Decision Tree Learning 2*

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Sommersemester 2011

05.05.2011

*Folien basieren teilweise auf Material von Tom Mitchell

Verbesserungen des Basisalgorithmus

Information Gain (IG)

Definition

Information Gain (IG)

Definition

$$Gain(S, A) \equiv Entropy(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Information Gain (IG)

Definition

$$Gain(S, A) \equiv Entropy(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Problem

- IG zielt auf möglichst große **Reinheit** der entstehenden Unterknoten ab.
- die **Anzahl der Unterknoten spielt keine Rolle**
- ein Datensplit mit einer Instanz pro Knoten ist rein, aber nicht unbedingt gut (keine Generalisierung)

Gain Ratio (GR)

Gain Ratio

bestraft kleinteilige Splits, durch Einbezug der **SplitInformation** (SI)

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$

$$SplitInformation(S, A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

wobei S_i die Teilmenge von S ist, für die A den Wert v_i hat, und c die Anzahl der Teilmengen, die durch Anwendung von A entstehen.

Beispiel Gain Ratio

Datum	Dozent	Thema	Wochentag	Wetter	Max da?
1.4.2011	B	A	D	S	N
4.4.2011	A	D	F	R	N
8.4.2011	B	D	D	R	J
11.4.2011	A	A	F	S	N
15.4.2011	B	A	D	R	J

Beispiel Gain Ratio

Datum	Dozent	Thema	Wochentag	Wetter	Max da?
1.4.2011	B	A	D	S	N
4.4.2011	A	D	F	R	N
8.4.2011	B	D	D	R	J
11.4.2011	A	A	F	S	N
15.4.2011	B	A	D	R	J

$$IG(S, Datum) = 0.971 - 0 = 0.971$$

$$\begin{aligned} SI(S, Datum) &= -\sum_{i=1}^{5} \frac{1}{5} \log_2 \frac{1}{5} \\ &= -\sum_{i=1}^{5} \frac{1}{5}(-2.322) \\ &= -2.322 \end{aligned}$$

$$GR(S, Datum) = \frac{0.971}{-2.322} = -0.418$$

$$GR(S, Wetter) = \frac{0.42}{0.971} = 0.433$$

Sonderfälle

Was passiert, wenn $|S_i| = |S|$?

Sonderfälle

Was passiert, wenn $|S_i| = |S|$?

- $SplitInformation(S, A) = -\sum_{i=1}^{c} 1 \log_2 1 = 0$
- $GainRatio(S, A) = undef$

Sonderfälle

Was passiert, wenn $|S_i| = |S|$?

- $SplitInformation(S, A) = -\sum_{i=1}^{c} 1 \log_2 1 = 0$
- $GainRatio(S, A) = undef$

Heuristische Lösung

Gain Ratio wird als eine Art Filter angewendet:

- berechne zuerst IG für alle A
- wende GR nur auf A an, deren IG überdurchschnittlich ist

ID3 vs. C4.5

Algorithmen zum Erlernen von Entscheidungsbäumen

ID3 (Quinlan, 1986)

einer der bekanntesten Algorithmen.

- nur diskrete Attribut-Werte
- kein Pruning

C4.5 (Quinlan, 1993)

Weiterentwicklung von ID3.

- diskrete und kontinuierliche Attribut-Werte
- Pruning gegen Overfitting
- fehlende Attribut-Werte möglich
- Attribute können mit Kosten versehen werden

Kontinuierliche Werte

Diskretisierung durch Schwellenwert

- $Temperature = 82.5$
- $(Temperature > 72.3) = t, f$

Temperature:	40	48	60	72	80	90
Temperature>*72.3*:	No	No	No	No	Yes	Yes

Fehlende Werte

Heuristiken während des Trainings

Wenn der Wert v des Attributs A bei einer Instanz x nicht bekannt ist, dann

- weise x den häufigsten Wert von A zu **oder**
- weise x den Wert von A zu, der bei anderen Instanzen mit der gleichen Ausgabeklasse am häufigsten vorkommt

Die gleichen Heuristiken können auch bei fehlenden Werten bei Testinstanzen angewendet werden.

Werte mit verschiedenen Kosten

Kosten können bei IG/GR berücksichtigt werden, z.B.:

$$\frac{Gain^2(S, A)}{Cost(A)}$$

Bibliographie I

J. R. Quinlan (1986):
Induction of Decision Trees.
Machine Learning 1:1, 81-106.

J.R. Quinlan (1993):
C4.5: Programs for Machine Learning.
Morgan Kaufmann Publishers.