Wissensentdeckung in Datenbanken: eine Herausforderung für das maschinelle Lernen

Katharina Morik
Universität Dortmund
Lehrstuhl Informatik VIII

* Was ist anders als bisher?
* Ansätze
* Lernaufgaben
* ILP
* Erfahrungen

Bisher... Datenanalyse

Statistische Datenanalyse:

strukturiert erhobene Daten

wenige Beobachtungen

wenige Merkmale

Hypothesen werden vom Anwender erstellt

Statistische Lerntheorie

Gegeben: numerische Daten und ihre Funktionswerte
Finde: Funktion, deren Vorhersage bei neuen Daten am wenigsten Fehler macht

Empirische Risiko-Minimierung:

$$R(a) = \frac{1}{n} \sum_{i=1}^{n} L(z_i, f(x_i, a))$$

L gibt den Abstand zwischen dem wahren Wert z und dem Wert der gelernten Funktion f an

Strukturelle Risiko-Minimierung:
Funktionen nach ihrer VC-Dimension geordnet
In jeder Teilmenge von Funktionen wird diejenige gefunden, die das Risiko einer falschen Vorhersage minimiert

Die einzige Lernaufgabe bei Vapnik ist die Funktionsapproximation!

Bisher... Maschinelles Lernen

Maschinelles Lernen zur Unterstützung des Wissenserwerbs:
Lernen von Entscheidungsbäumen (sichere Klassifikation)
fallbasiertes und induktives Lernen (fallbasierte Klassifikation)
statistisches Lernen (statistische Klassifikation)
induktives Begrifflernen (heuristische oder überdeckende Klassifikation)

Maschinelles Lernen zu Unterstützung der Wissenswartung:
Revision
Restrukturierung

Maschinelles Lernen zur Performanzsteigerung
erklärungsbasiertes Lernen
Reinforcement oder Q-learning

insbesondere

für das Lernverfahren geeignete Repräsentationssprache

wenige Beobachtungen

wenige Merkmale bzw. Relationen

klar spezifizierte Zielgröße:
Begriff, dessen Charakterisierung gelernt werden soll

Herausforderung

vorgegebene Beispielrepräsentation, die für andere Zwecke als das Lernen entworfen ist

große Anzahl von Beobachtungen:
Tupelanzahl mehrere 100 000

große Anzahl von Datenbankattributen in mehreren Tabellen

unbekanntes Lernziel "etwas Interessantes"

Ansätze

weg von der Datenbank:
Merkmale selektieren
eine Datenbanktabelle erstellen
in günstige Repräsentation transformieren

weg von der Datenmasse:
Stichproben ziehen
iteratives Lernen
(bootstrapping)

Vereinfachung der Lernaufgabe:
stark eingeschränkte Form der Lernergebnisse
(Abhängigkeit einer Zielgröße von einer Einflußgröße)

Geht es auch anders?

Verwendung der Datenbank:
Hypothesentest über SQL

Lernen aus allen Tupeln

Regellernen

in eingeschränkter Prädikatenlogik (ILP)

Lernaufgaben: Begriffslernen

Gegeben:
positive und negative Beispiele E = E+, E-
in einer Repräsentationssprache LE und
Hintergrundwissen B in einer Repräsentationssprache LB

Finde:
eine Hypothese H in einer Repräsentationssprache LH, die eine eingeschränkte Prädikatenlogik ist, so daß gilt:

Konsistenz: $B, H, E+ \not\models \perp$
Vollständigkeit: $B, H \models E+$
Korrektheit: $\forall e- \in E- : B, H \not\models e-$

Lernaufgabe: Regellernen

Gegeben:
Beobachtungen E in einer Repräsentationssprache LE und
Hintergrundwissen B in einer Repräsentationssprache LB

Finde:
die Menge H von Hypothesen in der Repräsentationssprache LH, die eine eingeschränkte Prädikatenlogik ist, so daß gilt:

Gültigkeit: $M+ (B \cup E) \subseteq M (H)$
Notwendigkeit: für alle $h \in H$ existiert $e \in E$, so daß $B, E - \{e\} \not\models e$, aber $B, E - \{e\}, h \models e$
Vollständigkeit: für alle $h \in LH$, die gültig und notwendig sind, gilt $H \models h$
Minimalität: Es gibt keine echte Teilmenge von H, die gültig und vollständig ist.

Beispiel

Potential Customer

Person	Income	Customer
Ann Smith	10 000	+
Joan Gray	1000000	+
Mary Blyte	20 000	-
Jane Brown	20 000	+
Bob Smith	100 000	+
Tom Blyte	10 000	-
Jack Brown	200 000	+

Married_to

Husband	Wife
Bob Smith	Ann Smith
Tom Blyte	Mary Blyte
Jack Brown	Jane Brown

Regellernen findet:
i) married(Person, Spouse) & customer(Person) → customer(Spouse)
ii) married(Person, Spouse) & income(Person, ≥100 000) → customer(Spouse)
iii) income(Person, ≥ 100 000) → customer(Person)

1. Fall

Potential Customer

Person	Income	Customer
Ann Smith	10 000	+
Joan Gray	1000000	+
Mary Blyte	20 000	-
Jane Brown	20 000	+
Bob Smith	100 000	+
Tom Blyte	10 000	-
Jack Brown	200 000	+

Married_to

Husband	Wife
Bob Smith	Ann Smith
Tom Blyte	Mary Blyte
Jack Brown	Jane Brown

Regellernen:
Jane Brown ist dann nicht in allen minimalen Modellen Kundin, also können i), ii) nicht gelernt werden. Nur iii) bleibt davon unberührt.
Begriffslernen über customer:
Jane Brown wird als Kundin vorhergesagt. Es brauchen nicht alle drei Regeln für die Vollständigkeitsbedingung des Begrifflernen gefunden zu werden. i), ii) oder i), iii)
Mit CWA werden i) und ii) falsch und zurückgewiesen. Dann reicht iii) nicht aus, um alle positiven Beispiele abzuleiten (Ann Smith) und Begrifflernen schlägt fehl.

2. Fall

Potential Customer		
Person	Income	Customer
Ann Smith	10 000	+
Joan Gray	1000000	+
Mary Blyte	20 000	-
Jane Brown	20 000	+
Bob Smith	100 000	+
Tom Blyte	10 000	-
Jack Brown	200 000	+

Married_to	
Husband	Wife
Bob Smith	Ann Smith
Tom Blyte	Mary Blyte
Jack Brown	Jane Brown

Regellernen:
i), ii), iii) werden gelernt, denn es wird nichts über die Höhe von Jack Browns Einkommen abgeleitet und Regel ii) betrifft Jane Brown nicht mehr.
Begriffslernen:
i) leitet ab, daß Jane Brown und Ann Smith Kundinnen sind, iii) leitet ab, daß Bob, Joan, Jack Kunden sind. Das sind alle KundInnen.
Die CWA ändert hier nichts.

3. Fall

Potential Customer		
Person	Income	Customer
Ann Smith	10 000	+
Joan Gray	1000000	+
Mary Blyte	20 000	-
Jane Brown	20 000	+
Bob Smith	100 000	+
Tom Blyte	10 000	-
Jack Brown	200 000	+

Married_to	
Husband	Wife
Bob Smith	Ann Smith
Tom Blyte	Mary Blyte
Jack Brown	Jane Brown

Regellernen:
iii) bleibt als einzige Regel in allen Modellen gültig. Falls Jane Brown keine Kundin ist und Jack Brown ein hohes Einkommen hat, gelten i) und ii) nicht.
Begriffslernen:
i), ii), iii) können nicht ableiten, daß Jack Brown Kunde ist. Da nicht alle positiven Beispiele abgedeckt werden können, schlägt das Begriffslernen fehl.

Unterschied Regel-/Begriffslernen

Potential Customer		
Person	Income	Customer
Ann Smith	10 000	+
Joan Gray	1000000	+
Mary Blyte	20 000	-
Jane Brown	20 000	+
Bob Smith	100 000	+
Tom Blyte	10 000	-
Jack Brown	200 000	+

Married_to	
Husband	Wife
Bob Smith	Ann Smith
Tom Blyte	Mary Blyte
Jack Brown	Jane Brown

	Regellernen	Begriffslernen mit CWA
Jane ?	iii)	i), ii) oder i), iii) ---
Jack ?	i), ii), iii)	i), iii) i), iii)
Jane, Jack ?	iii)	--- ---

Regellernen ist ...

- dichter an den Daten -- Begriffslernen hat mehr Vorhersagekraft
- schwieriger als Begriffslernen:
  Es gibt Lernaufgaben, die vom Regellernen noch gelöst werden, vom Begriffslernen hingegen nicht.
  Begriffslernen kann durch CWA die Gültigkeitsforderung des Regellernens erfüllen, die Vollständigkeitsbedingung jedoch nicht.
  Wendet Begriffslernen die CWA an, dann führt seine Vollständigkeitsbedingung leicht dazu, daß gar nichts mehr gelernt wird.
  Beweis von Stephen Muggleton und Luc De Raedt für definite Hornklauseln (haben genau ein minimales Modell) zeigt, daß Begriffslernen eine Aufgabe ist, die in Regellernen echt enthalten ist.
  Die Schwierigkeit liegt natürlich in der Vollständigkeitsbedingung des Regellernens: finde alle gültigen und notwendigen Regeln!

KDD mit ILP und direktem Datenbankzugriff

Rule Discovery Tool (Kietz, Wrobel 1992)
Regelschemata in Allgemeinheitsordnung
Benutzer geben an, welche Form von Regeln sie interessant finden

m1(C,P, Q): P(X,C) μ Q(X)
m2(P1, P2, Q): P1(X,Y) & P2(X) μ Q(Y)
m3(C, P1, P2, Q): P1(X, Y) & P2(X, C) μ Q(Y)

Regelschemata werden mit allen Prädikaten instantiiert, c mit allen vorhandenen Werten (Intervallen). Sicheres pruning!
Jede Instanz ist eine Regel. Ihre Gültigkeit wird mithilfe von SQL-Anfragen über der Datenbank getestet.

Akzeptanzkriterien

Benutzer geben Kriterium für die Akzeptanz einer Hypothese an.
Beispiele:

$$\frac{pos(H)}{concl(H)} - \frac{neg(H)}{concl(H)} \geq 0.8$$

a posteriori a priori bei 2 Klassen

$$\frac{pos(H)}{pos(H)+ neg(H)} \geq \frac{concl(H)}{concl(H)+ negconcl(H)}$$

pos(H): Prämisse und Konklusion kommen gemeinsam vor neg(H): Prämisse und Negation der Konklusion kommen gemeinsam vor	concl(H): Konklusion kommt vor negconcl(H): Konklusions-prädikat ist anwendbar, kommt aber nicht vor

Abbildung von Datenbank auf Prädikate

Jede Datenbanktabelle ist ein Prädikat, ihre Attribute sind die Argumente.
customer (Person, Income, Customer)

Auswahl interessanter Attribute:
Vorauswahl bei sehr vielen Datenbankattributen

Jedes Datenbank-Attribut wird ein Prädikat, der Schlüssel und der Attributwert die Argumente.
income (Person, Income) .., wife (Husband, Wife)

Jeder Wert eines Datenbankattributs wird Prädikat, Schlüssel das Argument
customer (Person), inc_10_20 (Person)
Sinnvolle Intervalle in Attributwerten vorher finden!

Größe des Hypothesenraums

$$r\,(p\,i^{c})^{k}$$

r Anzahl der Regelschemata
p Anzahl der Prädikate
k max. Anzahl von Literalen in einem Regelschema
Bei Konstantenlernen:
c Anzahl der zu lernenden Konstanten
i max. Anzahl von Werten für eine Konstante

Je nach gewählter Abbildung ist die Größe des Hypothesenraums sehr unterschiedlich.

Experimente

Daimler Benz AG
2,6 Gigabyte Datenbank: alle Fahrzeuge und ihre Garantiefälle
40 Tabellen mit je bis zu 40 Attributen

1. **Unterschiedlich großer Hypothesenraum:**
   4913 ≤ Größe ≥ 2,8 E 41
2. **Unterschiedlich großer Datenbankauszug:**
   max. 23 Tabellen
   max. 750 000 Tupel
3. **Verwendung von Hintergrundwissen**
4. **Vergleich mit anderen ILP-Verfahren**

Verwendung von Hintergrundwissen

Gegeben:
elektronisch verfügbares Werkstattbuch für PKW mit allen Fahrzeugteilen
Finde:
Gruppen von Teilen, die räumlich, funktional oder bzgl. ihrer Schadensart zusammenhängen

Umformen der Datei in einstellige Fakten, wobei das Fahrzeugteil als Argument, der Zusammenhang als Prädikat ausgedrückt ist
Verwendung von STT (Kietz 1989) zum Finden einer Subsumtionshierachie

Klassen von Fahrzeugteilen der Datenbank hinzufügen

Lernergebnisse

rel_niveauregulierung (FIN) → beanstandet (FIN)

motor_e-typ (FIN, Typ) & mobr_cyl (Typ, 6) → beanstandet (FIN)

rel_garantie (X1, X2, RB, X4, X5, X6, X7, Konfig, Teil) &
rel_garantie_benzman (X1, X2, RB, X4, X5, X6, X7, FIN) &
rel_motor_typ (FIN, 206) & italien(RB) → class_419 (Konfig, Teil)

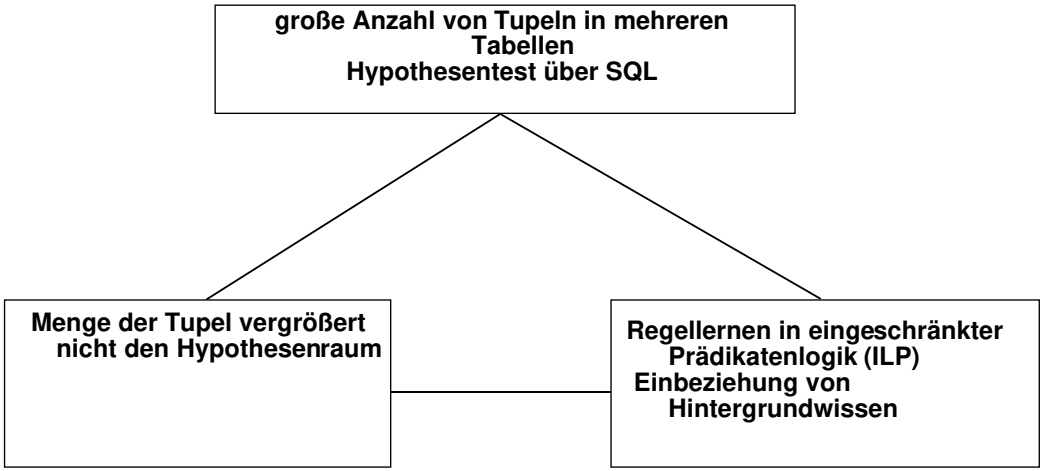
rel_garantie (X1, X2, X3, X4, X5, X6, X7, Konfig, Teil) & class_35 (Konfig, Teil)
→ kostbean_0_500 (X1, X2, X3, X4, X5, X6, X7)
a priori Wahrscheinlichkeit: 0,84 a posteriori Wahrscheinlichkeit: 0,89

Abhängigkeit der Lernzeiten

- **Größte Menge von Tupeln in 16 Tabellen,**
  größter Hypothesenraum
  längste Lernzeit (< 50 Stunden)
- **Mittlere Menge von Tupeln in 23 Tabellen,**
  mittlerer Hypothesenraum
  3 Minuten ≤ Lernzeit ≤ 3 Stunden 2 Minuten
- **Größte Menge von Tupeln in 3 Tabellen,**
  kleiner Hypothesenraum
  mittlere Lernzeit (≤ 14 Stunden)

Hypothesentest über einzelne SQL-Anfragen
ist zu aufwendig!
Sobald eine Tupelmenge in den Hauptspeicher paßt, muß sie dort gehalten werden!

Es geht!



Infrastruktur

moderierte Mailing-Liste:
KDD nuggets:
http://www.kdnuggets.com
gps@kdnuggets.com
subscription:
kdd-request@gte.com
1.Zeile:
subscribe kdd-nuggets

Software:
S*i*ftware:
http://www.intelligent-dir.com

Internationale Konferenz:
KDD-97
http://www-aig.jpl.nasa.gov/kdd97

Buch:
U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy
"Advances in Knowledge Discovery and Data Mining"
The MIT Press, 1996