Diagnosing Cancerous Abnormalities with Decision Tree Learning

Autoren: Thomas Hecker, Jörg Mennicke
Betreuer: Prof. Dr. Ute Schmid
In Kooperation mit dem Fraunhofer IIS (Erlangen)

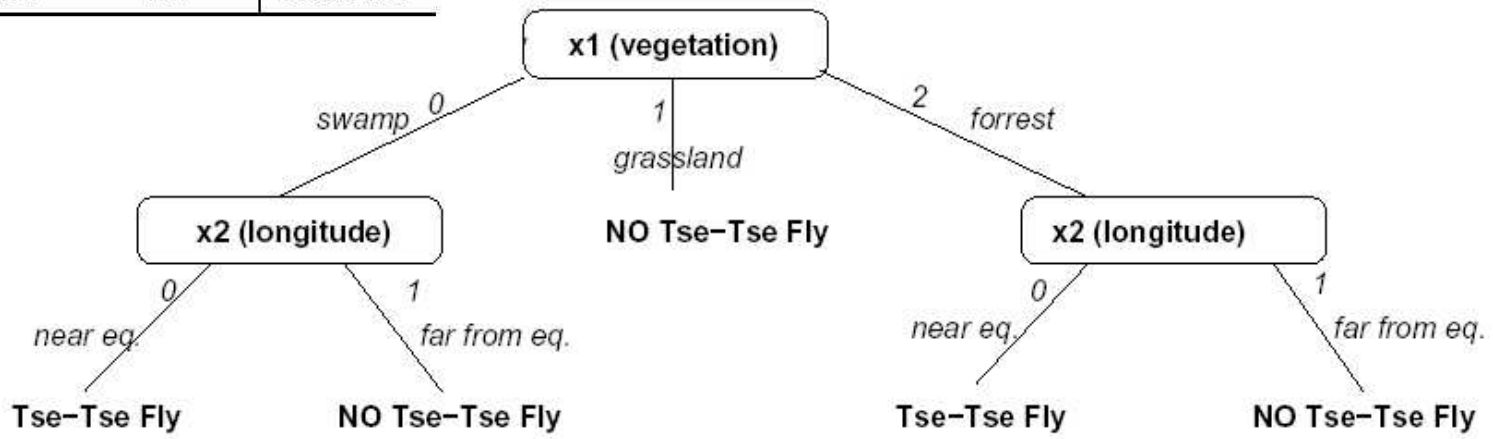
Gliederung

- Lernen mit Entscheidungsbäumen (DT)
- DTs für kontinuierliche Attribute
- K-nearest Neighbour vs. DT Learning
- Ergebnisse
- Generierte Hypothesen
- Mögliche Verbesserungen

Lernen mit Entscheidungsbäumen

- Verfahren: C4.5, CAL5
- Hypothese = Entscheidungsbaum (DT)

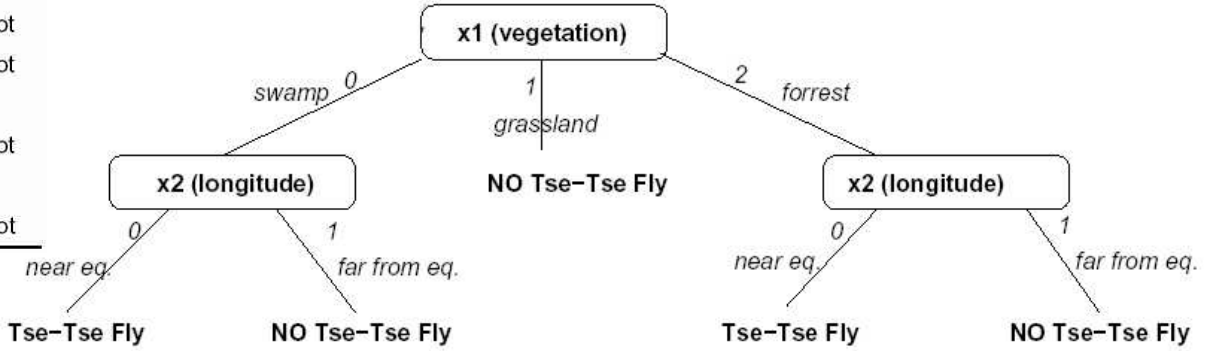
Nr	Vegetation	Longitude	Humidity	Altitude	Tse-Tse fly
1	swamp	near eq.	high	high	occurs
2	grassland	near eq.	high	low	occurs not
3	forrest	far f. eq.	low	high	occurs not
4	grassland	far f. eq.	high	high	occurs not
5	forrest	near eq.	high	low	occurs
6	grassland	near eq.	low	low	occurs not
7	swamp	near eq.	low	low	occurs
8	swamp	far f. eq.	low	low	occurs not



Lernen mit Entscheidungsbäumen

- Vom DT zu Entscheidungsregeln:

Nr	Vegetation	Longitude	Humidity	Altitude	Tse-Tse fly
1	swamp	near eq.	high	high	occurs
2	grassland	near eq.	high	low	occurs not
3	forrest	far f. eq.	low	high	occurs not
4	grassland	far f. eq.	high	high	occurs not
5	forrest	near eq.	high	low	occurs
6	grassland	near eq.	low	low	occurs not
7	swamp	near eq.	low	low	occurs
8	swamp	far f. eq.	low	low	occurs not



IF-THEN Regeln:

- **x1 = grassland → occurs**
- **x1 = forrest AND x2 = far → occurs_not**
- **...**

Lernen mit Entscheidungsbäumen

- Generalisierung:
  - Mustererkennung in Daten
  - Betrachtet nur relevante Attribute

- Probleme:
  - Wahl geeigneter Attributtests
    (relevante Muster identifizieren)

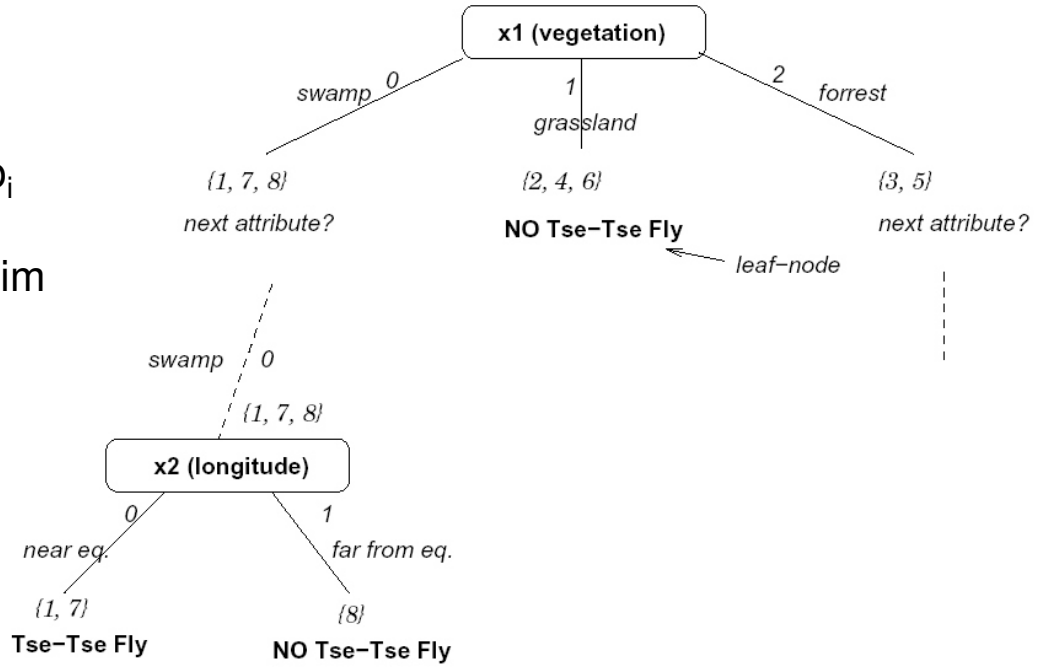
  - Abbruchbedingung / Overfitting
    (zufällige Muster ignorieren)

  - Klassen mit vielen Instanzen werden bevorzugt
    (-> Balancing)

Lernen mit Entscheidungsbäumen

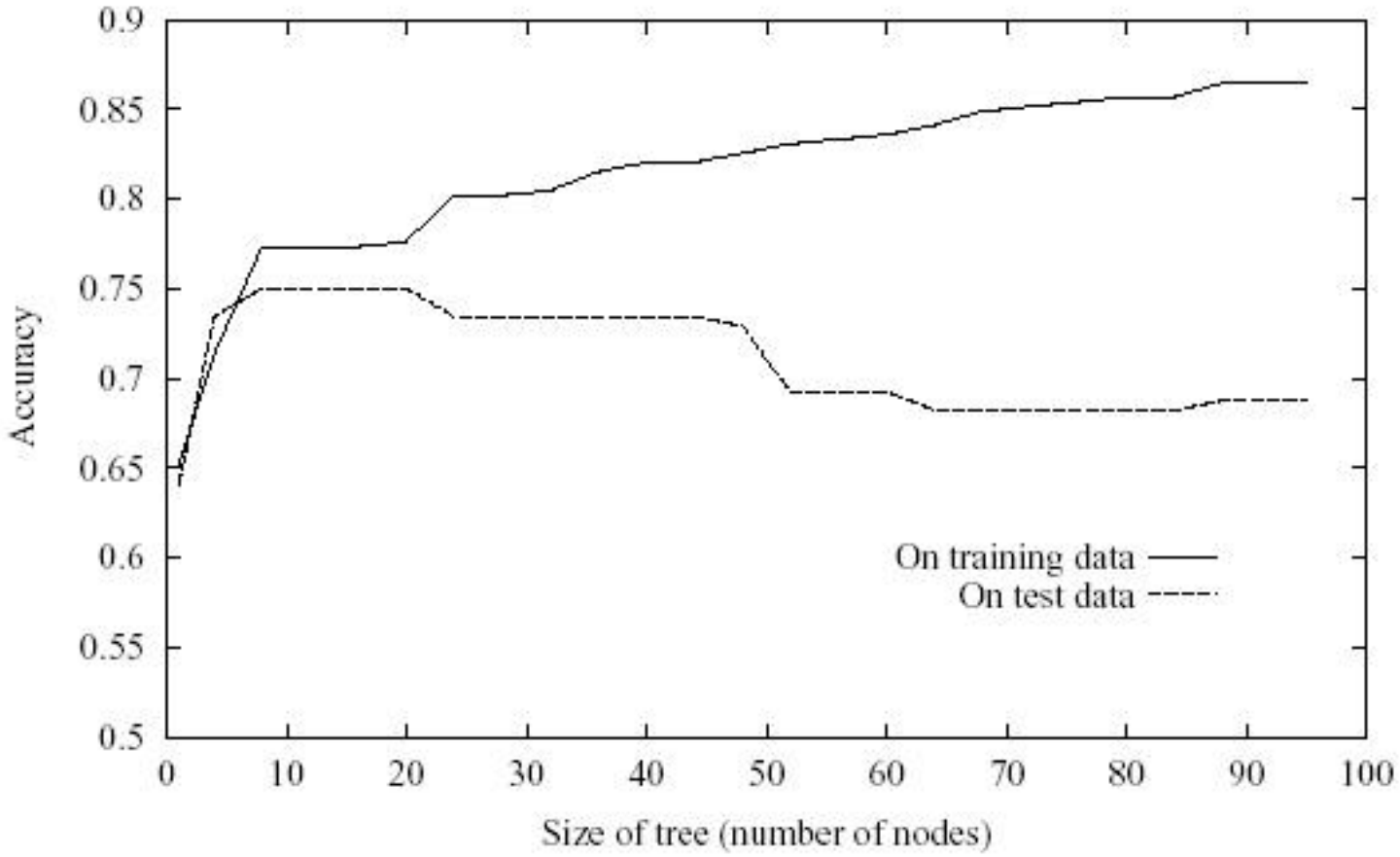
- Wahl geeigneter Attributtests:
  - In jedem Knoten, wähle das Attribut mit dem höchstem Informationsgehalt bezüglich der Zielfunktion auf Basis der verbleibenden Instanzen:

- Entropy: $H(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$
- S = Menge aller Instanzen im betrachteten Knoten
- p_i = Anteil der zu Klasse i gehörenden Instanzen



Lernen mit Entscheidungsbäumen

- Abbruchbedingung/Overfitting:



Lernen mit Entscheidungsbäumen

- Abbruchbedingung/Overfitting:
  - Pre-Pruning:
    - Stoppe während der Konstruktion des DT
    - z.B.: Wähle keine Attribute mit Informationsgehalt nahe 0
  - Post-Pruning:
    - Lerne kompletten Baum (d.h. jeder Endknoten enthält nur noch Instanzen einer einzelnen Klasse i)
    - Post-Prune den Baum durch nachträgliches Entfernen irrelevanter Tests
  - Statistische Kriterien:
    - Messe die „Reinheit“ eines Endknotens (leaf) i.d.R. auf Trainingsdaten
    - Prune bei „hinreichender Reinheit“
  - Klassifizierungsfehler als Kriterium:
    - Messe den Klassifizierungsfehler (i.d.R. auf testset)
    - Prune, wenn sich der Klassifizierungsfehler nach Entfernen eines Tests nicht (wesentlich) verschlechtert.
  - Tree- vs. Rule-Post-Pruning:
    - Postprune basierend auf dem DT selber
    - Postprune basierend auf Entscheidungsregeln (i.d.R. Bessere Ergebnisse)

Lernen mit Entscheidungsbäumen

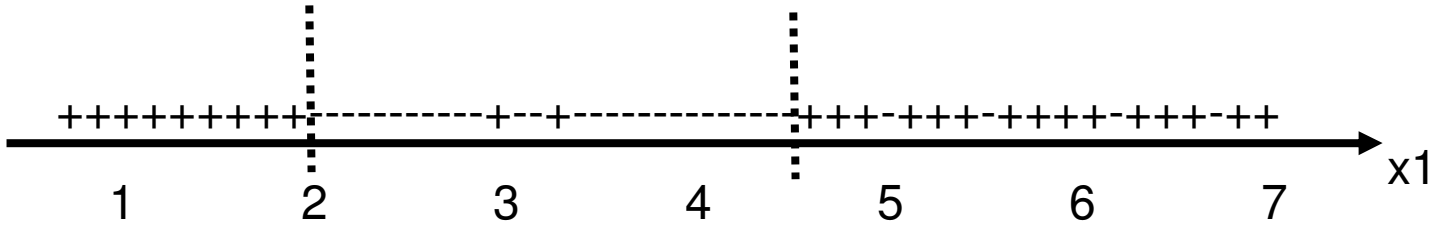
- Bevorzugung „großer“ Klassen:
  → Balancing

  - Under-sampling:
    - Angleichen der Anzahl von Instanzen verschiedener Klassen, durch Entfernen von Instanzen in überrepräsentierten Klassen
    - Probleme:
      - Instanzen sind unterschiedlich wichtig für Generalisierung
      - Kleinste Klasse bestimmt Größe des Trainingssets

  - Over-sampling:
    - Dupliziere Instanzen von unterrepräsentierten Klassen
    - Füge synthetische Instanzen nach geeigneten Kriterien ein

DTs für kontinuierliche Attribute

- Test kontinuierlicher Attribute:
  - Problem: Attributtests von DTs erwarten diskrete Attribute
  - Idee: Bilde geeignete Intervalle entlang der Attributachse
  - Behandle jedes Intervall als diskrete Merkmalsausprägung:



(x1 < 2) → +

(2 <= x1 < 4,5) → - (?)

(x1 >= 4,5) → + (?)

  - Binärer Split vs. Mehrere Intervalle

K-nearest neighbour vs DT Learning

- Nachvollziehbarkeit der Entscheidungen anhand der Hypothese/DT:

DT Learning:	kNN Learning:
Entscheidungsunterstützung	Klassifizierungsvorschlag
Problem: Pre-processing umkehrbar?	
Fuzzy Regeln?	

K-nearest neighbour vs DT Learning

- Höhere Genauigkeit bei multidimensionalen Daten?

DT Learning:	kNN Learning:
Extraktion relevanter Attribute	„Curse of dimensionality“
Lokale Selektion der Attribute	Globale Gewichtung der Attribute
Zur Klassifizierung herangezogene Attribute abhängig von der betrachteten Region im mehrdimensionalen Raum	

K-nearest neighbour vs DT Learning

- Schnelle Klassifizierung ungesehener Instanzen:

DT Learning:	kNN Learning:
Eager Learning	Lazy Learning
Klassifizierung aufgrund einiger weniger Entscheidungsregeln	Klassifizierung basierend auf allen gesehenen Instanzen
Schnelles Matching von Instanz und passender Regel	Retrieval der k nearest neighbours ist zeitintensiv

Ergebnisse

- Generalisierungsfehler:

	Size	CAL5	C4.5	k-NN
I186P102	300	30.7%	30.3%	23%
I233P1017	300	29.3%	22.7%	16%
I023P891	320	47.8%	17.8%	5%
I183P102	482	30.5%	26.1%	19%
I183P1035	482	26.7%	20.5%	15%
I001P003	499	29.4%	20.6%	12%
I003P886	749	17.1%	14.8%	15%
I499P892	9610	20.5%	20.0%	20%

- C4.5 erzielte die besseren Ergebnisse
  - 20% auf größter Datenbank

Ergebnisse

- Hypothesen auf kleineren Datenbanken sind vermutlich „nicht ausgelernt“:

Dataset	Size	CAL5	C4.5
I499P892L	2074	20.9%	21.8%
I499P892S	306	34.6%	33.0%
$\Delta error$		**13.7%**	**11.2%**

Ergebnisse

- Auf der größten DB wurden überrepräsentierte Klassen stark bevorzugt

C4.5							
Average Error:		20.0%					
Average Classification Matrix							
Actual Class	Avg. Cases per Class	Classified as					Misses in %
		1	3	4	5	6	
1	561.1	527.4	8.2	3.9	0.1	21.5	6.0%
3	121.4	5.7	65.5	3.4	0.2	46.6	46.0%
4	56.3	6.4	6.1	5.3	0.2	38.3	90.6%
5	24.7	2.1	2.6	0.6	0.2	19.2	99.2%
6	197.5	11.7	9.9	3.7	1.9	170.3	13.8%

- Balancing brachte keine Verbesserung

Ergebnisse

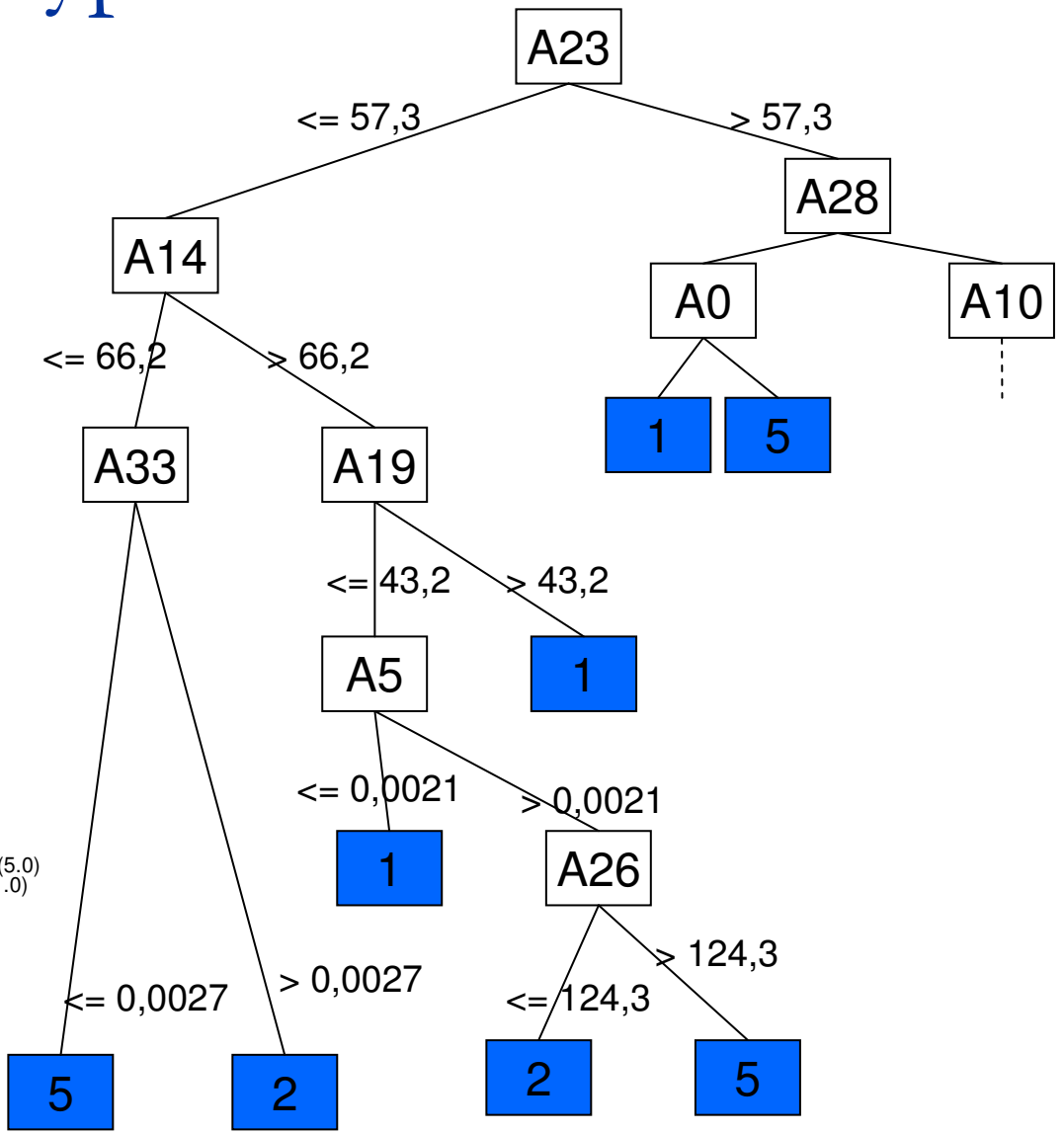
- Großer Unterschied im Generalisierungsfehler von C4.5 (zweitbestes Ergebnis) und CAL5 (schlechtestes Ergebnis) auf DB I023P891:
- Schlechteste Generalisierungsfehler auf DB I186P102 und I183P102:

Dataset	Size	CAL5	C4.5	$\Delta error$
I183P102	482	30.5%	26.1%	**4.4%**
I183P1035	482	26.7%	20.5%	**6.2%**
I186P102	300	30.7%	30.3%	0.4%
I233P1017	300	29.3%	22.7%	**6.6%**
I023P891	320	47.8%	17.8%	**30.0%**
I003P886	749	17.1%	14.8%	**2.3%**
I001P003	499	29.4%	20.6%	**8.8%**
I499P892	9610	20.5%	20.0%	0.5%

Generierte Hypothesen

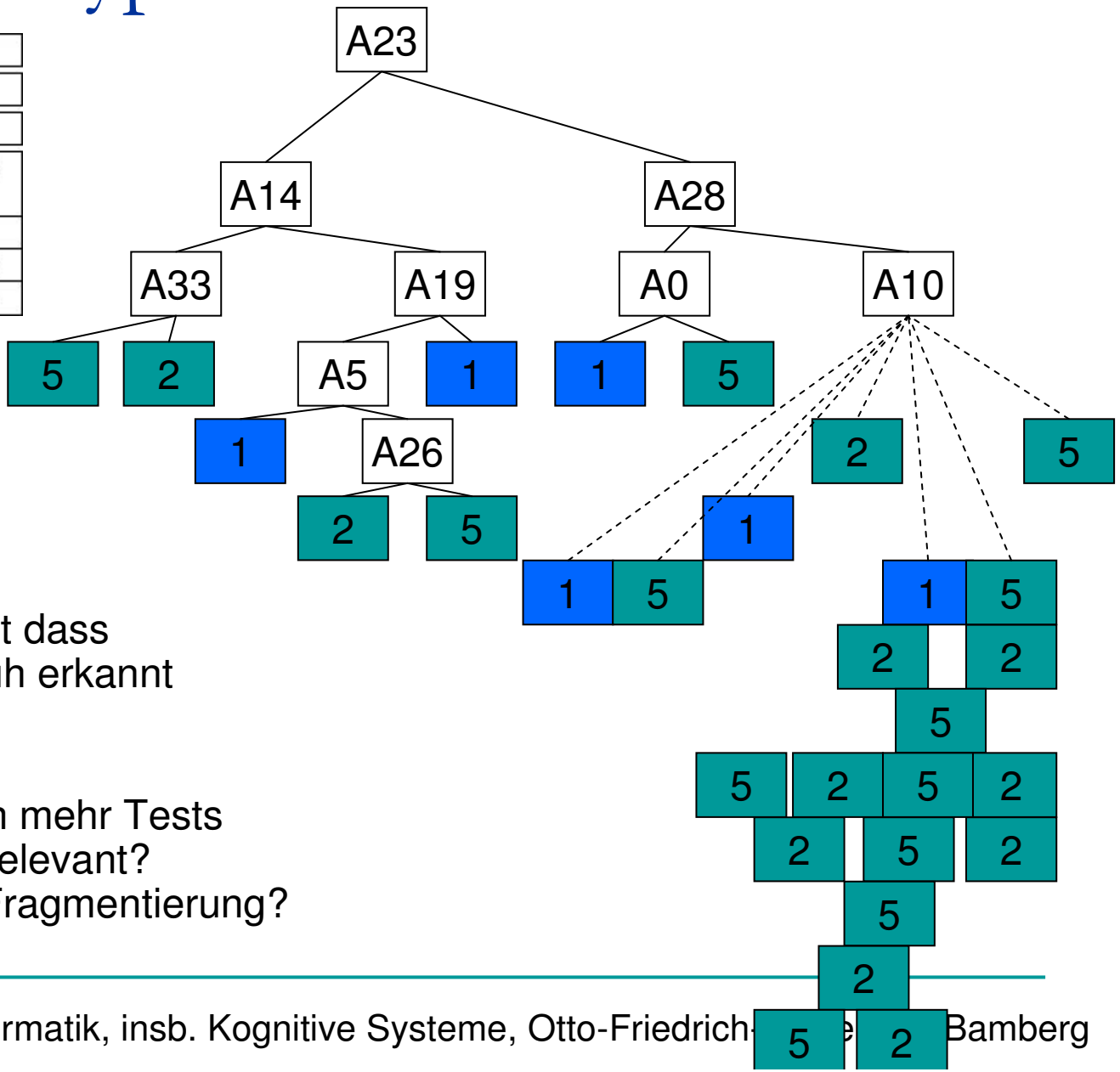
```
A23 <= 57.2952 [56.4453,57.6224]:
| A14 <= 66.1893 [65.2445,66.3912]:
| | A33 <= 0.002738 [0.00238255,0.00289972]: 5 (4.0)
| | A33 > 0.002738 [0.00238255,0.00289972]: 2 (2.0)
| A14 > 66.1893 [65.2445,66.3912]:
| | A19 > 43.188 [43.1534,43.56]: 1 (42.0)
| | A19 <= 43.188 [43.1534,43.56]:
| | | A5 <= 0.002103 [0.00199276,0.00212145]: 1 (6.0)
| | | A5 > 0.002103 [0.00199276,0.00212145]:
| | | | A26 <= 124.293 [122.883,124.329]: 2 (5.0)
| | | | A26 > 124.293 [122.883,124.329]: 5 (3.0)
A23 > 57.2952 [56.4453,57.6224]:
| A28 <= 3.04717 [3.04366,3.15709]:
| | A10 <= 125.884 [125.88,125.962]:
| | | A26 > 133.009 [132.867,134.018]: 5 (19.0)
| | | A26 <= 133.009 [132.867,134.018]:
| | | | A46 <= 159.872 [156.455,169.082]:
| | | | | A18 <= 115.102 [114.078,115.209]:
| | | | | | A13 > 0.620552 [0.595613,0.632738]: 2 (3.0)
| | | | | | A13 <= 0.620552 [0.595613,0.632738]:
| | | | | | | A38 > 123.812 [123.533,123.856]: 5 (21.0)
| | | | | | | A38 <= 123.812 [123.533,123.856]:
| | | | | | | | A16 <= 0.012077 [0.0109658,0.0122701]: 2 (6.0)
| | | | | | | | A16 > 0.012077 [0.0109658,0.0122701]:
| | | | | | | | | A18 > 131.894 [129.159,132.384]: 2 (5.0)
| | | | | | | | | A18 <= 131.894 [129.159,132.384]:
| | | | | | | | | | A8 <= 1.50253 [1.43635,1.50653]: 5 (11.0)
| | | | | | | | | | A8 > 1.50253 [1.43635,1.50653]:
| | | | | | | | | | | A34 <= 131.757 [129.351,132.069]: 2 (6.0)
| | | | | | | | | | | A34 > 131.757 [129.351,132.069]:
| | | | | | | | | | | | A5 <= 0.00222 [0.00202893,0.00228922]: 5 (5.0)
| | | | | | | | | | | | A5 > 0.00222 [0.00202893,0.00228922]: 2 (1.0)
| | | | A46 > 159.872 [156.455,169.082]:
| | | | | A37 <= 0.001619 [0.001445,0.00168812]: 1 (1.0)
| | | | | A37 > 0.001619 [0.001445,0.00168812]: 5 (11.0)
| | A10 > 125.884 [125.88,125.962]:
| | | A30 <= 101.506 [88.8001,102.476]: 2 (2.0)
| | | A30 > 101.506 [88.8001,102.476]:
| | | | A16 <= 0.014211 [0.0122189,0.0148288]: 1 (6.0)
| | | | A16 > 0.014211 [0.0122189,0.0148288]:
| | | | | A47 <= 34.2094 [24.2631,36.1087]: 1 (2.0)
| | | | | A47 > 34.2094 [24.2631,36.1087]: 5 (3.0)
| A28 > 3.04717 [3.04366,3.15709]:
| | A0 <= 0.017337 [0.0149768,0.0175913]: 1 (8.0)
| | A0 > 0.017337 [0.0149768,0.0175913]: 5 (2.0)
```

I233P1017
300 Instanzen
48 Attribute
3 Klassen



Generierte Hypothesen

C4.5				
Average Error:	22.7%			
Average Classification Matrix				
Actual Class	Classified as 1	2	5	Misses in %
1	8.9	0.3	0.5	8.2%
2	0.2	5.2	2.6	35.0%
5	1.1	2.1	9.1	26.0%

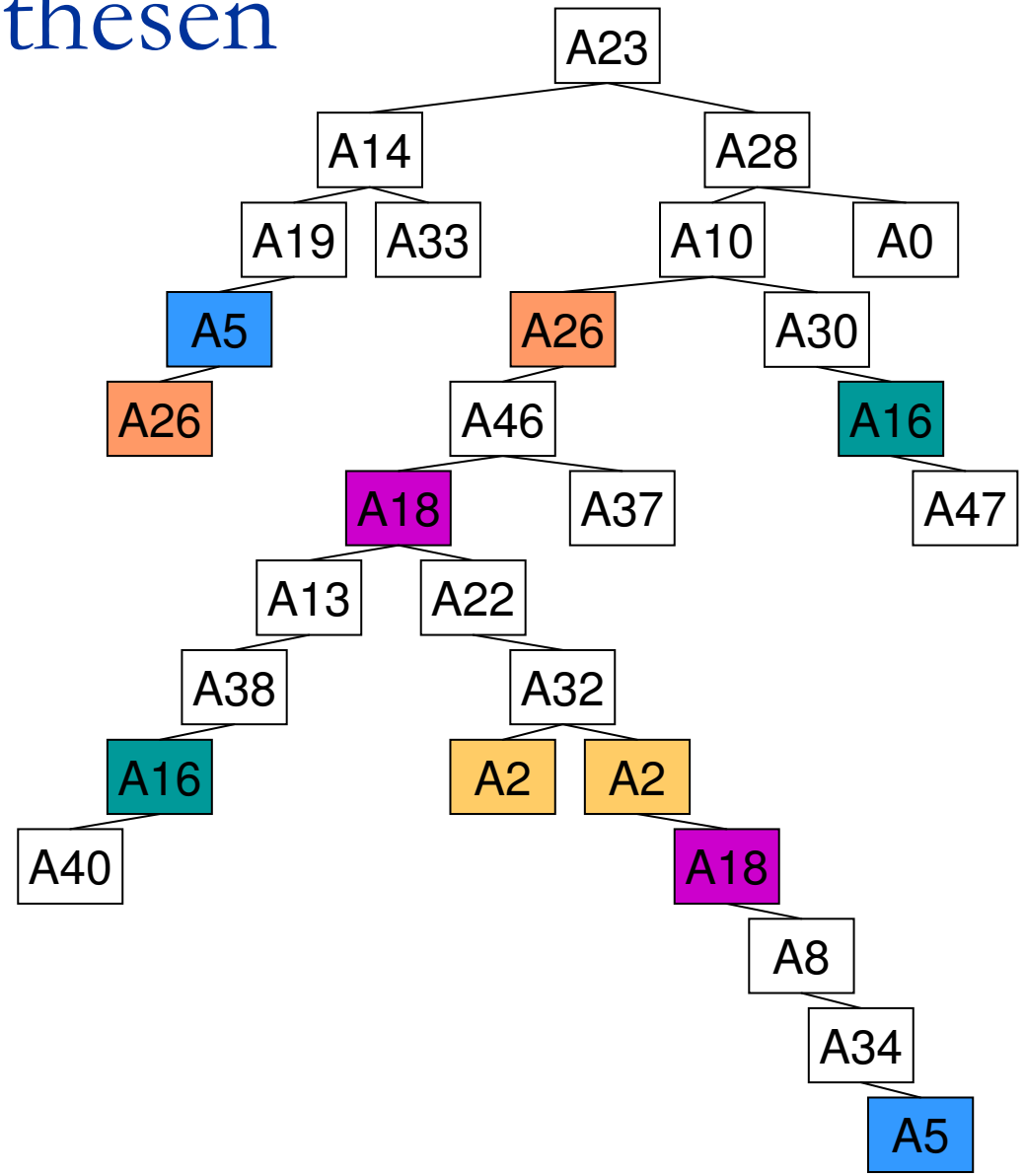


- Attibuttests so gewählt dass Klasse 1 möglichst früh erkannt
  → geringster Fehler
- Klassen 2/5 benötigen mehr Tests
  → vorherige Tests irrelevant?
  → Overfitting durch Fragmentierung?

Generierte Hypothesen

- A23 <= 57.2952 [56.4453,57.6224]:
  | A14 <= 66.1893 [65.2445,66.3912]:
  | | A33 <= 0.002738 [0.00238255,0.00289972]: 5 (4.0)
  | | A33 > 0.002738 [0.00238255,0.00289972]: 2 (2.0)
  | A14 > 66.1893 [65.2445,66.3912]:
  | | A19 > 43.188 [43.1534,43.56]: 1 (42.0)
  | | A19 <= 43.188 [43.1534,43.56]:
  | | | A5 <= 0.002103 [0.00199276,0.00212145]: 1 (6.0)
  | | | A5 > 0.002103 [0.00199276,0.00212145]:
  | | | | A26 <= 124.293 [122.883,124.329]: 2 (5.0)
  | | | | A26 > 124.293 [122.883,124.329]: 5 (3.0)
- A23 > 57.2952 [56.4453,57.6224]:
  | A28 <= 3.04717 [3.04366,3.15709]:
  | | A10 <= 125.884 [125.88,125.962]:
  | | | A26 > 133.009 [132.867,134.018]: 5 (19.0)
  | | | A26 <= 133.009 [132.867,134.018]:
  | | | | A46 <= 159.872 [156.455,169.082]:
  | | | | | A18 <= 115.102 [114.078,115.209]:
  | | | | | | A13 > 0.620552 [0.595613,0.622738]: 2 (3.0)
  | | | | | | A13 <= 0.620552 [0.595613,0.622738]:
  | | | | | | | A38 <= 0.012682 [0.0119716,0.0127312]:
  | | | | | | | | A16 <= 0.012682 [0.0119716,0.0127312]: 5 (2.0)
  | | | | | | | | A16 > 0.012682 [0.0119716,0.0127312]:
  | | | | | | | | | A40 <= 0.012682 [0.0119716,0.0127312]: 2 (2.0)
  | | | | | | | | | A40 > 0.012682 [0.0119716,0.0127312]: 5 (3.0)
  | | | | | | | A38 > 0.012682 [0.0119716,0.0127312]:
  | | | | | | | | A2 <= 48.743 [48.743,49.334]: 5 (1.0)
  | | | | | | | | A2 > 48.743 [48.743,49.334]: 2 (11.0)
  | | | | | | | A32 > 0.012682 [0.0119716,0.0127312]:
  | | | | | | | | A2 > 80.0364 [78.7698,84.4793]: 5 (11.0)
  | | | | | | | | A2 <= 80.0364 [78.7698,84.4793]:
  | | | | | | | | | A18 > 131.894 [129.159,132.384]: 2 (5.0)
  | | | | | | | | | A18 <= 131.894 [129.159,132.384]:
  | | | | | | | | | | A8 <= 1.50253 [1.43635,1.50653]: 5 (11.0)
  | | | | | | | | | | A8 > 1.50253 [1.43635,1.50653]:
  | | | | | | | | | | | A34 <= 131.757 [129.351,132.069]: 2 (6.0)
  | | | | | | | | | | | A34 > 131.757 [129.351,132.069]:
  | | | | | | | | | | | | A5 <= 0.00222 [0.00202893,0.00228922]: 5 (5.0)
  | | | | | | | | | | | | A5 > 0.00222 [0.00202893,0.00228922]: 2 (1.0)
  | | | | A46 > 159.872 [156.455,169.082]:
  | | | | | A37 <= 0.001619 [0.001445,0.00168812]: 1 (1.0)
  | | | | | A37 > 0.001619 [0.001445,0.00168812]: 5 (11.0)
  | | A10 > 125.884 [125.88,125.962]:
  | | | A30 <= 101.506 [88.8001,102.476]: 2 (2.0)
  | | | A30 > 101.506 [88.8001,102.476]:
  | | | | A16 <= 0.014211 [0.0122189,0.0148288]: 1 (6.0)
  | | | | A16 > 0.014211 [0.0122189,0.0148288]:
  | | | | | A47 <= 34.2094 [24.2631,36.1087]: 1 (2.0)
  | | | | | A47 > 34.2094 [24.2631,36.1087]: 5 (5.0)
  A28 > 3.04717 [3.04366,3.15709]:
  | A0 <= 0.017337 [0.0149768,0.0175913]: 1 (8.0)
  | A0 > 0.017337 [0.0149768,0.0175913]: 5 (2.0)

I233P1017
300 Instanzen
48 Attribute
23 verwendete Attribute



Generierte Hypothesen

Rule 1:
A1<=0.003012
A2<=104.079
A4>0.019506
A4<=0.032075
A5>0.001513
A10>119.203
A17>0.0022
A17<=0.002943
A18<=103.537
A32<=0.006944
A36>0.024735
A36<=0.037225
A40>0.555074
A45<=0.291982
A47>10.2432

-> class 4
[93.3%]

Rule 2:
A0 > 0.0122
A3 <= 57.6479
A4 <= 0.019506
A9 <= 0.60684
A14 > 78.4029
A17 <= 0.002069
A19 > 54.0669
A19 <= 58.1569
A21 <= 0.002421
A32 > 0.003396
A43 > 71.5114
A47 <= 27.3152

-> class 4 [92.2%]

Rule 3:
A2 <= 92.4308
A4 <= 0.022887
A7 > 10.733
A16 > 0.012894
A17 > 0.001678
A19 > 52.6954
A30 > 64.8675
A37 <= 0.003324

-> class 4 [90.6%]

Rule 4:
A2 > 103.912
A4 <= 0.019506
A8 > 1.09764
A17 <= 0.001444
A23 <= 14.3383
A33 > 0.000686
A44 > 1.87136
A45 <= 0.213256

-> class 4 [89.9%]

Mögliche Verbesserungen

- Größere Datenbanken
- Heuristiken für Parameteroptimierung
- Over-sampling statt under-sampling
- DT Ensembles durch Bagging & Boosting
- Wrapper Techniken zum Entfernen irrelevanter Attribute
- Post-Pruning auf seperatem Validationset
- Misklassifizierungskosten beachten
- Lazy Decision Trees
- Relevante Attribute für kNN-Distanzmetrik auf Basis von DTs bestimmen