

Information Mining

Übungen 16.04.2008

Organisatorisches

❖ Zielgruppe:

- ❑ DAI Hauptstudium mit 8 Kreditpunkten:
Bereich „D“
- ❑ ISE Master mit 8 Kreditpunkten

❖ Ablauf:

- ❑ Übungsblätter (mehr oder weniger regelmäßig, ohne Abgabe, werden in den Übungen besprochen)
- ❑ Vertiefung des Stoffes aus der VL

Leistungsnachweis

- ❖ Mündliche Prüfung
- ❖ Termin mit Prof. Fuhr ausmachen

Kommunikation

- ❖ Per Email:

`ingo.frommholz@uni-due.de`

- ❖ Sprechstunde: nach Vereinbarung

- ❖ Wichtig: unser RSS-Feed!

`http://www.is.inf.unidue.de/
news.rss`

U.A. Ankündigung von Übungsblättern

Einführung

Was sind die wesentlichen
Problemstellungen des Data Mining?

Problemstellungen

❖ Klassifikation:

```
if outlook = overcast  
then play = yes
```

❖ Assoziation:

```
if windy = false and  
   play = no  
then outlook = sunny and  
   humidity = high
```

Problemstellungen

❖ Numerische Vorhersagen:

$$\begin{aligned} \text{RPR} = & -55.9 + \\ & 0.0489 \text{ MYCT} + \\ & 0.0153 \text{ MMIN} + \\ & 0.0056 \text{ MMAX} + \\ & 0.641 \text{ CACH} - \\ & 0.27 \text{ CHMIN} + \\ & 1480 \text{ CHMAX} \end{aligned}$$

Problemstellungen

- ❖ Clustering:
Gruppen von zusammengehörigen
Beispielen suchen



Generalisierung als Suche

- ❖ Induktives Lernen: Suche nach einer Konzeptbeschreibung, die zu den Daten passt
- ❖ Beispiel: Regelmenge als Beschreibungssprache
 - ❑ Riesiger, aber endlicher Suchraum
- ❖ Einfache Lösung:
 - ❑ Aufzählen der Elemente des Konzeptraums
 - ❑ Eliminieren aller Beschreibungen, die nicht zu den Beispielen passen
 - ❑ Verbleibende Beschreibung stellt das gesuchte Konzept dar

Aufzählen der Elemente des Konzeptraums

- ❖ Suchraum für das Wetterproblem:
 - ❑ $4 \times 4 \times 3 \times 3 \times 2 = 288$ mögliche Regeln
 - ❑ Beschränkung auf maximal 14 Regeln in der Beschreibung $\Rightarrow 2.7 \times 10^{34}$ mögliche Regelmengen
- ❖ Möglicher Ausweg: Algorithmus zur Eliminierung von Kandidaten
- ❖ Weitere praktische Probleme:
 - ❑ Mehr als eine Beschreibung kann übrig bleiben
 - ❑ Keine Beschreibung bleibt übrig
 - Beschreibungssprache ist ungeeignet, um das Zielkonzept zu beschreiben
 - Daten können verrauscht sein

Der Versionsraum

- ❖ Raum von konsistenten Konzeptbeschreibungen
- ❖ Komplette bestimmt durch 2 Mengen:
 - ❑ L: spezifischste Beschreibungen, die alle positiven und keine negativen Beispiele abdecken
 - ❑ G: generellste Beschreibungen, die keine negativen und alle positiven Beispiele abdecken
- ❖ Nur L und G müssen verwaltet und aktualisiert werden
- ❖ Aber: immer noch hoher Berechnungsaufwand
- ❖ Und: löst die anderen praktischen Problem nicht

Versionsraum: Beispiel

❖ Gegeben: Rote oder grüne Kühe oder Hühner

$$L = \{\}$$

$$G = \{ \langle *, * \rangle \}$$

$\langle \text{green}, \text{cow} \rangle$: positive

$$L = \{ \langle \text{green}, \text{cow} \rangle \} \quad G = \{ \langle *, * \rangle \}$$

$\langle \text{red}, \text{chicken} \rangle$: negative

$$L = \{ \langle \text{green}, \text{cow} \rangle \} \quad G = \{ \langle \text{green}, * \rangle, \langle *, \text{cow} \rangle \}$$

$\langle \text{green}, \text{chicken} \rangle$: positive

$$L = \{ \langle \text{green}, * \rangle \} \quad G = \{ \langle \text{green}, * \rangle \}$$

Algorithmus zur Kandidaten-Eliminierung

```
Initialize  $L$  and  $G$ 
For each example  $e$ :
  If  $e$  is positive:
    Delete all elements from  $G$  that do not cover  $e$ 
    For each element  $r$  in  $L$  that does not cover  $e$ :
      Replace  $r$  by all of its most specific generalizations
        that 1. cover  $e$  and
           2. are more specific than some element in  $G$ 
    Remove elements from  $L$  that
      are more general than some other element in  $L$ 
  If  $e$  is negative:
    Delete all elements from  $L$  that cover  $e$ 
    For each element  $r$  in  $G$  that covers  $e$ :
      Replace  $r$  by all of its most general specializations
        that 1. do not cover  $e$  and
           2. are more general than some element in  $L$ 
    Remove elements from  $G$  that
      are more specific than some other element in  $G$ 
```

Beispiel

Wasserstand	Wind	Segeln?
hoch	stark	ja
hoch	normal	ja
hoch	schwach	nein
mittel	stark	ja
mittel	normal	ja
mittel	schwach	nein
niedrig	stark	ja
niedrig	normal	ja
niedrig	schwach	nein

if wasserstand = ? and wind = ? then segeln = ja

Bias

Welche 3 Arten von Bias kennt ihr?

Bias (systematische Fehler)

- ❖ Die wichtigsten Entscheidungen in Lernsystemen:
 - Konzept-Beschreibungssprache
 - Reihenfolge, in der der Raum durchsucht wird
 - Vermeidung der Überadaptation an die Trainingsdaten
- ❖ Diese Eigenschaften bestimmen des "Bias" der Suche
 - Beschreibungssprachen-Bias
 - Such-Bias
 - Überadaptions-Vermeidungs-Bias

Beschreibungssprachen- Bias

- ❖ Wichtigste Frage:
 - ❑ Ist die Sprache universell oder beschränkt sie das zu Lernende?
- ❖ Universelle Sprache kann beliebige Teilmengen der Beispiele beschreiben
- ❖ Wenn die Sprache die Oder-Verknüpfung von Aussagen zulässt, ist sie universell
- ❖ Domänenwissen kann benutzt werden, um einige Konzeptbeschreibungen von vornherein von der Suche auszuschließen

Such-Bias

❖ Such-Heuristik

- ❑ "Greedy"-Suche: wähle jeweils den besten Einzelschritt aus
- ❑ "Beam"-Suche: Behalte mehrere Alternativen im Auge
- ❑ ...

❖ Richtung der Suche

- ❑ Vom Allgemeinen zum Speziellen
 - Z.B. Spezialisieren einer Regel durch Hinzufügen von Bedingungen
- ❑ Vom Speziellen zum Allgemeinen
 - Z.B. Generalisierung einer einzelnen Instanz zu einer Regel

Überadaptions- Vermeidungs-Bias

- ❖ Kann als Teil des Such-Bias gesehen werden
- ❖ Modifiziertes Bewertungskriterium
 - ❑ Z.B. Balance zwischen Einfachheit und Fehleranzahl
- ❖ Modifizierte Suchstrategie
 - ❑ Z.B. Pruning (Vereinfachen einer Beschreibung)
 - Pre-Pruning: Stoppt bei einer einfachen Beschreibung, bevor übermäßig komplexe Beschreibungen generiert werden
 - Post-Pruning: Generiert zunächst eine komplexe Beschreibung, die anschließend vereinfacht wird