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 $\text{entropie}(T)$, $\text{entropie}(T_i^A)$ für $i \in \{1, \ldots, m_A\}$ sowie $\text{informationsgewinn}(T, A)$. Interpretieren Sie Ihr Ergebnis!

$$
\begin{aligned}
\text{entropie}(T) &= -\sum_{i=1}^{k} p_i \log_2(p_i) = -\sum_{i=1}^{k} \frac{|C_i \cap T|}{|T|} \log_2 \left(\frac{|C_i \cap T|}{|T|} \right) \\
&= -\sum_{i=1}^{k} \frac{1}{k} \log_2 \left(\frac{1}{k} \right) = -\log_2 \left(\frac{1}{k} \right) = \log_2(k) \\
\text{entropie}(T_i^A) &= -\sum_{j=1}^{k} \frac{|C_j \cap T_i^A|}{|T_i^A|} \log_2 \left(\frac{|C_j \cap T_i^A|}{|T_i^A|} \right) \\
&= -\sum_{j=1}^{k} \frac{1}{k} \log_2 \left(\frac{1}{k} \right) = -\log_2 \left(\frac{1}{k} \right) = \log_2(k) \\
\text{entropie}_A(T) &= \sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} \text{entropie}(T_i^A) \\
&= \log_2(k) \sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} = \log_2(k) \\
\text{informationsgewinn}(T, A) &= \text{entropie}(T) - \mathit{entropie}_A(T) \\
&= \log_2(k) - \log_2(k) = 0
\end{aligned}
$$

   Durch die Gleichverteilung der Klassenzugehörigkeit und die Unabhängigkeit von den Merkmalsausprägungen von A ist der Informationsgewinn gleich Null. Denn aus den Ausprägungen von A läßt sich nicht auf die Klassenzugehörigkeit schließen.

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 $\text{informationsgewinn}(T, A)$ 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!

   $$\text{informationsgewinn}(T, A') = \text{informationsgewinn}(T, A) - \frac{|T^A_{m_A+1}|}{|T|} \log_2(k)$$

   Der Informationsgewinn wird kleiner, denn die zusätzliche Ausprägung kann nicht zum Lernen der Klassenzugehörigkeit genutzt werden.

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?

   Der Entscheidungsbaum lernt die Trainingsmenge auswendig, d. h. jedes Blatt enthält genau ein Trainingsobjekt. Das hat zur Folge, dass neue Objekte nicht korrekt klassifiziert werden können, wenn sie nicht exakt die gleichen Merkmalsausprägungen wie eines der Trainingsobjekte haben.

2 Entscheidungsbäume

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

   Weder Breite noch Tiefe sind ein qualitatives Maß für einen Entscheidungsbaum. Ziel ist eine einfache Klassenbeschreibung. Die Form des Entscheidungsbaumes ist zusätzlich abhängig vom Verfahren (binäre Splits ergeben weniger breite Bäume). Große Tiefe und Breite können im Extremfall zu Überspezialisierung (Overfitting) führen.

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: $\log_2(x) = 1 - 1/x$.

Folgende Formeln sind hier wichtig:

$$\text{informationsgewinn}(X) = \text{entropie}(T) - \text{entropie}_X(T) \tag{1}$$

$$\text{entropie}(T) = -\sum_{i=1}^{k} \frac{|C_i \cap T|}{|T|} \log_2 \left(\frac{|C_i \cap T|}{|T|} \right) \tag{2}$$

$$\text{entropie}_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \text{entropie}(T_i) \tag{3}$$

$$\text{entropie}(T_i) = -\sum_{j=1}^{k} \frac{|C_j \cap T_i|}{|T_i|} \log_2 \left(\frac{|C_j \cap T_i|}{|T_i|} \right) \tag{4}$$

Dabei ist n die Anzahl der unterschiedlichen Ausprägungen des Attributs X und T_i die Menge der Objekte für die Attribut X die Ausprägung i hat. Des weiteren ist k die Anzahl der unterschiedlichen Ausprägungen des Klassifikationsattributes C und C_i die Menge der Objekte, die als i klassifiziert wurden. Schließlich ist T die Menge aller Objekte.

Im konkreten Beispiel gilt nun:

$$\text{info(T)} = -2 * \frac{4}{8} * \log_2(\frac{4}{8}) = -\log_2(\frac{1}{2}) = -(1-2) = 1$$

Für X = Heart Rate (HR) gilt nun :

$$\text{info}_{\text{HR}}(T) = (\frac{|T_{regular}|}{|T|} * \text{info}(\text{T}_{\text{regular}}) + \frac{|T_{irregular}|}{|T|} * \text{info}(\text{T}_{\text{irregular}}))$$

dabei :

$$\text{info}(\text{T}_{\text{regular}}) = -(\frac{|C_{ill} \cap T_{regular}|}{|T_{regular}|}) * \log_2(\frac{|C_{ill} \cap T_{regular}|}{|T_{regular}|}) + \frac{|C_{healthy} \cap T_{regular}|}{|T_{regular}|} * \log_2(\frac{|C_{healthy} \cap T_{regular}|}{|T_{regular}|}))$$

$$= -(\frac{1}{5} * \log_2(\frac{1}{5}) + \frac{4}{5} * \log(\frac{4}{5})) = -(\frac{1}{5})(1-5) - \frac{4}{5}(1-\frac{5}{4}) = \frac{4}{5} + \frac{4}{20} = \frac{4}{5} + \frac{1}{5} = 1$$

$$\text{info}(\text{T}_{\text{irregular}}) = -(\frac{|C_{ill} \cap T_{irregular}|}{|T_{irregular}|}) * \log_2(\frac{|C_{ill} \cap T_{irregular}|}{|T_{irregular}|}) + \frac{|C_{healthy} \cap T_{irregular}|}{|T_{irregular}|} * \log_2(\frac{|C_{healthy} \cap T_{irregular}|}{|T_{irregular}|}))$$

$$= -(\frac{3}{3} * \log_2(\frac{3}{-}) + \frac{0}{3} * \log_2(\frac{0}{3})) = 0$$

Und folglich :

$$\text{info}_{\text{HR}}(T) = (\frac{5}{8} * 1 + \frac{3}{8} * 0) = \frac{5}{8}$$

Weiter für X = Blood Pressure(BP) :

$$\text{info}_{\text{BP}}(T) = (\frac{|T_{normal}|}{|T|} * \text{info}(\text{T}_{\text{normal}}) + \frac{|T_{abnormal}|}{|T|} * \text{info}(\text{T}_{\text{abnormal}}))$$

dabei :

$$\text{info}(\text{T}_{\text{normal}}) = -(\frac{|C_{healthy} \cap T_{normal}|}{|T_{normal}|}) * \log_2(\frac{|C_{healthy} \cap T_{normal}|}{|T_{normal}|}) + \frac{|C_{ill} \cap T_{normal}|}{|T_{normal}|} * \log_2(\frac{|C_{ill} \cap T_{normal}|}{|T_{normal}|}))$$

$$= -(\frac{4}{6} * \log_2(\frac{4}{6}) + \frac{2}{6} * \log(\frac{2}{6})) = -(\frac{4}{6})(1-\frac{6}{4}) - \frac{2}{6}(1-\frac{6}{2}) = \frac{8}{24} + \frac{8}{12} = \frac{8}{24} + \frac{16}{24} = 1$$

$$\text{info}(\text{T}_{\text{abnormal}}) = -(\frac{|C_{ill} \cap T_{abnormal}|}{|T_{abnormal}|}) * \log_2(\frac{|C_{ill} \cap T_{abnormal}|}{|T_{abnormal}|}) + \frac{|C_{healthy} \cap T_{abnormal}|}{|T_{abnormal}|} * \log_2(\frac{|C_{healthy} \cap T_{abnormal}|}{|T_{abnormal}|}))$$

$$= -(\frac{2}{2} * \log_2(\frac{2}{2}) + \frac{0}{2} * \log_2(\frac{0}{2})) = 0$$

Und somit :

$$\text{info}_{\text{BP}}(T) = (\frac{6}{8} * 1 + \frac{2}{8} * 0) = \frac{6}{8}$$

Ingesamt erhalten wir also :

$$\text{gain(HR)} = \text{info(T)} - \text{info}_{\text{HR}}(T) = 1 - \frac{5}{8} = \frac{3}{8}$$

$$\text{gain(BP)} = \text{info(T)} - \text{info}_{\text{BP}}(T) = 1 - \frac{6}{8} = \frac{2}{8}$$

Folglich ist HeartRate das bessere (d.h. stärker diskriminierende) Attribut und sollte im Entscheidungsbaum vor BloodPressure stehen.

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

   Overfitting ist die Überanpassung des gelernten Modells (Entscheidungsbaum) an die Trainingsdaten und daraus resultierendes schlechtes Abschneiden des Klassifikators auf unbekannten Daten.
   Strategien zur Vermeidung von Overfitting sind neben der Wahl geeigneter Parameter (Größe der Trainingsmenge, minimaler Support, minimale Konfidenz) vor allem das Entfernen fehlerhafter Trainingsdaten, nachträgliches Pruning (Abschneiden) von Ästen des Baumes und Überkreuz-Validierung.

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

   Durch getrennte Attributlisten an jedem Knoten des Entscheidungsbaumes lässt sich der Algorithmus parallelisieren. Dazu wird die Datenmenge an jedem Knoten gesplittet und auf die Äste verteilt. Jeder Ast kann dann parallel berechnet werden.

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.