MIDOS

Interesting Subgroups

Daniel Wiese

Was sind Subgruppen?

„Das statistische Auffinden von Subgruppen gehört zu den am meisten populären und einfachsten Formen des Wissens“[Klösgen]

Die Arbeitslosenrate ist überproportional hoch bei **jungen Männern mit niedrigen Ausbildungsgrad**

Die Todesrate bei Lungenkrebs ist bei **Frauen** signifikant in den letzten 10 Jahren gestiegen

Junge arme Frauen sind stärker mit *AIDS* infiziert als ihre männliche Vergleichsgruppe

Überblick-Ziele

Visualisierungstools und Kepler

Wie werden Subgruppen mit Kepler gesucht und visualisiert?

Hypothesensprache

Welche Sprache wird zur Beschreibung von Subgruppen verwendet?

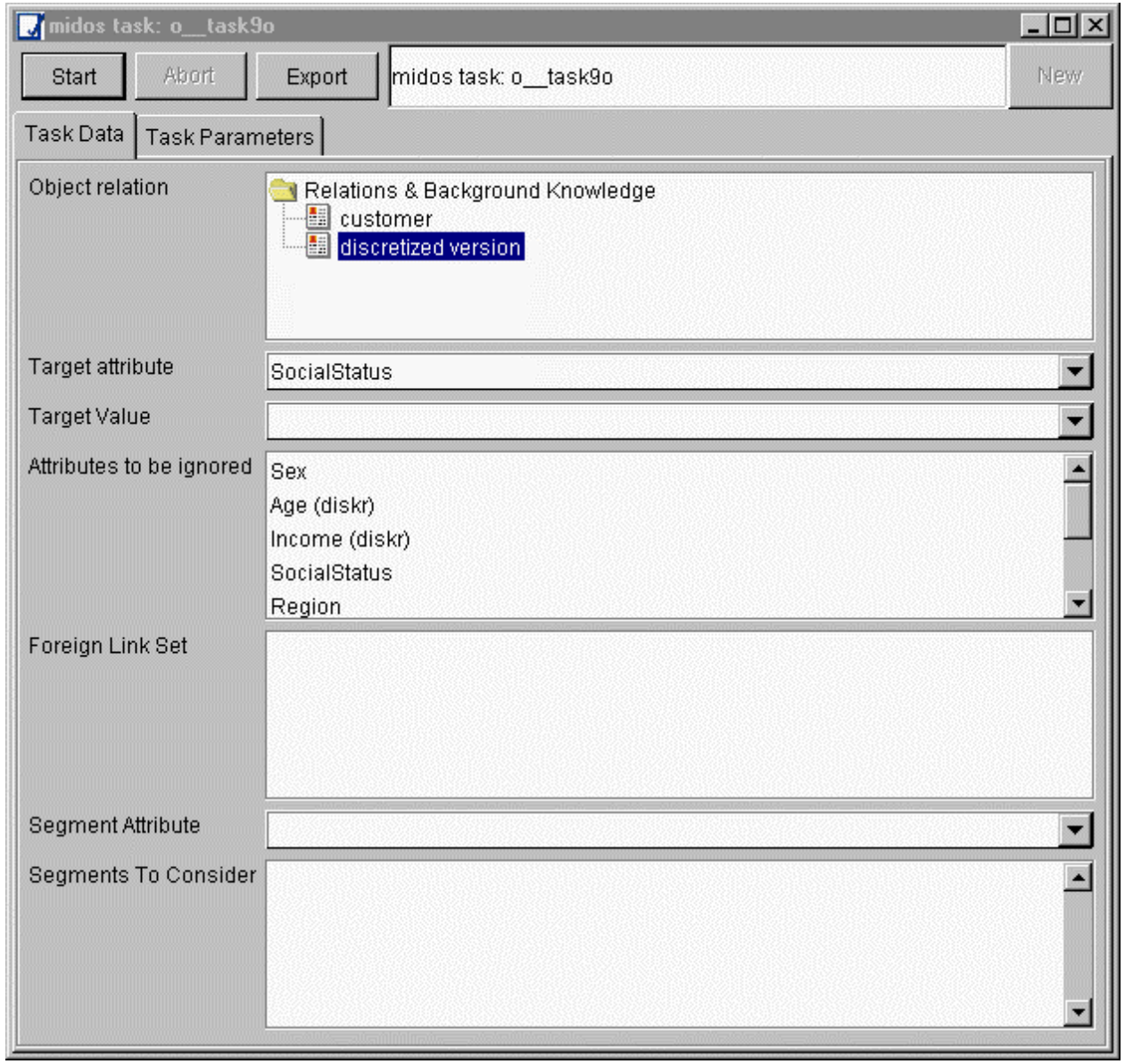
Evaluationsfunktion

Wie wird das Interesse gemessen?

Suchstrategien

Wie wird der mögliche Hypothesenraum durchsucht?

Subgruppenerkennung mit Kepler

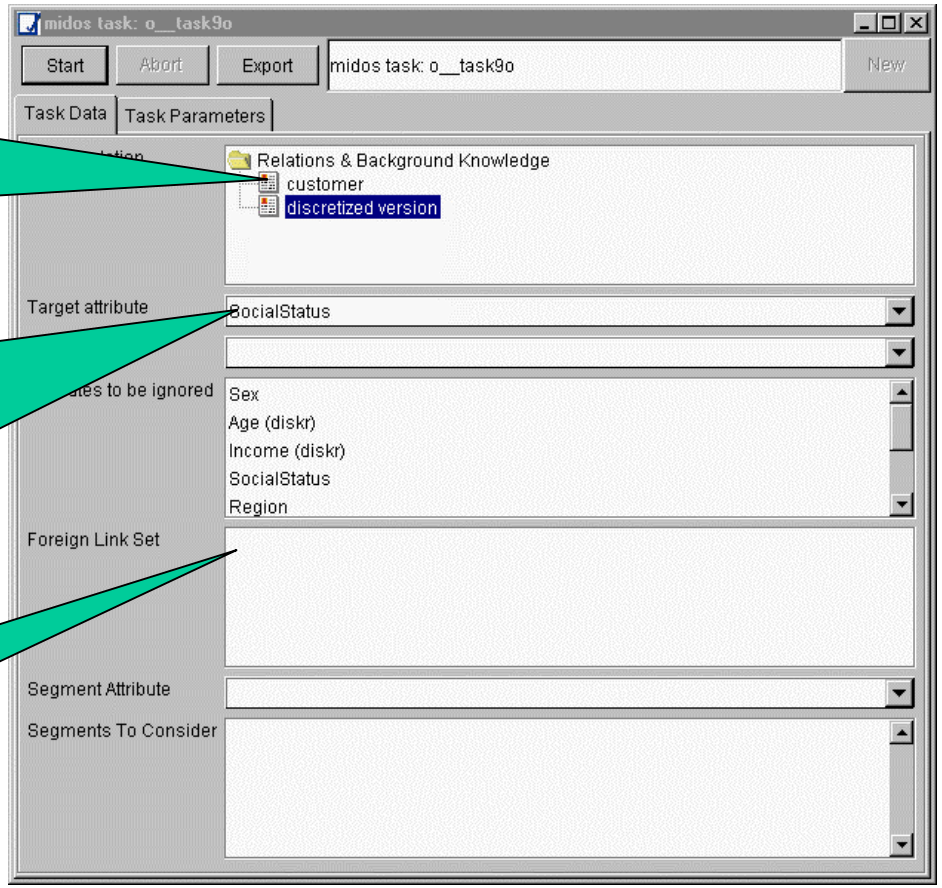


Subgruppenerkennung mit Kepler

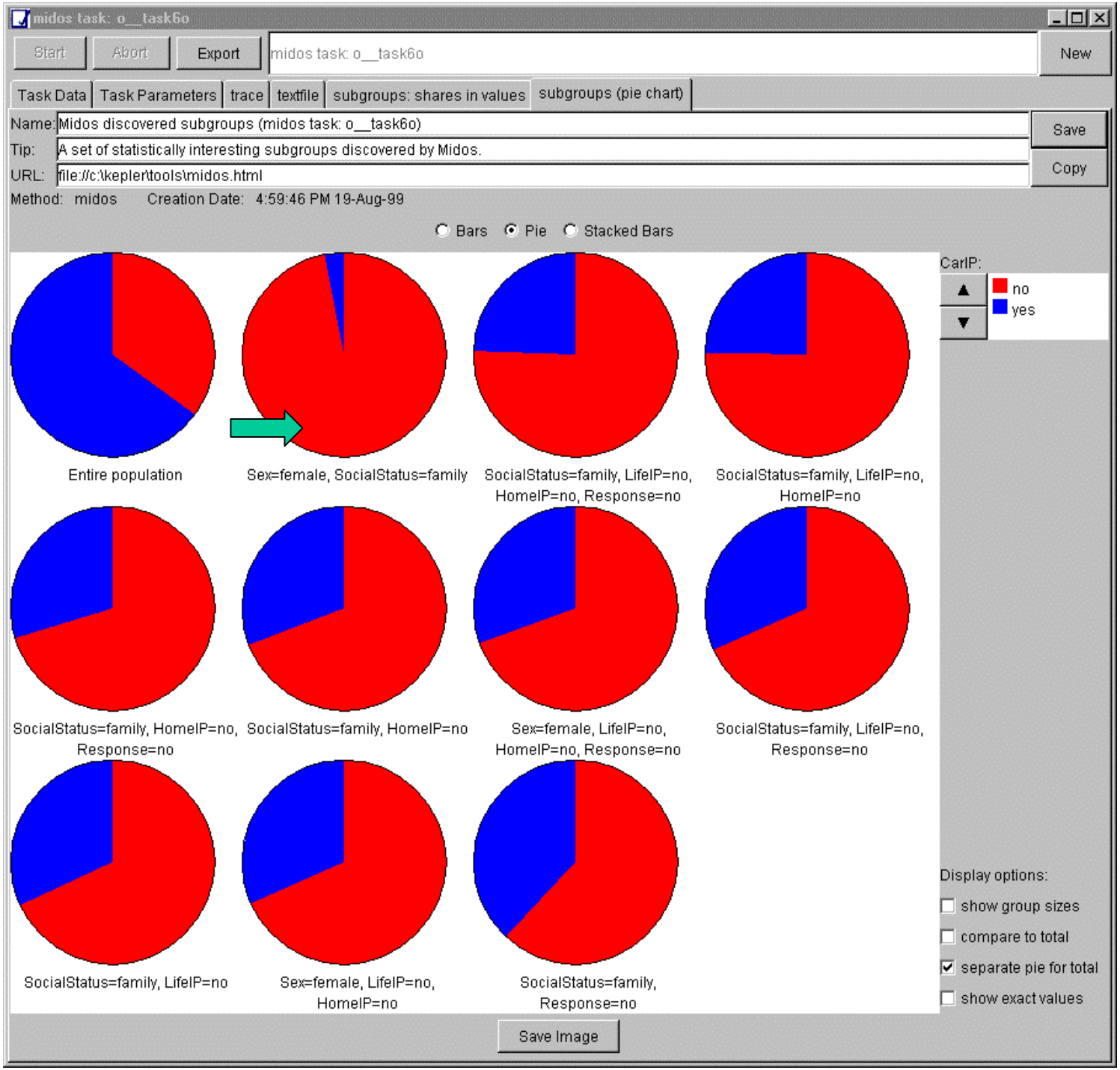
"*Objektrelation*" Für die multirelationale Analyse wird hier die zu analysierende Relation bestimmt. (vs. einrelational siehe Klösgen. Explora)

„*Zielattribut*" Die Werteverteilung dieses Zielattributes in der Subgruppe wird mit der Verteilung in der gesamten Population verglichen. Eine Subgruppe wird als interessant eingeschätzt wenn die Verteilung von der Gesamtverteilung abweicht.

„*Fremdschlüsselmenge*" Hier wird das Hintergrundwissen bestimmt. Durch die Fremdschlüssel werden zusammenhängende Relationen bestimmt.



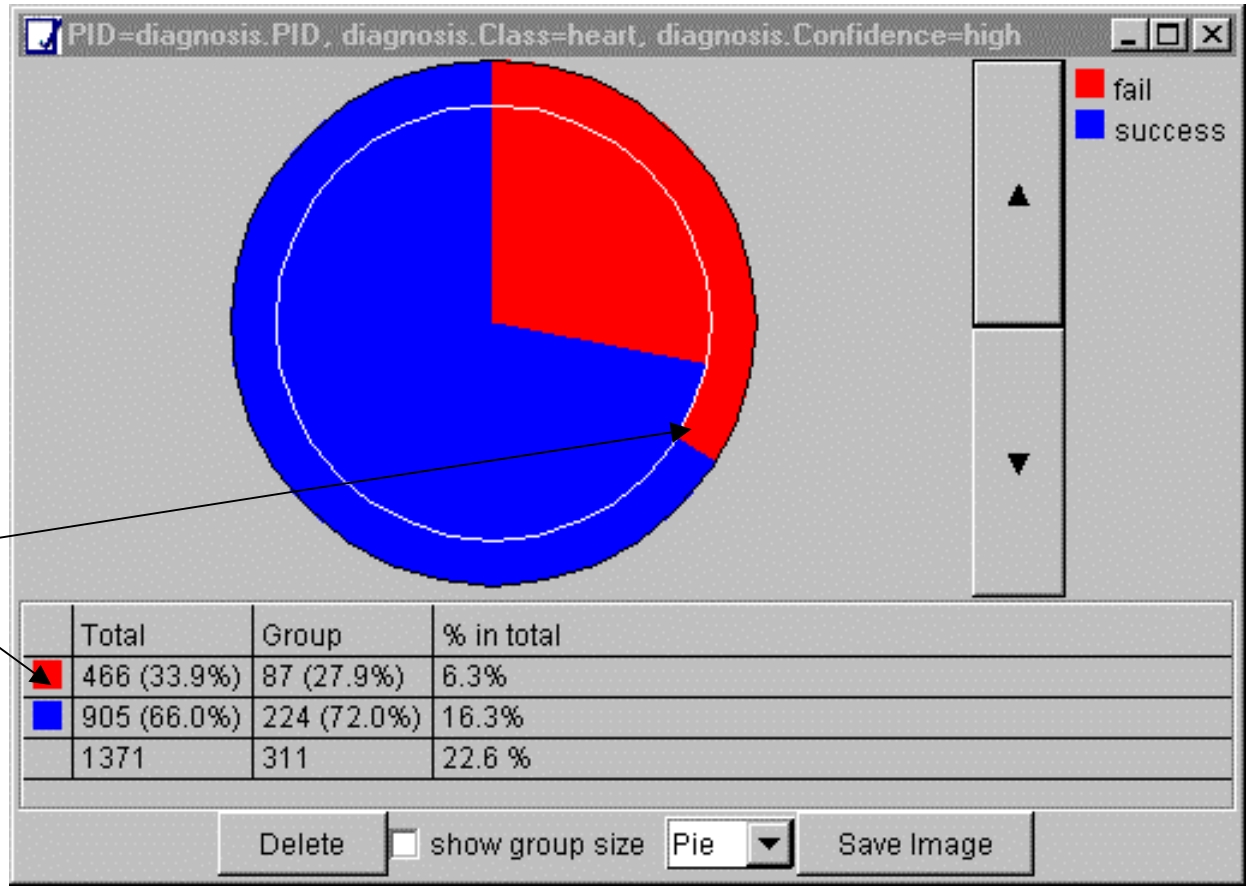
Quelle: „Kepler Walkthrough Handbuch“



Beispiel (Kepler)

9 gefundene Subgruppen-auffälliges Verhalten bzgl. Autobesitz (CarIP(yes/no))

Subgruppenerkennung mit Keppler



Gesammt-
population

Welche Hypothesensprache kann verwendet werden?

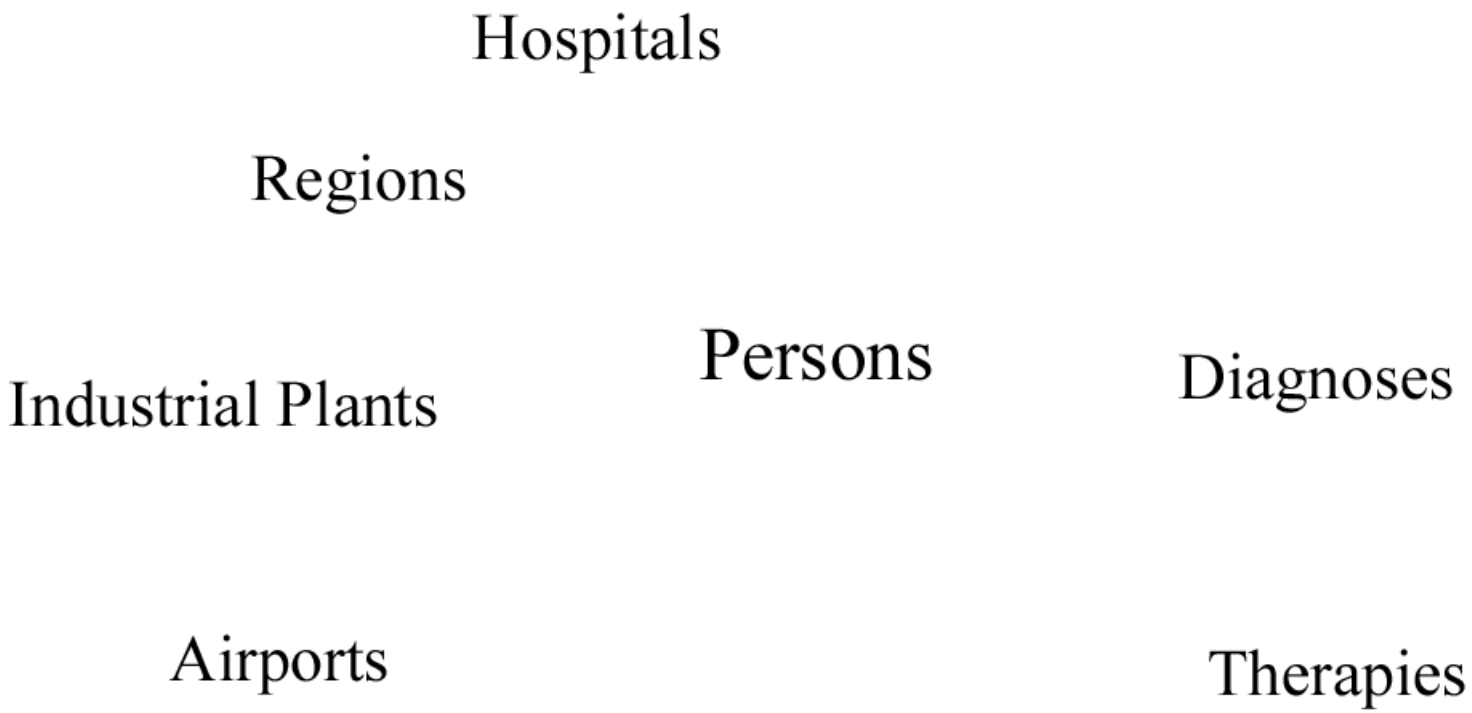
- Ein-relationale Hypothesensprache
- **Multi-relationale Hypothesensprache**
- Hypothesensprache die räumliche Nachbarschaftsbeziehungen darstellt (Topologie, Richtung, Entfernung)

Multi-Relationale-Subgruppen

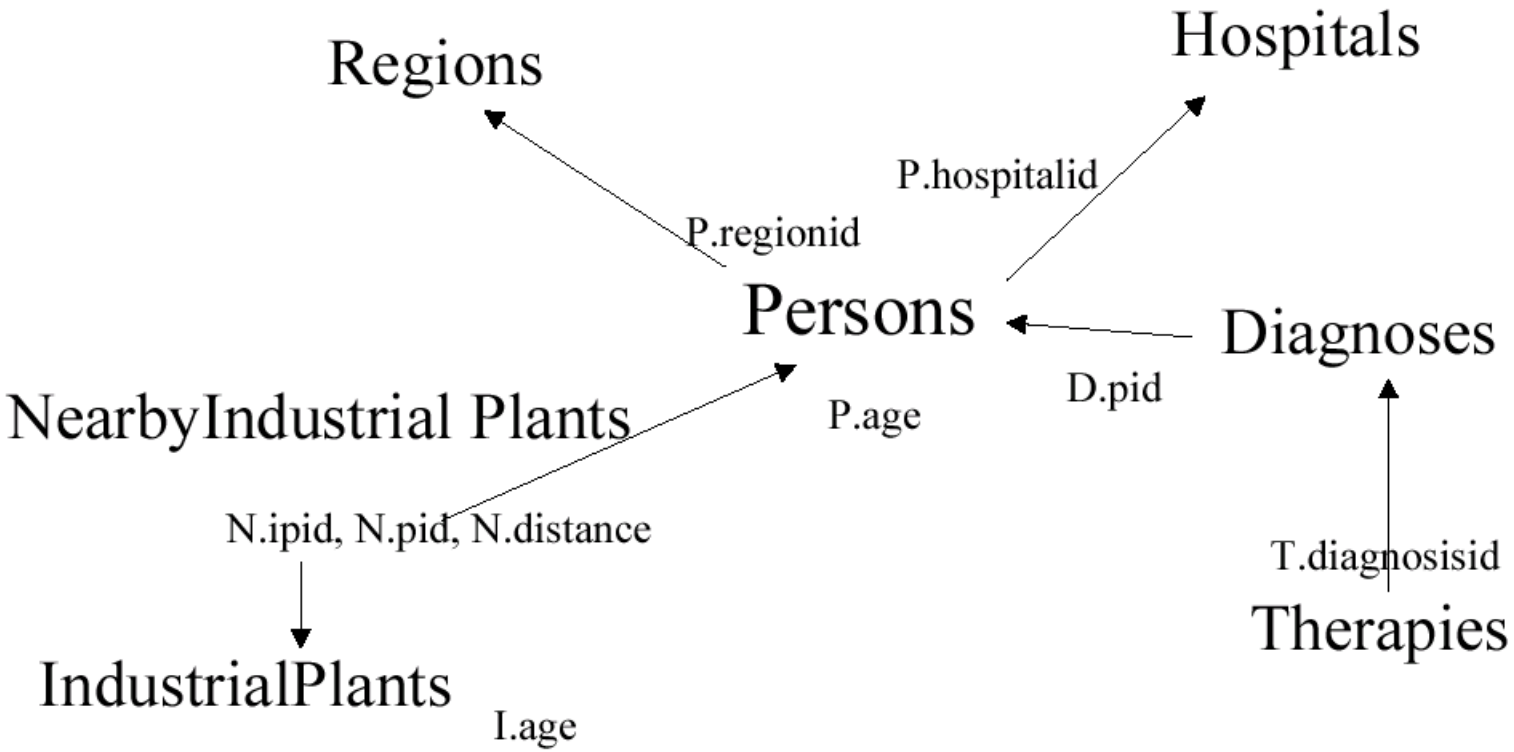
Beispielanwendung: Daten eines Krankenhauses
Relationale Datenbank mit 7 Relationen

R1: Krankenhäuser und Abteilungen [300 Tupel]
R2: Patienten [6.000 Tupel]
R3: Diagnosen [25.000 Tupel]
R4: Therapien [43.000 Tupel]
usw..

Multirelationale Hypothesensprache

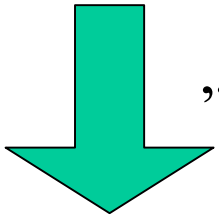


Multirelationale Hypothesensprache



Anwendungsbeispiel der multi-relationalen Subgruppenerkennung

- patient(PatientID,Name,Age,Sex, ...)
- diagnose(PatientID,DiagnosisID,Date,HospitalID)
- therapie(PatientID,TherapyID,Dosage,Date,HospitalID)
- krankenhaus(HospitalID,Name,Location,Size,Owner,Class)



„entdeckte“ Subgruppe

„Patienten älter als 65 Jahre, die in einem kleinen Krankenhaus behandelt werden haben eine überdurchschnittlich hohe Mortalitätsrate.“

Diese „Subgruppe“ wird mit dem folgenden Ausdruck beschrieben:

patient(ID,N,A,S) & A > 65 & diagnose(ID,D_ID,Dt,H) & krankenhaus(H,_,_,small,_,_).

Forderungen an die Hypothesensprache

- vorbestimmte Zielrelation

es werden nur Subgruppen von Objekten gebildet, die die Zielrelation beinhalten *z.B. Subgruppen von Personen*

- Links zwischen Relationen

implizit durch die Fremdschlüssel, explizit durch Linkattribute *z.B. patient_id*

Links dürfen nur entlang vorher spezifizierten Pfaden gebildet werden!

Definition des Problems

Multi-relationale Subgruppe kann beschrieben werden als:

Gegeben:

- eine relationale Datenbank D mit Relationen $R = \{r_1; :::::: ; r_m\}$
- eine Sprache für Hypothesen L_H (um die Subgruppen zu beschreiben)
- eine Qualitätsfunktion, um die Qualität der gefundenen Subgruppe zu beurteilen d: $h \in L_H$, D -> [0..1]
- eine Integerzahl k>0

Finde:

- eine Teilmenge $H \subseteq L_H$ von Hypothesen; mindestens k-viele ;
- so, dass für jede gefundene Hypothese gilt: $h \in H$, $d(h, D) > 0$ (die Qualität ist größer null)

und für jedes $h' \in L_H \setminus H$ gilt: $\min_{h \in H} d(h,D) \geq d(h',D)$ (die anderen Hypothesen h' sind schlechter)

Definition der Hypothesensprache

Gegeben ist eine Datenbank D mit Relationen R und Fremdschlüssel-Links F. Die Hypothesensprache besteht aus einer Menge von verknüpften Prädikaten

$$C = A1 \& ... \& An$$

der folgenden Form:

- $A_i = r_i (V_{i1}, ..., V_{ia})$ ← jede Relation wird durch ein korrespondierendes Prädikat beschrieben
- $A_i = V_j [<|=|>]$ ←Prädikate mit booleschen Ausdrücken

Es gilt:

- Die Hypothese startet immer von einer vorher bestimmten Objektrelation r_0
- Zwei Prädikate teilen sich eine Variable nur dann, wenn ein korrespondierender Link (Fremdschlüssel) existiert.

Beispiel:

$D=\{r_0, r_1, r_2, r_3, ...\}$, $F = \{r_0[2]->r_1[1], r_0[3]->r_2[1], r_1[2]->r_3[1], r_2[2]->r_3[2]\}$

$H = r_0(X,Y,Z) \& r_1(Y,U) \& r_2(Z,R) \& r_3(U,R) \& (X = x_0) \& (U=a)$

→Literal Y,R ist in zwei Prädikaten, da ein Link existiert

Beispiele

Erlaubte Hypothese

$r_0(X,Y,Z) \,\&\, r_1(Y,U) \,\&\, r_2(Z,R) \,\&\, r_3(U,R) \,\&\, (X = x_0) \,\&\, (R \geq \text{medium})$

Wenn folgende Bedingungen gelten

- die Objektrelation ist r_0,
- die Fremdschlüssellinks sind

$\{r_o[2] \rightarrow r_1[1],\ r_0[3] \rightarrow r_2[1],\ r_1[2] \rightarrow r_3[1],\ r_2[2] \rightarrow r_3[2]\}$.

Verbotenen Hypothesen

$r_0(X,Y,Z) \,\&\, r_3(U,R)$ [nicht verlinkt]

$r_0(X,Y,Z) \,\&\, r_3(X,R)$ [link nicht erlaubt/nicht definiert]

$r_1(Y,U) \,\&\, r_2(Z,R) \,\&\, r_3(U,R)$ [r_1 ist nicht die Objektrelation]

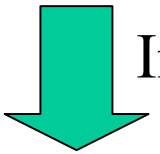
Evaluation von Subgruppen

Was ist eine interessante Subgruppe?

Ziel-Attribut: Behandlungserfolg (binär, ja | nein)

Referenzpopulation:

Wahrscheinlichkeitsverteilung [61% | 31%]



Interessant da abweichend

„Patienten älter als 65 Jahre, die ihre Erstdiagnose in einem kleinen Krankenhaus erhielten“

Wahrscheinlichkeitsverteilung [43% | 57%]

„Weibliche Patienten des Dortmunder Krankenhauses“

Wahrscheinlichkeitsverteilung [75% | 25%]

Evaluation von Subgruppen

-Was ist eine abweichende Subgruppe?

-Was ist die Referenz zu der Abweichung?

-Wann ist eine Abweichung signifikant?

Evaluation von Subgruppen

- $c(h) := \pi_{[K]}(\{\sigma \mid h\sigma \in D\})$, die Abdeckung von h
- $T := \{t \in r_o \mid t[A_g] = 1\}$, das Zielattribut (nur die wahren „Tupel“)
- $g(h) := |c(h)| / |r_0|$, Generalität (→ wie viele „Tupel“ werden im Verhältnis zur Gesamtmenge durch die Hypothese abgedeckt?)
- $p_0 := |T| / |r_o|$, die Referenzwahrscheinlichkeit der Zielattributes (binär) in D
- $p(h) := |c(h) \cap T| / |c(h)|$, Wahrscheinlichkeit des Zielattributes (binär) in c(h)

Die Evaluationsfunktion ist

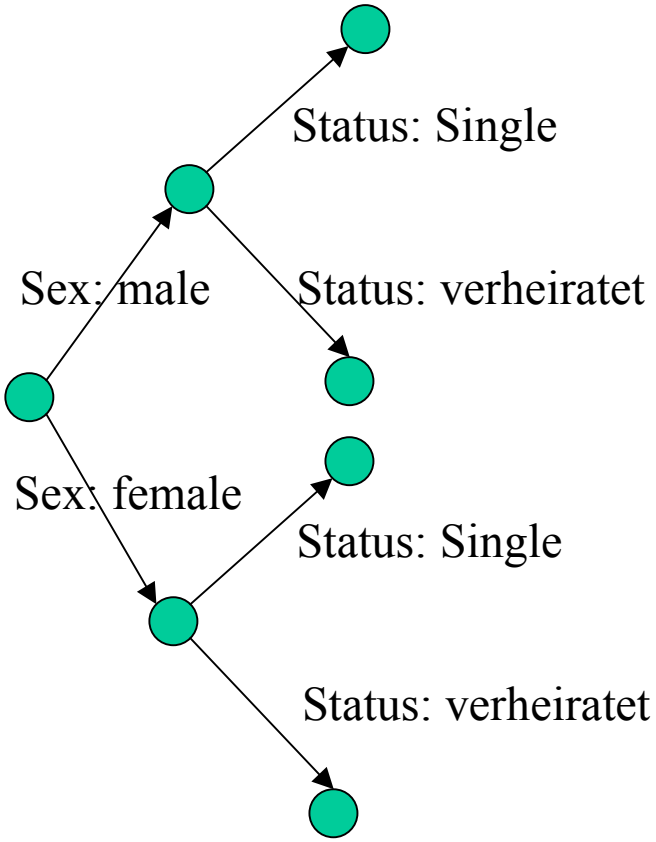
$d(h) := g(h) \ |(p(h) - p_0)|$

Wurden genügend Tupel abgedeckt?

Weicht die Subgruppe Genügend von der Referenzwahrscheinlichkeit ab?

Suchstrategien im Hypothesenraum

Theoretisch müssten wir die Menge aller Hypothesen vollständig durchsuchen!



Erster Ansatz:

Man beginnt mit der allgemeinsten Hypothese und verfeinert diese zu immer kleineren Subgruppen.

Wie können wir den Suchaufwand verkleinern?

1. Möglichkeit: Systematisches und geordnetes Durchsuchen des Hypothesenraumes

Ziel: spezifiziere einen Verfeinerungsoperator ρ, so dass jede Hypothese nur einmal generiert werden muss

Lösungsansatz:

•Konstruiere eine Ordnung o im Verfeinerungsoperator ρ mit folgender Eigenschaft: für alle $h' \in \rho(h)$ soll gelten $o(h') > o(h)\}$.

Vorteile

•Duplikate können dadurch vermieden werden, indem man sich einfach nur die aktuelle Hypothese und die Ordnung der Verfeinerung merkt

•vorteilhaft für die parallele Abarbeitung

Wie können wir den Suchaufwand verkleinern?

2. Möglichkeit: Beschneiden des Hypothesenraumes

Top-down Suche: generell → spezifisch

•Wir beschneiden den Hypothesenraum, wenn die Anzahl der Tupel in unserer Subgruppe eine bestimmte Anzahl von Elementen unterschreitet

•Wir verwenden eine optimistische Schätzfunktion d_{max}:
Unser Algorithmus sucht die k-besten Hypothesen. Wenn die Vorhersage für unsere aktuelle Hypothese h_0 schlechter ist, als die schlechteste bisher gefundene können wir den Hypothesenraum ohne Gefahr beschneiden
$d_{max}(h_c) < \min_{h \text{ in } H} d(h)$

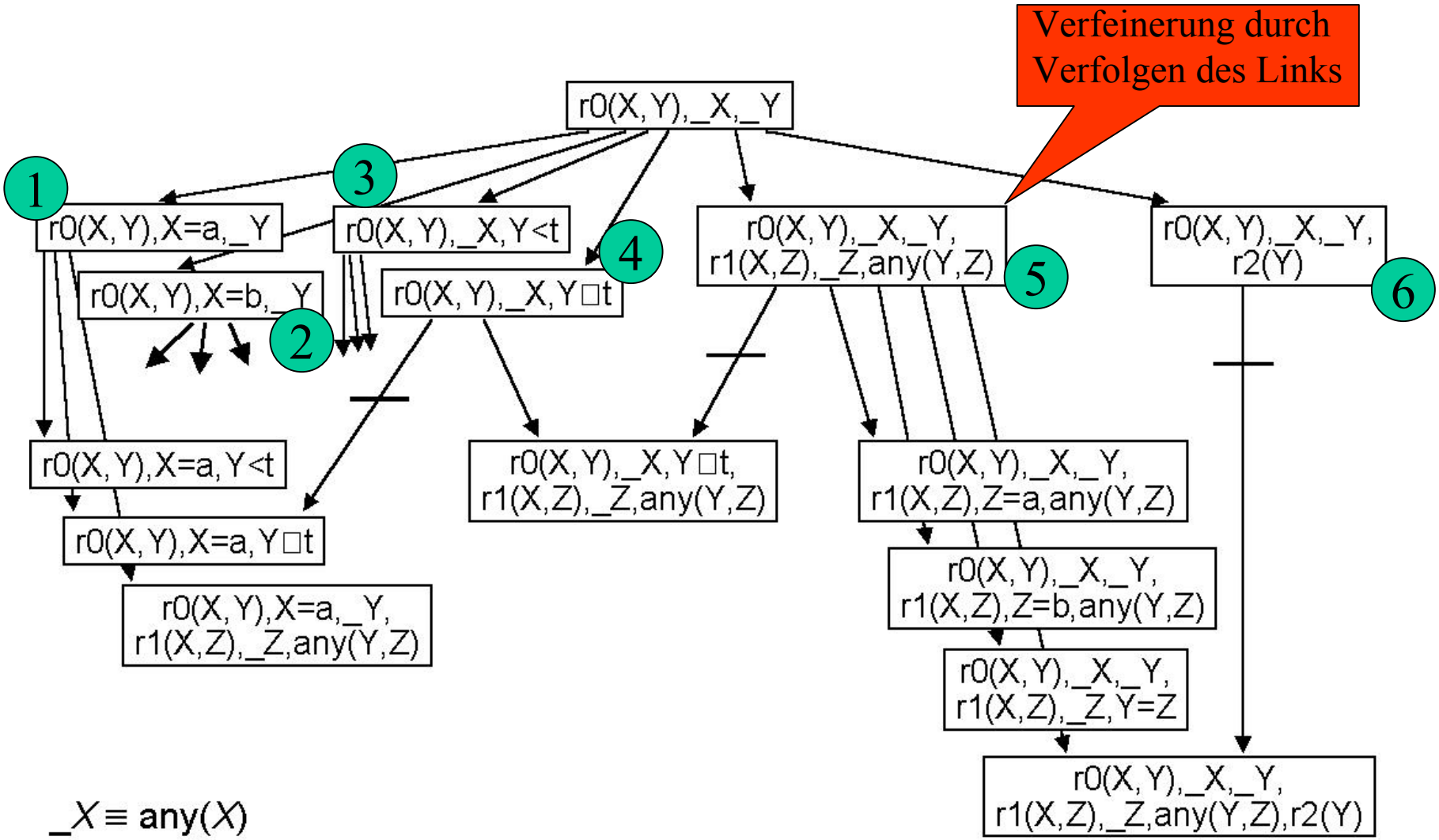
Verfeinerung der Hypothese

Des Hypothesenraum wird top-down durchsucht. Zunächst beginnt man mit generellen Hypothesen die immer weiter spezialisiert werden z.B. „*Gebiet mit großen Arbeitslosigkeit und einer großen Anzahl von medizinischen Einrichtungen*“ ist spezieller als „*Gebiet mit großen Arbeitslosigkeit*“

Verwenden eines Spezialisierungsoperators p:
p:L_H→2 hoch L_H

Der Spezialisierungsoperator liefert zu einer Hypothese h **alle** unmittelbar speziellen Nachfolger p(h) ←Partialordnung für alle Hypothesen!!!

Verfeinerung der Hypothese



$\_X \equiv \text{any}(X)$

Fremdschlüssel: (1) $r_0[1] \rightarrow r_1[1]$, (2) $r_0[2] \rightarrow r_2[1]$, ...

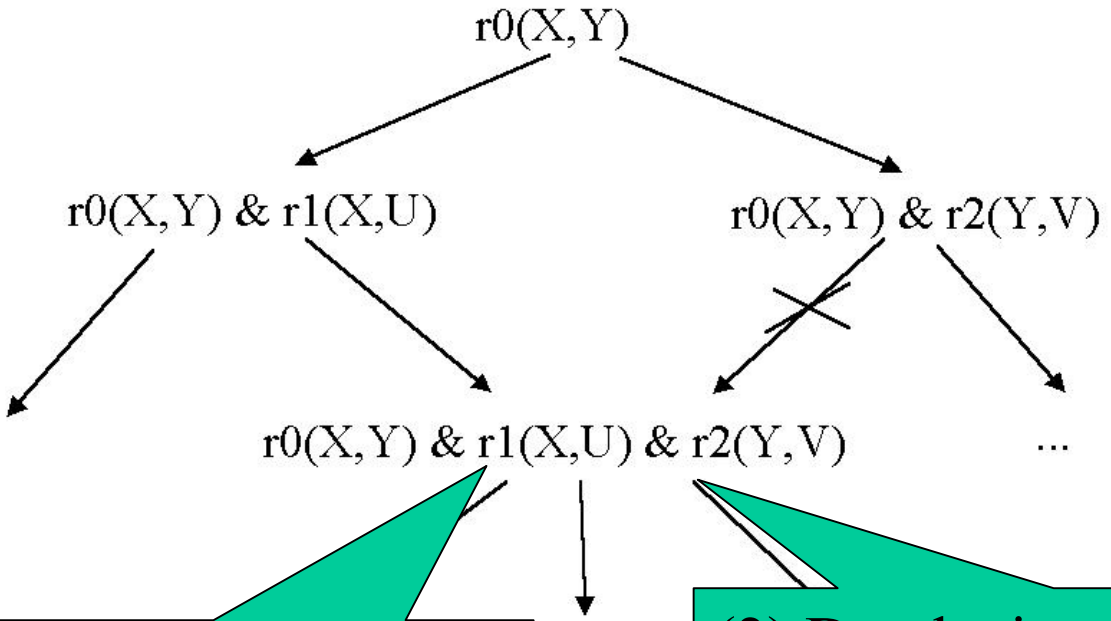
Verfeinerung der Hypothese

Vorgehen bei der Verfeinerung

1. Verfeinere ein Literal i durch Spezialisierung
   - 1.1 ersetze any(X) durch X=a oder X=b ... (abhängig von der Wertehierarchie)
   - 1.2 ersetze any(X,Y) durch X=Y
   - usw...

2. Verfeinere ein Literal *i* durch verfolgen des Links
   - füge ein Literal nur mit Variablen (korrespondierend zur Zielrelation) hinzu z.B das Literal r2(X,Y) hinzufügen, wenn es einen Link von r1→r2 gibt

Wie können wir den Suchaufwand verkleinern?



(1) Durch die Partialordnung können wir kontrollieren ob diese Hypothese bereits generiert wurde.

(2) Durch eine optimistische Vorhersagefunktion könnten wir den Hypothesenraum beschneiden

Eine einfache optimistische Schätzfunktion

Wahrscheinlichkeit in der Subgruppe

$$d(h) := g(h)\ (p(h) - p_0) \Rightarrow\ d_{max} := g(h)\ (1 - p_0)$$

Wahrscheinlichkeit in der gesamten Population

$g(h) := |c(h)| / |r_0|$, Generalität
(→ wie viele „Tupel“ werden im Verhältnis
zur Gesamtmenge durch die Hypothese abgedeckt?)

Der MIDOS Algorithmus

Überblick

- Ausnutzen, dass die Hypothesen partiell geordnet sind

- iterative Suche von der generellen zur spezifischen Hypothese

- 2 Phasen:
  1. Generierung der verfeinerten Hypothesen,
  2. Evaluation der neuen Hypothesen

- Beschneiden des Hypothesenraumes
  - Durch eine „minimale“ Subgruppe
  - Durch Vorhersage der Qualität

MIDOS-Algorithmus

z.B. könnten die Hypothesen mit der besten Qualität zuerst verfeinert werden

$Q:=\{V_1, ..., V_{a(r0)}\}$, ← start mit der gesamten Objektrelation→allgemeinste Hypothese;
H:=0 ← k-besten Hypothesen
while nicht fertig
- wähle eine Teilmenge C aus Q gem. Suchstrategie
-p(C):={p(h)|h ∈ C}← Menge aller unmittelbar spezielleren Hypothesen
-Teste jede Hypothese auf ihre Qualität (berechne d(h) mit h ∈ p(C))
-if d(h)=0→Hypothesenraum abschneiden
-else if d_{max} (h) schlechter ist als die bisher schlechteste Hypothese in H→Hypothesenraum abschneiden
-else
-if d(h) besser als die bisher schlechteste Hypothese→schlechteste Hypothese entfernen
-füge h zu H hinzu
-füge h zu Q hinzu

Literatur

•W. Klösgen. Explora: A multipatern and multistrategy discovery assistant. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, ch. 10 s. 249-271. AAAI/MIT Press, Cambridge, USA, 1996.

•Stefan Wrobel. An algorith for multi-relational discovery of subgroups. In J. Komorowski and J. Zydkow, Symposium proceedings/PKDD´97, s. 78-87. Springer Verlag, Berlin, New York, 1997.

•Stefan Wrobel, Katharina Morik, Thorsten Joachims. Maschinelles Lernen und Data Mining. In Hadbuch der Künstlichen Intelligenz, 2001.

Ende

Verfahren der Subgruppenerkennung

- **Abweichende Subgruppen ← MIDOS**

Abweichende Verteilung einer Eigenschaft in einer Zielvariable der Subgruppe z.B. „ *Die Arbeitslosenrate ist überproportional hoch bei jungen Männern mit niedrigen Ausbildungsgrad*“

- **Paare von assoziierten Subgruppen**

Binäre Relation zwischen Objekten z. B. „*Windeln werden nach 18.00h oft zusammen mit Bier gekauft*“

- **Partiell geordnete Menge von Subgruppen**

Zeitlich geordnete Objekte → aufdecken von Sequenzregeln z.B. „*die Fehlersequenzen in einem Netzwerk*„

Beispiel Sampling (Weglassen?)