JProf. Dr. Uwe Ligges, Prof. Dr. Katharina Morik
Julia Schiffner, Felix Jungermann

Dortmund, 25. Mai
Abgabe: bis zum 01.06. um 10:00 Uhr an
`felix.jungermann@tu-dortmund.de`

Übungen zur Vorlesung
**Wissensentdeckung in Datenbanken**
Sommersemester 2010

Blatt 6

**Aufgabe 6.1 (10 Punkte)**

| ID | Play | Outlook | Temperature | Humidity | Wind |
|---|---|---|---|---|---|
| 1 | no | sunny | 85.0 | 85.0 | false |
| 2 | no | sunny | 80.0 | 90.0 | true |
| 3 | yes | overcast | 83.0 | 78.0 | false |
| 4 | yes | rain | 70.0 | 96.0 | false |
| 5 | yes | rain | 68.0 | 80.0 | false |
| 6 | no | rain | 65.0 | 70.0 | true |
| 7 | yes | overcast | 64.0 | 65.0 | true |
| 8 | no | sunny | 72.0 | 95.0 | false |
| 9 | yes | sunny | 69.0 | 70.0 | false |
| 10 | yes | rain | 75.0 | 80.0 | false |
| 11 | yes | sunny | 75.0 | 70.0 | true |
| 12 | yes | overcast | 72.0 | 90.0 | true |
| 13 | yes | overcast | 81.0 | 75.0 | false |
| 14 | no | rain | 71.0 | 80.0 | true |

Tabelle 1: Golf-Datensatz

(a) Stellen Sie mit Hilfe des ID3-Algorithmus den Entscheidungsbaum für den in Tabelle 1 angegebenen Golf-Datensatz manuell auf! Benutzen Sie als Gütemaß den aus der Vorlesung bekannten Informationsgewinn auf Basis der Entropie.
Für die Attribute 'Temperature' und 'Humidity' muss die Anzahl der Attribute vor Anwendung des Algorithmus eingeschränkt werden. Entwickeln Sie für beide Attribute jeweils drei sinnvolle Wertebereiche, in die Sie die vorhandenen Einträge einordnen.
(Beachten Sie: das Attribut 'Play' ist das Label. 'ID' und 'Play' werden nicht zum Lernen benutzt!)

(b) Es sind weitere Gütemaße anwendbar. Informieren Sie sich (z.B. im Internet bei `wikipedia.de`) über den sogenannte Gini-Index und benutzen Sie diesen als Gütemaß.

Stellen Sie den Entscheidungsbaum nun noch einmal mit Hilfe dieses Gütemaßes auf! Natürlich sollen Sie auch hier den Datensatz wie in der ersten Teilaufgabe hinsichtlich 'Temperature' und 'Humidity' vorverarbeiten.

(c) Laden Sie den Golf-Datensatz in RapidMiner ein. Benutzen Sie dann den Operator `ID3`, um einen Entscheidungsbaum zu erstellen. (Achtung: da dieser Operator nur nominale Werte verarbeiten kann, müssen Sie auch hier wieder die Werte für 'Temperature' und 'Humidity' in drei Bereiche einschränken.) Welchen Operator nimmt man hierfür am besten? (`Data Transformation/Type Conversion/Discretization` enthält hierfür nützliche Operatoren.)

(d) Sie haben nun drei Entscheidungsbäume vorliegen. Wenden Sie alle drei auf das noch nicht klassifizierte Beispiel in Tabelle 2 an! Sollte laut der drei Entscheidungsbäume bei

| ID | Play | Outlook | Temperature | Humidity | Wind |
|---|---|---|---|---|---|
| 15 | ? | sunny | 75.0 | 75.0 | true |

Tabelle 2: Golf-Wetter?

diesem Wetter Golf gespielt werden oder nicht?