Maschinelles Lernen und Data Mining

Übungsblatt für den 24.11.2005

Aufgabe 1

Gegeben sei das Golf-Spiel Datenset aus der Vorlesung.

```
@relation weather.symbolic
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Die positive Klasse sei die Klasse `yes`.

1. Führen Sie eine Iteration des Batch-FindG Algorithmus aus der Vorlesung durch. Woran erkennen Sie, daß dieses Problem nicht mit diesem Algorithmus lösbar ist?

2. Wenden Sie den Covering-Algorithmus an

   - mit dem Maß Precision
   - mit dem Maß Accuracy, wobei jede Regel solange verfeinert wird, bis sie nur mehr positive Beispiele abdeckt.

- mit dem Maß Accuracy, wobei jeweils die Regel mit der höchsten Bewertung verwendet wird.

Diskutieren Sie die Ergebnisse. Welche Regelmenge sieht am besten aus?

3. Wiederholen Sie 1.2, indem sie die Rolle der Klassen vertauschen (also die positive Klasse sei jetzt `no`).

4. Eine Bottom-Up (also Specific-To-General) Lern-Strategie zur Batch-Induktion einzelner Regeln könnte so aussehen, daß ein positives Beispiel zufällig ausgewählt wird, und dann sukzessive generalisiert wird. Simulieren Sie diese Strategie an diesen Trainings-Beispielen.

5. Eine alternative Strategie wäre, alle Beispiele in Regeln zu verwandeln, zwei beliebige Regeln auszuwählen, das lgg dieser Beispiele zu finden, und dann die beiden alten Regeln durch diese neue zu ersetzen. Wieso wird diese Strategie i.a. nicht funktionieren? Wie könnte man sie verbessern (z.B. durch Auswahl der Regeln, Abbruchbedingungen, etc.)?

6. Überlegen Sie sich, wie dieser Algorithmus mit numerischen bzw. hierarchischen Attributen umgehen könnte.

Aufgabe 2

Versuchen Sie, eine möglichst einfache Regelmenge zu finden oder zu lernen, die folgende Beispiele erklärt.

```
@relation x
@attribute a1 {0,1}
@attribute a2 {0,1}
@attribute a3 {0,1}
@attribute a4 {0,1}
@attribute x {yes, no}
@data
1,0,0,0,yes
1,1,0,1,yes
0,0,1,1,no
1,0,0,1,no
1,1,1,0,no
0,0,1,0,yes
0,0,0,1,no
1,1,0,0,no
0,1,1,1,yes
1,0,1,0,yes
0,1,0,1,yes
0,1,1,0,no
```

Aufgabe 3

1. Wo im Coverage Space liegen all jene Klassifizierer, die für ein Beispiel—unabhängig von seinen konkreten Attribut-Werten—mit einer Wahrscheinlichkeit von $p(+)$ die positive Klasse vorhersagen?

2. Overfitting aufgrund von fehlerhaften Trainings-Beispielen äußert sich oft, indem Regeln mit geringer Coverage gelernt werden. Identifizieren Sie den für Overfitting ausschlaggebenden Bereich im Coverage Space und überlegen Sie sich die Eigenschaften der in der Vorlesung besprochenen Maß e bezüglich Overfitting. Z.B. welch ein Maß neigt eher zu Overfitting, Precision oder Accuracy?