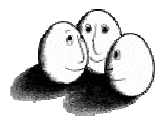
Merkmalsauswahl und -generierung

- Merkmalsauswahl als Aufgabe der Parameteroptimierung
  - –Filteransatz
  - –Wrapperansatz (John, Kohavi, Pfleger 1994)



Lernaufgabe (Wiederholung)

Gegeben:

- Beispiele X in LE
  - die anhand einer Wahrscheinlichkeitsverteilung P auf X erzeugt wurden und
  - Wobei jedes x mit einem Funktionswert y = t(x) versehen ist.
- H die Menge von Funktionen in LH.

Ziel: Eine Hypothese h(X) $\in$ H, die das erwartete Fehlerrisiko R(h) minimiert.

Klassifikation à la Bayes

Gegeben:

- Eine Wahrscheinlichkeitsverteilung P
- Berechne für jedes Beispiel $\vec{x} \in X$ die Klassenzugehörigkeit y aus Y $p(Y = y \mid \vec{x})$

$$= p(\vec{x} \mid Y = y) \frac{p(Y = y)}{p(\vec{x})}$$

- Wähle die wahrscheinlichste Klasse für ein Beispiel.

Merkmalsauswahl

- Würden wir die Wahrscheinlichkeitsverteilung kennen, hätte Merkmalsauswahl keinen Sinn. Überflüssige Attribute könnten den Klassifikationsfehler nicht verringern.
- Maschinelles Lernen soll eine Klassifikationsfunktion aus Beispielen erwerben. Probleme:
  - Bias verringern (mehr Attribute) vs. Varianz verringern (Attributwerte genauer schätzen)
  - Komplexität: Finden des optimalen Entscheidungsbaums ist NP-vollständig (Hyafil, Rivest 1976).
- Das Finden der Klassifikation wird approximiert und dies wird durch die richtigen Merkmale leichter.

Was sind "richtige" Merkmale?

- Merkmale mit hohem Informationsgewinn (Quinlan 1986)
- Quadratische Residuensumme für alle möglichen Regressionen (Furnival, Wilson 1974)
- Primäre Merkmale sind solche, bei denen sich die bedingte Wahrscheinlichkeit ändert, wenn das Merkmal einen bestimmten Wert hat. Ein Kontextmerkmal ist ein nicht primäres Merkmal, das aber zusammen mit anderen Merkmalen die bedingte Wahrscheinlichkeit ändert. (Turney 1996)

Relevante Merkmale

- Relevante Merkmale: Sei X_i ein Merkmal und S_i die Menge der Merkmale ohne X_i und s_i bezeichne eine Wertzuweisung zu allen Merkmalen S_i , dann ist X_i stark relevant, gdw. es x_i , y und s_i gibt mit $p(X_i = x_i, S_i = s_i) > 0$, so dass

$$p(Y = y | S_i = s_i, X_i = x_i) \neq p(Y = y | S_i = s_i)$$

Probleme mit der Definition

Seien die Beispiele derart, dass $X_2 = \neg X_4$ und $X_3 = \neg X_5$. Die 8 möglichen Beispiele sind gleichwahrscheinlich. Die zu lernende Funktion ist $Y = X_1$ XOR $X_2 = X_1$ XOR $\neg X_4$.

\+ X_3 und X_5 sind irrelevant.

\- Gemäß der Definition sind X_2 und X_4 auch irrelevant, weil sie S_2 und S_4 keine Information hinzufügen.

\- Alle Merkmale sind primär.

X_1	X_2	X_3	X_4	X_5
1	0	1	1	0
0	1	1	0	0
1	0	0	1	1
0	1	0	0	1
0	0	1	1	0
1	1	1	0	0
0	0	0	1	1
1	1	0	0	1

Achtung: Kodierung LE

Oft werden nominale Werte als Binärzahlen kodiert:

x_1 =rot, x_2 =grün, x_3 =blau

Wenn ein Objekt nur eine Farbe hat, ergibt sich {001, 010, 100}.

Jedes Merkmal ist aus den beiden übrigen abzuleiten.

$x_1\, x_2\, x_3$ ergeben keine zusätzliche Information zu $S_1\, S_2\, S_3$.

Damit werden alle Farbinformationen irrelevant.

Ausweg

- Schwache Relevanz: Ein Merkmal X_i ist schwach relevant, gdw. es nicht stark relevant ist und es gibt eine Menge von Merkmalen $S'_i \subset S_i$ für die es x_i , y und s'_i gibt mit $p(X_i = x_i, S'_i = s'_i) > 0$, so dass

$$p(Y = y | S'_i = s'_i, X_i = x_i) \neq p(Y = y | S'_i = s'_i)$$

- Ein Merkmal ist relevant, wenn es stark oder schwach relevant ist, sonst irrelevant.

Relevanz ist nicht Optimalität: LH

- Es gilt nicht notwendigerweise: alle relevanten Merkmale sind in der optimalen Merkmalsmenge.
- Sei L_H die Menge der Ausdrücke mit nur einer binären Variablen. Sei L_E die Menge der Ausdrücke mit 3 binären Variablen. Die zu lernende Funktion ist $(x_1 \wedge x_2) \vee x_3$.
- Alle drei Eingangsvariablen sind relevant.
- Die optimale Merkmalsmenge ist $\{x_3\}$.

Optimalität ist nicht Relevanz: Algorithmus

- Es gilt nicht notwendigerweise: Irrelevante Merkmale kommen in der optimalen Merkmalsauswahl nicht vor.
- Sei ein Merkmal immer gleich 1. Es ist also irrelevant. Sei Klassifikator1 so, dass für festes $\theta = 0$
  Y=1 gdw. $\theta < \sum_{Merkmale} w * m$

  Sei Klassifikator2 so, dass θ irgendein Wert sein kann.
- Mit dem irrelevanten Merkmal ist Klassifikator1 so lernfähig wie Klassifikator2. Für Klassifikator1 kommt es in der optimalen Auswahl vor.

Filteransatz

- Filteransätze sind solche, die Merkmale anhand der Beispiele und ihrer Verteilung auf die Klassen auswählen.
- Der Lernalgorithmus wird nicht beachtet.
- Man kann den Filteransatz als Vorverarbeitung unabhängig von einem Lerner anwenden.
- Der Filteransatz kann große Datenmengen verarbeiten.

Relief

- Ziel: alle (schwach und stark) relevanten Merkmale finden!
- Vorgehen:
  - zufällig Beispiele ziehen,
  - den nächsten Nachbarn derselben Klasse finden (near hit),
  - den nächsten Nacharn der anderen Klasse finden (near miss)
  - Die Relevanzwerte für verschiedene Ausprägungen bei near miss erhöhen.
  - Merkmale mit genügend hohem Relevanzgewicht werden ausgewählt.

(Kira, Rendell 1992), für Mehrklassenprobleme (Kononenko 1994)

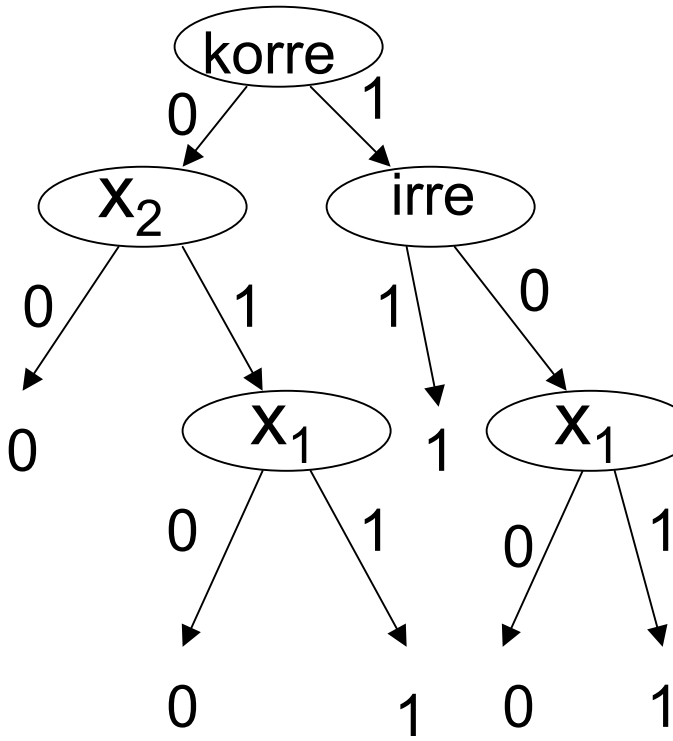
Entscheidungsbaum

- Entscheidungsbäume wählen Merkmale nach Informationsgewinn aus.
- C4.5 beschneidet den gelernten Baum, indem ein Vorgängerknoten durch seinen sehr viel besseren Nachfolger erstzt werden kann. ID3 tut dies nicht.
- Ihre Auswahl kann auch für andere Lerner genutzt werden.
- Es können die n-obersten Merkmale des Baums gewählt werden.

Problemfall

- L_E: 6 binäre Merkmale. Merkmal "irre" ist zufällig, Merkmal "korre" stimmt in 75% der Fälle mit Y überein. X besteht aus nur 32 Beispielen.
- Die gesuchte Funktion ist: $(x_1 \wedge x_2) \vee (x_3 \wedge x_4)$.
- Entscheidungsbäume wählen "korre" aus.

  C4.5 korrigiert dies durch pruning.



Suche

- Suchraum
  - Verband der 2^m Teilmengen von m Merkmalen
- Suchoperatoren
  - ein Merkmal löschen oder hinzufügen
  - Mehrere erfolgreiche Operatoren zusammenfassen
- Suchalgorithmus
  - Bergsteigen
  - Bestensuche
- Suchstrategie
  - Forward selection
  - Backward elimination
- Bewertungsfunktion

Wrapper-Ansatz

- Die Bewertungsfunktion ist die Performanz desjenigen Lernalgorithmus, der zum Lernen optimiert werden soll.
- Beispiele werden aufgeteilt:
  Kreuzvalidierung 1
  - Trainingsdaten Merkmalsauswahl
  - Testdaten Merkmalsauswahl

  Kreuzvalidierung 2
  - Trainingsdaten Lernen
  - Testdaten Lernen

Suchstrategien und Suchalgorithmen

- Squential backward elimination (Marill, Green 1963)
- Plus ℓ – take away r (Kittler 1978)
- Statistische monotone Bewertungsmaße: Sequenz geschachtelter Mengen $F_1 \supseteq F_2 \supseteq ... \supseteq F_n$ befolgt $f(F_1) < f(F_2) < ... < f(F_n)$
  - Bergsteigen kann bei monotoner Berwertungsfunktion und Auswahl von 1 Merkmal (löschen, hinzufügen) je Iteration nicht die richtige Merkmalsmenge finden. (Cover, Campenhout 1977)
- Genetische Algorithmen (Vafai, De Jong 1992, 1993)

Bergsteigen

1. Sei v der initiale Zustand.
2. Expandiere v zu den Kindern von v
3. Bewerte jedes Kind w von v gemäß f(w)
4. Sei v' das Kind mit höchstem f(w)
5. Wenn f(v') > f(v) dann v:=v' GOTO 2
6. Gib v aus.

Ergebnisse

- ID3, Naive Bayes als Lernalgorithmen
- 8 echte, 5 künstliche Datensätze
- Forward selection wegen teilweise großer Anzahl von Merkmalen (180).
- ID3 sucht selbst Merkmale aus, aber der Wrapper-Ansatz sucht weniger aus, ohne dass die accuracy abstiege.
- Naive Bayes wird durch die Merkmalsauswahl nicht schlechter, die Ergebnisse aber verständlicher.
- Die ausgewählten Merkmalsmengen für ID3, Naive Bayes überlappen sich, sind aber verschieden.

Bestensuche

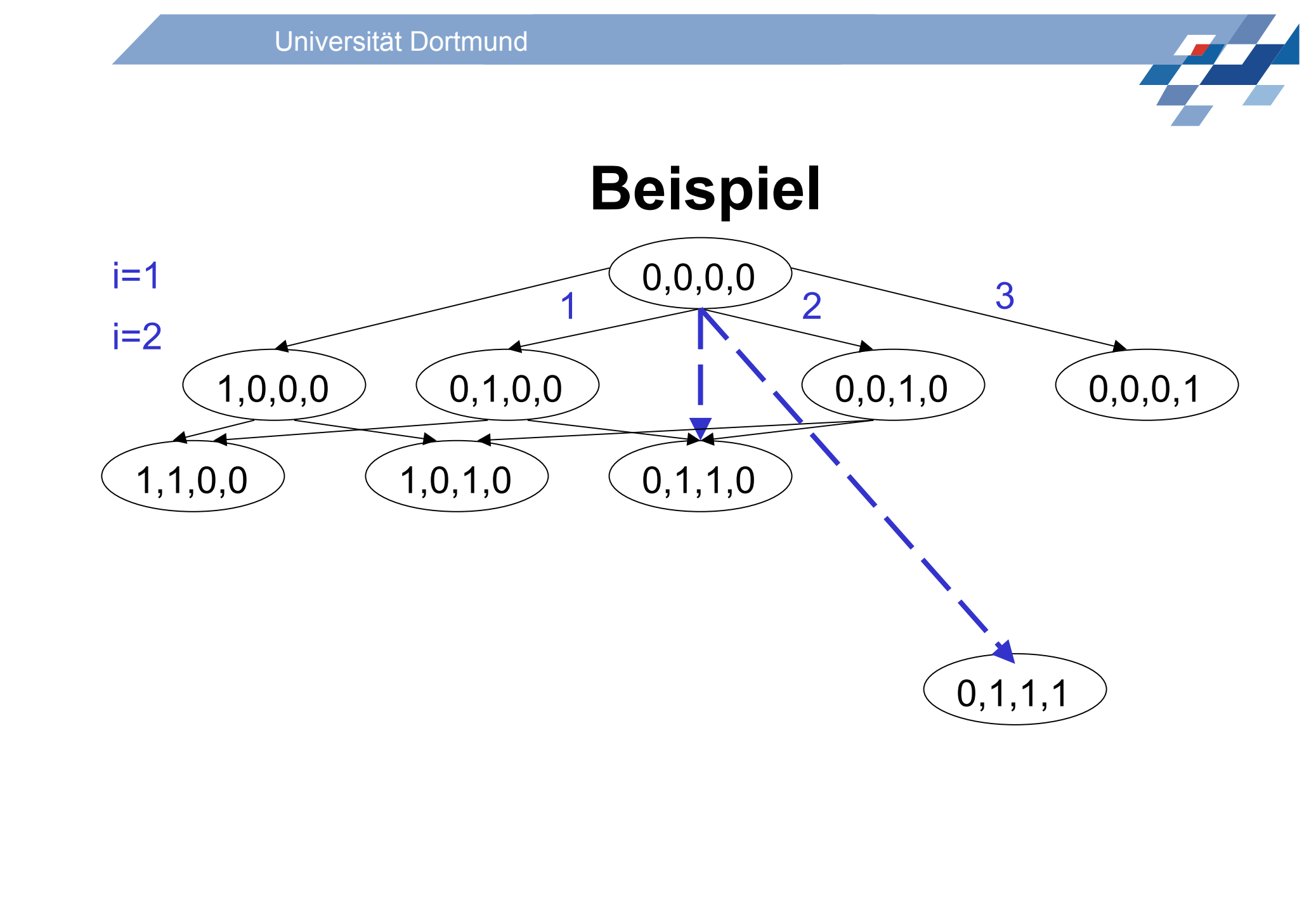
1. Sei der initiale Zustand in der Liste OPEN und sei der initiale Wert von BEST. Die Liste CLOSED ist leer.
2. v:=argmax f(w), w aus OPEN
3. Lösche v aus OPEN, trage v ein in CLOSED.
4. Wenn f(v)-ε > f(BEST), dann BEST:=v
5. Expandiere v.
6. Jedes Kind, das nicht in OPEN oder CLOSED ist, wird in OPEN eingetragen und bewertet.
7. Wenn BEST sich in den letzten k Iterationen geändert hat, GOTO 2.
8. Gib BEST aus.

Ergebnisse

- Bei den echten Datensätzen macht der Suchalgorithmus keinen Unterschied aus.
- Bei den künstlichen Datensätzen findet der Wrapperansatz für ID3 in 3 Fällen den optimalen Merkmalssatz, die accuracy steigt.
- Bei Naive Bayes findet der Wrapperansatz die Merkmale, die zu besserer accuracy führen, darunter ist ein eindeutig korreliertes.
  Ohne korreliertes Merkmal: 87,5% accuracy,
  mit korreliertem Merkmal: 90,62% accuracy.

Mehrere Operatoren zusammenfassen

1. Expansion von v mit üblichen Operatoren.
2. Bewertung der Nachfolger.
3. Kombination der die i+1 besten Nachfolger generierenden Operatoren zu einem KombiOperator c_i.
4. Anwendung von c_i auf v.
5. Bewertung der Nachfolger w.
6. Solange f(w) besser wird, i:=i+1 und GOTO 3.



Beispiel

KombiOperatoren

- Kürzen die Suche ab. Backward elimination wird damit überhaupt erst möglich (Rechenzeit).
- Können sich schneller überanpassen an die Daten.
- Für Naive Bayes ist backward elimination mit KombiOperatoren günstiger als forward selection mit Bestensuche. Für ID3 nicht.
- Für C4.5 (pruning) ist backward elimination mit KombiOperatoren günstig: z.B. mussten zur Auswahl von 12 aus 180 Merkmalen nur 3555 statt (180-12)180=30 240 expandiert werden!

Vergleich C4.5, Wrapper, Relief

Daten	C4.5	RLF	BFS
Breast cancer	95,42	94,42	95,28
Cleve	72,30	74,95	**77,88**
Crx	85,94	84,06	85,8
DNA	92,66	92,75	**94,44**
Horse colic	85,05	85,88	84,77
Pima	**71,6**	64,18	70,18
Euthyroid	97,73	97,73	97,91
Soybean large	91,35	91,35	91,93

C4.5 "pur",

BFS: C4.5 Wrapper backward elimination Bestensuche mit KombiOperatoren,

RLF: Filter Relief vor C4.5 Anwendung

Obwohl C4.5 Merkmalsauswahl und pruning hat, wird es durch den Wrapper noch besser!

Merkmalsauswahl

Daten	Original	RLF	C4.5	BFS	NB-BFS
Breast cancer	10	5,7	7	3,9	5,9
Cleve	13	10,5	9,1	5,3	7,9
Crx	15	11,5	9,9	7,7	9,1
DNA	180	178	46	12	48
Horse colic	22	18,2	5,5	4,3	6,1
Pima	8	1,2	8	4,8	4,4
Euthyroid	25	24	4	3	3
Soybean large	35	34,8	22	17,1	16,7
Reduktion	0	30%	37%	40%	28%

Was wissen Sie jetzt?

- Merkmalsauswahl ist eine Art der Modellselektion.
- Sie kann als Suche im Raum der Merkmalsmengen aufgefasst werden:

  Suchraum, Suchoperatoren (expandieren einen Zustand zu den Nachfolgern), Suchalgorithmus, Suchstrategie, Bewertungsfunktion
- Der Wrapper-Ansatz nimmt einen Lerner als Bewertungsfunktion. Durch die Kreuzvalidierung wird probabilistisch bewertet.
- Der Filteransatz wählt unabhängig vom Lerner, meist deterministisch.
- Sie kennen mindestens die Definition für starke und schwache Relevanz von Kohavi.