

DKFZ Heidelberg

Maschinelles Lernen, Klassifikation und diagnostische Tests (4)

Falk Schubert
Intelligente Bioinformatiksysteme
DKFZ Heidelberg

f.schubert@dkfz.de

Klassifikatoren

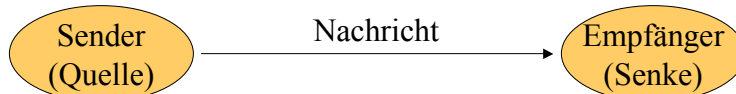
- Support vector machines (SVM)
- K-nächster-Nachbar (k-NN)
- Neural network (KNN, ANN)
- Entscheidungsbäume, Gruppen von Bäumen (forests)

Informationstheorie

Informationstheorie

- Shannon (1948)
- Information ist eine Grundgröße
- Information Maß für die Unsicherheit, die durch Eintreffen eines Zeichens oder Ereignisses beseitigt wird

Modell Nachrichtensystem



Nachrichten bestehen aus Zeichenfolgen
aus dem Alphabet A .

Für jedes Zeichen a aus A sei dessen
Sendewahrscheinlichkeit p_a bekannt.

Definition Entropie

- Informationsgehalt eines Zeichens a :

$$I(a) = \text{ld}(1/p_a) \text{ [Bit]}$$

- Entropie, mittlere Entscheidungsgehalt
eines Zeichenvorrates A :

$$H(A) = \sum_{a \in A} p_a \text{ld} \frac{1}{p_a} \text{ [Bit]}$$

- Entropie ist maximal bei Gleichverteilung
der Zeichen

Motivation des Logarithmus

- Motivation für Logarithmus:
 - Null für sichere Ereignisse: $I(a)=0$ für $p_a=1$
 - Über unabhängige Ereignisse summierbar
 $I(a,b)=0$ für $p_a=1$
- $I(a)=\text{ld}(1/p_a)$ [Bit]
 - Logarithmus zur Basis 2: binary units (bits)
- $I(a)=\ln(1/p_a)$ [Nats]
 - Logarithmus zur Basis e: natural units (nats)

Beispiel

- Alphabet = {T, G, C, A}
- $p_A = 0.3$
- $p_T = 0.3$
- $p_G = 0.2$
- $p_C = 0.2$

- $H = 1.97$

Redundanz

- Entscheidungsgehalt H_0
 - Maß für die Größe des Alphabets
 - $H_0 = \text{ld } |A|$
- Redundanz
 - $R = H_0 - H$

Analytische Bewertung eines Klassifikators

- AIC
 - Akaike information criterion
- MDL
 - Minimum description length
- VC-Dimension

Transinformation (mutual information)

- $T(\text{Attribut 1, Attribut 2}) = H(\text{Attribut 1}) + H(\text{Attribut 2}) - H(\text{Attribut 1, Attribut 2})$
- Anwendung u.a. zur Merkmalsselektion

Exkurs: Minimum description length

- Rissanen, 1978
- Länge des kleinsten binären Programmes zur Berechnung (Erzeugung) eines Datensatzes
- Nutze Regularitäten in einem Datensatz zu seiner optimalen Beschreibung
- Zur Evaluierung von Hypothesen
 - $MDL = \text{Länge des Modells} + \text{Länge der Ausnahmen}$
 - $MDL = L[\text{Modell}] + L[\text{Trainingsdaten} | \text{Modell}]$
 - Keine Testdaten zur Evaluierung nötig

Zusammenfassung

- Information ist eine Grundgröße und gibt die Unsicherheit an, die durch Eintreffen eines Zeichens oder Ereignisses beseitigt wird
- Informationsgehalt eines Zeichens a:
 $I(a) = \text{ld}(1/p_a)$ [Bit]
- Entropie, mittlere Entscheidungsgehalt eines Zeichenvorrates A:

$$H(A) = \sum_{a \in A} p_a \text{ld} \frac{1}{p_a} \text{ [Bit]}$$

■

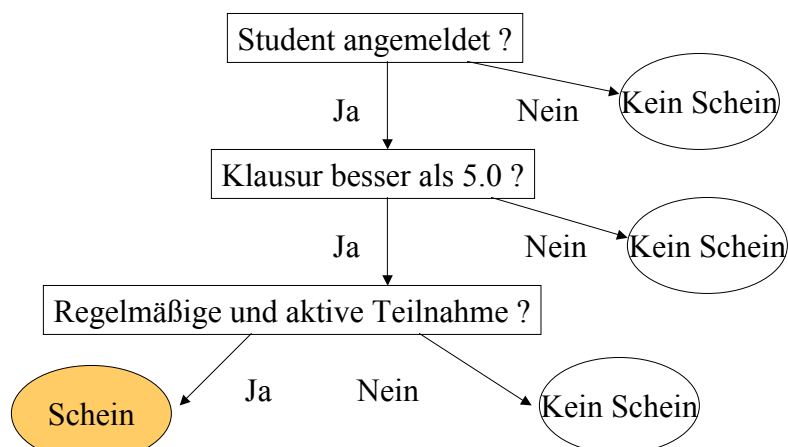
Klassifikation mit Bäumen

Klassische Entscheidungsbäume

- Aufgrund einer Sequenz von Fragen wird eine Alternative ausgewählt
- Darstellung der Abfolge der Fragen als Baum
- Blattknoten sind Klassifikationen (Entscheidungen)
- Anwendungen:
 - Fehlersuche
 - Betriebsanweisungen

Beispiel klassischer Entscheidungsbaum

- Scheinvergabe



Entscheidungsbäume

- Nominale, kategoriale und quantitative Attribute
 - Je ein Vergleichsknoten für eine nominale Merkmalsausprägung
 - Vergleiche mit größer, gleich oder kleiner zu einer Konstanten bei allen anderen Merkmalsausprägungen
- Blattknoten enthalten Klassenzuordnung

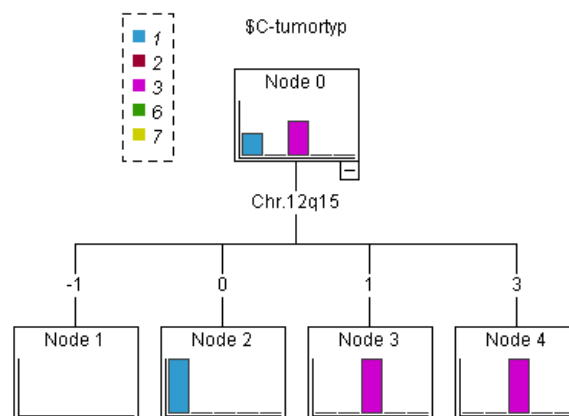
DKFZ Heidelberg

17

Falk Schubert

Entscheidungsbaum (einfach)

- Pict



DKFZ Hei

k Schubert

Divide and conquer

- Teile und Herrsche
- Basisprinzip in der Informatik

- 1. Rekursive Zerlegung eines komplexen Problems in leichtere Teilprobleme (Divide)
- 2. Lösung der Teilprobleme (Conquer)
- 3. Kombination der Lösungen der Teilprobleme

Beispiel Divide and Conquer

- Sortierung einer ungeordneten Liste mit „Mergesort“: -5,8,312,12,0,-4,283,2

- Zerlege die Liste in zwei Hälften
 - Bis jede Teilliste nur noch 2 Elemente enthält
- Sortiere jede Teilliste
- Kombiniere die sortierten Teillisten

- Sortierung von zwei Listenelementen trivial

Rekursion

- Basisprinzip der Programmierung
- Programm oder Funktion, ruft sich selbst auf
- Beispiel Fakultät:
 - $\text{Fakultaet}(n) = n * \text{Fakultaet}(n-1)$
 - $\text{Fakultaet}(1) = 1$
 - $\text{Fakultaet}(0) = 1$

Maschinelles Lernen von Entscheidungsbäumen

- CART
 - Classification and Regression Trees
 - Breiman
- C4.5, kommerzieller Nachfolger C5.0
 - Quinlan
- Random Forests
 - Breiman

Basisprinzip

- Rekursive Zerlegung des (Eingabe-)Raumes zum Erreichen einer maximalen Klassentrennung
- Für jede Ebene (beginnend mit der Wurzel) wähle ein Attribut und eine Aufteilung
 - Mit einem Schwellwert für ordinale und quantitative Merkmale
 - Mit einer Mengenaufteilung für nominale Merkmale

Stop des Algorithmus

- Jeder Knoten enthält nur noch Datensätze einer Klasse
- Eine Mindestanzahl von Datensätzen pro Aufteilung wurde unterschritten.

Auswahl der Merkmale

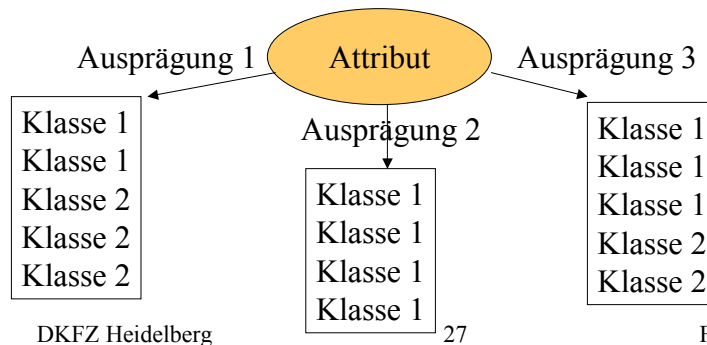
- Ziele
 - gute Auftrennung der beiden Klassen
 - Kleine Bäume (wenige Knoten)
- Wähle das Attribut mit dem höchsten Informationsgewinn (gain)
 - Gewichtete Entropie vor der Aufteilung
 - Gewichtete Entropie nach der Aufteilung

Informationsgewinn

- Erwartete Entropiereduktion durch eine Aufteilung an einem Attribut
- $\text{Gain}(\text{Attribut}|\text{Aufteilung}) =$
 - + $\text{info}[\text{Klassenaufteilung}]$
 - $g_1 * \text{info}[\text{Klassenaufteilung Knoten 1}]$
 - $g_2 * \text{info}[\text{Klassenaufteilung Knoten 2}]$
- $g_i = \frac{|\text{Klassenaufteilung Knoten } i|}{|\text{Klassenaufteilung}|}$
- $\text{info}[a,b] = H(a/(a+b)) + H(b/(a+b))$

Beispiel Informationsgewinn

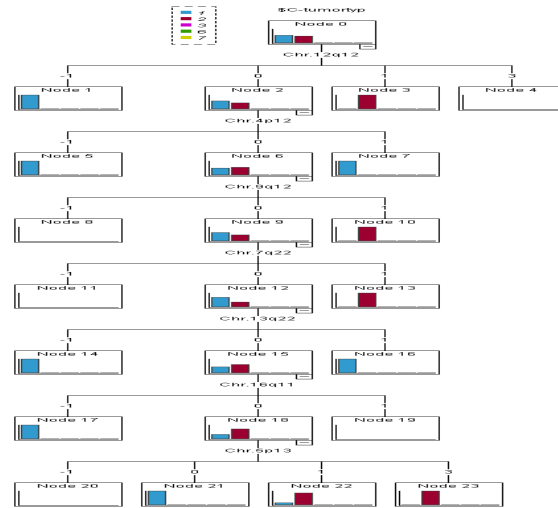
- $\text{gain}(\text{Attribut}) = \text{info}([9,5])$
 - $\text{info}([2,3], [4,0], [3,2])$
 - = $0.940 - 0.693$ bits



Quantitative Attribute

- Berechne Informationsgewinn für jeden möglichen Splitpunkt und wähle den besten
- Splitpunkte werden in der Mitte zwischen zwei beobachteten Werten gesetzt

Entscheidungsbaum (komplex)



DKFZ Heidelberg

29

Falk Schubert

Beispielregeln

Rules for 1 - contains 4 rule(s)

Rules for 2 - contains 3 rule(s)

Rule 1 for 1

Rule 1 for 2

if Chr.9p22 = -1
and Chr.5q23 = 0

if Chr.5q23 = 1

then 1

then 2

...

Rule 2 for 1

if Chr.11p15 = -1

then 1

...DKFZ Heidelberg

30

Falk Schubert

Vorteile von Bäumen

- Einfach zu verstehen und zu erklären
- Etablierte und einfach zu verwendende Methode
- Im medizinischen Bereich:
 - Für Ärzte intuitiv nicht einfach zu verstehen.

Nachteile von Bäumen

- Fehlende Stabilität
 - Baum ändert sich meist bei Verwendung einer geringfügig anderen Trainingsmenge
- Achsenparallele Trennfunktion

Pruning

- Beschneiden von Bäumen
 - Während der Baumerstellung
 - Am Ende (Postpruning)
 - Unterbaumersetzung
 - Unterbaumhochsetzen
- Abschätzung der Fehlerrate, ggf. statist. Test
- Ziel: Generalisierungsfähigkeit des Baumes erhöhen

Zusammenfassung

- Entscheidungsbäume sind erzeugen einfach zu verstehende, stückweise axen-parallele Klassifikatoren
- Training umfasst die rekursive Zerlegung des (Eingabe-)Raumes zum Erreichen einer maximalen Klassentrennung
- Klassentrennung wird nach informationstheoretischen Größen („gain“) maximiert.
- Beispiele für Entscheidungsbäume: CART, C4.5, C5.0