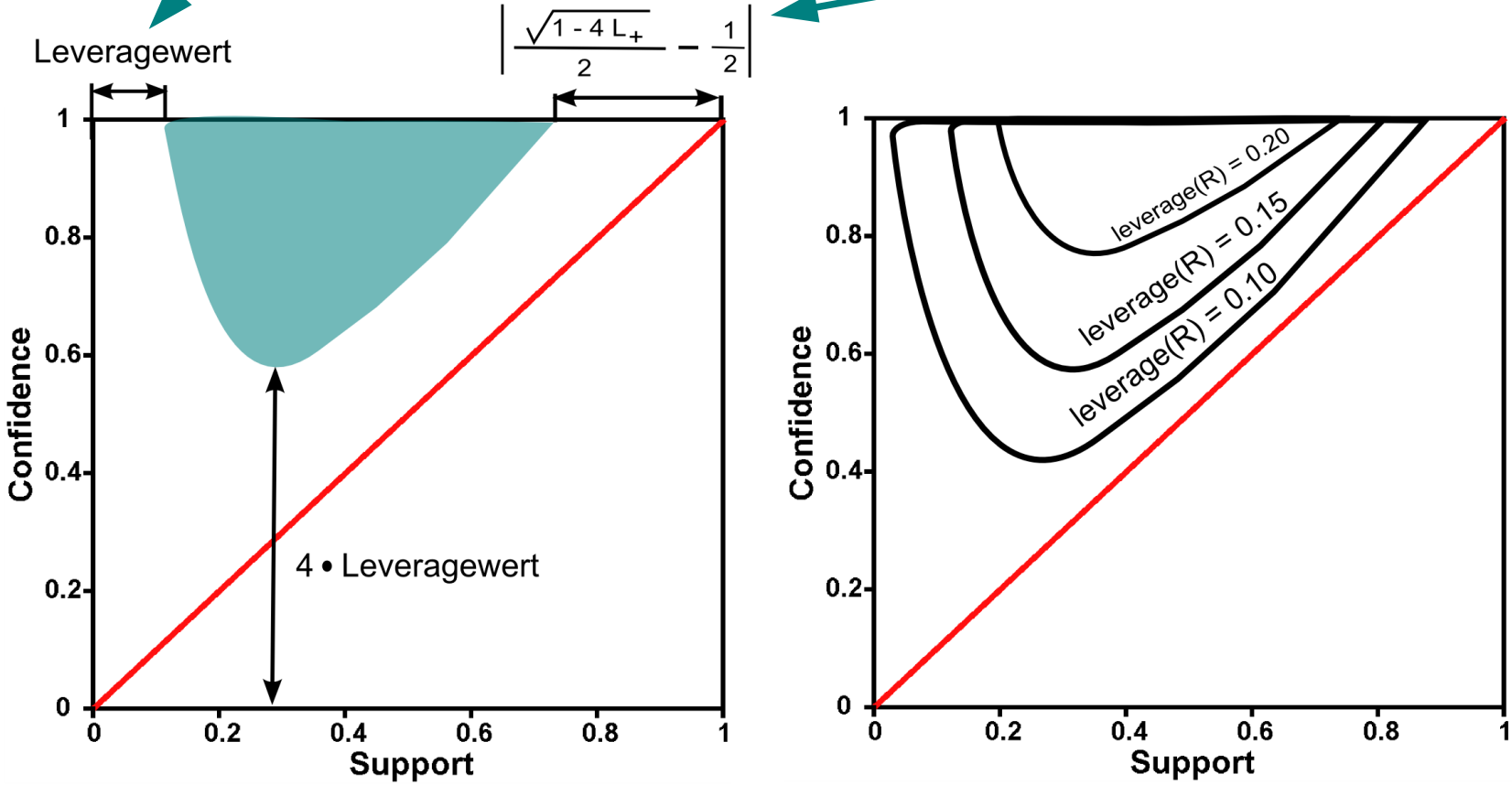
Nattermann

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

$$max\{supp(R)\} \rightarrow max\{conf(R)recall(R)\}$$

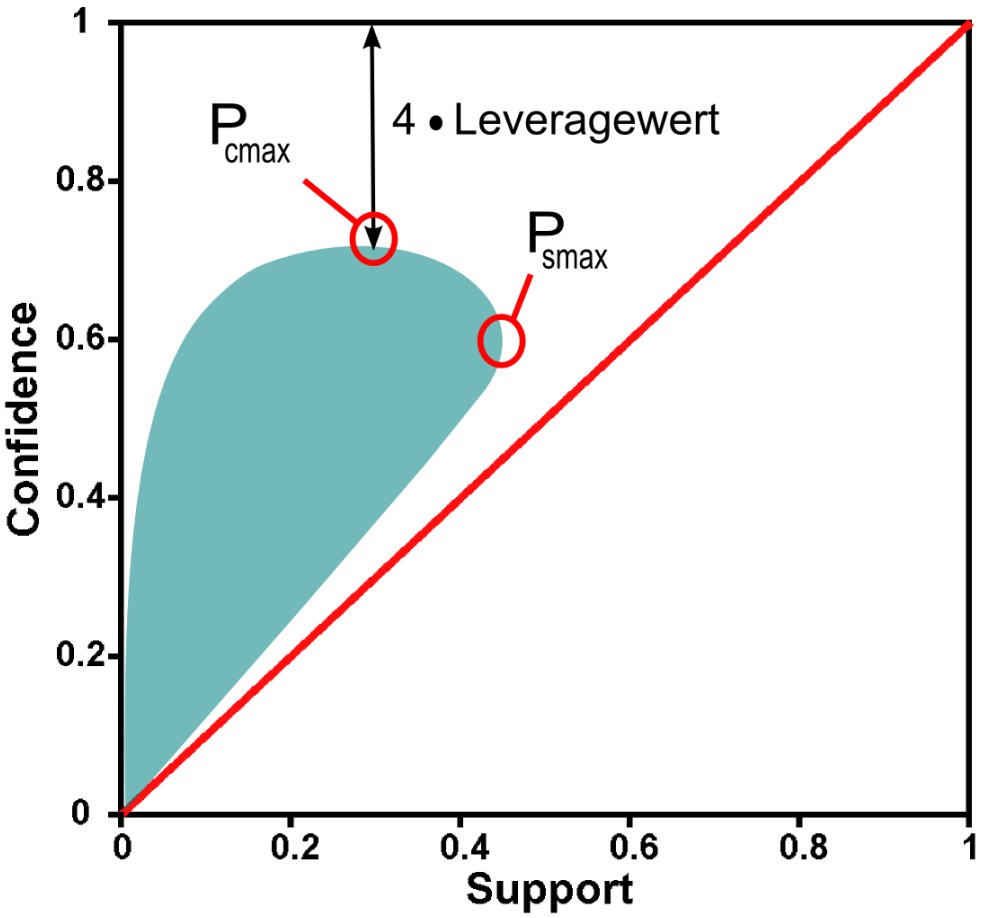
$$supp(R) = \frac{conf(R)recall(R)}{2}\left(1 \pm \sqrt{\left(1 - \frac{4\,leverage(R)}{conf(R)recall(R)}\right)}\right)$$

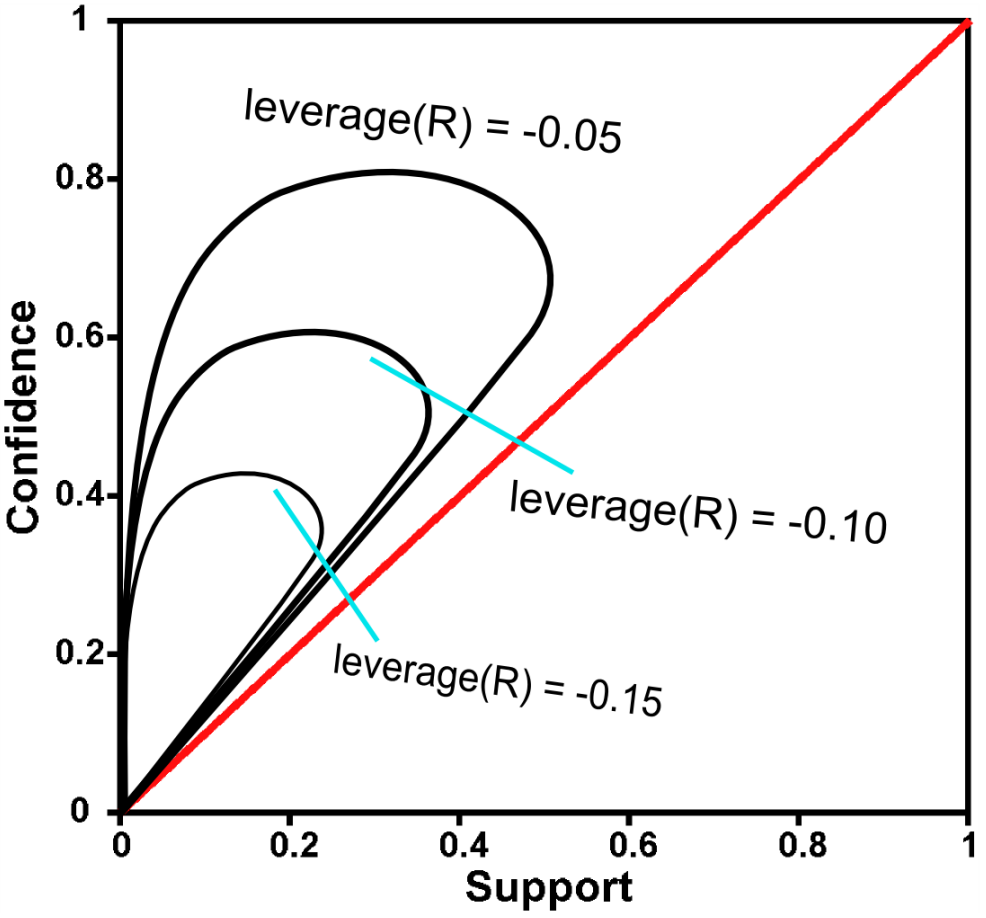
$$max\{supp(R)\} \rightarrow \frac{1}{2}\left(1 + \sqrt{1 - 4L_+}\right)$$



Negativer Leveragewert

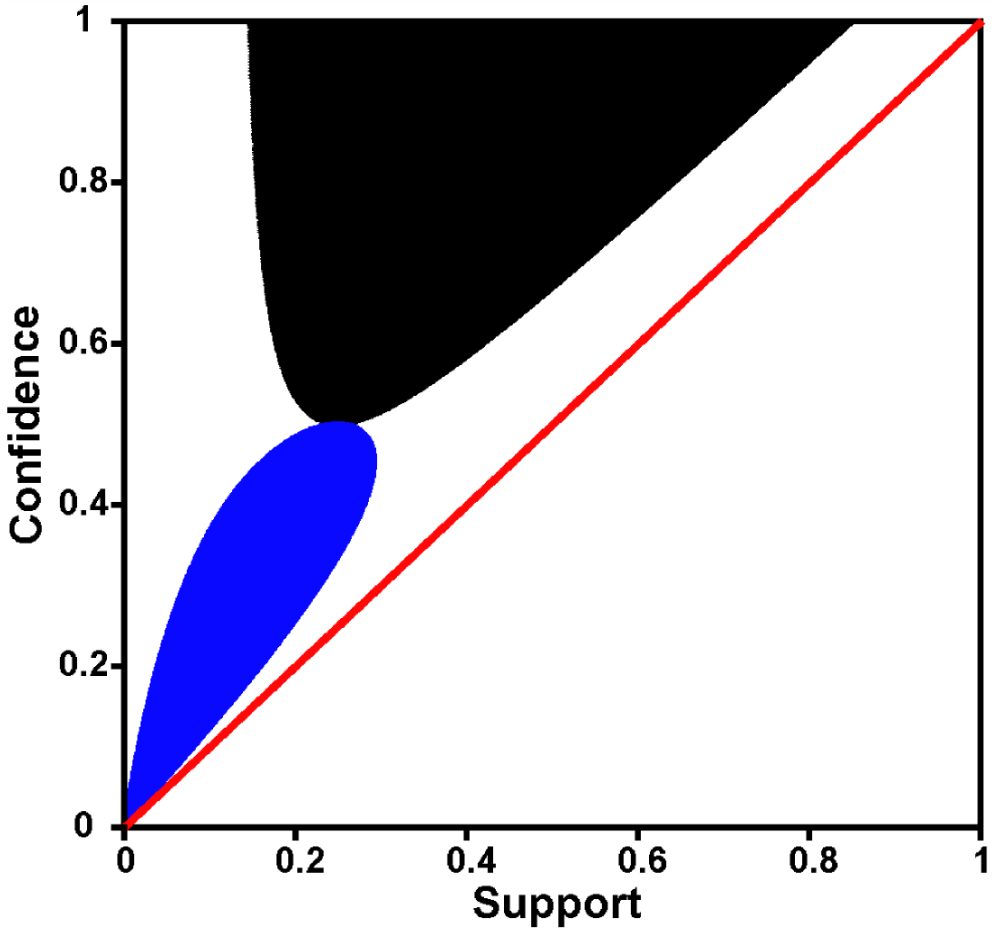
Nattermann

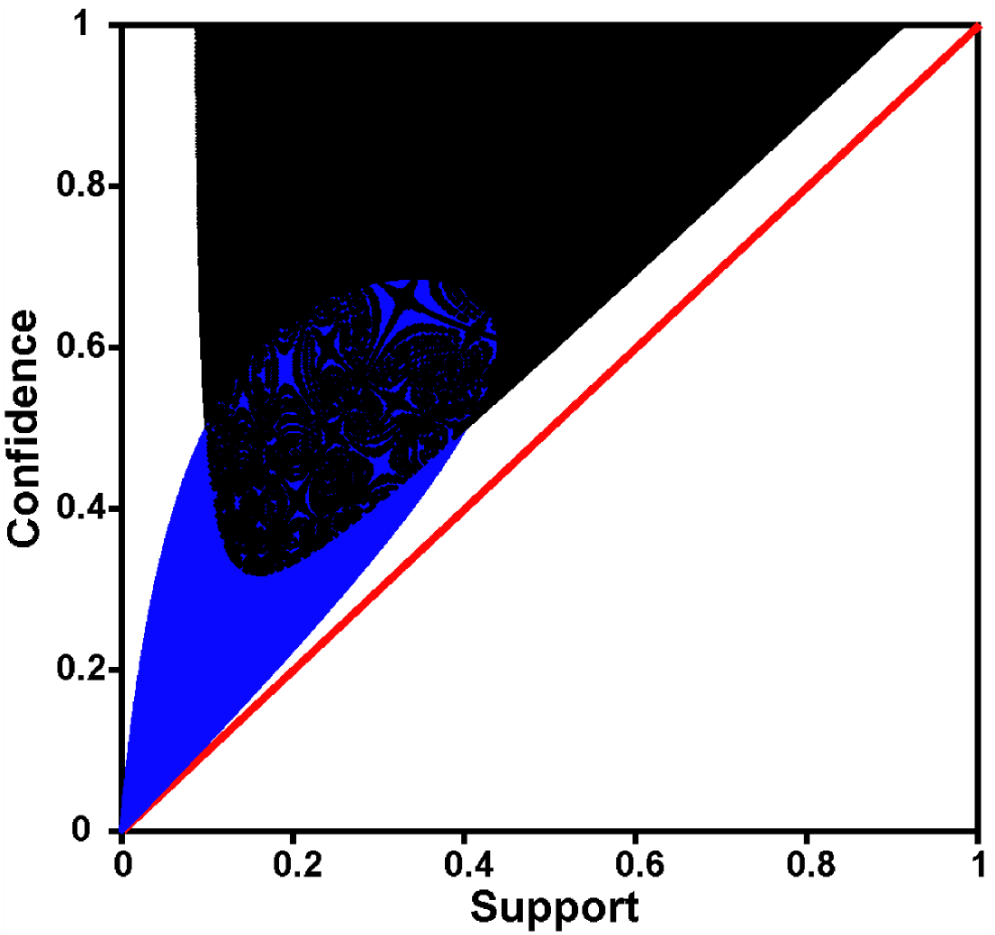




Negativer Leveragewert

Nattermann





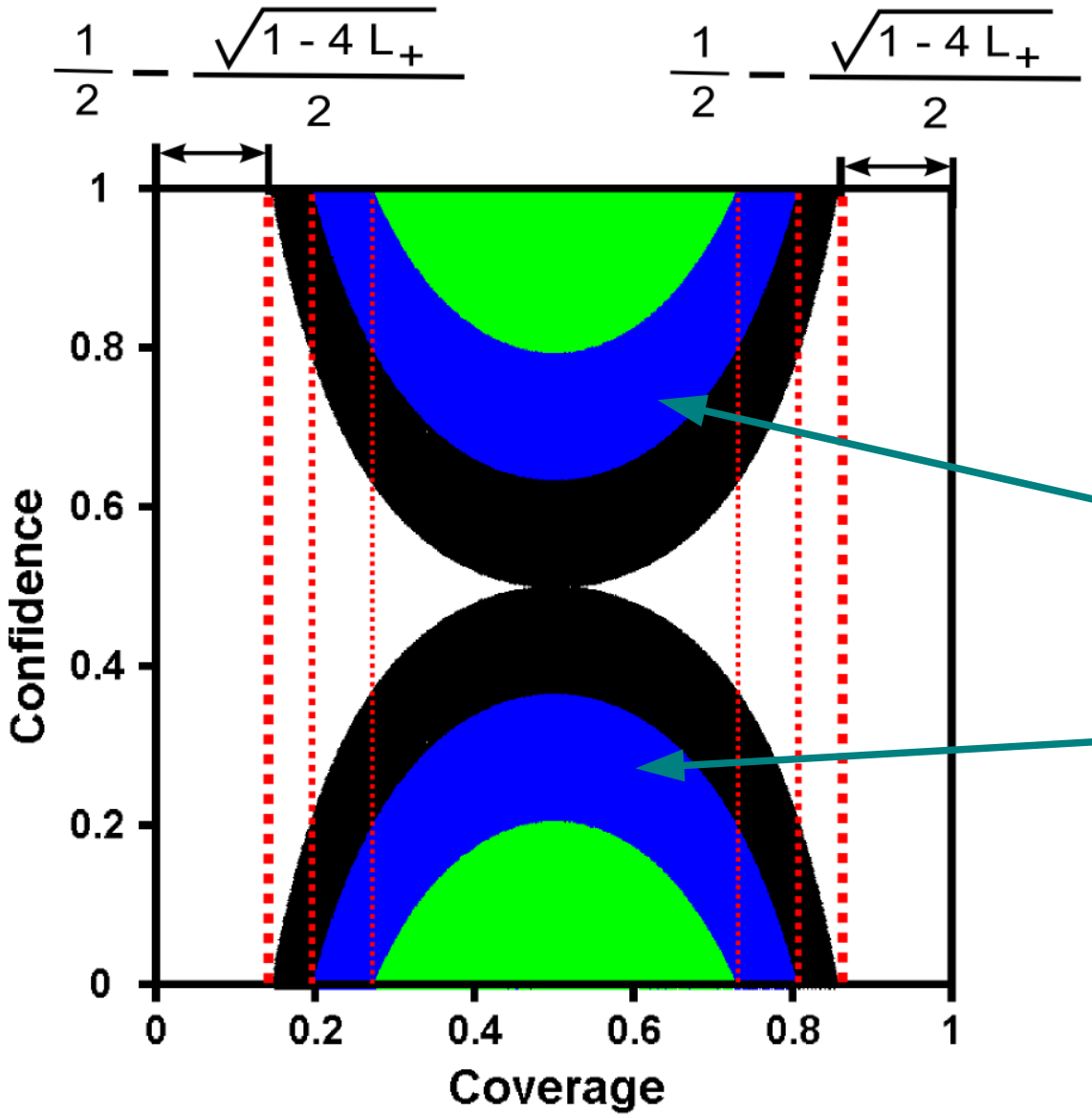
Leverage

$$leverage(R) = \left[\frac{conf(R)-1}{4} , \frac{conf(R)}{4}\right]$$

- Dieses Intervall hat eine konstante Breite von 0.25
- Der Punkt P_{smax} liefert nur Informationen über die Punktwolke falls man weiß, dass man sich auf der Grenzlinie befindet.
- Das Intervall lässt sich durch die Analyse des Coverage-Confidence-Raums noch einschränken.

Leverage im Coverage-Confidence-Raum

Nattermann



- Grün = +/- 0.20
- Blau = +/- 0.16
- Schwarz = +/- 0.125

Positive Leveragewerte

Negative Leveragewerte

Die Phi-Koeffizienten-Heuristik

Phi-Koeffizient

Nattermann

Der Phi-Koeffizient:

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

	B	$\bar{B}$
A	C0	C1
$\bar{A}$	C2	C3

P.-N. Tan, V. Kumar and J. Srivastava: Selecting the right interestingness measure for association patterns.

Phi-Koeffizient - Chi-Quadrat-Unabhängigkeitstest:

- $$phi(R) = \frac{P(A,B)\ P(\bar{A},\bar{B}) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$$

Der allgemeine Phi-Koeffizient beachtet nicht nur die Wahrscheinlichkeit der richtigen Klassifikation sondern auch die Vermeidung der Falschen!

Phi-Koeffizient

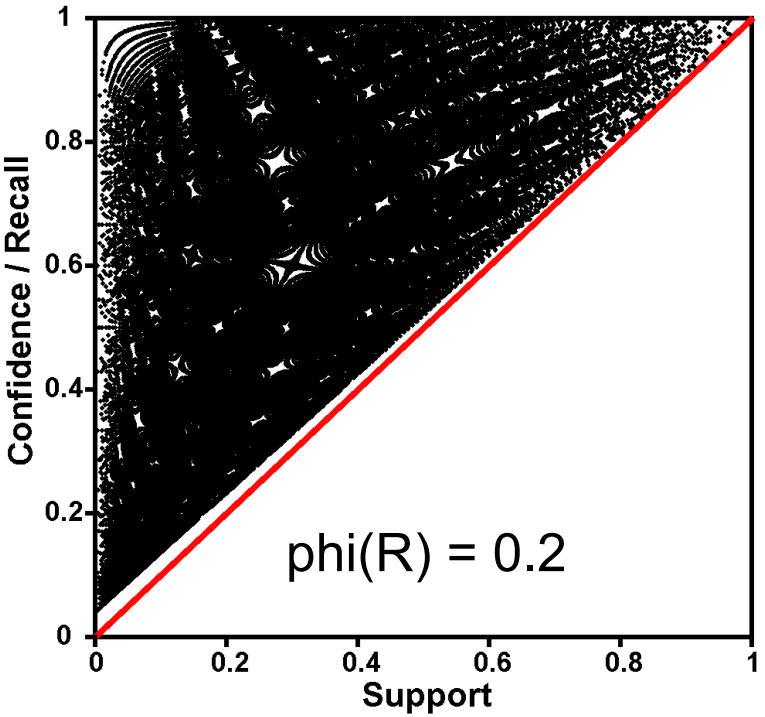
- $$phi(R) = \frac{leverage(R)}{\sqrt{\frac{1}{lift(R)^2} \cdot (conf(R) - supp(R)) \ (recall(R) - supp(R))}}$$

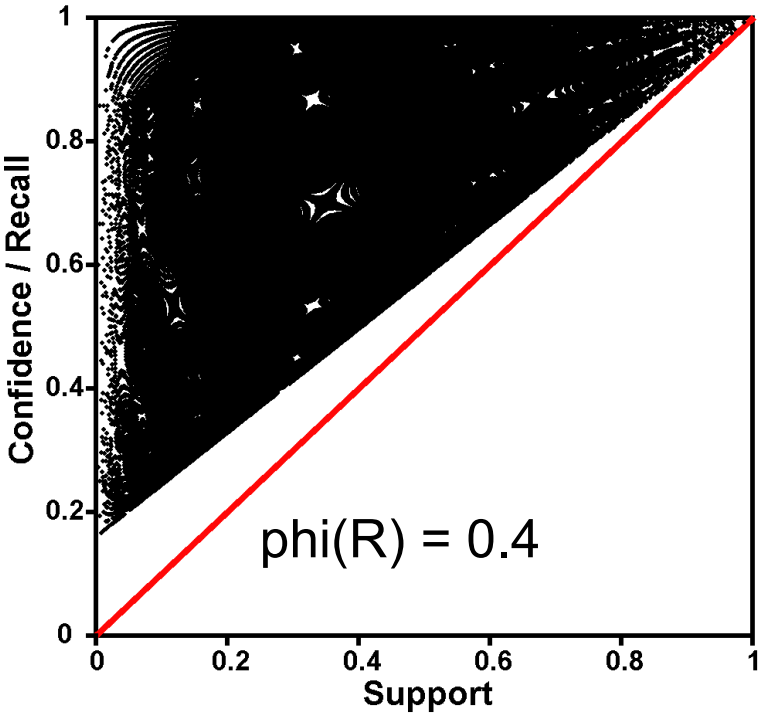
- $$phi(R) = \frac{leverage(R) \ lift(R)}{\sqrt{(conf(R) - supp(R)) \ (recall(R) - supp(R))}}$$

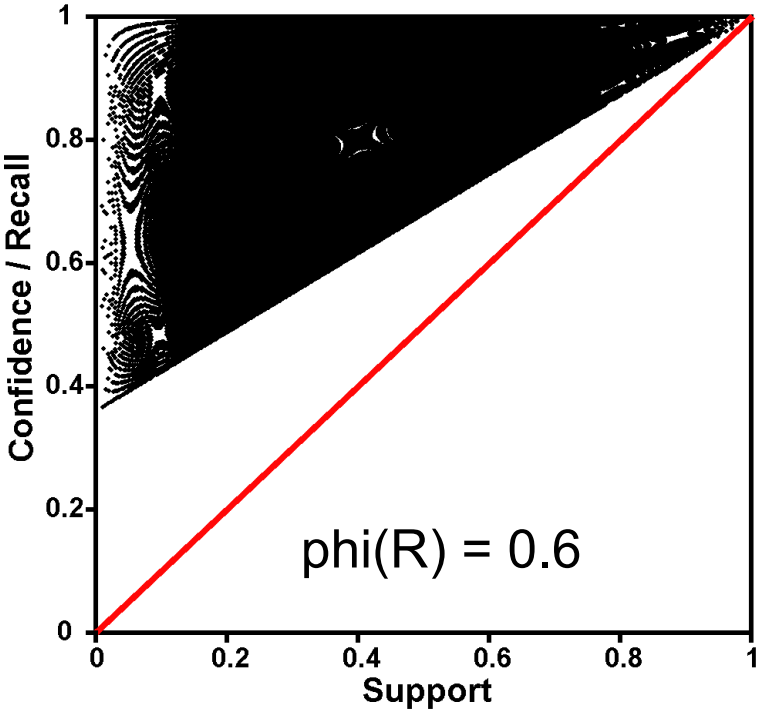
- $$phi(R) = \frac{recall(R) \ conf(R) - supp(R)}{\sqrt{(conf(R) - supp(R)) \ (recall(R) - supp(R))}}$$

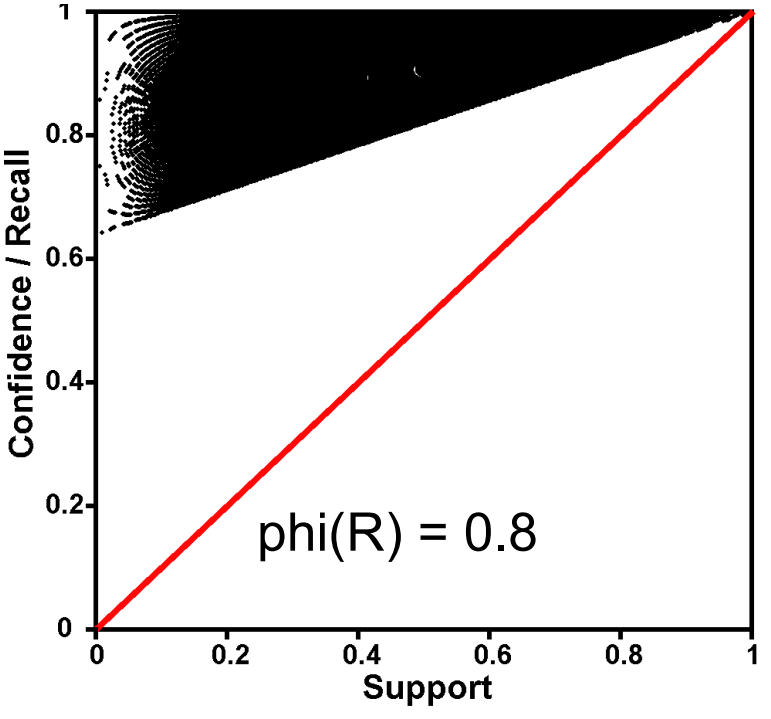
Positiver Phi-Koeffizientenwert

Nattermann



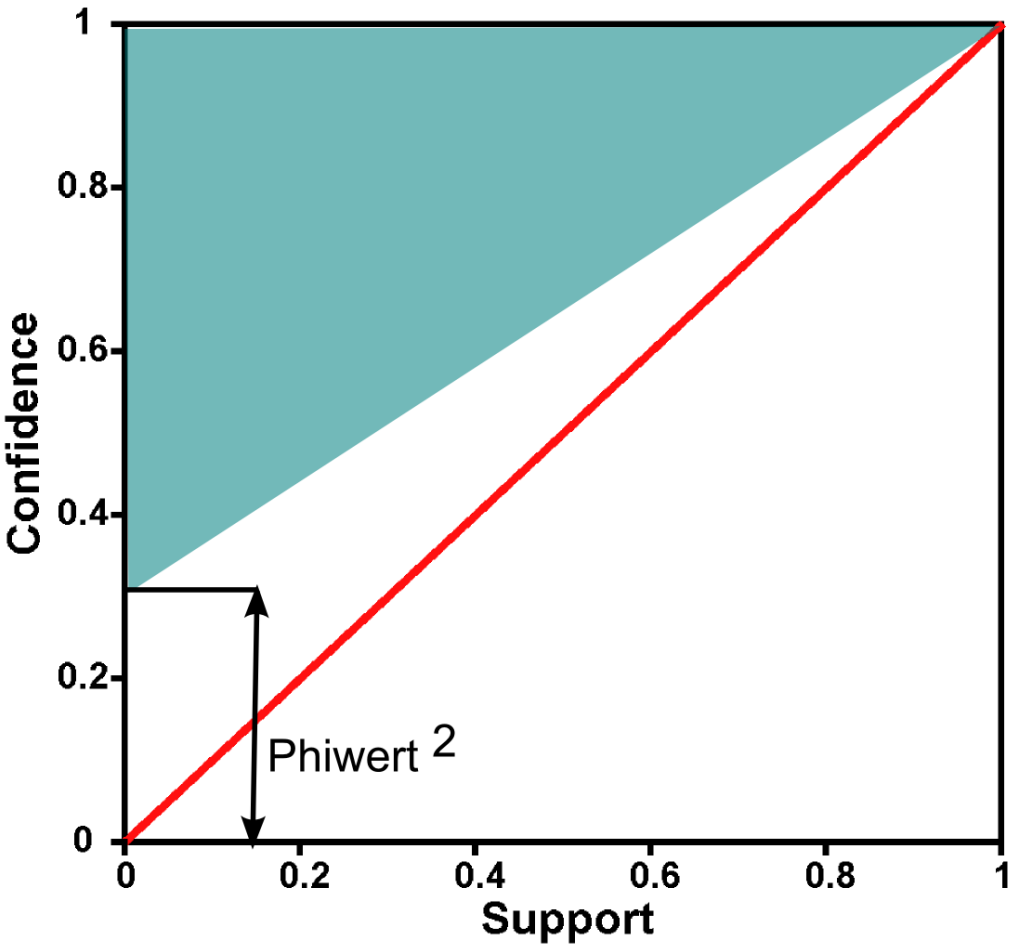


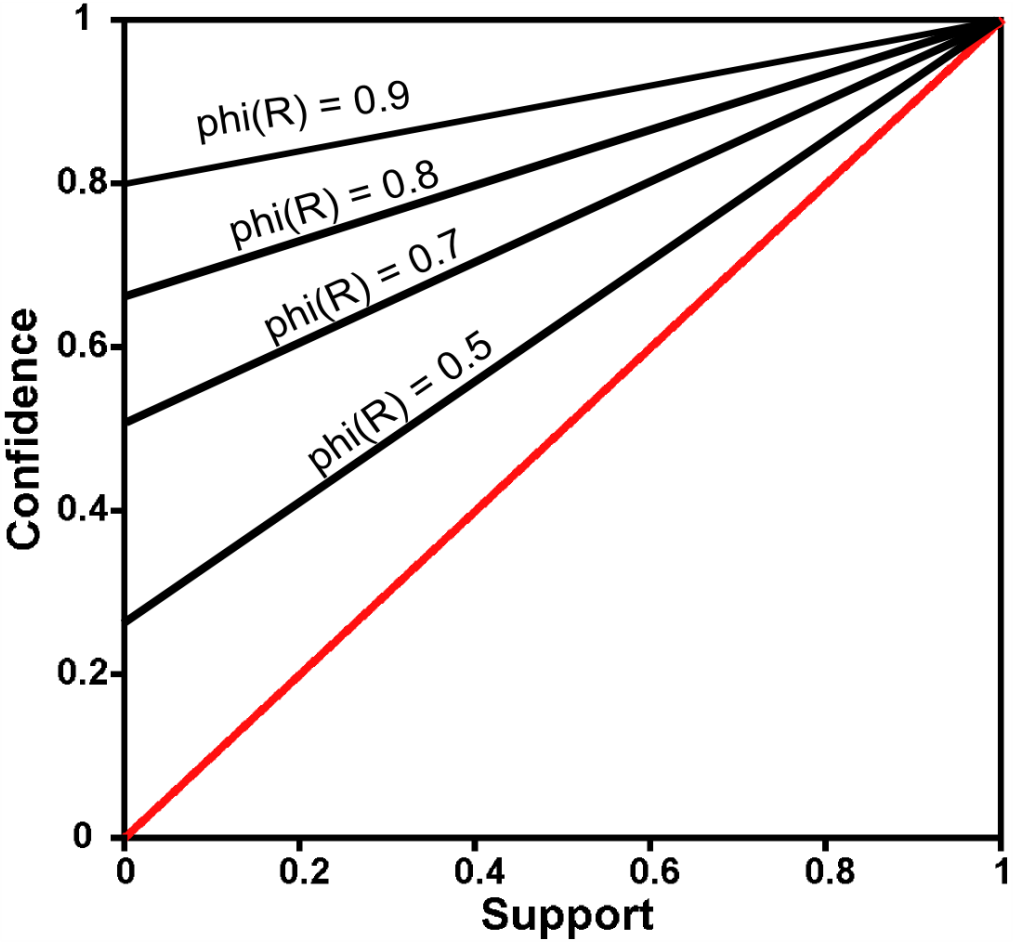




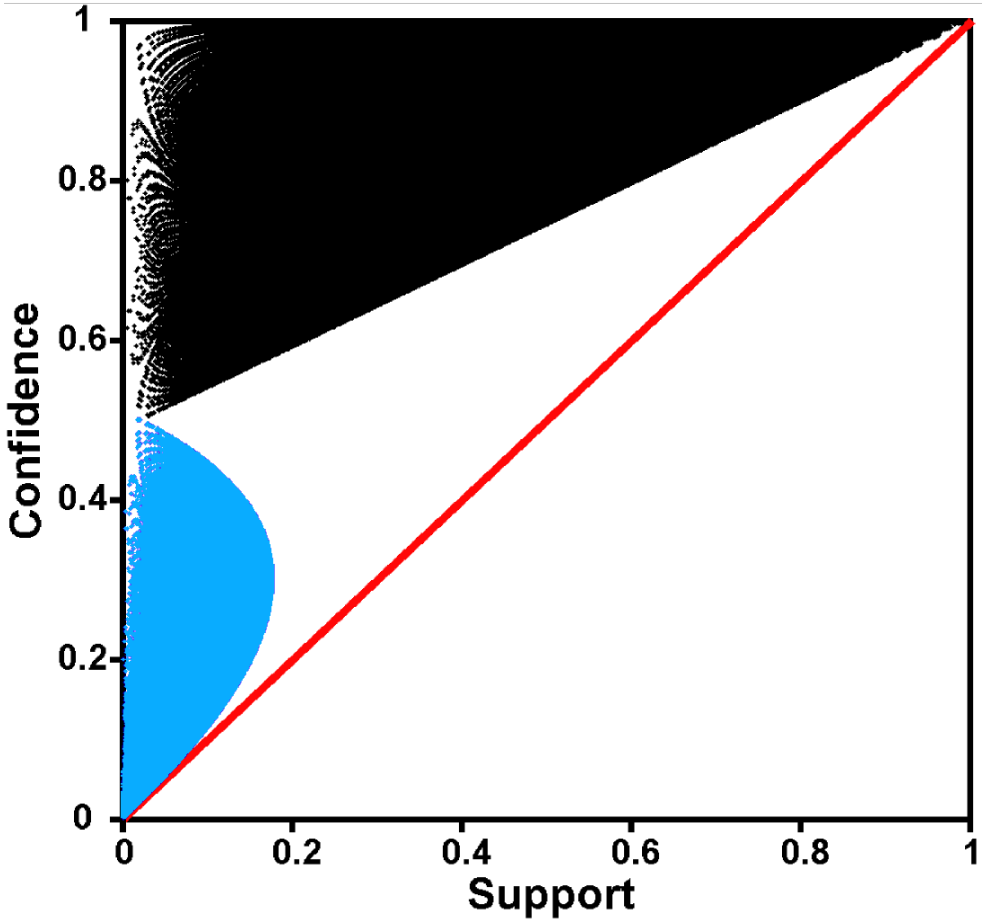
Positiver Phi-Koeffizientenwert

Nattermann

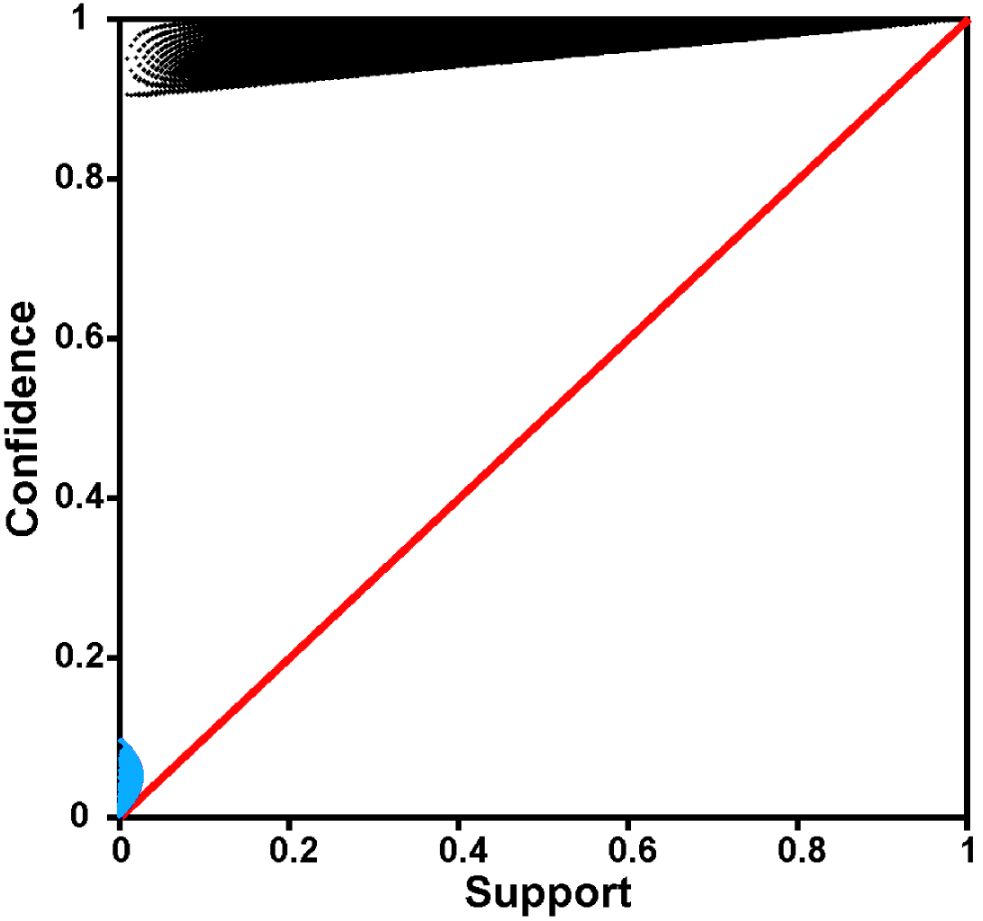




Phi-Koeffizientenwert



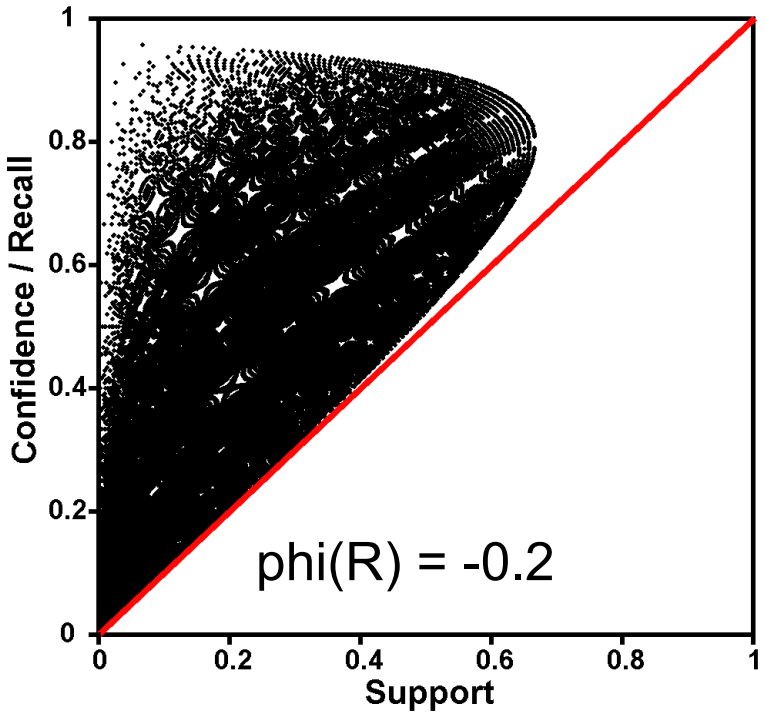
$$phi(R) = 0.25$$

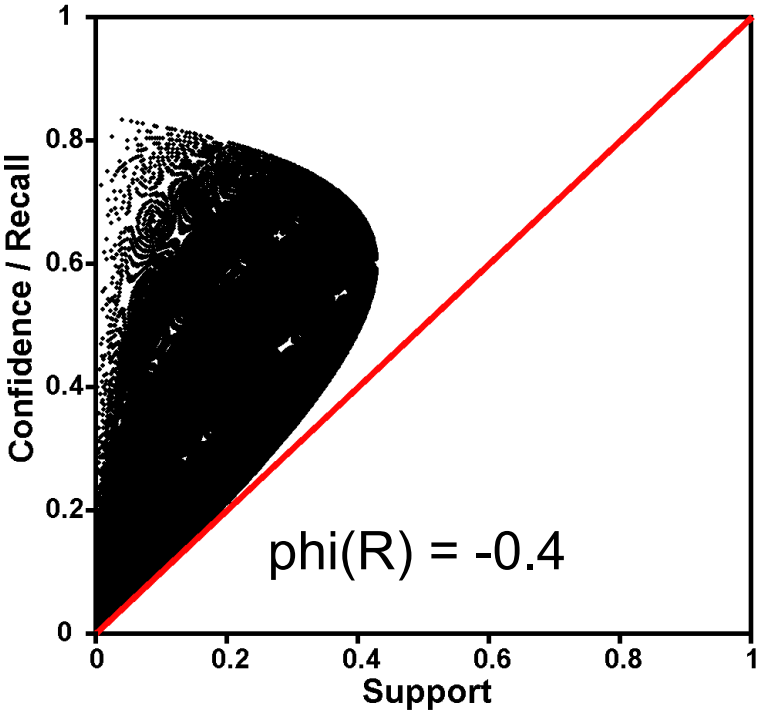


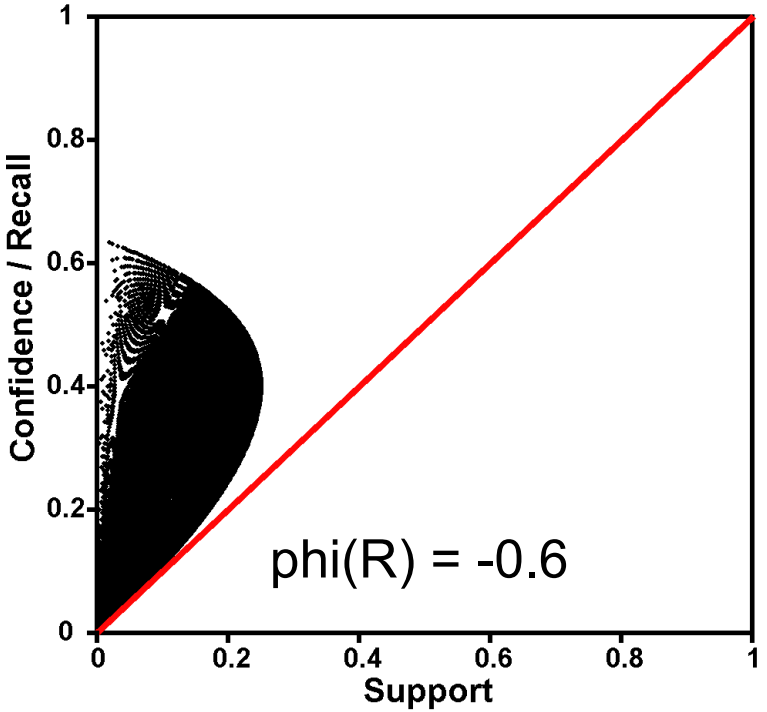
$$phi(R) = 0.95$$

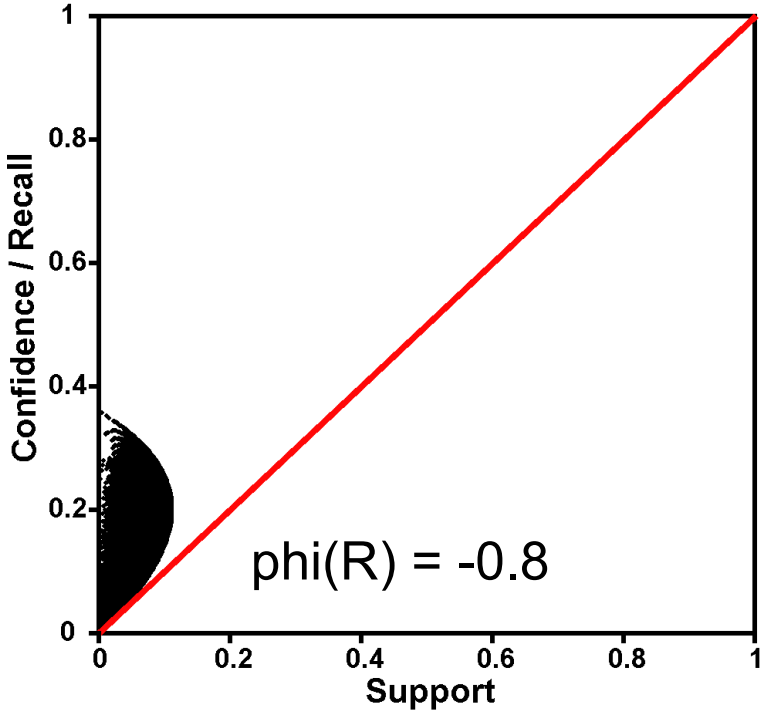
Negativer Phi-Koeffizientenwert

Nattermann



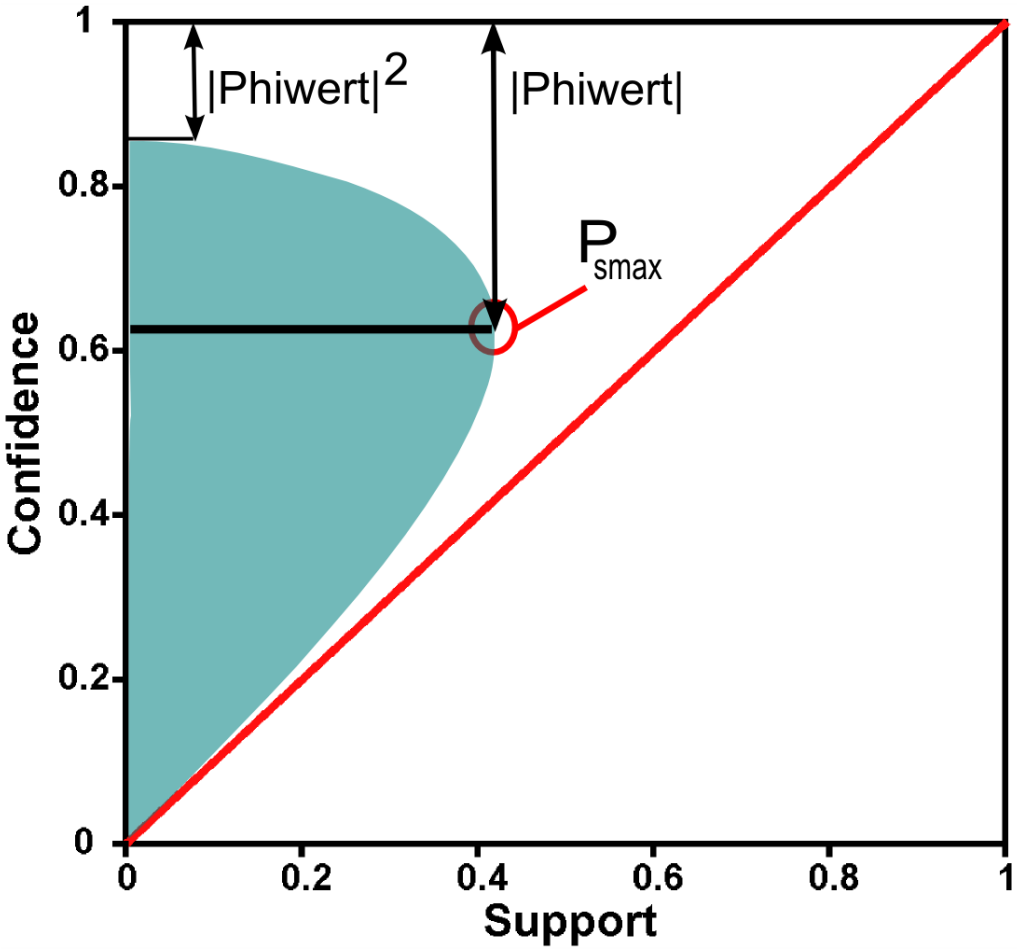


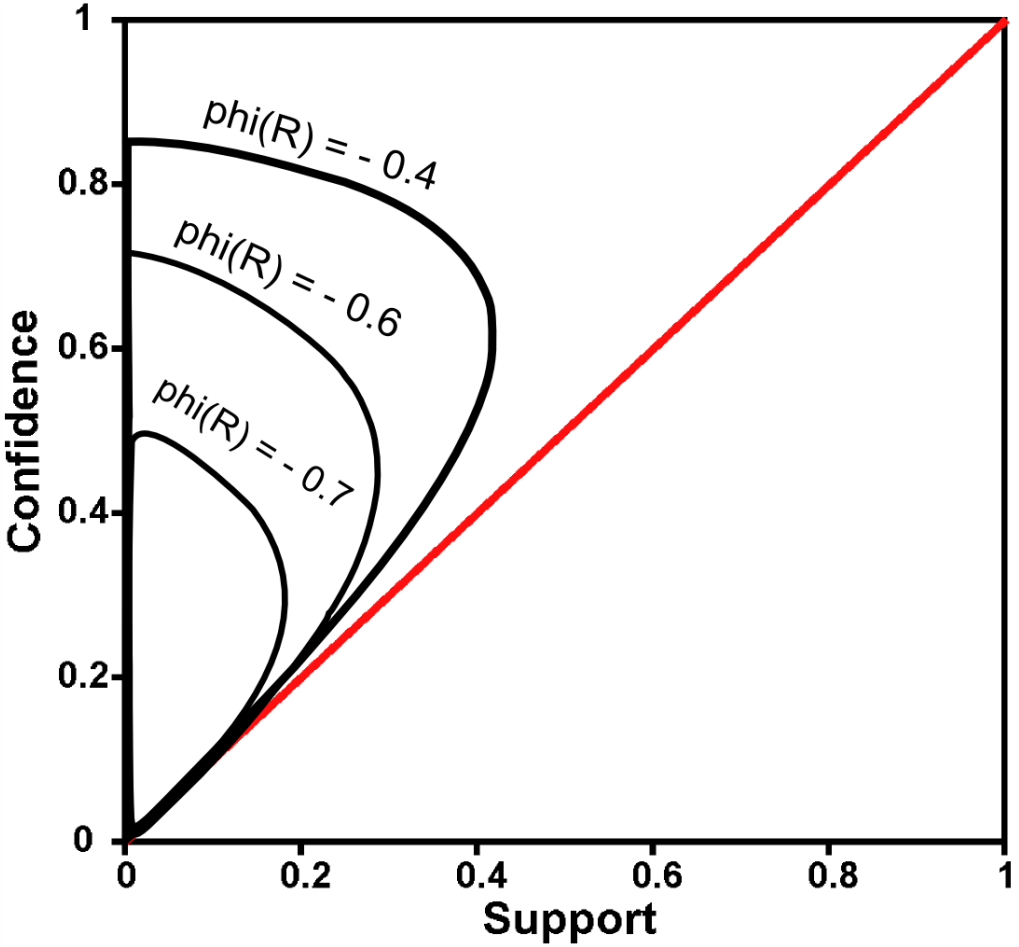




Negativer Phi-Koeffizientenwert

Nattermann





Negativer Phi-Koeffizientenwert

Nattermann

$$phi(R) = [\ \ .... \ \ , \ \ .... \ \ ]$$

$$max\{-\sqrt{1-conf(R)}\ \ , \ \ \frac{supp(R)\ -\ 1}{1\ +\ supp(R)}\}$$

$$min\{\sqrt{conf(R)}\ \ , \ \ \sqrt{\frac{supp(R)-conf(R)}{supp(R)-1}}\}$$

Vorläufiges Fazit

Fazit

- Es ist möglich Heuristiken im Assoziationsraum darzustellen.
- Man benötigt bei der Darstellung nicht die Größe N der Datenbank.
- Mit Support-, Confidence- und Recall-Wert kann man den jeweiligen Heuristikwert entsprechend berechnen
- Nur mit Support- und Confidence-Wert ist es lediglich möglich den jeweiligen Heuristikwert einzuschränken.

Referenzen

[1] P.-N. Tan, V. Kumar and J. Srivastava: *Selecting the right interestingness measure for association patterns.* In Proceedings of the Eight A CM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 2002.

[2] Johannes Fürnkranz and Peter Flach: *ROC 'n' rule learning – towards a better understanding of covering algorithms.* Machine Learning, 58(1):39-77, 2005

[3] Roberto J. Bayardo Jr. and Rakesh Agrawal: *Mining the Most Interesting Rules* Appears in Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 145-154, 1999.

[4] Tianyi Wu, Yuguo Chen, Jiawei Han: *Association Mining in Large Databases: A Re-examination of Its Measures.* Proceedings PKDD 2007: 621-628

[5] Frederik Janssen: *Eine Untersuchung des Trade-Offs zwischen Precision und Coverage bei Regel-Lern-Heuristiken* Diplomarbeit TU Darmstadt 2006

[6] Frederik Janssen and Johannes Fürnkranz: *On Trading Off Consistency and Coverage in Inductive Rule Learning* TU Darmstadt 2006

[7] Rakesh Agrawal, Tomasz Imielinski and Arun Swami: *Mining Association Rules between Sets of Items in Large Databases* IBM Almaden Research Center

[8] Bart Goethals: *Survey on Frequent Pattern Mining* University of Helsinki

[9] Alípio M. Jorge, Paulo J. Azevedo: *An experiment with association rules and classification: post-bagging and conviction* Universität von Porto

[10] Véronique Tietz: *Verallgemeinerung von Assoziationsregeln*

[11] J.Lehn, H.Wegmann: *Einführung in die Statistik*

Vielen Dank !!!

Anhang !!!

Die Klösgenheuristik

Klösgenheuristik

$$klosgen(R) = coverage(R)^w \; precgain(R)$$

$$klosgen(R) = supp(A)^w \; (conf(R) - supp(B))$$

$$klosgen(R) = \left(\frac{supp(R)}{conf(R)}\right)^w \left(conf(R) - \frac{supp(R)}{recall(R)}\right)$$

$$klosgen(R) = supp(A)^w \; \frac{leverage(R)}{supp(A)}$$

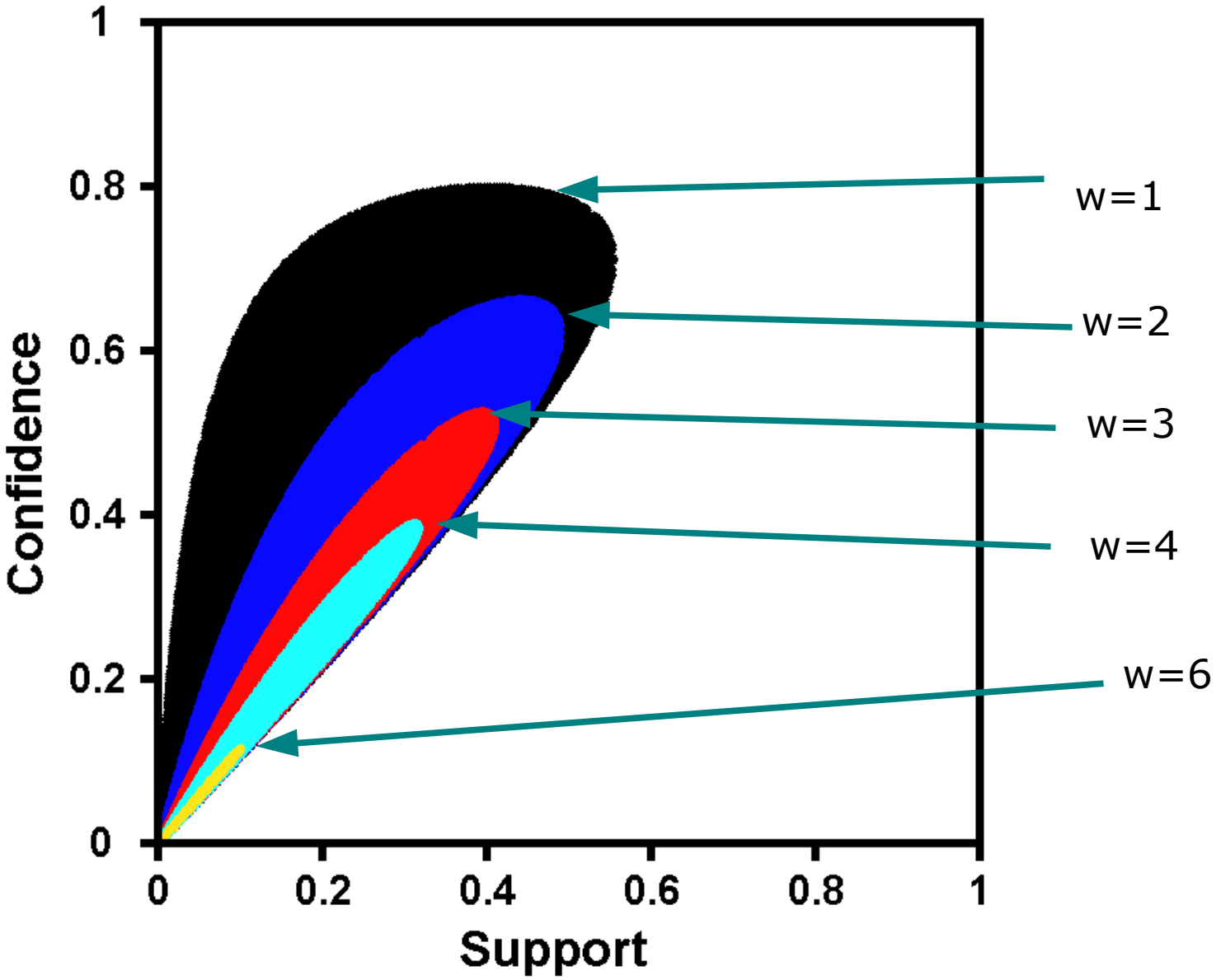
$$x = w - 1$$

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

Frederik Janssen: „Eine Untersuchung des Trade-Offs zwischen Precision und Coverage bei Regel-Lern-Heuristiken“

Negative Klösgenwerte

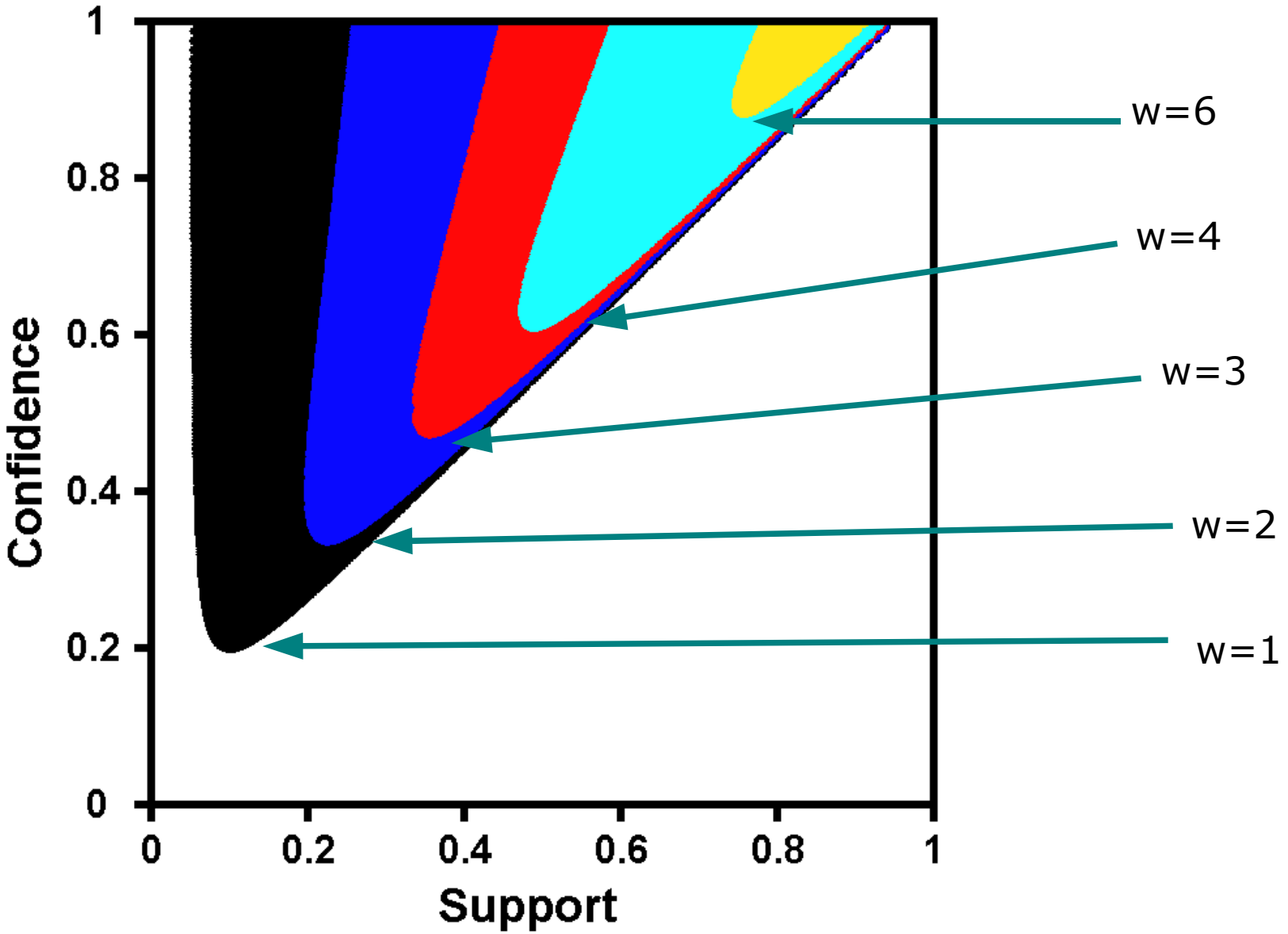
Nattermann



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

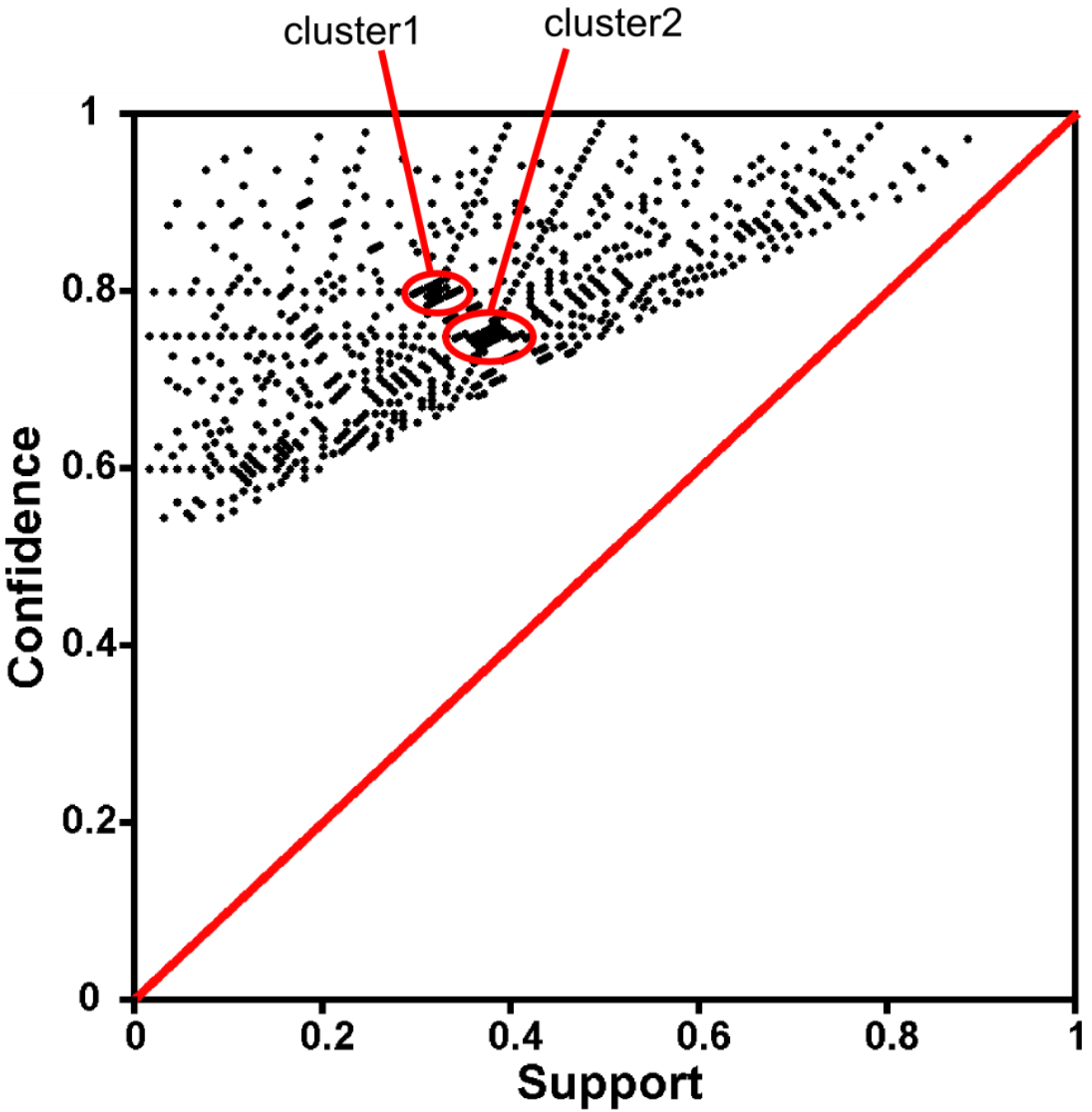
Positive Klösgenwerte

Nattermann

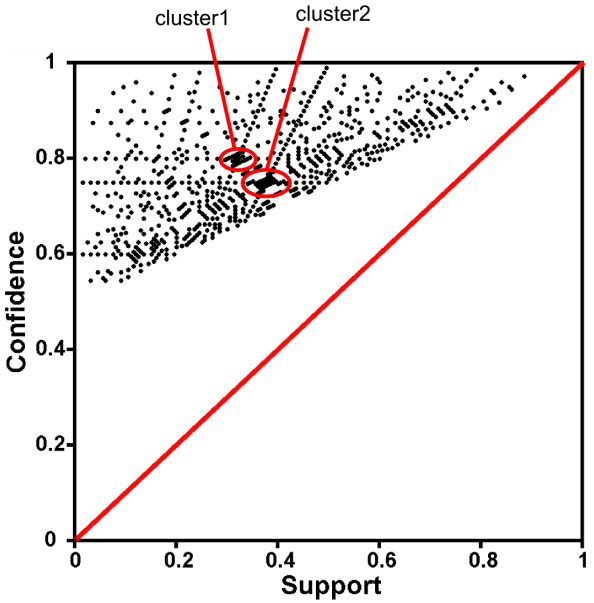


$$klosgen(R) = supp(A)^{x} \; leverage(R)$$

Cluster



Cluster



$$min\{\Delta C0\} = 1$$

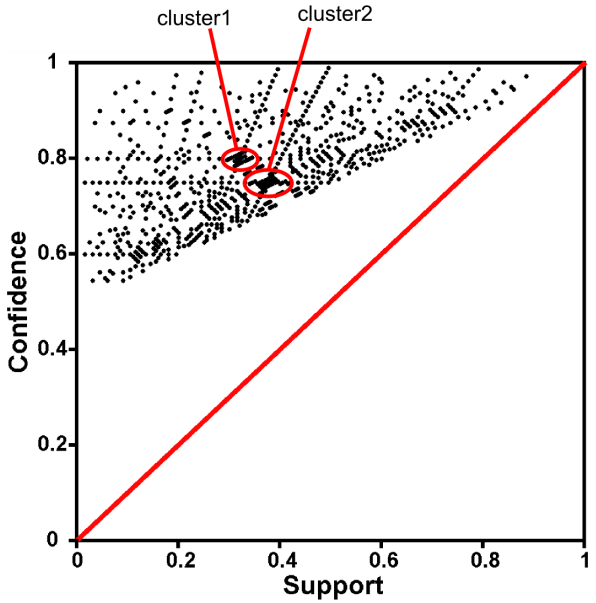
$$\Rightarrow \quad \Delta supp(R) = \frac{1}{N}$$

Analog zum Supportwert geht man beim Confidencewert vor. Hinzu kommt die Information das der neue Confidencewert entscheidend vom gerade ermittelten C0-Differenzwert abhängt

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

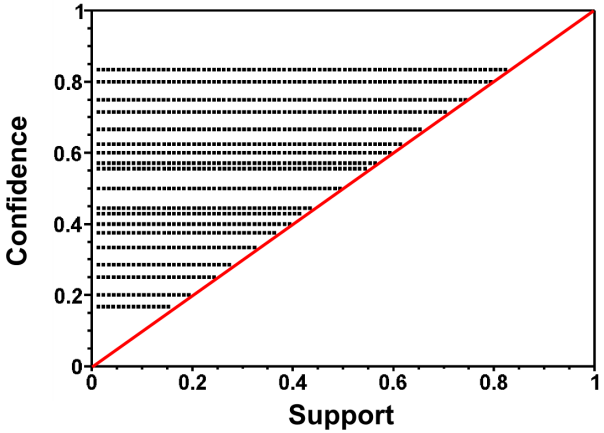
$$\Delta C1 = \frac{\Delta C0 \quad C1}{C0}$$

Cluster



$$\Delta C1 = \frac{\Delta C0 \;\; C1}{C0}$$

Da Delta C1 ganzahlig ist, muss der Zähler ein vielfaches von C0 sein. Es ergeben sich eine Fülle von konstanten Confidencefunktionen, auf denen die Cluster liegen können.

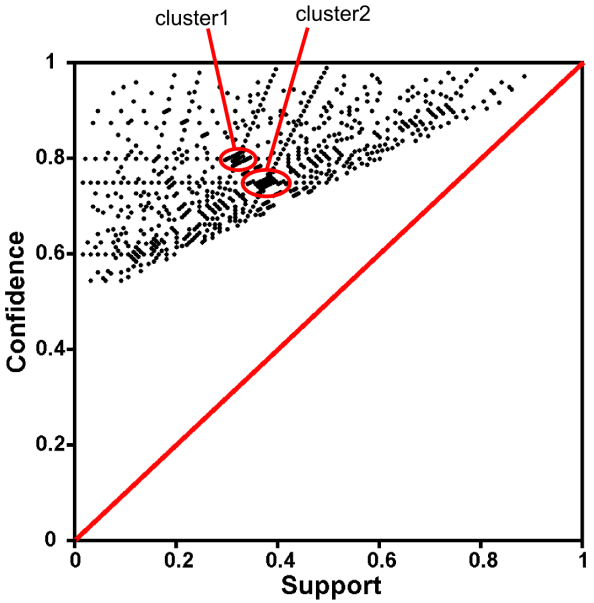


$$C1 = \frac{\Delta C1 \;\; C0}{\Delta C0} \rightarrow conf(R) = \frac{\Delta C0}{\Delta C1 + \Delta C0}$$

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

Wobei C0 und C1 das Zentrum des Clusters definieren und die Werte Delta C0 und Delta C1 die Abweichung der einzelnen Clusterpunkte widerspiegeln.

Cluster



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

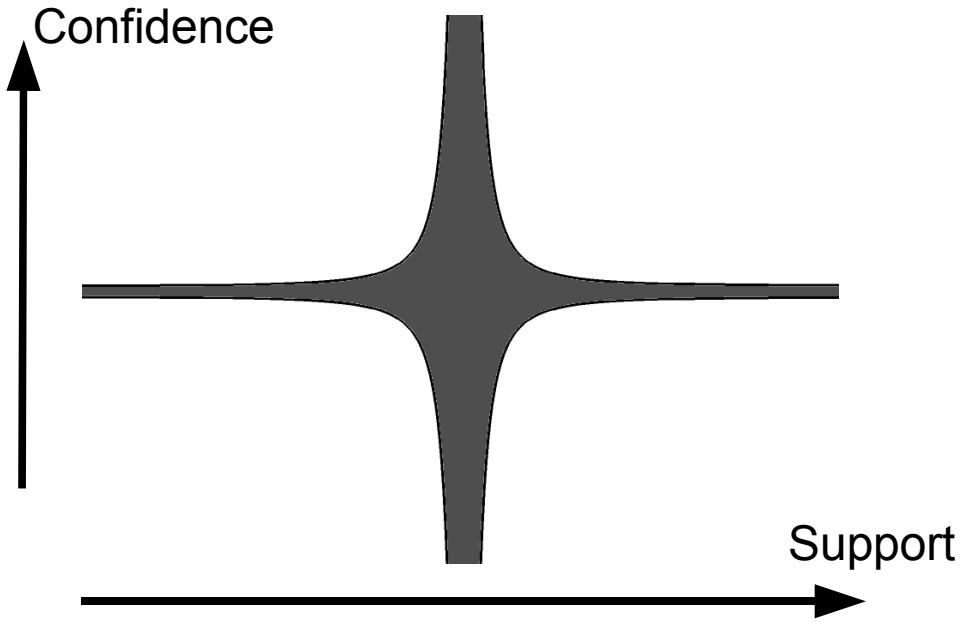
Wenn der Wert Delta C0 steigt, dann muss der Wert Delta C1 proportional steigen.

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

$$c0nf(R) = \frac{\Delta C0}{\Delta C1 + \Delta C0}$$

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

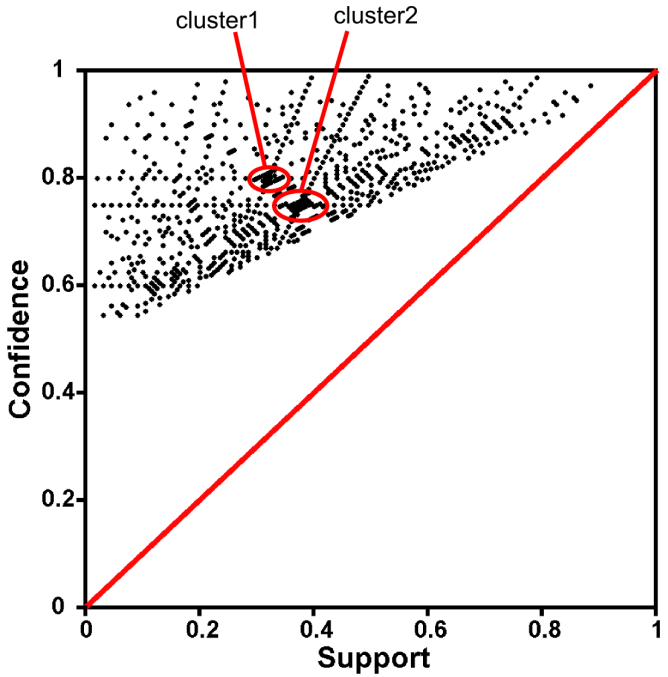
Cluster

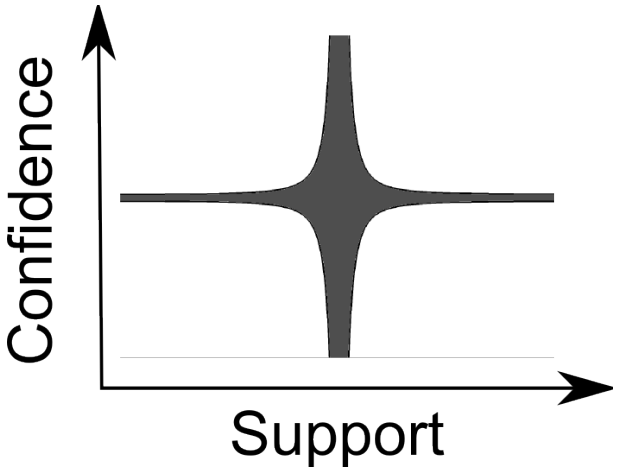


$$c0nf(R) = \frac{\Delta C0}{\Delta C1 + \Delta C0}$$

- 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.

Cluster





Dieses Verhalten ist unabhängig vom Ort und bewirkt ein ungefähr ähnliches aussehen aller Cluster.