Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungsprofessur
Wilhelmshöher Allee 73
34121 Kassel

Email: hotho@cs.uni-kassel.de
Tel.: ++49 561 804-6252

Dr. Andreas Hotho
Prof. Gerd Stumme

15.06.04

Knowledge Discovery

Lösungsblatt 4

Sommersemester 2004

Aufgabe 1: Induktive Logische Programmierung

a) Was ist der Unterschied zwischen einer extensionalen und einer intensionalen Begriffsbeschreibung?

Bei der extensionalen Beschreibung eines Begriffes, führt man explizit alle Elemente auf, die den Begriff ausmachen, z.B.

AIFB_Doktorand={Peter,Sudhir,Philipp,Nenad,Andreas,Christoph,Jens,Marc,Chriszoph,Julien,Daniel}.

Bei einer intensionalen Beschreibung wird der Begriff abstrakt (eventuell unter Zuhilfenahme von Zusatzprädikaten) in einer bestimmten Beschreibungssprache umschrieben, z.B.

$$AIFB\_Doktorand(Y) \leftarrow Doktorand(Y) \wedge arbeitet\_in(Y, AIFB)$$

b) Was ist das Ziel von ILP? Nehmen Sie dabei Bezug auf ihre Antwort in a).

Das Ziel von ILP ist, eine **intensionale** Beschreibung des Zielprädikates zu induzieren.

c) Warum braucht man bei ILP überhaupt negative Beispiele? Was macht man typischerweise wenn man keine expliziten negativen Beispiele hat?

Man braucht negative Beispiele, um nicht zu generelle Begriffsbeschreibungen zu induzieren.
Wenn man keine expliziten negativen Beispiele hat, dann nimmt man durch die sogennante **Closed World Assumption** (CWA) an, dass alle nicht explizit positiven Beispiele negativ sind.

d) Gegeben sei das Zielprädikat *online_Kunde* mitsamt der unten aufgeführten Beispielsmenge. Verwenden Sie den FOIL-Algorithmus, um eine Beschreibung dieses Zielprädikates in Bezug auf die Beispielsmenge und das gegebene Hintergrundwissen zu induzieren. Verwenden Sie das Information-Gain Kriterium, um die beste Spezialisierung auszuwählen.

Beispielsmenge:

online_Kunde(Paul) +
online_Kunde(Sonja) +
online_Kunde(Marcus) +
online_Kunde(Hilke) -
online_Kunde(Susanne) -

Hintergrundwissen:

handy(Paul)
handy(Hilke)
handy(Marcus)

ISDN(Paul)
ISDN(Hilke)
ISDN(Sonja)
ISDN(Marcus)

alter20-35(Marcus)
alter20-35(Paul)
alter20-35(Susanne)

1. Durchlauf des covering-Algorithmus:

$$c := online\_Kunde(x) \leftarrow$$

1. Durchlauf des specialization-Algorithmus:

Beispielmenge:

online_Kunde(Paul) +
online_Kunde(Sonja) +
online_Kunde(Marcus) +
online_Kunde(Hilke) -
online_Kunde(Susanne) -

Es stehen 6 verschiedene Literale zur Verfügung:

L_i	n_{i+1}^{+}	n_{i+1}^{-}	n_i^{++}	I_{L_i}	$Gain(L_i)$
handy	2	1	2	1/2	1/3
ISDN	3	1	3	1/3	**1**
$alter_{20-35}$	2	1	2	1/2	1/3
$\neg handy$	1	1	1	1	-1/3
$\neg ISDN$	0	1	0	$-\log_2(0)$	0
$\neg alter_{20-35}$	1	1	1	1	-1/3

Also wird in dieser Runde das Attribut **ISDN** ausgewählt. Damit erhalten wir

$$c_1 := online\_Kunde(x) \leftarrow ISDN(x)$$

Die neue Beispielsmenge ist:

online_Kunde(Paul) +
online_Kunde(Sonja) +
online_Kunde(Marcus) +
online_Kunde(Hilke) -

Hier stehen nun vier mögliche Attribute zur Spezialisierung zur Verfügung:

L_i	n_{i+1}^{+}	n_{i+1}^{-}	n_i^{++}	I_{L_i}	$Gain(L_i)$
handy	2	1	2	1/2	-1/3
$alter_{20-35}$	2	0	2	0	**2/3**
$\neg handy$	1	1	1	1	-2/3
$\neg alter_{20-35}$	1	1	1	1	-2/3

Also wird das Attribut alter20-35 als Spezialisierung ausgewählt. Da es keine negativen Beispiele mehr gibt, endet die Spezialisierungsschleife mit

$$H=\{ c' := online\_Kunde(x) \leftarrow ISDN(x), alter_{20-35}(x) \}$$

2. Durchlauf des covering-Algorithmus:

$$c := online\_Kunde(x) \leftarrow$$

Nach Entfernen der erklärten negativen Beispiele erhält man die neue Beispielsmenge:

online_Kunde(Sonja) +
online_Kunde(Hilke) -
online_Kunde(Susanne) -

1. Durchlauf des specialization-Algorithmus.

Folgende Attribute kommen hier in Frage:

L_i	n_{i+1}^{+}	n_{i+1}^{-}	n_i^{++}	I_{L_i}	$Gain(L_i)$
handy	0	1	0	$-\log_2(0)$	0
ISDN	1	1	1	1/2	**1/6**
$alter_{20-35}$	0	1	0	$-\log_2(0)$	0
$\neg handy$	1	1	1	1/2	**1/6**
$\neg ISDN$	0	1	0	$-\log_2(0)$	0
$\neg alter_{20-35}$	1	1	1	1/2	**1/6**

Da die Attribute $ISDN, \neg handy$ und $\neg alter_{20-35}$ gleichwertig sind, wählen wir einfach ein beliebiges aus, z.B. ISDN. Damit ergibt sich:

$$c_1 := online\_Kunde(x) \leftarrow ISDN(x)$$

mit der neuen Beispielsmenge:

online_Kunde(Sonja) +
online_Kunde(Hilke) -

Nun ergeben sich folgende Möglichkeiten zur Verfeinerung:

L_i	n_{i+1}^{+}	n_{i+1}^{-}	n_i^{++}	I_{L_i}	$Gain(L_i)$
handy	0	1	0	$-\log_2(0)$	0
$alter_{20-35}$	0	0	0	$.\log_2(0/0)$	0
$\neg handy$	1	0	1	0	**2/3**
$\neg alter_{20-35}$	1	1	1	1	-1/3

Also erhalten wir folgende Spezialisierung:

$$c' := online\_Kunde(x) \leftarrow ISDN(x), \neg handy(x)$$

Ingesamt erhalten wir also folgende Hypothese:

$$H = \{online\_Kunde(x) \leftarrow ISDN(x), alter_{20-35}(x); online\_Kunde(x) \leftarrow ISDN(x), \neg handy(x)\}$$

e) Wenn nur einstellige Prädikate verwendet werden, gilt immer $\varepsilon_{i+1} \subseteq \varepsilon_i$ und damit dann auch $n_{i+1}^{+} = n_i^{++}$.