Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

22.1.2009

10. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 Informationsgewinn

Im folgenden betrachten wir die Menge T von n Trainingsobjekten, mit den Attributen $A_1, \ldots, A_a$ und den k Klassen C_1 bis C_k.
Sei $\{T_i^A \mid i \in \{1, \ldots, m_A\}\}$ die disjunkte, vollständige Partitionierung von T, die durch einen Split auf dem Attribut A erzeugt wird (wobei m_A die Anzahl von Ausprägungen von A ist).

1. *Gleichverteilung*

   Berechnen Sie unter der Annahme, dass die Klassenzugehörigkeiten in T gleichverteilt und unabhängig von den Ausprägungen von A sind $\mathrm{entropie}(T)$, $\mathrm{entropie}(T_i^A)$ für $i \in \{1, \ldots, m_A\}$ sowie $\mathrm{informationsgewinn}(T, A)$. Interpretieren Sie Ihr Ergebnis!

2. *Zusätzliche gleichverteilte Ausprägung*

   Wir wollen untersuchen, inwieweit die Anzahl der Ausprägungen den Informationsgewinn beeinflusst.

   Betrachten wir dazu ein *beliebiges* Attribut A mit seinen m_A Ausprägungen. Wie ändert sich der $\mathrm{informationsgewinn}(T, A)$ wenn wir A durch A' mit $m_{A'} = m_A + 1$ Ausprägungen ersetzen, wobei die relativen Häufigkeiten in den Ausprägungen 1 bis m_A von A' identisch zu A sind und in der Ausprägung $m_{A'}$ eine Gleichverteilung der Klassen herrscht? Interpretieren Sie Ihr Ergebnis!

3. *Attribute mit sehr vielen Ausprägungen*

   Sei A ein Attribut mit zufälligen, nicht mit der Klasse der Objekte korrelierten Werten. Weiterhin verfüge A über so viele Ausprägungen, dass keine zwei Objekte der Trainingsmenge zu derselben Ausprägung in A gehören. Was geschieht in dieser Situation beim Aufbau das Entscheidungsbaumes? Was ist daran problematisch?

2 Entscheidungsbäume

1. Welche Form sollte ein Entscheidungsbaum haben? Möglichst breit oder möglichst tief? Warum?

2. Ein Krankenhaus möchte die Diagnosefähigkeit seiner Ärzte unterstützen. Dazu wurden Daten über gesunde und kranke Patienten gesammelt. Die Krankenhausleitung hat erfahren, dass man mit einem Entscheidungsbaumverfahren anhand vorhandener Beispieldaten ein Modell generieren kann, welches die Entscheidung eines Arztes simuliert. Berechnen Sie mittels der folgenden Daten einen Entscheidungsbaum und zeichnen Sie diesen auf.

Patient Nr.	Heart Rate	Blood Pressure	Klasse
1	irregular	Normal	Ill
2	regular	Normal	Healthy
3	irregular	Abnormal	Ill
4	irregular	Normal	Ill
5	regular	Normal	Healthy
6	regular	Abnormal	Ill
7	regular	Normal	Healthy
8	regular	Normal	Healthy

   Nutzen Sie zum Erstellen des Entscheidungsbaumes das Kriterium des *Informationsgewinns*. Ohne Taschenrechner nähern Sie bitte den Logarithmus mittels folgender Formel an: $\log_2(x) = 1 - 1/x$.

3. Definieren Sie den Begriff Overfitting. Schlagen Sie eine Strategie zur Vermeidung vor.

4. Beschreiben Sie das prinzipielle Vorgehen, um das Entscheidungsbaumlernen zu parallelisieren.

3 Vorbereitung der letzten Übung

Überlegen Sie sich für die letzte KDD-Übung am 12.2.2009 Verfahren, zu denen Sie gerne noch einmal eine Übungsaufgabe rechnen würden. Schicken Sie ihre Vorschläge bis zum 8.2.2008 an jaeschke@cs.uni-kassel.de.