Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungslehrstuhl
Wilhelmshöher Allee 73
34121 Kassel

Email: hotho@cs.uni-kassel.de
Tel.: ++49 561 804-6252

Dr. Andreas Hotho
Prof. Gerd Stumme

08.06.04

Knowledge Discovery

Lösungsblatt 2

Sommersemester 2004

Aufgabe 1: Preprocessing und Datenbeschaffenheit

a) Welche zwei Typen des Preprocessing gibt es? Diskutieren Sie in diesem Zusammenhang den Begriff Datenverständnis?

1. Vorbereitung der Daten
2. Vorbereitung des Miners (oder auch des Geschäftsmannes)

Das Verständnis der Daten ist ein wesentlicher Schritt für die korrekte Ableitung von Preprocessing-Schritten. Nur mit dem nötigen Verständnis der Daten ist der Miner in der Lage die richtigen Schritte so abzuschätzen, dass der Daten für die Verarbeitung durch den Algorithmus in adäquater Form vorliegen. Mit Hilfe von statistischen Datencharakteristiken oder einer explorativen Analyse/Visualisierung kann man Preprocessingschritte wie Reduktion, Ableitung und Transformation für einen entsprechenden Datensatz bestimmen und deren Erfolg wiederum überprüfen. Eine klare Trennung von Datenverständnis und Preprocessing ist daher eher schwierig.

b) In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Nennen Sie die drei wichtigsten Skalenniveaus und beschreiben Sie sie kurz. Was bedingt ein Skalenniveau bei der Untersuchung von Daten?

Skalenniveaus

Nominalskalierte Merkmale: Ausprägungen sind qualitativ, keine Ordnung möglich (rot,grün)

Ordinalskalierte Merkmale: Ausprägungen können geordnet, aber Abstände nicht interpretiert werden. (Tafelwein, Qualitätswein, prämierter Qualitätswein

Kardinalskalierte Merkmale: Ausprägungen sind Zahlen, Interpretation der Abstände möglich (metrisch)

Auswirkungen:

- Informationsgehalt der Daten
- sinnvolle Anwendbarkeit von Rechenoperationen

c) Welche Ziele werden in Bezug auf die spätere Anwendung von Algorithmen mit Preprocessing im Data Mining verfolgt. Nennen Sie in diesem Zusammenhang zwei Beispiele, in denen Algorithmen bestimmte Preprocessingschritte erzwingen.

Die Daten müssen so vorverarbeitet werden, dass die in ihnen enthaltenen Informationen bestmöglich dem anzuwendenden Verfahren zur Verfügung stehen.

- **Neuronale Netze** brauchen z.B. numerische Daten
- **Entscheidungsbäume** funktionieren besser mit kategorischen Daten

Aufgabe 2: Entscheidungsbäume

a) 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.

b) 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 *gain-ratio*(x) Kriterium:

Ohne Taschenrechner nähern Sie bitte den Logarithmus mittels folgender Formel an: $log_2(x) = 1 - \frac{1}{x}$

Folgende Formeln sind hier wichtig:

$$\text{gain\_ratio(x)} = \frac{\text{gain(x)}}{\text{split\_info(x)}}$$

$$\text{gain(x)} = \text{info(T)} - \text{info}_x\text{(T)}$$

$$\text{split\_info(x)} = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} * \log_2(\frac{|T_i|}{|T|}),$$

$$\text{info(T)} = -\sum_{i=1}^{k} \frac{|C_i \cap T|}{|T|} * \log_2(\frac{|C_i \cap T|}{|T|}),$$

$$\text{info}_x\text{(T)} = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * \text{info}(T_i)$$

$$\text{info}(T_i) = -\sum_{i=1}^{k} \frac{|C_i \cap T_i|}{|T_i|} * \log_2(\frac{|C_i \cap T_i|}{|T_i|})$$

Dabei ist n immer die Anzahl der unterschiedlichen Ausprägungen des Attributs X und T_i dann die Menge der Objekte für die Attribut X den die Ausprägung i hat.
K ist dann die Anzahl der Unterschiedlichen Ausprägungen des Klassifikationsattributes C und C_i dann die Menge der Objekte, die als i klassifiziert wurden.
T ist 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 :

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

dabei :

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

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

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

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

Und folglich :

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

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 :

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

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

Für die split_infos gilt nun :

$$\text{split\_info(HR)} = -(\frac{|T_{regular}|}{|T|} * \log_2(\frac{|T_{regular}|}{|T|}) + \frac{|T_{irregular}|}{|T|} * \log_2(\frac{|T_{irregular}|}{|T|})) = -(\frac{5}{8} * \log_2(\frac{5}{8}) + \frac{3}{8} * \log_2(\frac{3}{8}))$$

$$= -(\frac{5}{8} * (1-\frac{8}{5}) + \frac{3}{8} * (1-\frac{8}{3})) = \frac{5}{8} * \frac{3}{5} + \frac{3}{8} * \frac{5}{3} = \frac{15}{40} + \frac{15}{24} = \frac{3}{8} + \frac{5}{8} = \frac{8}{8} = 1$$

$$\text{split\_info(BP)} = -(\frac{|T_{normal}|}{|T|} * \log_2(\frac{|T_{normal}|}{|T|}) + \frac{|T_{abnormal}|}{|T|} * \log_2(\frac{|T_{abnormal}|}{|T|})) = -(\frac{6}{8} * \log_2(\frac{6}{8}) + \frac{2}{8} * \log_2(\frac{2}{8}))$$

$$= -(\frac{6}{8} * (1 - \frac{8}{6}) + \frac{2}{8} * (1 - \frac{8}{2})) = \frac{6}{8} * \frac{2}{6} + \frac{2}{8} * \frac{6}{2} = \frac{12}{48} + \frac{12}{16} = \frac{1}{4} + \frac{3}{4} = 1$$

Also gain_ratio(HR) = gain(HR) und gain_ratio(BP) = gain(BP)

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