Was wissen Sie jetzt?

- Sie haben drei Prinzipien für die Regelbewertung kennen gelernt:
  - Unabhängige Mengen sollen mit 0 bewertet werden.
  - Der Wert soll höher werden, wenn die Regel mehr Belege hat.
  - Der Wert soll niedriger werden, wenn die Mengen weniger Belege haben.
- Sie haben drei Maße kennen gelernt, die den Prinzipien genügen:
  - Einfaches Maß,
  - statistisches Maß und
  - Sicherheitsmaß.

Verbesserungen von Apriori

- Bessere Kriterien als support und Konfidenz
- Kondensierte Repräsentationen
- Anfrageoptimierung im Sinne induktiver Datenbanken durch constraints

- Die erste Verbesserung haben wir schon gesehen.
- Hier sehen wir die zweite Verbesserung.
- Die Konferenzen KDD, PKDD und ICDM sind aber voll von Beiträgen zu „frequent itemsets“!

Kondensierte Repräsentationen

Ersetzen der Datenbank bzw. der Baumstruktur durch eine kondensierte Repräsentation,

- die kleiner ist als die ursprüngliche Repräsentation und
- aus der wir alle häufigen Mengen und ihre Häufigkeit ableiten können, ohne noch mal die Daten selbst anzusehen.

Kondensierte Repräsentationen für Assoziationsregeln:

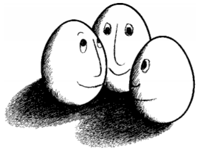
- Closed item sets
- Free sets

Operator, der die Menge aller Assoziationsregeln ableitet:

- Cover operator

Wir erinnern uns...

- Hypothesen werden in einem Verband angeordnet.
- Ein Versionenraum gibt die möglichen Hypothesen an, die zu den gegebenen Daten passen – durch weitere Daten wird der Versionenraum weiter eingeschränkt:
  - Wenn ein positives Beispiel nicht abgedeckt ist, wird die Menge der speziellsten Hypothesen generalisiert,
  - Wenn ein negatives Beispiel abgedeckt ist, wird die Menge der generellsten Hypothesen spezialisiert.



In anderen Worten:

Wir hätten gern einen Versionenraum!
Der Versionenraum ist kleiner als der Hypothesenraum.
Außerhalb des Versionenraums kann das Lernziel nicht liegen.

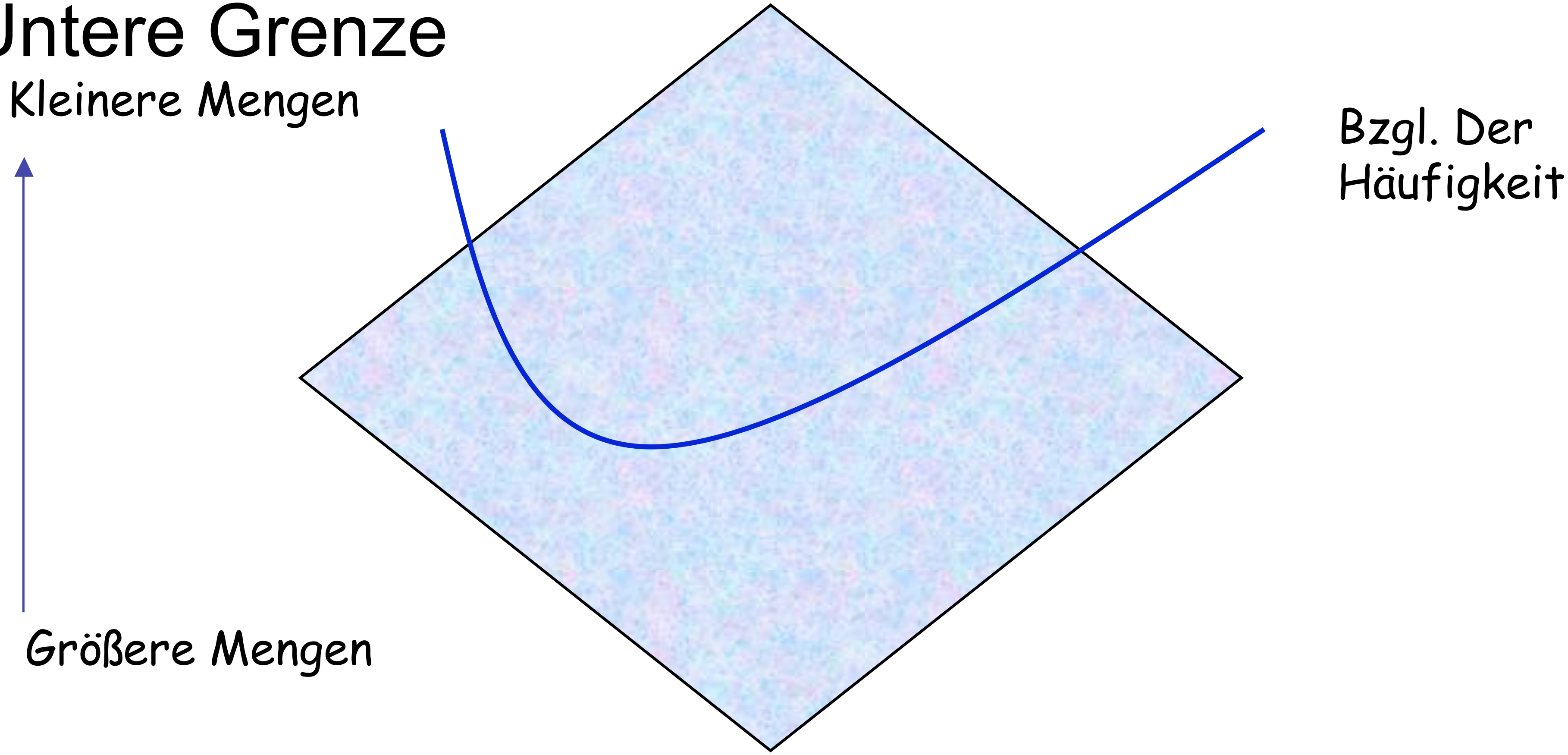
Wir müssen also aus den Beispielen

- eine untere Grenze und
- eine obere Grenze konstruieren.

Eine Halbordnung bzgl. Teilmengenbeziehung haben wir schon.

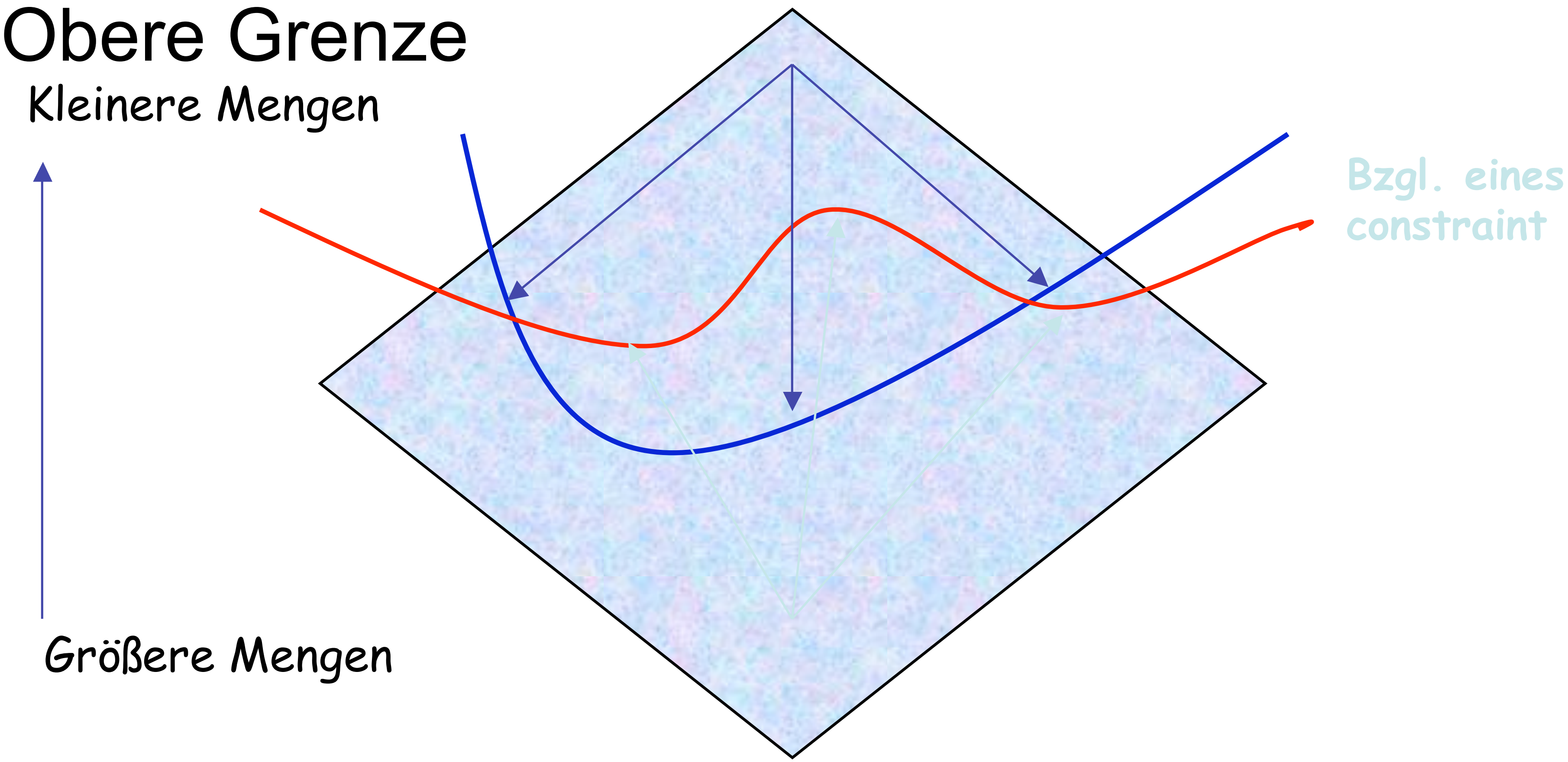
Die Grenzen haben wir auch.
    Gemerkt?

Untere Grenze



- Wenn eine Menge häufig ist, so auch all ihre Teilmengen. (Anti-Monotonie)
- Beschneiden der Ausgangsmengen für die Kandidatengenerierung gemäß dieser Grenze!

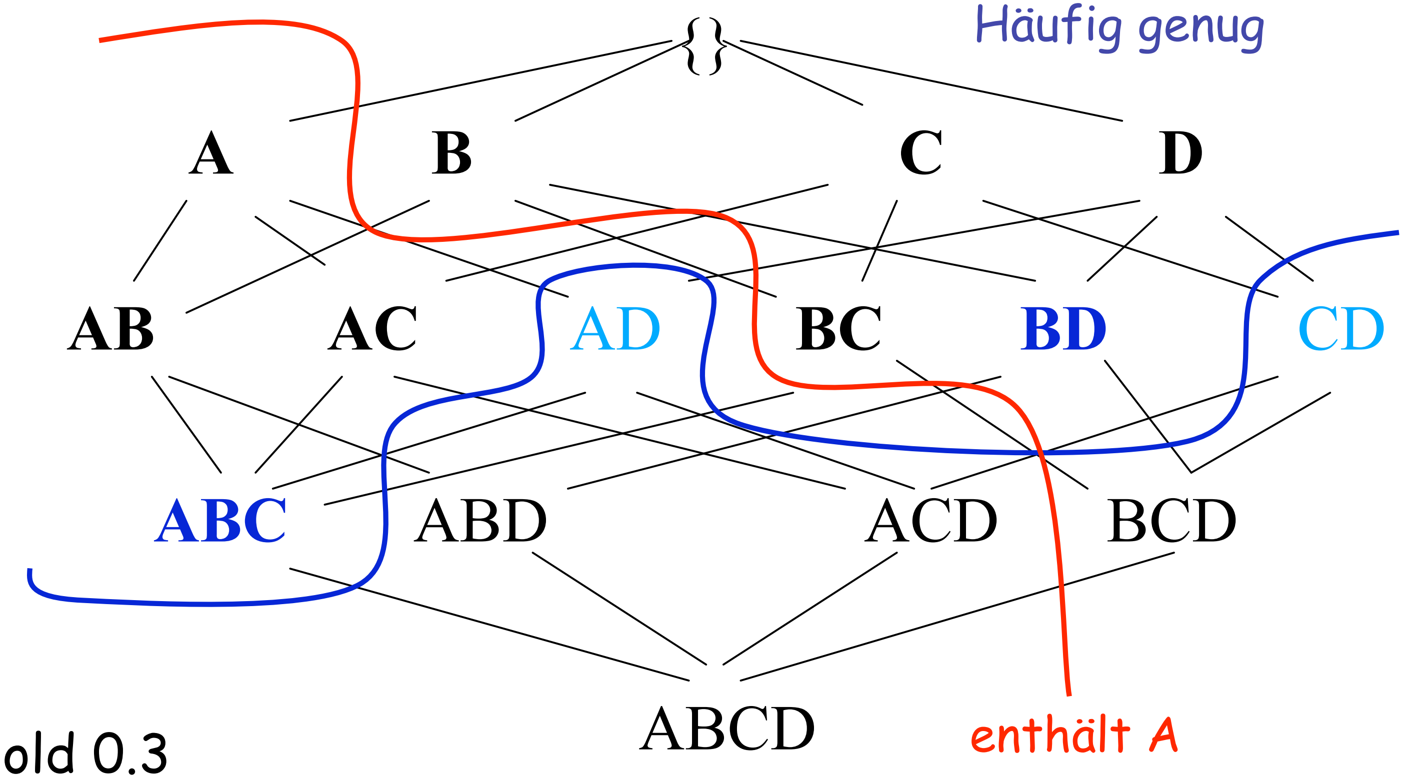
Obere Grenze



- Monotonie der Seltenheit: Wenn eine Teilmenge selten ist, so auch jede Menge, die sie enthält. Seltenheit ist ein constraint.
- Beschneidung der Kandidatengenerierung nach der Monotonie.

Beispiel

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0



Frequency threshold 0.3

Dank an Jean-Francois Boulicaut!

Closed Item Sets

A	B	C	D
1	1	1	1
0	1	1	0
1	0	1	0
1	0	1	0
1	1	1	1
1	1	1	0

- closure(S) ist die maximale Obermenge (gemäß der Teilmengenbeziehung) von S, die noch genauso häufig wie S vorkommt.
- S ist ein *closed item set*, wenn closure(S)=S.
- Bei einem Schwellwert von 0,2 sind alle Transaktionen häufig genug.
- Closed sind: C, AC, BC, ABC, ABCD
  keine Obermenge von C kommt auch 6 mal vor;
  A kommt 5 mal vor, aber auch die Obermenge AC und keine Obermenge von AC
  ...

Kondensierte Repräsentation und Ableitung

Closed item sets sind eine kondensierte Repräsentation:

- Sie sind kompakt.
- Wenn man die häufigen closed item sets C berechnet hat, braucht man nicht mehr auf die Daten zuzugreifen und kann doch alle häufigen Mengen berechnen.

Ableitung:

- Für jede Menge S prüfen wir anhand von C:
  Ist S in einem Element X von C enthalten?
  - Nein, dann ist S nicht häufig.
  - Ja, dann ist die Häufigkeit von S ungefähr die von X.
    Wenn es in mehreren Elementen von C vorkommt, nimm die maximale Häufigkeit!

Freie Mengen (free sets)

- Eine Menge S ist frei, wenn es keine Regel mit Konfidenz=1 zwischen ihren Elementen gibt, d.h.

$$\neg \exists X, Y \,|\, S = X \cup Y, Y \neq \{\,\}, X \Rightarrow Y$$

- Eine Menge S ist **d**-frei, wenn es keine Regel mit weniger als **d** Ausnahmen zwischen ihren Elementen gibt.

- Die closed sets sind die closure der freien Mengen!
  Man kann die closed sets aus den freien Mengen berechnen.
- Freiheit ist eine anti-monotone Eigenschaft von Mengen.
  Deshalb kann man die freien Mengen effizient berechnen.

Beispiel

A	B	C	D
1	1	1	1
0	1	1	0
1	0	1	0
1	0	1	0
1	1	1	1
1	1	1	0

5 4 6 2

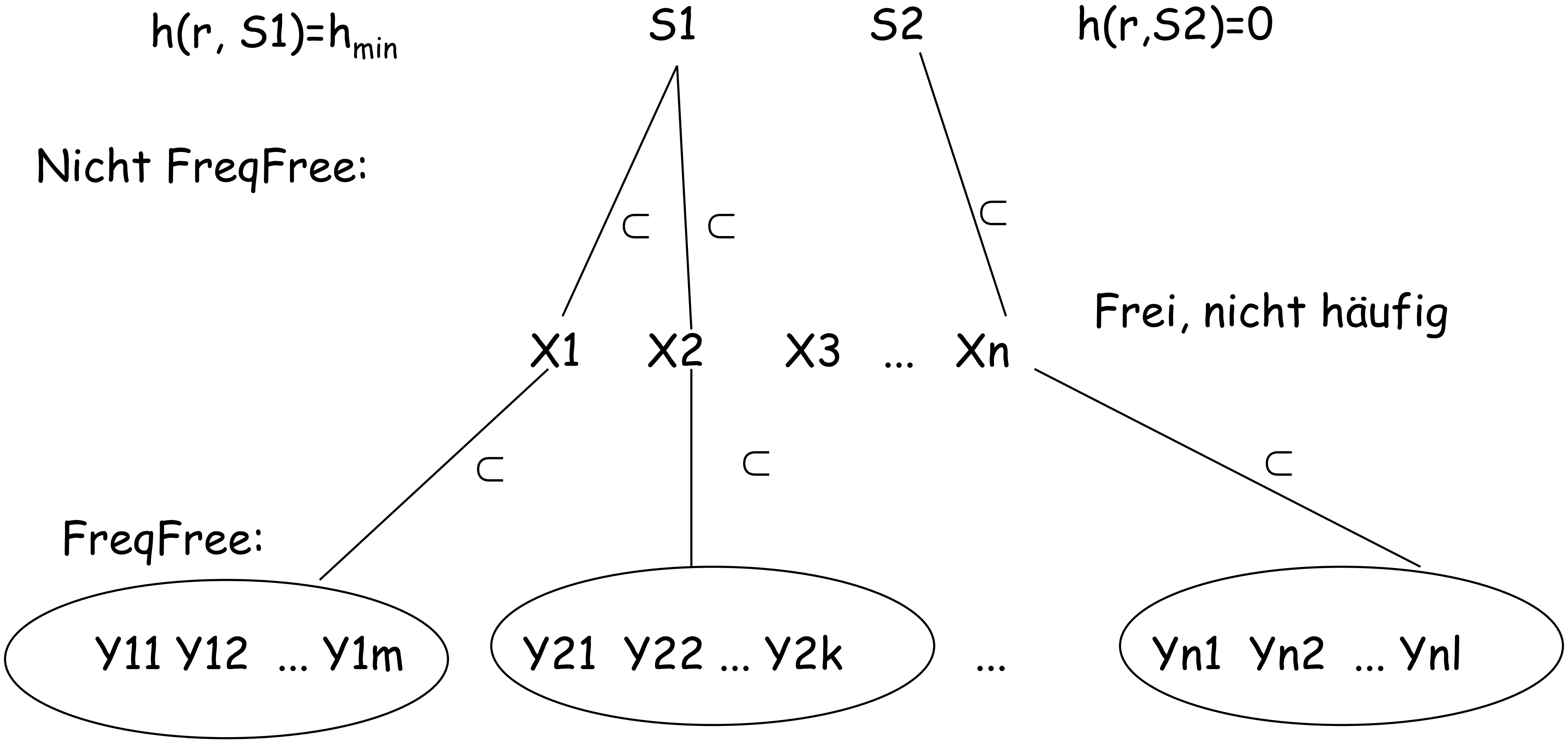
- Bei einem Schwellwert von 0,2 sind die häufigen freien Mengen:
  {}, A,B,D,AB
- Closed sind: C, AC, BC, ABCD, ABC
- Closure({})=C
  closure(A)=AC
  closure(B)= BC
  closure(D)=ABCD
  closure(AB)=ABC

"Unfreie" Mengen: AD: D $\Rightarrow$ A, BD: D $\Rightarrow$ B, ABD: D $\Rightarrow$ AB
C:{} $\Rightarrow$ C, AC: A $\Rightarrow$ C, BC: B $\Rightarrow$ C, CD: D $\Rightarrow$ C, ABC, ADC, BCD, ABCD

Arbeiten mit freien Mengen

- Free(r, δ): Eine Menge X ist δ-frei, wenn es in r keine Regel zwischen ihren Elementen mit weniger als δ Ausnahmen gibt.
- Freq(r, σ): $\{X \mid X \subseteq R, |X \in r| / |r| \geq \sigma\}$
- FreqFree(r, σ, δ): Freq (r, σ) $\cap$ Free(r, δ)
- Negative Grenze Bd-(r, σ, δ): $\{X \mid X \subseteq R, X \notin \text{FreqFree}(r, \sigma, \delta)$ und $\forall Y \subset X, Y \in \text{FreqFree}(r, \sigma, \delta)\}$
  Also die kürzesten Mengen, die gerade nicht häufig und frei sind, deren Teilmengen aber häufig und frei sind.
- Wir schätzen die Häufigkeit einer Menge S so ab:
  $\exists X \subseteq S$ und X ist δ-frei, aber nicht σ–häufig, dann
  nimm 0 als Häufigkeit von S.
  Sonst nimm die kleinste Anzahl im Vorkommen der Teilmengen X als Häufigkeit von S.

Abschätzung



$\min(\{h(r,Y) \mid Y \subset X\}) = h_{min}$

MinEx

- Statt alle häufigen Mengen zu suchen, brauchen wir nur noch alle FreqFree(r, σ, δ) zu suchen.
- Bottom-up Suche im Halbverband der Mengen
  beginnt beim leeren Element, nimmt dann alle 1-elementigen Mengen,...
  endet bei den größten Mengen, die noch FreqFree(r, σ, δ) sind.
- Der Test, ob Mengen frei sind, erfordert das Bilden von strengen Regeln und erlaubt das Pruning der Mengen, in denen solche gefunden wurden.

Algorithmus von Jean-Francois Boulicaut

Algorithmus (abstrakt)

Gegeben eine binäre Datenbasis r über Objekten R und
die Schwellwerte σ und δ,

Gebe FreqFree(r, σ, δ) aus.

1. C_0:={ {} }
2. i:=0
3. **While** $C_i \neq \{\}$ **do**
4. $FreqFree_i := \{X \mid X \in C_i$, X ist σ-häufig und δ-frei}
5. $C_{i+1} := \{X \mid X \subseteq R, \forall Y \subset X, Y \in FreqFree_j(r, \sigma, \delta), j \leq i\} \setminus \bigcup_{j \leq i} C_j$
6. i:=i+1 **od**
7. **Output** $\bigcup_{j < i} FreqFree_j$

Pruning

- In der i-ten Iteration werden die δ–starken Regeln der Form X → {A} berechnet, wobei
  X häufig und frei ist auf der i-ten Ebene und
  $A \subseteq R \backslash X$.
- Das Ergebnis wird verwendet, um alle nicht δ -freien Mengen zu entfernen – sie sind keine Kandidaten mehr in der i+1-ten Iteration.

Eigenschaften von MinEx

- Der Algorithmus ist immer noch aufwändig, aber schneller als APRIORI und schneller als die Verwendung von closed sets.
- Der Algorithmus ist exponentiell in der Menge R.
- Der Algorithmus ist linear in der Menge der Datenbanktupel, wenn δ im selben Maße steigt wie die Zahl der Tupel.
  Wir verdoppeln δ, wenn wir die Tupelzahl verdoppeln.
- Der Algorithmus approximiert das „wahre“ Ergebnis.
  In der Praxis ist eine Abweichung von 0,3% aber kein Problem.

Was wissen Sie jetzt?

- Es gibt zwei Repräsentationen, die weniger Elemente für eine Suche nach häufigen Mengen ausgeben als eben alle häufigen Mengen. Aus diesen Repräsentationen können alle häufigen Mengen hergeleitet werden.
  - Die closed sets sind maximale Obermengen von S mit derselben Häufigkeit wie S.
  - Die free sets sind Mengen, aus denen man keine Assoziationsregeln machen kann.
- Wenn man die häufigen freien Mengen berechnet, hat man die untere Grenze im Versionenraum für Assoziationsregeln gefunden.
- Der Algorithmus MinEx findet diese Grenze.