Vorlesung Wissensentdeckung

MinEx

Katharina Morik, Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

14.4.2010

Gliederung

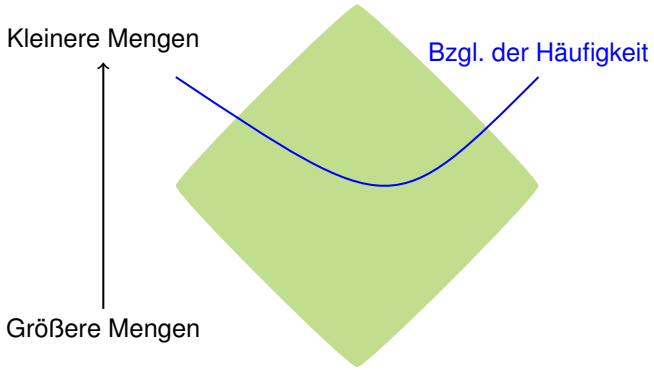
Wir erinnern uns...

- 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:

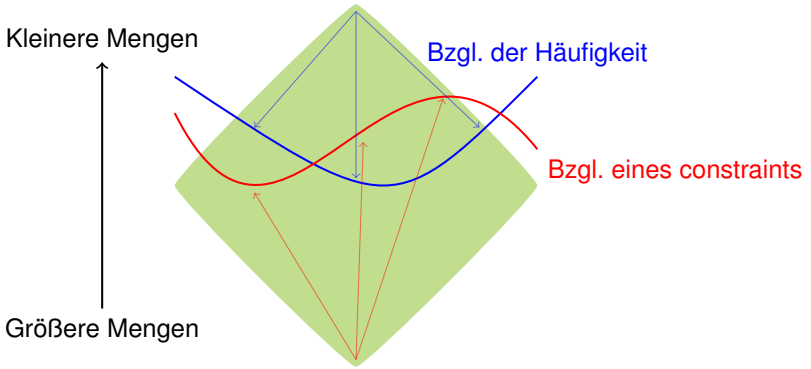
- Wir müssen also aus den Beispielen
  - eine untere Grenze und
  - eine obere Grenze konstruieren.
- Eine Halbordnung bzgl. Teilmengenbeziehung haben wir schon.
- Die Grenzen haben wir auch.
- Gemerkt?

Untere Grenze



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

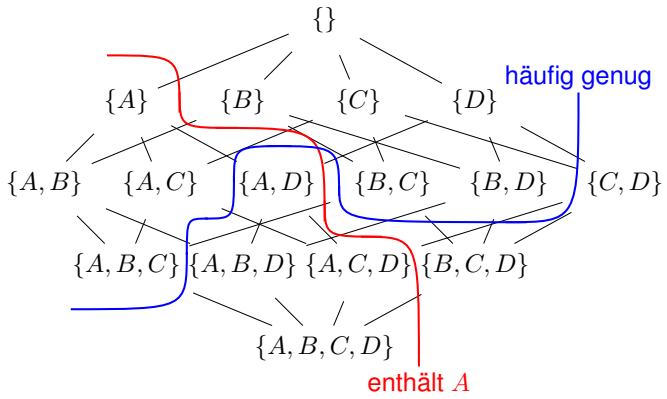
Obere Grenze



- 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 mit Frequency threshold 0.3

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



Dank an Jean-Francois Boulicaut!

Kondensierte Repräsentationen

- Statt Suche nach allen häufigen Mengen: Suche nach einer kondensierten 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

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 $closure(S) = S$
- $support(S) = support(closure(S))$ (für alle S)
- Bei einem Schwellwert von 0.1 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 genau die der kleinsten solchen Obermenge X.
    - Wenn es in mehreren Elementen von C vorkommt, nimm die maximale Häufigkeit!

Freie Mengen (free sets)

- Eine Menge S ist δ-frei, wenn es keine Regel mit δ oder weniger Ausnahmen zwischen ihren Elementen gibt.
- Beispiel
  - $support(\{a, c, f\}) = 128$
  - $support(\{a, c, f, g\}) = 125$

  Hier hat die Regel $acf \rightarrow g$ also 3 Ausnahmen, daher ist $\{a, c, f, g\}$ nicht 3-frei

Freie Mengen (free sets)

- Aus einer 0-freien Menge lässt sich also keine exakte Regel bilden
- Also hat jede echte Teilmenge einer 0-freien Menge X höheren Support als X
- Also hat keine 0-freie Menge X eine echte Teilmenge Y mit:

$$closure(Y) = closure(X) \textbf{ und } support(Y) = support(X)$$

- Also sind die 0-freien Mengen die kleinsten Mengen, aus denen sich die closed item sets mit gleichem Support berechnen lassen!
- Weiterhin gilt: jede Menge M hat eine δ-freie Teilmenge, mit der sich der Support von M approximieren lässt (hier nicht behandelt)

Beispiel

- Bei einem Schwellwert von 0.2 sind die häufigen 0-freien Mengen:

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

$$\{\}, A, B, D, AB$$

- Closed sind: $C, AC, BC, ABC, ABCD$

$$closure(\{\}) = C, \qquad support(\{\}) = support(C)$$
$$closure(A) = AC \qquad \ldots$$
$$closure(B) = BC \qquad \ldots$$
$$closure(D) = ABCD \qquad \ldots$$
$$closure(AB) = ABC, \quad support(AB) = support(ABC)$$

Nicht 0-freie Mengen: $AC : A \to C, AD : D \to A, BC : B \to C, BD : D \to B, CD : D \to C, ABC, ABD, ACD, BCD, ABCD$

Arbeiten mit freien Mengen

- $Free(r, \delta)$: Eine Menge X ist δ-frei, wenn es in r keine Regel zwischen ihren Elementen mit weniger als δ Ausnahmen gibt $\to$ So eine Regel heiSSt δ-stark
- $Freq(r, \sigma) : \{X | X \subseteq R, (\frac{|\{t|t \in r, X \subseteq t\}|}{|r|}) \geq \sigma\}$
- $FreqFree(r, \sigma, \delta)$: $Freq(r, \sigma) \cap Free(r, \delta)$
- Antimonotonie: Für $Y \subseteq X \in Free(r, \delta)$ gilt $Y \in Free(r, \delta)$

Arbeiten mit freien Mengen

- Negative Grenze:

$FreeBd_-(r, \sigma, \delta) : \{X \subseteq R | X \in Free(r, \delta), X \notin Freq(r, \sigma)$ **und**

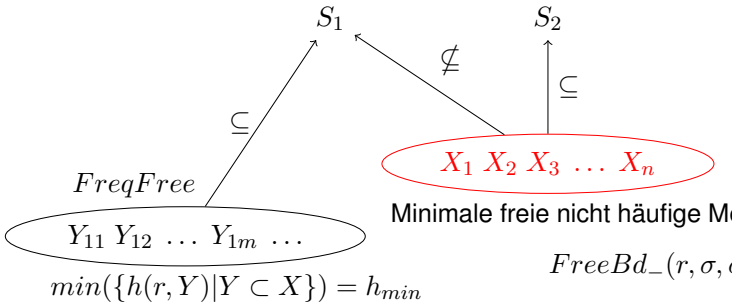
$$\forall Y \subset X : Y \in FreqFree(r, \sigma, \delta)\}$$

Also die kürzesten Mengen, die δ-frei sind, aber nicht häufig, und deren Teilmengen sowohl häufig als auch δ-frei sind.

- Wir schätzen die Häufigkeit einer Menge S so ab: Falls $\exists X$ mit $X \subseteq S$ und $X \in FreeBd_-(r, \sigma, \delta)$, so ist S nicht häufig. Sonst approximiere die Häufigkeit von S durch die kleinste Häufigkeit einer Teilmenge Z von S mit $Z \in FreqFree(r, \sigma, \delta)$.

Abschätzung

$h(r, S_1) = h_{min}$ $\qquad$ $h(r, S_2) = 0$



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

Minimale freie nicht häufige Mengen:

$FreeBd_-(r, \sigma, \delta)$

MinEx

- Statt alle häufigen Mengen zu suchen, brauchen wir nur noch alle Mengen aus $FreqFree(r, \sigma, \delta)$ 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 $FreqFree(r, \sigma, \delta)$ 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,
- **Ausgabe**: $FreqFree(r, \sigma, \delta)$

Listing 1: MinEx-Algorithmus

```
1 C_0 := {{}}
2 i := 0
3 While C_i ≠ {} do
4    FreqFree_i := {X|X ∈ C_i, X ist σ–häufig und δ–frei}
5    C_{i+1} := {X|X ⊆ R, ∀Y ⊂ X, Y ∈ FreqFree_j(r,σ,δ), j ≤ i} \ (∪_{j≤i} C_j)
6    i := i + 1
7 Output ∪_{j<i} FreqFree_j
```


Pruning

- In der i-ten Iteration werden die δ-starken Regeln der Form $X \to \{A\}$ berechnet, wobei X häufig und frei ist auf der i-ten Ebene und $A \subseteq \frac{R}{X}$.
- Das Ergebnis wird verwendet, um alle nicht δ-freien Mengen zu entfernen - sie sind keine Kandiaten 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, wenn also bei doppelter Tupelzahl auch δ verdoppelt wird.
- Für $\delta > 0$ liefern die δ-freien Mengen nur eine Approximation des tatsächlichen Supports. In der Praxis ist eine durchschnittliche Abweichung von 0.3% aber kein Problem.

Was wissen Sie jetzt?

- Sie kennen 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 größten häufigen freien Mengen berechnet, hat man die untere Grenze im Versionenraum für Assoziationsregeln gefunden.
- Der Algorithmus MinEx findet diese Grenze.