
17. Wissensentdeckung

Das Material dieses Kapitels basiert
auf den Arbeiten von Frau Gabriele
Kern-Isberner.

Wissensentdeckung

Konzepte beschreiben spezifische Teilmengen von Objekten etc.

Beispiele:

- "Auto"
- "Säugetier"
- "stetige Funktion"

Ein Konzept lässt sich damit als Boolesche Funktion auffassen, die Objekten etc. die Werte *true* bzw. *false* zuordnet, je nachdem, ob das Objekt zu der spezifischen Teilmenge gehört oder nicht.

Wissensentdeckung

Aufgabe des Konzeptlernens:

Aus einer Menge von positiven (und negativen) Beispielen für ein Konzept c soll eine allgemeine Definition von c generiert werden.

Unterschiede zum Entscheidungsbaumlernen:

- Konzepte können durch komplexere Darstellungen beschrieben werden;
- Konzeptlernverfahren sind meistens inkrementelle Lernverfahren.

Wissensentdeckung

Ein Konzept c ist eine einstellige Funktion

$$c : M \rightarrow \{0, 1\}$$

über einer Grundmenge M von Beispielen.

Für $x \in M$ gilt

- x gehört zum Konzept c (x ist ein positives Beispiel, c deckt x ab) gdw. $c(x) = 1$;
- x gehört nicht zum Konzept c (x ist ein negatives Beispiel) gdw. $c(x) = 0$.

Die Menge aller positiven Beispiele für c wird auch Extension von c genannt.

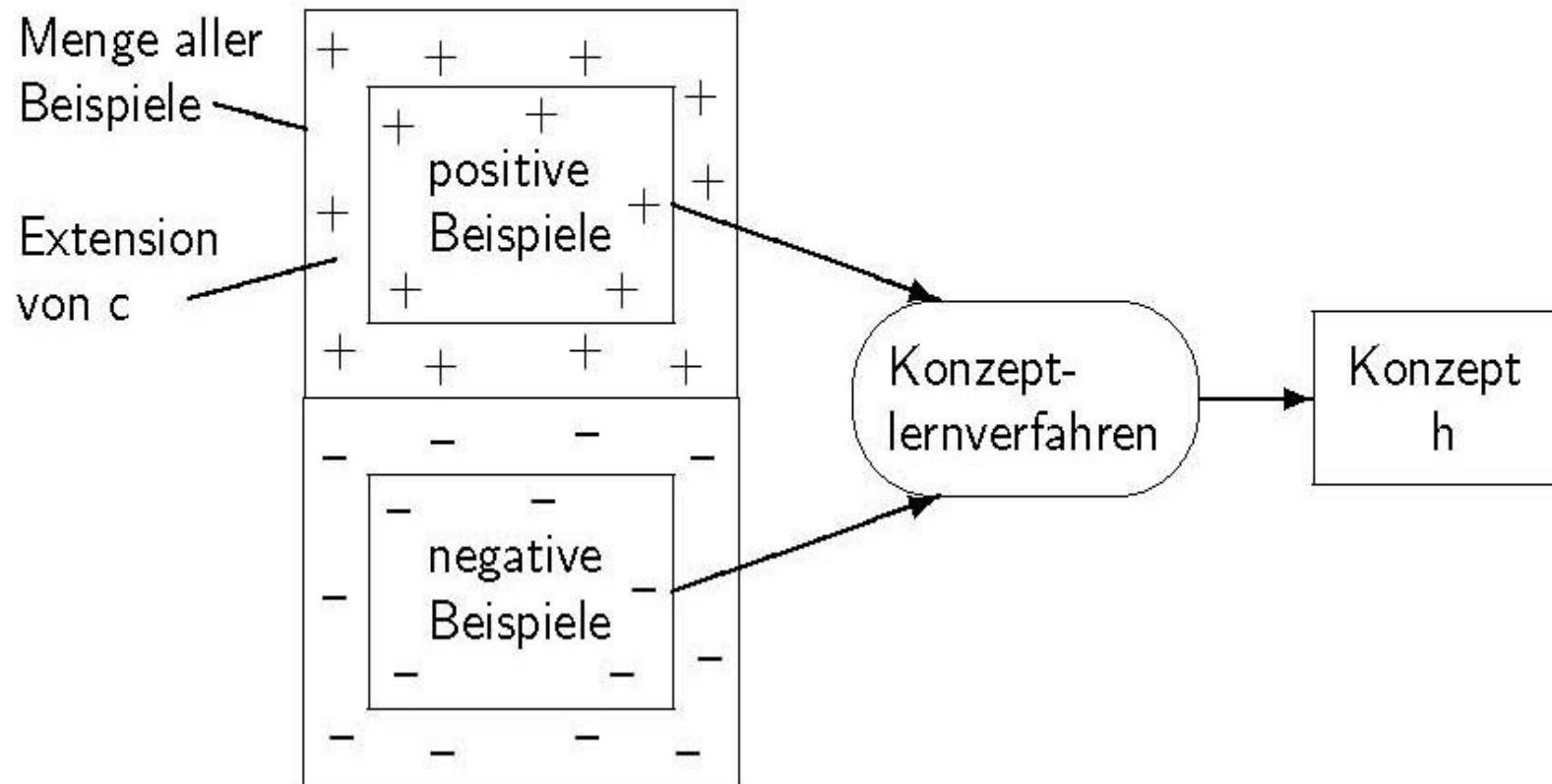
Wissensentdeckung

Aus einer gegebenen Menge von (positiven und negativen) Beispielen soll ein Konzept gelernt werden, dass zu diesen gegebenen Beispielen “passt”:

Sei $B \subseteq M$ eine Menge von Beispielen und c ein Konzept.

- c ist **vollständig** bzgl. B gdw. für alle $b \in B$ gilt: wenn b positiv ist, dann ist $c(b) = 1$. D.h., alle positiven Beispiele werden abgedeckt.
- c ist **korrekt** bzgl. B gdw. für alle $b \in B$ gilt: wenn b negativ ist, dann ist $c(b) = 0$. D.h., es wird kein negatives Beispiel abgedeckt.
- c ist **konsistent** bzgl. B gdw. c ist vollständig und korrekt bzgl. B .

Wissensentdeckung



Wissensentdeckung

- Die Güte der Klassifikation lässt sich als Prozentsatz der richtig klassifizierten Objekte der gesamten Grundmenge ausdrücken.
- Auf der negativen Seite kann man die Summe der Kosten aller Fehlklassifikationen über der gesamten Grundmenge berechnen.

Wissensentdeckung

Zu erlernendes Konzept:

“Sportsendungen, die sich Paul Trops im Fernsehen anschaut”

Menge von Trainingsbeispielen:

<i>Beisp.</i>	<i>Sport</i>	<i>Art</i>	<i>Ort</i>	<i>Ebene</i>	<i>Tag</i>	<i>Anschauen</i>
X_1	Fußball	Mannschaft	draußen	national	Samstag	+
X_2	Hockey	Mannschaft	draußen	national	Samstag	+
X_3	Bodenturnen	Einzel	drinnen	Welt	Samstag	–
X_4	Handball	Mannschaft	drinnen	national	Samstag	+
X_5	Zehnkampf	Einzel	draußen	Welt	Sonntag	–

Wissensentdeckung

Beispielsprache: Vektor mit 5 Komponenten

(Sport, Art, Ort, Ebene, Tag)

und entsprechenden Werten.

Konzeptsprache: Vektor mit 5 Komponenten – *Sport, Art, Ort, Ebene, Tag* (Constraints zur Einschränkung möglicher Werte); für jedes Attribut gibt die Hypothese einen der drei folgenden Fälle an:

- "?" zeigt an, dass jeder Attributwert akzeptabel ist.
- Ein bestimmter **Attributwert** (z.B. *Zehnkampf*) spezifiziert, dass nur dieser Wert akzeptabel ist.
- "∅" zeigt an, dass kein Wert akzeptabel ist.

Wissensentdeckung

Wenn ein Beispiel e alle Constraints einer Hypothese h erfüllt, dann deckt h e ab, also $h(e) = 1$, anderenfalls gilt $h(e) = 0$.

Die Hypothese

$\langle \text{Fußball}, ?, ?, \text{national}, ? \rangle$

besagt, dass Paul nur Sportsendungen über Fußballspiele auf nationaler Ebene anschaut.

Die **allgemeinste** Hypothese wird durch $\langle ?, ?, ?, ?, ? \rangle$ repräsentiert (deckt alle Beispiele ab),

während die **speziellste** Hypothese $\langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle$ kein Beispiel abdeckt.

Wissensentdeckung

Ein Konzeptlernproblem hat die folgenden Komponenten:

- eine Beispielsprache L_E , in der Beispiele beschrieben werden;
- eine Konzeptsprache L_C , in der Konzepte beschrieben werden. Jede Konzeptbeschreibung $k \in L_C$ definiert eine Abbildung

$$k : L_E \rightarrow \{0, 1\}$$

- ein zu erlernendes Zielkonzept $c \in L_C$;
- eine Menge $P \subseteq L_E$ von positiven Beispielen für das zu erlernende Konzept, d.h., für alle $p \in P$ gilt $c(p) = 1$;
- eine Menge $N \subseteq L_E$ von negativen Beispielen, die von dem zu erlernenden Konzept nicht erfasst werden sollen, d.h., für alle $n \in N$ gilt $c(n) = 0$.

Wissensentdeckung

Ziel der Konzeptlernaufgabe:

Bestimme ein Konzept h aus der Konzeptsprache, so dass $h(e) = c(e)$ für alle Beispiele $e \in L_E$ ist.

Idealisierende Annahmen:

- P und N enthalten keine Fehler;
- L_C ist genügend mächtig, um das gesuchte Konzept auch tatsächlich ausdrücken zu können.

Wissensentdeckung

- $e_1 = \langle \text{Hockey}, \text{Mannschaft}, \text{draußen}, \text{Welt}, \text{Mittwoch} \rangle \in L_E$
- $h_1 = \langle ?, \text{Mannschaft}, ?, \text{Welt}, ? \rangle \in L_C$
Für die durch h_1 definierte Abbildung $h_1 : L_E \rightarrow \{0, 1\}$ gilt
 $h_1(e) = 1$ genau dann, wenn $\text{Art} = \text{Mannschaft}$ und $\text{Ebene} = \text{Welt}$
ist, wobei die übrigen Attribute Sport , Ort und Tag beliebige Werte
annehmen können.
- Für das zu erlernende Konzept $c : L_E \rightarrow \{0, 1\}$ gilt $c(e) = 1$ genau
dann, wenn "Paul Trops sich die durch e beschriebene Sportsendung
im Fernsehen anschaut."
- Menge der positiven Beispiele: $P = \{X_1, X_2, X_4\}$;
Menge der negativen Beispiele: $N = \{X_3, X_5\}$.

Wissensentdeckung

Suchraum: Menge aller möglichen Hypothesen

Sport-Beispiel:

- Anzahl möglicher Beispiele: $5 \cdot 2 \cdot 2 \cdot 2 \cdot 7 = 280$;

- Anzahl syntaktisch verschiedener Hypothesen:

$$7 \cdot 4 \cdot 4 \cdot 4 \cdot 9 = 4032;$$

- Anzahl semantisch verschiedener Hypothesen:

$$1 + 6 \cdot 3 \cdot 3 \cdot 3 \cdot 8 = 1297;$$

Wichtig zur Bewältigung des Suchproblems: Der Hypothesensuchraum muss – z.B. durch eine **partielle Ordnung** (d.h. eine reflexive, transitive und antisymmetrische Relation) – strukturiert werden.

Wissensentdeckung

Seien $h_1, h_2 : L_E \rightarrow \{0, 1\}$ zwei Konzepte.

- h_1 ist spezieller oder gleich h_2 (h_2 ist allgemeiner oder gleich h_1),

$$h_1 \leq h_2$$

gdw. $\forall e \in L_E (h_1(e) = 1 \Rightarrow h_2(e) = 1)$, wenn also jedes bezgl. h_1 positive Beispiel auch von h_2 abgedeckt wird und umgekehrt jedes bezgl. h_2 negative Beispiel auch von h_1 nicht abgedeckt wird.

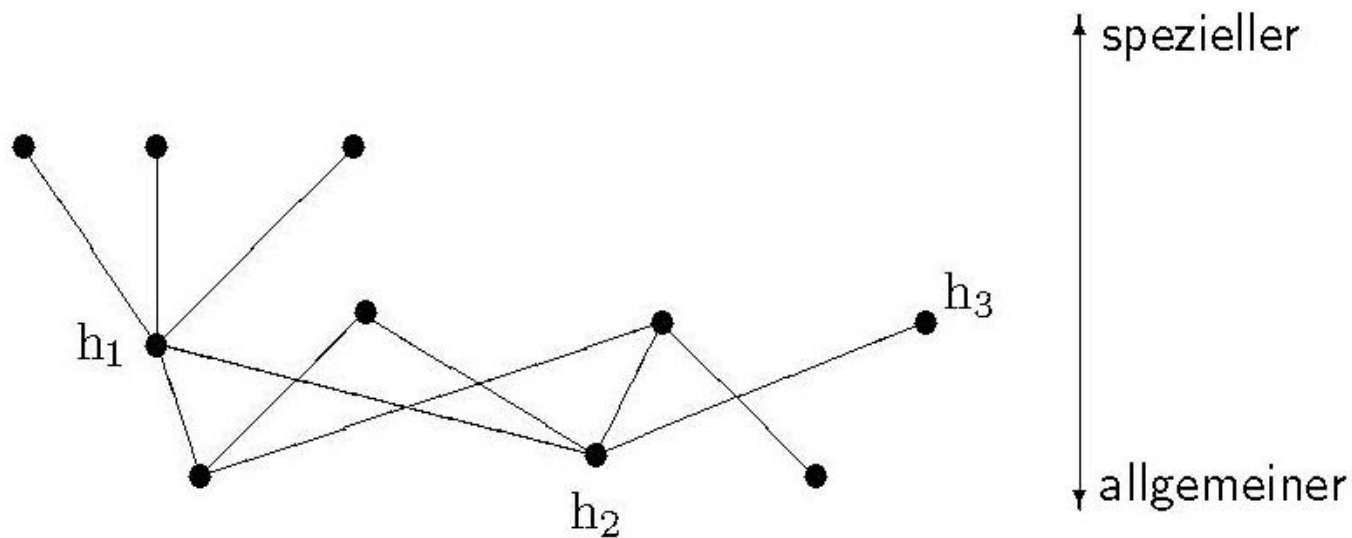
- h_1 ist (echt) spezieller h_2 bzw. h_2 ist (echt) allgemeiner als h_1

$$h_1 < h_2$$

gdw. $h_1 \leq h_2$ und $h_2 \not\leq h_1$.

Diese Definition ist unabhängig von der gewählten Sprache!

Wissensentdeckung



$h_1 = \langle \text{Fußball}, ?, ?, \text{national}, ? \rangle$

$h_2 = \langle \text{Fußball}, ?, ?, ?, ? \rangle$

$h_3 = \langle \text{Fußball}, ?, ?, ?, \text{Samstag} \rangle$

$h_1 \leq h_2, \quad h_3 \leq h_2$

h_1 und h_3 sind aber nicht miteinander vergleichbar!

Wissensentdeckung

Die Beispiele werden inkrementell zur Verfügung gestellt.

- Kandidaten-Eliminations-Methode:
Zu Beginn umfasst der Hypothesenraum H die ganze Menge L_C ; bei jedem neuen Beispiel e werden dann aus H alle Hypothesen entfernt, die nicht mit der vorgegebenen Klassifikation von e übereinstimmen.
- Suchrichtung speziell $\rightarrow$ allgemein:
Als initiale Hypothese h wird die speziellste Hypothese aus L_C genommen. h wird dann schrittweise bei jedem neuen positiven Beispiel e , das noch nicht von h abgedeckt wird, gerade soweit verallgemeinert, dass e mit abgedeckt wird. Als Ergebnis erhält man eine speziellste Hypothese, die konsistent mit den gegebenen Beispielen ist.

Wissensentdeckung

- Suchrichtung allgemein $\rightarrow$ speziell:
Als initiale Hypothese h wird die allgemeinste Hypothese genommen.
Bei jedem neuen negativen Beispiel e , das fälschlicherweise von h mit abgedeckt wird, muss h gerade so weit spezialisiert werden, dass e nicht mehr von h abgedeckt wird. Als Ergebnis erhält man eine allgemeinste Hypothese, die konsistent mit den angegebenen Beispielen ist.

Wissensentdeckung

Grundidee des Versionenraumverfahrens:

- sucht in beiden Richtungen “spezieller $\leftrightarrow$ allgemeiner” gleichzeitig;
- repräsentiert zu jedem Zeitpunkt die Menge *aller* Hypothesen, die konsistent mit der Trainingsmenge sind.

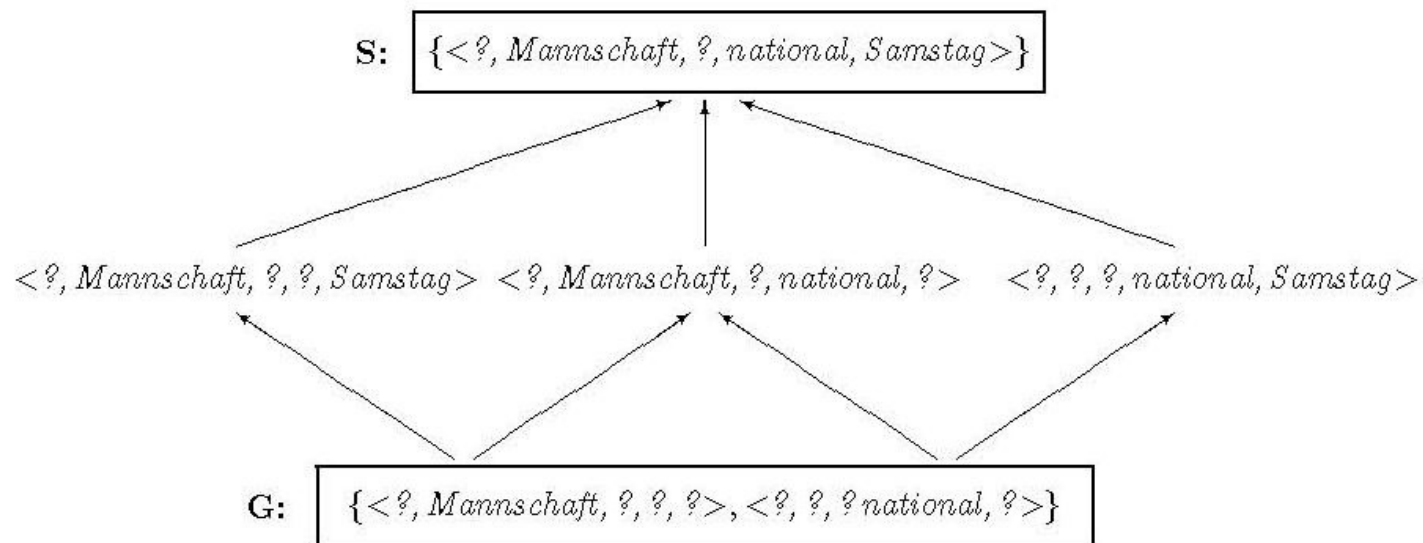
Sei B eine Menge von Trainingsbeispielen. Dann ist die Menge

$$V_B = \{h \in L_C \mid h \text{ ist konsistent bzgl. } B\}$$

der Versionenraum bzgl. der Beispielmenge B .

Das Versionenraum-Lernverfahren arbeitet prinzipiell nach der Kandidaten-Eliminations-Methode, wobei die einzelnen Hypothesen in kompakter Form repräsentiert werden, nämlich durch zwei Begrenzungsmengen: die speziellsten und die allgemeinsten Hypothesen.

Wissensentdeckung



Versionenraum mit 6 Hypothesen in partieller Ordnung, begrenzt durch die Mengen S und G .

Wissensentdeckung

Zu einer Beispielmengende B werden die Begrenzungsmengen wie folgt definiert:

- Ein Konzept h ist eine **speziellste** Generalisierung von B gdw. gilt:
 - ▶ h ist konsistent bzgl. B ;
 - ▶ es gibt kein Konzept h' , das konsistent bzgl. B ist und für das $h' < h$ gilt.

Ein Konzept h ist eine **allgemeinste** Generalisierung von B gdw. gilt:

- ▶ h ist konsistent bzgl. B ;
- ▶ es gibt kein Konzept h' , das konsistent bzgl. B ist und für das $h < h'$ gilt.

Wissensentdeckung

Als obere und untere Schranke eines Versionenraums für eine Beispielmengende B nehmen wir die Menge S der speziellsten und die Menge G der allgemeinsten Generalisierungen.

Das folgende Theorem präzisiert diese Darstellung:

Der Versionenraum enthält genau die Hypothesen, die in S , G oder "zwischen" S und G liegen.

Wissensentdeckung

Repräsentationstheorem für Versionenräume

Sei B eine Menge von Beispielen und

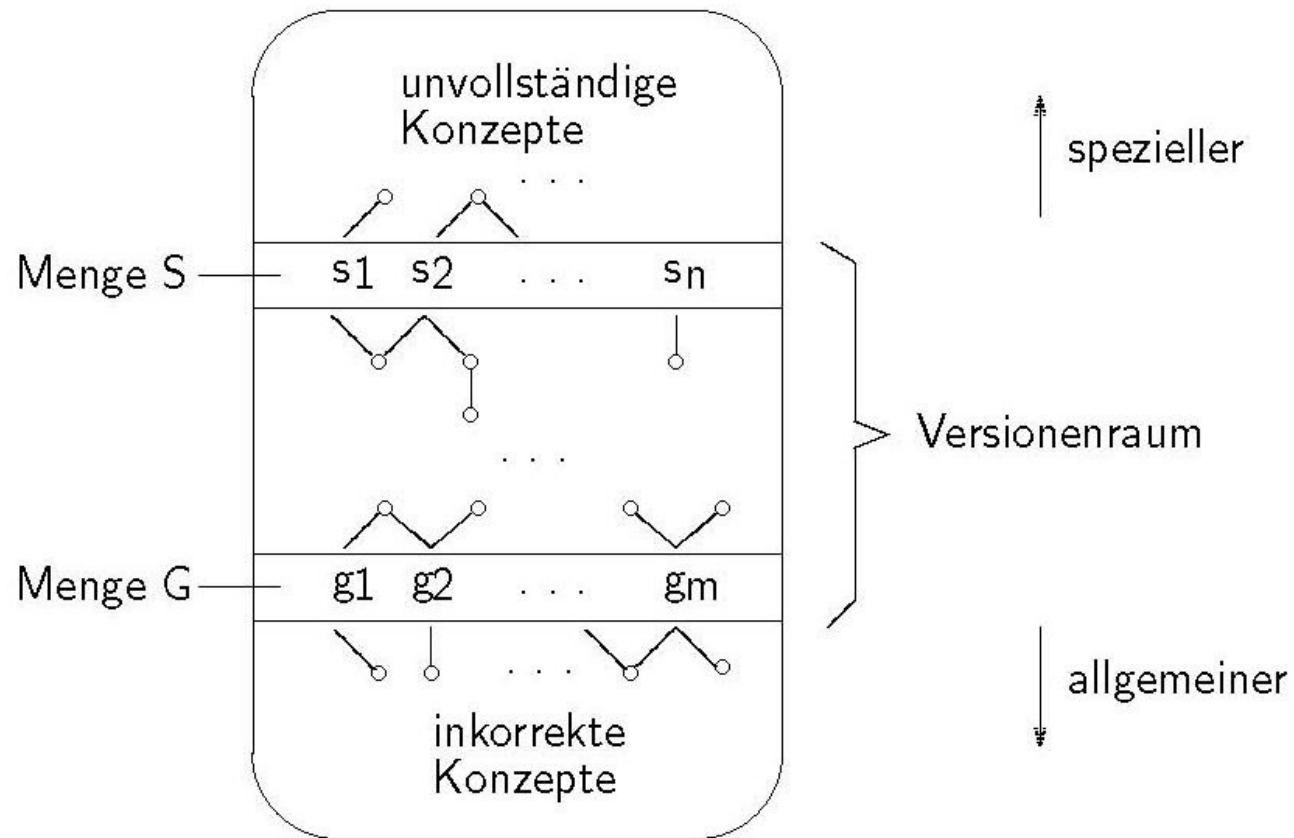
$$S = \{h \in L_C \mid h \text{ ist speziellste Generalisierung von } B\}$$

$$G = \{h \in L_C \mid h \text{ ist allgemeinste Generalisierung von } B\}$$

Für den Versionenraum V_B gilt:

$$V_B = \{h \in L_C \mid \exists s \in S \exists g \in G (s \leq h \leq g)\}$$

Wissensentdeckung



Wissensentdeckung

Mit jedem neuen Beispiel e müssen die Begrenzungsmengen S und G des aktuellen Versionenraums überprüft und gegebenenfalls angepasst werden.

Jede Hypothese $h \in S \cup G$ ist zu untersuchen:

- Stimmt h mit e über – d.h. es gilt $h(e) = 1$, falls e ein positives Beispiel ist, und $h(e) = 0$, falls e ein negatives Beispiel ist, – so ist nichts zu tun.
- Anderenfalls können im Prinzip zwei Fälle auftreten:
 - ① e ist für h fälschlicherweise negativ, d.h. $h(e) = 0$, obwohl e ein positives Beispiel ist.
 - ② e ist für h fälschlicherweise positiv, d.h. $h(e) = 1$, obwohl e ein negatives Beispiel ist.

Wissensentdeckung

- e ist für $s \in S$ fälschlicherweise positiv: d.h. s ist zu allgemein. Alle Hypothesen in S sind aber schon möglichst speziell, d.h. s kann nicht weiter spezialisiert werden, ohne die Vollständigkeit zu verlieren. s muss also aus S entfernt werden.
- e ist für $s \in S$ fälschlicherweise negativ: d.h. s ist zu speziell – s muss soweit verallgemeinert werden, dass e mit abgedeckt wird.
- e ist für $g \in G$ fälschlicherweise positiv: d.h. g ist zu allgemein – g muss soweit spezialisiert werden, dass e nicht mehr mit abgedeckt wird.
- e ist für $g \in G$ fälschlicherweise negativ: d.h. g ist zu speziell. Alle Hypothesen in G sind aber schon möglichst allgemein, d.h. g kann nicht weiter verallgemeinert werden, ohne die Korrektheit zu verlieren. g muss also aus G entfernt werden.

Wissensentdeckung

function *VS*

Eingabe: Konzeptlernaufgabe mit den Sprachen L_E und L_C
und Folge von Trainingsbeispielen

Ausgabe: Versionenraumrepräsentation aller Konzepte, die vollständig
und korrekt bzgl. der eingegebenen Beispiele sind

Initialisiere S zu der Menge der speziellsten Hypothesen aus L_C

Initialisiere G zu der Menge der allgemeinsten Hypothesen aus L_C

Wissensentdeckung

for each neues Trainingsbeispiel e **do**

- **if** e ist ein positives Beispiel **then**

- entferne aus G alle Hypothesen g mit $g(e) = 0$

- **for each** $h \in S$ mit $h(e) = 0$ **do**

- entferne h aus S

- füge zu S alle Hypothesen h' hinzu mit:

- h' ist *minimale Verallgemeinerung* von h bzgl. e und

- es gibt eine Hypothese $g \in G$ mit $h' \leq g$

- entferne aus S jede Hypothese, die (echt) allgemeiner als eine andere Hypothese in S ist

Wissensentdeckung

- **if** e ist ein negatives Beispiel **then**
 - entferne aus S alle Hypothesen s mit $s(e) = 1$
 - **for each** $h \in G$ mit $h(e) = 1$ **do**
 - entferne h aus G
 - füge zu G alle Hypothesen h' hinzu mit:
 - h' ist *minimale Spezialisierung* von h bzgl. e und
 - es gibt eine Hypothese $s \in S$ mit $s \leq h'$
 - entferne aus G jede Hypothese, die (echt) spezieller als eine andere Hypothese in G ist

Wissensentdeckung

Wichtig: Minimale Verallgemeinerungen (Spezialisierungen) erfolgen immer so, dass jede neue Hypothese h' in S (h' in G) immer noch spezieller (allgemeiner) oder gleich einer Hypothese aus G (S) ist.

VS terminiert in jedem Fall; dann liegt eine der folgenden drei Situationen vor:

- S ist leer und/oder G ist leer: In diesem Fall ist der Versionenraum ebenfalls zur leeren Menge kollabiert, d.h. es gibt keine konsistente Hypothese für die Trainingsbeispiele in dem vorgegebenen Hypothesenraum L_C .

Wissensentdeckung

- S und G sind identische einelementige Mengen: $S = G = \{h\}$. Das bedeutet, dass die Hypothese h das einzige Konzept aus L_C ist, das konsistent bzgl. der Trainingsmenge ist.
- Alle Beispiele sind bearbeitet, S und G sind beide nicht leer und enthalten unterschiedliche Hypothesen. In diesem Fall sind alle in dem durch S und G bestimmten Versionenraum liegenden Hypothesen konsistent bzgl. der Trainingsmenge.

Wissensentdeckung

Es gelten folgende allgemeine Beobachtungen:

- Jede Hypothese, die allgemeiner oder gleich einer Hypothese aus S ist, deckt alle bisherigen positiven Beispiele ab.
- Jede Hypothese, die spezieller oder gleich einer Hypothese aus G ist, deckt keines der bisherigen negativen Beispiele ab.

Wissensentdeckung

Initialisierung des Versionenraums für das Sportbeispiel:

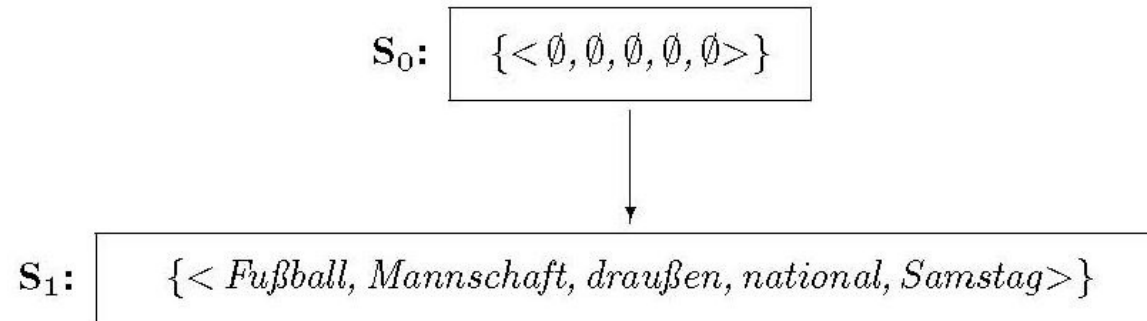
$$\begin{aligned} S = S_0 &= \{ \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle \} \\ G = G_0 &= \{ \langle ?, ?, ?, ?, ? \rangle \} \end{aligned}$$

$$S_0: \boxed{\{ \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle \}}$$

$$G_0: \boxed{\{ \langle ?, ?, ?, ?, ? \rangle \}}$$

Diese beiden Begrenzungsmengen umfassen den gesamten Versionenraum $V_\emptyset = L_C$.

Wissensentdeckung



$G_0, G_1:$ $\{ \langle ?, ?, ?, ?, ? \rangle \}$

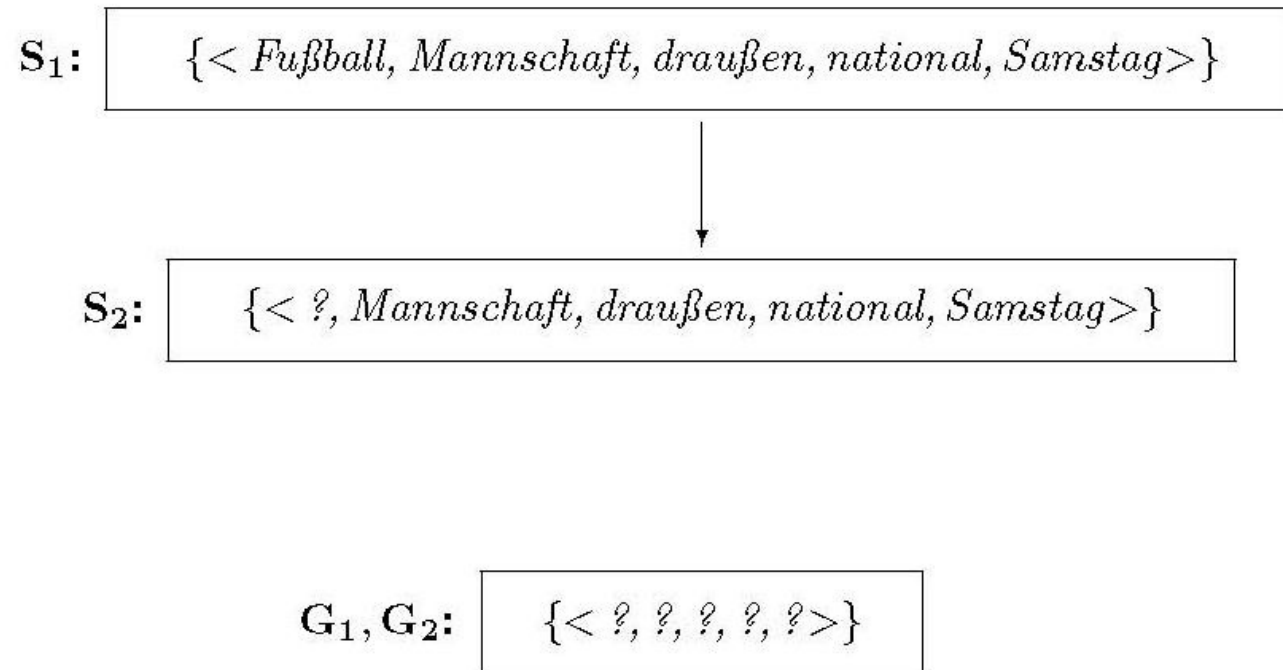
$X_1 = \langle \text{Fußball}, \text{Mannschaft}, \text{draußen}, \text{national}, \text{Samstag} \rangle, +$

Wissensentdeckung

$X_2 = \langle \text{Hockey, Mannschaft, drau\ss en, national, Samstag} \rangle, +$

wird von G_1 abgedeckt, aber nicht von S_1 .

Wissensentdeckung



Wissensentdeckung

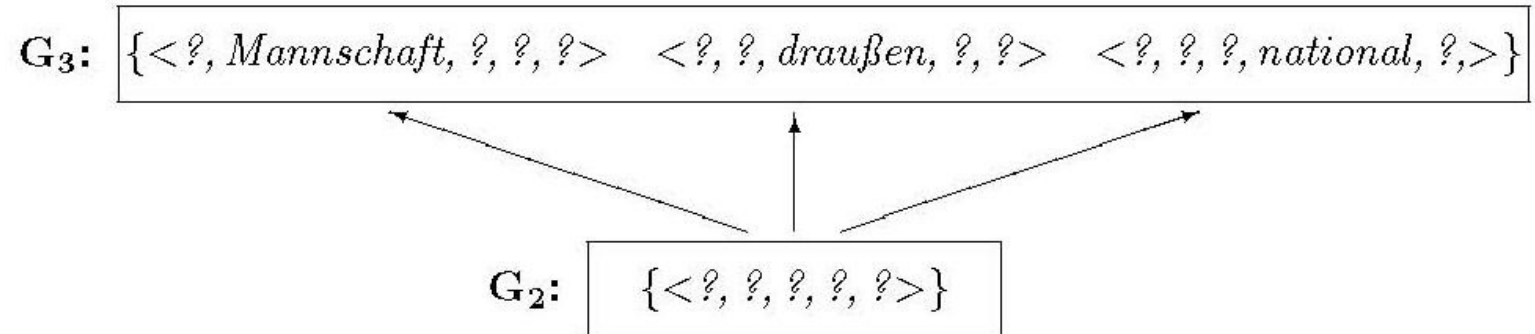
$X_3 = \langle \text{Bodenturnen, Einzel, drinnen, Welt, Samstag} \rangle, \quad -$

X_3 wird von S_2 nicht abgedeckt, aber fälschlicherweise von der in G_2 enthaltenen Hypothese $h_0 = \langle ?, ?, ?, ?, ? \rangle$

→ diese Hypothese muss spezialisiert werden.

Wissensentdeckung

$S_2, S_3: \{ \langle ?, Mannschaft, draußen, national, Samstag \rangle \}$



Wissensentdeckung

$X_4 = \langle \text{Handball}, \text{Mannschaft}, \text{drinnen}, \text{national}, \text{Samstag} \rangle, +$

Die in S_3 enthaltene Hypothese

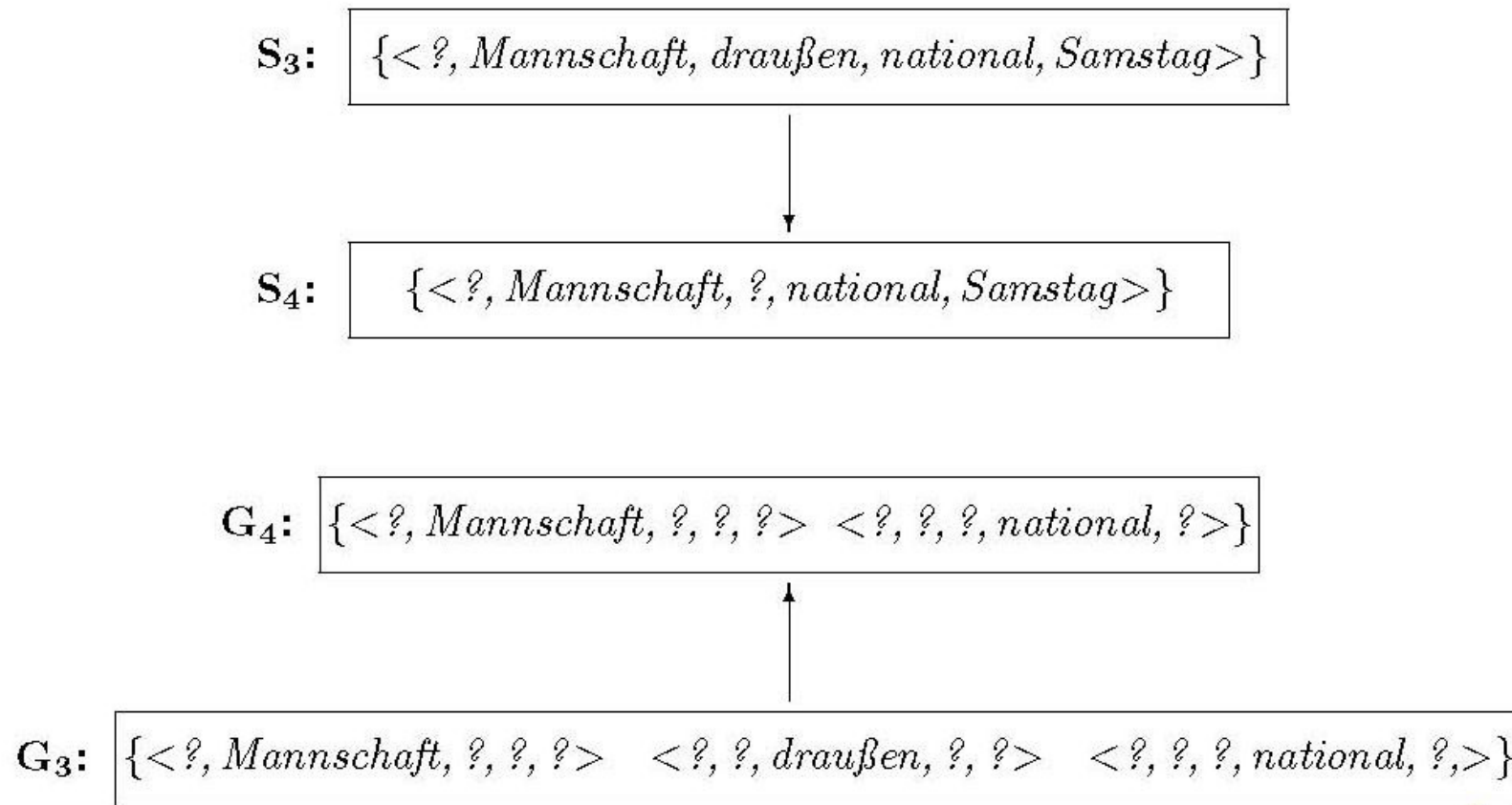
$\langle ?, \text{Mannschaft}, \text{draußen}, \text{national}, \text{Samstag} \rangle$

muss zu

$\langle ?, \text{Mannschaft}, ?, \text{national}, \text{Samstag} \rangle$

verallgemeinert werden, damit X_4 ebenfalls mit abgedeckt wird. Aus G_3 muss die Hypothese $\langle ?, ?, \text{draußen}, ?, ? \rangle$ entfernt werden, da sie X_4 nicht mit abdeckt.

Wissensentdeckung

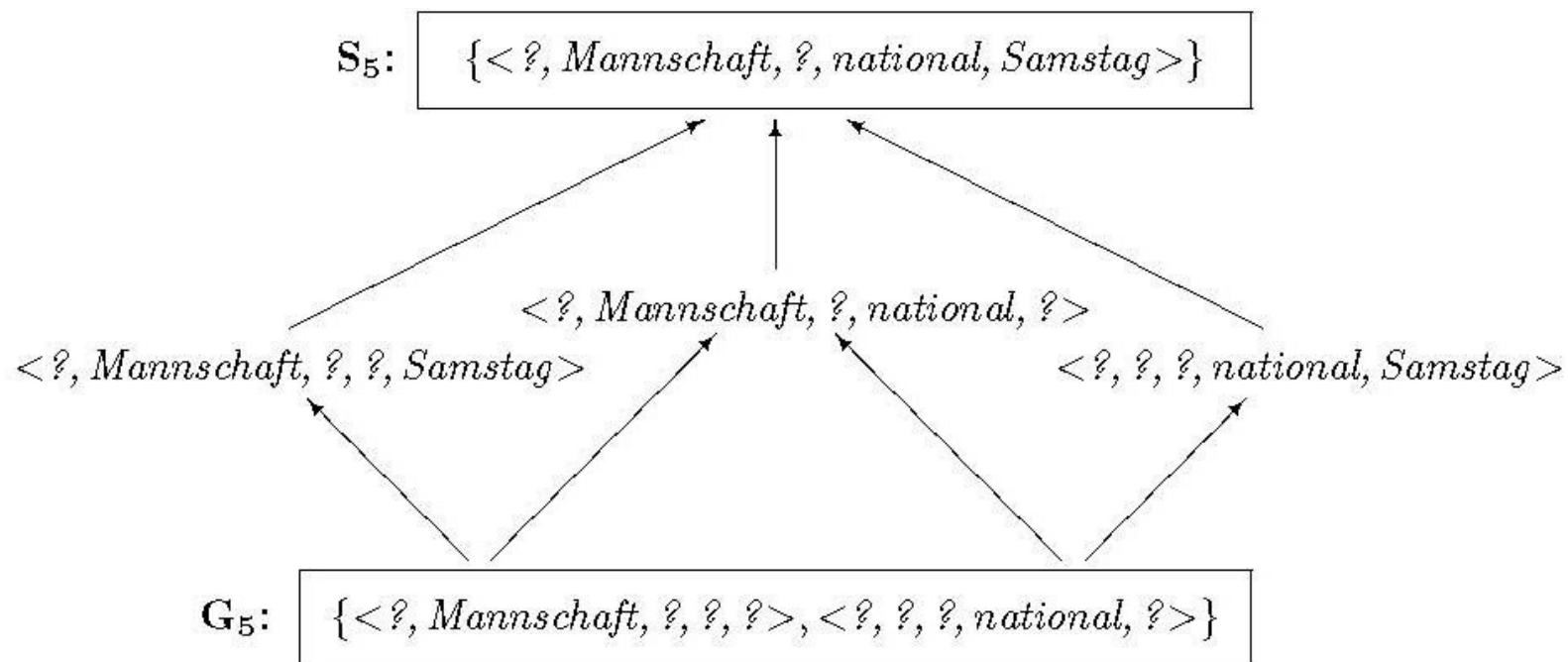


Wissensentdeckung

$X_5 = \langle \text{Zehnkampf}, \text{Einzel}, \text{draußen}, \text{Welt}, \text{Sonntag} \rangle, -$

X_5 ist negativ und wird von keiner der Hypothesen in S_4 oder G_4 abgedeckt, die Begrenzungsmengen bleiben also unverändert.

Wissensentdeckung



Wissensentdeckung

Wichtig ist, dass die Konzeptlernsprache L_C mächtig genug ist, um die gesuchte Hypothese auszudrücken.

Beispiel: Im Rahmen des Sportsendungenproblems sei die folgende Trainingsmenge gegeben:

<i>Beisp.</i>	<i>Sport</i>	<i>Art</i>	<i>Ort</i>	<i>Ebene</i>	<i>Tag</i>	<i>Anschauen</i>
X_1	Fußball	Mannschaft	draußen	national	Samstag	+
X_2	Fußball	Mannschaft	draußen	Europa	Samstag	+
X_3	Fußball	Mannschaft	draußen	Welt	Samstag	—

Wissensentdeckung

Die speziellste Hypothese, die konsistent mit den ersten beiden Beispielen X_1 und X_2 ist, ist

$$h = \langle \text{Fußball}, \text{Mannschaft}, \text{draußen}, ?, \text{Samstag} \rangle$$

h ist aber bereits zu allgemein, da sie X_3 fälschlicherweise auch abdeckt. Es wird also **keine Hypothese** gefunden, die konsistent bzgl. dieser drei Beispiele ist, denn es können keine Disjunktionen wie z.B.

$$\text{Ebene} = \text{national} \vee \text{Ebene} = \text{Europa}$$

ausgedrückt werden.

Wissensentdeckung

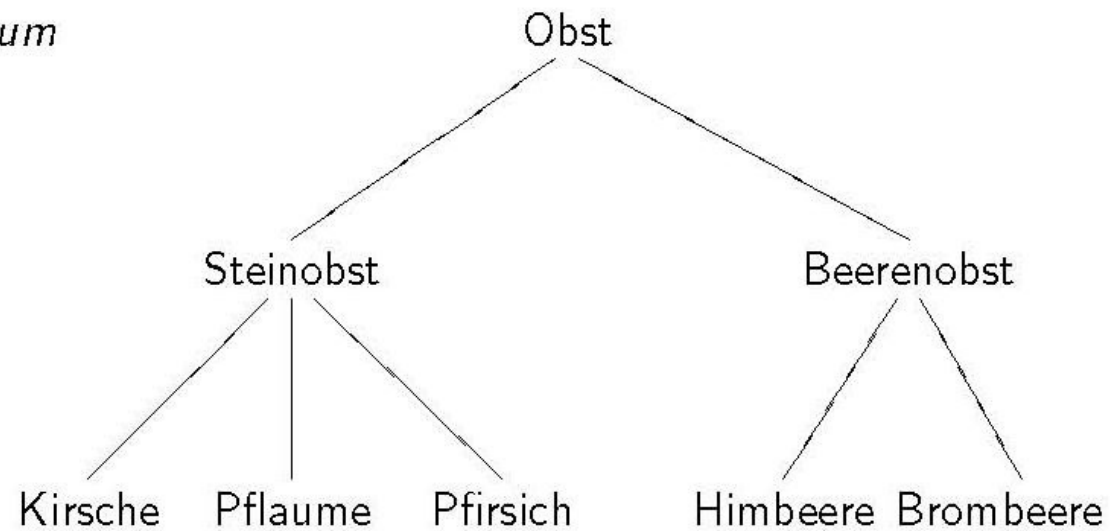
Jedes induktive Konzeptlernverfahren braucht irgendeine induktive Hypothese (inductive bias), um überhaupt ein Lernverhalten zu erzielen, das über die Klassifikation der vorgegebenen Trainingsbeispiele hinausgeht. In diesem Fall bestand die induktive Hypothese darin, nur konjunktive Hypothesen zu berücksichtigen.

Alternative:

Merkmalsbäume sind ein Mittel zur hierarchischen Strukturierung von Attributwerten, die eine geschicktere Verallgemeinerung beim Versionenraumlernen zulassen.

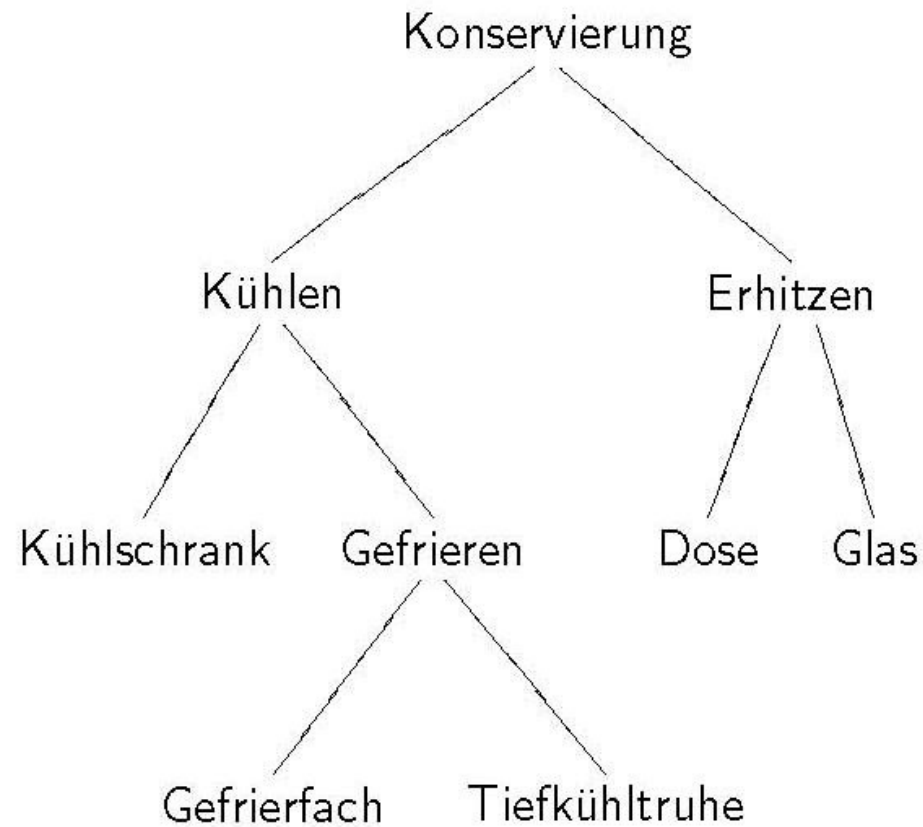
Wissensentdeckung

Merkmalsbaum
"Obst":



Wissensentdeckung

*Merkmalsbaum
"Konservierung":*



Wissensentdeckung

Beispielsprache: alle Paare (a, b) mit mit

$a \in \{Kirsche, Pflaume, Pfirsich, Himbeere, Brombeere\}$

$b \in \{Kühlschrank, Gefrierfach, Tiefkühltruhe, Dose, Glas\}$

Konzeptsprache: alle Paare $[A, B]$, wobei A ein Knoten aus dem "Obst"-Merkmalbaum und B ein Knoten aus dem "Konservierung"-Merkmalbaum ist, z.B. $[Steinobst, Glas]$, $[Pfirsich, Dose]$; außerdem $[\emptyset, \emptyset]$.

Wissensentdeckung

→ induzierte Abbildung:

$$[A, B] : \text{Beispielsprache} \rightarrow \{0, 1\}$$

$$[A, B](a, b) = 1 \text{ gdw.}$$

- A liegt in dem Merkmalsbaum "Obst" über a oder ist gleich a , und
- B liegt in dem Merkmalsbaum "Konservierung" über b oder ist gleich b .

Z.B. gilt

$$[\text{Steinobst}, \text{Glas}](\text{Kirsche}, \text{Glas}) = 1$$

$$[\text{Steinobst}, \text{Glas}](\text{Himbeere}, \text{Dose}) = 0$$

Wissensentdeckung

Trainingsmenge:

<i>Nr.</i>	<i>Beispiel</i>	<i>Klassifikation</i>
X_1	<i>(Himbeere, Kühlschrank)</i>	—
X_2	<i>(Kirsche, Glas)</i>	+

Initialisierung:

$$S_0 = \{[\emptyset, \emptyset]\}$$

$$G_0 = \{[Obst, Konservierung]\}$$

Wissensentdeckung

1. Trainingsbeispiel:

$$X_1 = (\textit{Himbeere}, \textit{Kühlschrank}) \text{ (negativ)}$$

Die Menge $S_0 = S_1$ muss nicht verändert werden.

$g_0 = [\textit{Obst}, \textit{Konservierung}]$ muss spezialisiert werden. Es gibt vier verschiedene minimale Spezialisierungen von g_0 bzgl. X_1 , die X_1 nicht mehr mit abdecken:

$$h_1 = [\textit{Steinobst}, \textit{Konservierung}]$$

$$h_2 = [\textit{Brombeere}, \textit{Konservierung}]$$

$$h_3 = [\textit{Obst}, \textit{Erhitzen}]$$

$$h_4 = [\textit{Obst}, \textit{Gefrieren}]$$

Wissensentdeckung

$S_0, S_1:$ $\{[\emptyset, \emptyset]\}$

$G_1:$ $\{[Steinobst, Konservierung], [Brombeere, Konservierung], [Obst, Erhitzen], [Obst, Gefrieren]\}$

$G_0:$ $\{[Obst, Konservierung]\}$

Trainingsbeispiel: $X_1 = (Himbeere, K\u00fchlschrank)$, *Klassifikation* = –

Wissensentdeckung

2. Trainingsbeispiel:

$$X_2 = (Kirsche, Glas) \text{ (positiv)}$$

S_1 muss verallgemeinert werden:

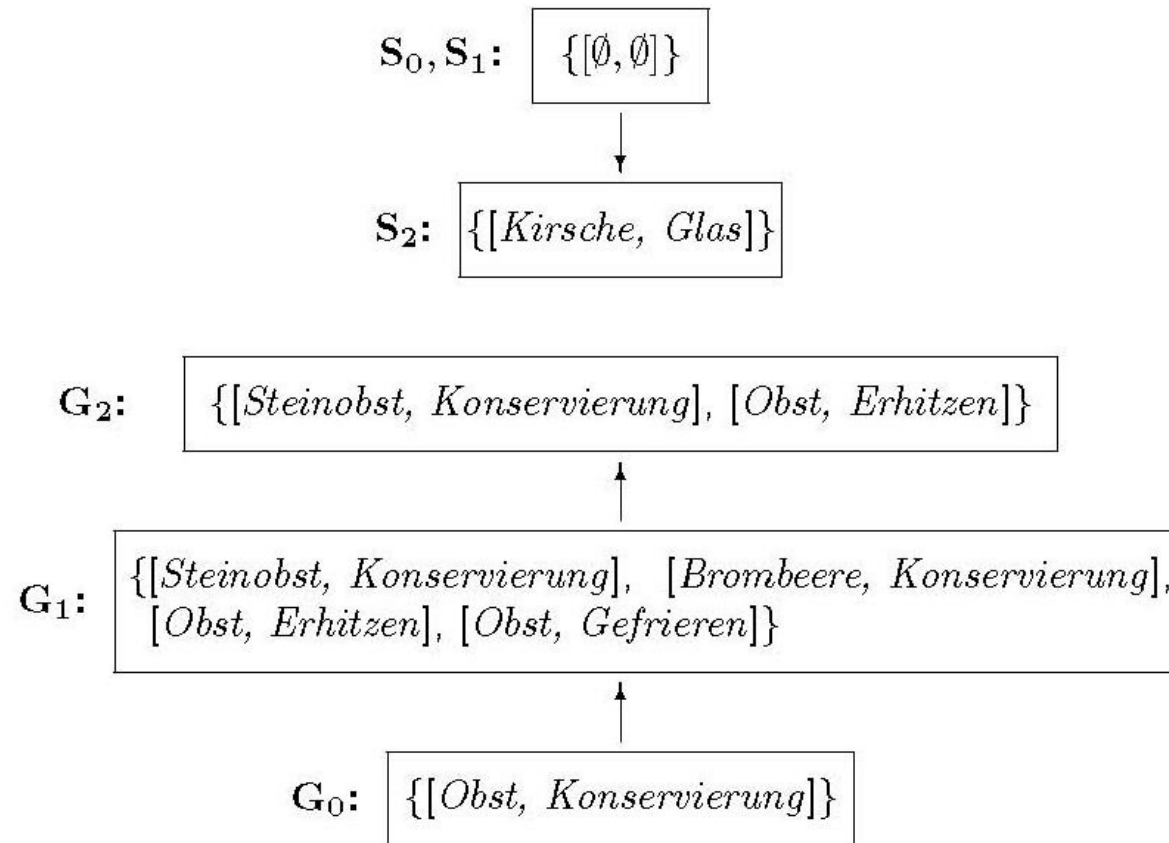
$$S_2 = \{[Kirsche, Glas]\}$$

und aus der Begrenzungsmenge G_1 müssen alle Hypothesen entfernt werden, die nicht allgemeiner sind als eine der Hypothesen in S_2 :

$$\begin{array}{llll} h_1 & = & [Steinobst, Konservierung] & \geq s_2 \\ h_2 & = & [Brombeere, Konservierung] & \not\geq s_2 \\ h_3 & = & [Obst, Erhitzen] & \geq s_2 \\ h_4 & = & [Obst, Gefrieren] & \not\geq s_2 \end{array}$$

h_2, h_4 müssen also aus G_1 entfernt werden.

Wissensentdeckung



Wissensentdeckung

Bisher sind wir nicht in der Lage, disjunktive Information zu repräsentieren:

Die beiden positiven Beispiele

X_1 : *Fußball Mannschaft draußen national Samstag*

X_2 : *Fußball Mannschaft draußen Europa Samstag*

haben als speziellste Verallgemeinerung bereits

$h = \langle \text{Fußball, Mannschaft, draußen, ?, Samstag} \rangle$

Wissensentdeckung

Besser wäre es, wenn man (unter Erweiterung der Sprache) Mengen statt einzelner Werte oder Symbole in den Vektorkomponenten zulassen würde, also z.B.

$$h' = \langle \{Fußball\}, \{Mannschaft\}, \{draußen\}, \\ \{national, Europa\}, \{Samstag\} \rangle$$

mit der logischen Repräsentation

$$h' = (Sport = Fußball) \wedge (Art = Mannschaft) \wedge (Ort = draußen) \\ \wedge (Ebene = national \vee Ebene = Europa) \wedge (Tag = Samstag)$$

Hypothesen dieser Art lassen sich also als Konjunktionen von Disjunktionen von Literalen auffassen

→ Klauseln, logische Programmierung

Wissensentdeckung

In logischen Programmen werden Regeln

$$B_1, \dots, B_n \Rightarrow H$$

in der Form

$$H \leftarrow B_1, \dots, B_n$$

geschrieben und entsprechen definiten Klauseln

$$[H, \neg B_1, \dots, \neg B_n] \equiv H \vee \neg B_1 \vee \dots \vee \neg B_n$$

(Klassische) Logische Programme entsprechen Mengen von (definiten) Klauseln.

Wissensentdeckung

Syntax für Beispiele und Konzepte: Klauseln und Klauselmengen
Zielsetzung:

Induktive Inferenz logischer Programme aus Beispielen $E = E^+ \cup E^-$ und Hintergrundwissen B

Induktion = Lernen aus Beispielen, Verallgemeinern.
Über die Bedeutung der Induktion schreibt W. Poundstone, Im Labyrinth des Denkens [Rowohlt 2000]:

Ohne sie hätten wir nichts als Trillionen von Einzelerfahrungen, von denen jede so isoliert und bedeutungslos wäre wie Konfetischnipsel. [...] Das Zusammenspiel von Induktion und Deduktion ist die Grundlage der wissenschaftlichen Methode.

Wissensentdeckung

$$B \wedge H \models E$$

- Deduktion: Erschließen von E aus B und H ;
- Abduktion = Generierung brauchbarer Hypothesen aus Beobachtungen: Erschließen von H aus B und E (= Beobachtung);
- Induktion = Generierung verallgemeinernder Hypothesen aus Beispielen: Erschließen von H aus B und E (= Beispielmenge);

Induktion = Abduktion + Begründung
 = Inversion der Deduktion

Wissensentdeckung

Berühmte induktive Logiker (u.a.):

Aristoteles, Carnap, Popper

Literatur zu ILP:

G.D. Plotkin, A note on inductive Generalization, 1970.

S. Muggleton and L. de Raedt, Inductive logic programming: theory and methods. *Journal of Logic Programming*, 19,20:629–679, 1994.

Wissensentdeckung

Hintergrundwissen B :

$grandfather(x, y) \leftarrow father(x, z), parent(z, y).$
 $father(henry, jane).$
 $mother(jane, john).$
 $mother(jane, alice).$

Positive Beispiele E^+ :

$grandfather(henry, john).$
 $grandfather(henry, alice).$

Negative Beispiele E^- :

$\leftarrow grandfather(john, henry).$
 $\leftarrow grandfather(alice, john).$

Zu lernendes Konzept: $parent$

Wissensentdeckung

Hintergrundwissen B :

$haswings(x) \leftarrow bird(x).$
 $hasbeak(x) \leftarrow bird(x).$
 $bird(x) \leftarrow vulture(x).$
 $carnivore(x) \leftarrow vulture(x).$

Positive Beobachtungen E^+ :

$haswings(Tweety).$
 $hasbeak(Tweety).$

Zu findende Hypothese: Was ist *Tweety* für ein Tier?

Wissensentdeckung

Gegeben seien

B Hintergrundwissen

E Evidenz/Beispiele/klassifizierte Beobachtungen:

E^+ Menge von positiven Beispielen

E^- Menge von negativen Beispielen

Ein weiterer **wichtiger Parameter** für ein ILP-Verfahren (ebenso wie für jedes ML- bzw. Konzeptlernverfahren) ist die **syntaktische Hypothese** (*syntactic bias*), die die syntaktische Form der zu findenden Hypothesen (oder generell den syntaktischen Rahmen) festlegt.

Typische syntaktische Form für

Beispiele : Fakten

Hypothesen: Mengen von definiten Klauseln (logische Programme)

Wissensentdeckung

Gesucht ist eine Hypothese H , so dass folgende Bedingungen erfüllt sind (wobei $\square$ der leeren Klausel bzw. einem elementaren Widerspruch entspricht):

- priori Erfüllbarkeit: $B \wedge E^- \not\models \square$, d.h. die negativen Beispiele stehen nicht im Widerspruch zu B ;
- posteriori Erfüllbarkeit: $B \wedge H \wedge E^- \not\models \square$, d.h. die negativen Beispiele stehen nicht im Widerspruch zu $B \wedge H$ ($\rightarrow$ Konsistenz);
- priori Notwendigkeit: $B \not\models E^+$, d.h. B genügt nicht, um die positiven Beispiele zu erklären;
- posteriori Hinlänglichkeit: $B \wedge H \models E^+$, d.h. zusammen mit B erklärt H die positiven Beispiele ($\rightarrow$ Vollständigkeit).

Wissensentdeckung

$B :$ $grandfather(x, y) \leftarrow father(x, z), parent(z, y).$
 $father(henry, jane).$
 $mother(jane, john).$
 $mother(jane, alice).$

$E^+ :$ $grandfather(henry, john).$
 $grandfather(henry, alice).$

$E^- :$ $\leftarrow grandfather(john, henry).$
 $\leftarrow grandfather(alice, john).$

Hypothese: $H = parent(x, y) \leftarrow mother(x, y).$

Wissensentdeckung

Überprüfung der formalen Bedingungen:

- $B \wedge E^- \not\models \square$, d.h. die negativen Beispiele stehen nicht im Widerspruch zum Hintergrundwissen (priori Erfüllbarkeit);
- $B \not\models E^+$, d.h. es ist priori Notwendigkeit gegeben;
- $B \wedge H \wedge E^- \not\models \square$ – die Hypothese ist konsistent mit Hintergrundwissen und negativen Beispielen (posteriori Erfüllbarkeit);
- $B \wedge H \models E^+$ – die Hypothese erklärt die positiven Beispiele (posteriori Hinlänglichkeit).

Wissensentdeckung

Modelltheoretisch lassen sich die obigen Bedingungen folgendermaßen formulieren (wobei $M^+(T)$ das minimale Herbrandmodell eines klassischen logischen Programms T bezeichnet):

- priori Erfüllbarkeit – $B \wedge E^- \not\models \square$: alle negativen Beispiele $e \in E^-$ sind falsch in $M^+(B)$;
- posteriori Erfüllbarkeit – $B \wedge H \wedge E^- \not\models \square$: alle negativen Beispiele $e \in E^-$ sind falsch in $M^+(B \wedge H)$;
- priori Notwendigkeit – $B \not\models E^+$: einige positive Beispiele $e \in E^+$ sind falsch in $M^+(B)$;
- posteriori Hinlänglichkeit – $B \wedge H \models E^+$: alle positiven Beispiele $e \in E^+$ sind wahr in $M^+(B \wedge H)$.

Wissensentdeckung

$B :$ $haswings(x) \leftarrow bird(x).$
 $hasbeak(x) \leftarrow bird(x).$
 $bird(x) \leftarrow vulture(x).$
 $carnivore(x) \leftarrow vulture(x).$

$E^+ :$ $haswings(Tweety).$
 $hasbeak(Tweety).$

Mögliche Hypothesen:

H_1 $bird(Tweety).$

H_2 $vulture(Tweety).$

Wissensentdeckung

Überprüfung der formalen Bedingungen:

- $E^- = \emptyset$ und $B \not\models \square$, also ist priori Erfüllbarkeit gegeben;
- $B \not\models E^+$, also ist priori Notwendigkeit gegeben;
- $B \wedge H_i \not\models \square$ – beide Hypothesen sind konsistent mit dem Hintergrundwissen (posteriori Erfüllbarkeit);
- $B \wedge H_i \models E^+$ – damit erklären beide Hypothesen die Beispiele (posteriori Hinlänglichkeit).

Beide Hypothesen erfüllen alle Bedingungen, die Hypothese
 $H_1 = \text{bird}(\text{Tweety})$.

ist jedoch plausibler, da weniger speziell.

Wie lassen sich brauchbare Hypothesen effizient finden bzw. effektiv konstruieren?

Wissensentdeckung

Beim Konzeptlernen hilft die “spezieller als”-Relation, den Hypothesenraum zu strukturieren. Beim ILP entspricht der Spezialisierung die Deduktion und der Generalisierung die Induktion:

Eine Hypothese G ist allgemeiner als eine Hypothese S (S ist dann spezieller als G), wenn gilt:

$$G \models S \text{ und } S \not\models G$$

Bemerkung. “Allgemeiner als” wird hier also nicht im logischen Sinn verstanden, sondern gerade entgegengesetzt: Im logischen Sinn folgt aus einer spezielleren Aussage die allgemeinere.

Wissensentdeckung

Mit Hilfe dieser Relation lässt sich der Hypothesenraum einschränken:

- Kriterium der posteriori Hinlänglichkeit:
Wenn $B \wedge H \not\models E^+$, dann ist die posteriori Hinlänglichkeit auch für alle Spezialisierungen H' von H verletzt (denn aus $B \wedge H' \models E^+$ und $H \models H'$ würde auch $B \wedge H \models E^+$ folgen);
- Kriterium der posteriori Erfüllbarkeit:
Wenn $B \wedge H \wedge E^- \models \square$, dann verletzen auch alle Generalisierungen H' von H die posteriori Erfüllbarkeit (denn $B \wedge H' \wedge E^- \models B \wedge H \wedge E^- \models \square$).

Wissensentdeckung

Es gibt (im Wesentlichen) folgende Möglichkeiten, um Hypothesen (logische Programme) zu spezialisieren:

- ① Ersetzen von Variablen durch Terme;
- ② Erweitern des Regelrumpfes um Literale;
- ③ Entfernen einer Klausel aus dem Programm.

Die inversen Operationen dienen der Verallgemeinerung von Hypothesen. Insbesondere gilt:

Sind C_1 und C_2 zwei Klauseln mit $C_1 \subseteq C_2$, so ist C_1 eine Verallgemeinerung von C_2 .

(wegen Punkt 2 oben)

Wissensentdeckung

Dies ist nur ein kurzer Einblick in ILP!

- Zur Beschränkung des Hypothesenraumes nach unten wird eine speziellste Klausel aus dem ersten Beispiel berechnet.
- Sog. *mode*-Deklarationen legen fest, welche Prädikatssymbole in Kopf und Rumpf der Hypothesen auftreten können, und in welcher Rolle Variablen vorkommen können.
- Im Allgemeinen werden Länge der Klauseln und die Tiefe der vorkommenden Variablen beschränkt.
- Die Güte der erzeugten Hypothesen wird numerisch bewertet; insbesondere kann man probabilistisch das Maß an Glauben quantifizieren, das man einer bestimmten Hypothese zuordnet.

Wissensentdeckung

Durch die fortschreitende Automatisierung fallen immense Mengen an Daten an ...

Knowledge Discovery in Databases (KDD) ...

... bezeichnet den Prozess, neues, nützliches und interessantes Wissen aus Daten herauszufiltern und in verständlicher Form zu präsentieren.

Data Mining (DM) ...

bezeichnet die konkrete Anwendung von Algorithmen zur Wissensfindung in Daten.

Wissensentdeckung

... umfasst die folgenden Schritte:

- Hintergrundwissen und Zielsetzung: Bereitstellung von bereichsspezifischem Wissen und Definition der Ziele des durchzuführenden KDD-Prozesses.
- Datenauswahl: Festlegung der zu untersuchenden Datenmenge.
- Datenbereinigung: Herausfiltern von Ausreißern und Rauscheffekten, Behandlung fehlender Daten.
- Datenreduktion und -projektion: Komprimierung der Datenmenge durch Reduktion und/oder Transformation der Variablen.

Wissensentdeckung

- Modellfunktionalität: Welchem Zweck dient das Data Mining? – Klassifikation, Clustering, Regressionsanalyse etc.
- Verfahrenswahl: Bestimmung des Data Mining-Verfahrens.
- Data Mining: der eigentliche Data Mining-Prozess.
- Interpretation: Aufbereitung der gefundenen Information.

Wissensentdeckung

- verständliche Form $\rightarrow$ Assoziationsregeln

$$r : A_1 \dots A_n \rightarrow B_1 \dots B_m$$

Beispiel:

"80 % der Kunden, die Bier und Sekt kaufen, kaufen auch Kartoffelchips"

- nützlich und interessant $\rightarrow$ Wahrscheinlichkeiten nahe 1
 - ▶ **Support** $\text{support}(r) = \text{Prob}(A_1 \dots A_n, B_1 \dots B_m)$
 - ▶ **Konfidenz** $\text{confidence}(r) = \text{Prob}(B_1 \dots B_m | A_1 \dots A_n)$

Finde Assoziationsregeln mit hohem Support und hoher Konfidenz!

Wissensentdeckung

- Item = Artikel, Ding, Gegenstand;
 $\mathcal{I}$ Menge aller betrachteten Items
- $X \subseteq \mathcal{I}$ Itemmenge; $|X| = k$: k-Itemmenge
- Transaktion $t \subseteq \mathcal{I}$ Itemmenge;
Datenbasis $\mathcal{D} = \{t_1, t_2, \dots\}$ Menge von Transaktionen
- Support einer Itemmenge

$$\text{support}(X) = \frac{|\{t \in \mathcal{D} \mid X \subseteq t\}|}{|\mathcal{D}|}$$

Wissensentdeckung

Anwendungsbereich: Verkaufsdatenanalyse eines Supermarktes

Items: Artikel des Marktsortimentes

Transaktionen: Kundeneinkäufe

Datenbasis $\mathcal{D}$: alle Verkaufstransaktionen eines bestimmten Zeitraums

Support der Itemmenge $\{Milch\}$: Prozentsatz der Kunden, die bei ihrem Einkauf auch Milch einkauften.

Wissensentdeckung

Eine Assoziationsregel hat die Form

$$X \rightarrow Y$$

wobei X und Y disjunkte Itemmengen sind.

Der Support einer Assoziationsregel ist definiert als der Support von Prämisse + Konklusion:

$$\text{support}(X \rightarrow Y) = \text{support}(X \cup Y)$$

Die Konfidenz einer Assoziationsregel $X \rightarrow Y$ ist der (relative) Anteil derjenigen der X enthaltenden Transaktionen, die auch Y enthalten:

$$\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \rightarrow Y)}{\text{support}(X)}$$

Wissensentdeckung

Finde alle Assoziationsregeln, die in der betrachteten Datenbasis mit einem Support von mindestens *minsupp* und einer Konfidenz von mindestens *minconf* gelten, wobei *minsupp* und *minconf* benutzerdefinierte Werte sind.

- Finde alle Itemmengen, deren Support über der *minsupp*-Schwelle liegt; diese Itemmengen werden häufige Itemmengen (frequent itemsets) genannt.
- Finde in jeder häufigen Itemmenge I alle Assoziationsregeln

$$I' \rightarrow (I - I') \text{ mit } I' \subset I,$$

deren Konfidenz mindestens *minconf* beträgt.

Wissensentdeckung

$|\mathcal{I}| = n \rightarrow 2^n$ Teilmengen sind zu untersuchen

Idee des Apriori-Algorithmus basiert auf:

Alle Teilmengen einer häufigen Itemmenge sind ebenfalls häufig, und alle Obermengen einer nicht-häufigen Itemmenge sind ebenfalls nicht häufig.

Dies ist klar, denn für zwei Itemmengen I_1, I_2 gilt:

$$I_1 \subseteq I_2 \text{ impliziert } \text{support}(I_2) \leq \text{support}(I_1)$$

Wissensentdeckung

Eingabe : Datenbasis $\mathcal{D}$

Ausgabe: Menge L_k häufiger Itemmengen I ,
: d.h. $\text{support}(I) \geq \text{minsupp}$

- Berechnung häufiger 1-Itemmengen L_1
- $k - 1 \rightarrow k$:
 - ▶ Berechnung von Kandidatenmengen C_k aus den (k-1)-häufigen Itemmengen mittels **AprioriGen**
 - ▶ Berechnung des tatsächlichen Supports der Kandidatenmengen
 - ▶ Aufnahme der Mengen mit genügend hohem Support (mindestens minsupp) in L_k
- Ausgabe von $\bigcup L_k$

Wissensentdeckung

... nutzt aus, dass jede $(k - 1)$ -Teilmenge einer Menge aus L_k in L_{k-1} enthalten sein muss; dabei wird angenommen, dass die Items geordnet sind:

Eingabe : Menge häufiger $(k - 1)$ -Itemmengen L_{k-1}

Ausgabe: Obermenge C_k der Menge häufiger k -Itemmengen

- ① Je zwei häufige $(k - 1)$ -Itemmengen $p, q \in L_{k-1}$ mit denselben ersten $(k - 2)$ Elementen werden zu einer k -Itemmenge vereinigt $\rightarrow C_k$
- ② Entferne aus C_k diejenigen Itemmengen, deren $(k - 1)$ -Teilmengen nicht alle in L_{k-1} liegen.

Wissensentdeckung

Items: $A < B < C < \dots$

$$L_3 = \{\{A, B, C\}, \{A, B, D\}, \{A, C, D\}, \{A, C, E\}, \{B, C, D\}\}$$

- ① $C_4: \{A, B, C, D\}, \{A, C, D, E\}$
- ② Entferne $\{A, C, D, E\}$, da $\{A, D, E\} \notin L_3$.

Ausgabe: $C_4 = \{\{A, B, C, D\}\}$

Beachten Sie: $\{A, B, C, E\}$ wird **nicht** in C_4 aufgenommen, da $\{A, B, C\}$ und $\{A, B, E\}$ nicht beide in L_3 enthalten sind.

Wissensentdeckung

Eingabe : Datenbasis $\mathcal{D}$

Ausgabe: Menge L_k häufiger Itemmengen I ,
: d.h. $\text{support}(I) \geq \text{minsupp}$

- Berechnung häufiger 1-Itemmengen L_1
- $k - 1 \rightarrow k$:
 - ▶ Berechnung von Kandidatenmengen C_k aus den (k-1)-häufigen Itemmengen mittels **AprioriGen**
 - ▶ Berechnung des tatsächlichen Supports der Kandidatenmengen
 - ▶ Aufnahme der Mengen mit genügend hohem Support (mindestens minsupp) in L_k
- Ausgabe von $\bigcup L_k$

Wissensentdeckung

... nutzt aus, dass jede $(k - 1)$ -Teilmenge einer Menge aus L_k in L_{k-1} enthalten sein muss; dabei wird angenommen, dass die Items geordnet sind:

Eingabe : Menge häufiger $(k - 1)$ -Itemmengen L_{k-1}

Ausgabe: Obermenge C_k der Menge häufiger k -Itemmengen

- ① Je zwei häufige $(k - 1)$ -Itemmengen $p, q \in L_{k-1}$ mit denselben ersten $(k - 2)$ Elementen werden zu einer k -Itemmenge vereinigt $\rightarrow C_k$
- ② Entferne aus C_k diejenigen Itemmengen, deren $(k - 1)$ -Teilmengen nicht alle in L_{k-1} liegen.

Wissensentdeckung

Aus den häufigen Itemmengen müssen noch die gesuchten Assoziationsregeln mit einer Konfidenz $\geq \text{minconf}$ bestimmt werden. Dabei nutzt man folgenden Zusammenhang aus:

Beträgt für Itemmengen X, Y mit $Y \subset X$ die Konfidenz einer Regel $(X - Y) \rightarrow Y$ mindestens minconf , so gilt dies auch für jede Regel der Form $(X - Y') \rightarrow Y'$ mit $Y' \subseteq Y$.

Erfüllt also eine Assoziationsregel das Konfidenzkriterium, so auch alle Regeln, die sich aus denselben Items und mit kürzerer Konklusion bilden lassen.

→ Bilde zuerst Assoziationsregeln mit möglichst kurzer Konklusion und erweitere die Konklusion schrittweise.

Wissensentdeckung

Es gilt nämlich für Itemmengen X, Y, Y' mit $Y' \subseteq Y \subset X$:

$$\text{confidence}((X - Y') \rightarrow Y') \geq \text{confidence}((X - Y) \rightarrow Y)$$

$$\text{wegen } \text{confidence}((X - Y') \rightarrow Y') = \frac{\text{support}(X)}{\text{support}(X - Y')}$$

$$\text{confidence}((X - Y) \rightarrow Y) = \frac{\text{support}(X)}{\text{support}(X - Y)},$$

$$\text{da } (X - Y) \cup Y = X = (X - Y') \cup Y',$$

und aus $(X - Y') \supseteq (X - Y)$ folgt

$$\text{support}(X - Y') \leq \text{support}(X - Y)$$

Wissensentdeckung

.. aus einer häufigen Itemmenge X :

- Berechne alle Assoziationsregeln mit genügend großer Konfidenz, deren Konklusion nur ein Item enthält.
- Ist H_m eine Menge von m -Item-Konklusionen von X , so setze

$$H_{m+1} := \text{AprioriGen}(H_m)$$

Für alle Konklusionen $h_{m+1} \in H_{m+1}$ überprüft man nun die Konfidenz der Regel

$$(X - h_{m+1}) \rightarrow h_{m+1};$$

liegt sie über der Schwelle minconf , so wird die Regel ausgegeben, andernfalls wird h_{m+1} aus H_{m+1} entfernt.

Wissensentdeckung

Einkaufstransaktionen in einem Drogeriemarkt:

<i>Label</i>	<i>Artikel</i>	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	<i>support</i>
A	Seife	•				•		•		•		0.4
B	Shampoo	•	•	•	•		•		•	•	•	0.8
C	Haarspülung		•	•	•		•		•	•		0.6
D	Duschgel	•			•		•	•		•	•	0.6
E	Zahnpasta	•		•		•		•				0.4
F	Zahnbürste			•		•						0.2
G	Haarfärbemittel		•		•				•			0.3
H	Haargel		•									0.1
J	Deodorant			•	•	•	•	•	•			0.6
K	Parfüm						•		•			0.2
L	Kosmetikartikel		•		•		•		•		•	0.5

Wissensentdeckung

Festlegung des minimalen Supports und der minimalen Konfidenz:

$$\text{minsupp} = 0.4, \quad \text{minconf} = 0.7$$

In realen Anwendungen wird *minsupp* in der Regel sehr viel kleiner gewählt (oft $< 1\%$).

Bestimmung der häufigen k -Itemmengen:

$k=1$:

$$L_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{J\}, \{L\}\}$$

$k=2$:

Berechnung der Kandidatenmenge C_2 für L_2 durch paarweise Kombinationen der Mengen in L_1 :

Wissensentdeckung

C_2 -Menge	Support	C_2 -Menge	Support	C_2 -Menge	Support
{A,B}	0.2	{B,D}	0.5	{C,L}	0.4
{A,C}	0.1	{B,E}	0.2	{D,E}	0.2
{A,D}	0.2	{B,J}	0.4	{D,J}	0.3
{A,E}	0.3	{B,L}	0.5	{D,L}	0.3
{A,J}	0.2	{C,D}	0.3	{E,J}	0.3
{A,L}	0.0	{C,E}	0.1	{E,L}	0.0
{B,C}	0.6	{C,J}	0.4	{J,L}	0.3

Bemerkung: Kein Teilmengencheck, da per Konstruktion alle betrachteten 1-Teilmengen auch in L_1 liegen.

Wissensentdeckung

Also ist

$$L_2 = \{\{B, C\}, \{B, D\}, \{B, J\}, \{B, L\}, \{C, J\}, \{C, L\}\}$$

k=3:

Berechnung der Kandidatenmenge C_3 (zum Vergleich mit und ohne Teilmengencheck, d.h. alle 2-Teilmengen müssen in L_2 liegen):

C_3 vor Teilmengencheck	C_3 nach Teilmengencheck	Support
{B,C,D}	{B,C,J}	0.4
{B,C,J}	{B,C,L}	0.4
{B,C,L}		
{B,D,J}		
{B,D,L}		
{B,J,L}		
{C,J,L}		

Wissensentdeckung

Damit ist

$$L_3 = \{\{B, C, J\}, \{B, C, L\}\}$$

Die einzig mögliche weitere Kombination $\{B, C, J, L\}$ ist nicht häufig, da (z. B.) $\{C, J, L\}$ nicht in L_3 enthalten ist; folglich ist

$$C_4 = L_4 = \emptyset$$

Berechnung der Assoziationsregeln aus den häufigen Itemmengen:

Es bezeichne (wie oben) H_m die Menge der m -Item-Konklusionen der jeweils betrachteten häufigen Itemmenge.

Wissensentdeckung

Regeln der Länge 2, d.h. aus

$$L_2 = \{\{B, C\}, \{B, D\}, \{B, J\}, \{B, L\}, \{C, J\}, \{C, L\}\}$$

gebildete Regeln:

<i>Regel</i>	<i>Konfidenz</i>	<i>Regel</i>	<i>Konfidenz</i>
$B \rightarrow C$	0.75	$C \rightarrow B$	1.00
$B \rightarrow D$	0.63	$D \rightarrow B$	0.83
$B \rightarrow J$	0.50	$J \rightarrow B$	0.67
$B \rightarrow L$	0.63	$L \rightarrow B$	1.00
$C \rightarrow J$	0.67	$J \rightarrow C$	0.67
$C \rightarrow L$	0.67	$L \rightarrow C$	0.80

Wissensentdeckung

Regeln der Länge 3, d.h. aus

$$L_3 = \{\{B, C, J\}, \{B, C, L\}\}$$

gebildete Regeln:

Aus $\{B, C, J\}$ (mit $H_1 = \{B, C, J\}$) entstehen die folgenden Regeln:

$$BC \rightarrow J [0.67], \quad BJ \rightarrow C [1.00], \quad CJ \rightarrow B [1.00]$$

Folglich wird J aus H_1 entfernt:

$$H_1 := \{B, C\}$$

$$H_2 := \text{AprioriGen}(H_1) = \{\{B, C\}\}$$

führt auf die Regel

$$J \rightarrow BC [0.67],$$

deren Konfidenz allerdings unterhalb der gewählten Schwelle liegt.

Wissensentdeckung

Aus $\{B, C, L\}$ erhält man die Regeln

$$BC \rightarrow L [0.67], \quad BL \rightarrow C [0.8], \quad CL \rightarrow B [1.00]$$

und durch Erweiterung der Konklusion noch

$$L \rightarrow BC [0.8]$$

die ebenfalls ausgegeben wird.

Durch den Apriori-Algorithmus werden insgesamt folgende Regeln berechnet:

Wissensentdeckung

<i>Regel</i>		<i>Support</i>	<i>Konfidenz</i>
Shampoo	→ Haarspülung	0.6	0.75
Haarspülung	→ Shampoo	0.6	1.00
Duschgel	→ Shampoo	0.5	0.83
Kosmetik	→ Shampoo	0.5	1.00
Kosmetik	→ Haarspülung	0.4	0.80
Shampoo, Deodorant	→ Haarspülung	0.4	1.00
Haarspülung, Deodorant	→ Shampoo	0.4	1.00
Shampoo, Kosmetik	→ Haarspülung	0.4	0.80
Haarspülung, Kosmetik	→ Shampoo	0.4	1.00
Kosmetik	→ Shampoo, Haarspülung	0.4	0.80

Wissensentdeckung

- Produktion: Informationen zur Prozessoptimierung und Fehleranalyse;
- Ökonomie: Warenkorbanalyse; Katalog-Design; Supermarkt Layout;
- Bankenwesen: Aufdeckung von Kreditkartenmissbrauch, Bonitätsanalyse und Devisenkursprognose;
- (Electronic) Commerce: Kundenprofile, Auswahl möglicher Kunden (Zielgruppen), Kundensegmentierung;
- Internet: Suchen nach relevanter Information (web mining);
- Wissenschaft: Gewinnung wichtiger Informationen über beobachtete Phänomene, Finden kausaler Zusammenhänge;
- Geologie: Auffinden verdächtiger seismographischer Strukturen, z.B. zur Vorhersage von Erdbeben.