Vorlesung Wissensentdeckung in Datenbanken

Merkmalsauswahl

Kristian Kersting, (Katharina Morik), Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

08.07.2014

Allgemeine Lernaufgabe — Wiederholung

Gegeben

- eine Menge $\mathbf{X} = \{\vec{x_1}, ..., \vec{x_N}\} \subset \mathcal{X}$ von Beobachtungen,
- die anhand einer Wahrscheinlichkeitsverteilung P auf $\mathcal{X}$ erzeugt wurden und
- wobei jedes Beispiel $vecx_i$ mit einem Funktionswert $y_i = f(\vec{x_i})$ (Klassifikation/Regression) versehen ist,
- und eine Menge H von Funktionen in einem Hypothesenraum.

Finde

- eine Hypothese $h(\mathbf{X}) \in H$, die "gute" den Funktionswert für Beispiele nach P bestimmt.

Klassifikation a la Bayes

Gegeben

- eine Wahrscheinlichkeitsverteilung P auf $\mathcal{X}$, dann
- berechne für jedes $\vec{x}_i \in \mathbf{X}$ die Klassenzugehörigkeit y nach

$$P(y|\vec{x}_i) = P(\vec{x}_i|y) \cdot \frac{P(y)}{P(\vec{x}_i)} \ .$$

- Wähle die wahrscheinlichste Klasse für das Beispiel.

Merkmalsauswahl

- Würden wir die Wahrscheinlichkeitsverteilung P kennen, hätte Merkmalsauswahl keinen Sinn. Überflüssige Attribute könnten den Klassifikationsfehler nicht verringern.
- Aber wir kennen die Wahrscheinlichkeitsverteilung P nicht! Data Mining soll "einfach nur" eine Klassifikationsfunktion aus den Beispielen $\mathbf{X}$ erwerben. Und das ist nicht einfach:
  1. Bias verringern (mehr Attribute) versus Varianz verringern (Attributwerte genauer schätzen)
  2. Komplexität: Finden des optimalen Entscheidungsbaums is NP-schwer (Hyafil, Rivest 1976).
- Das finden der Klassifikation wird approximiert und genaue das wird durch die richtigen Merkmale leichter!

Warum Merkmalsauswahl?

Neben der besseren Generalisierbarkeit sprechen die folgenden Punkte für eine Merkmalsauswahl:

- Geringerer Datenerfassungs- und Speicherkosten
- Geringerer Rechenaufwand
- Leichtere Iterpretierbarkeit

Gliederung

Irrelevante Merkmale

Irrelevante Merkmale sind solche Merkmale, die einen dem Ziel der guten Klassifizierung nicht näher bringen, egal unter welchen Umständen.

	Attribut A	Klasse
Instanz 0	1	0
Instanz 1	1	0
Instanz 2	1	1
Instanz 3	1	1
Instanz 4	1	1
Instanz 5	1	1

Ein Merkmal, das für alle Beispiel immer nur eine Ausprägung annimmt, ist irrelevant, weil es keinen Einfluss auf die Klassifizierung nimmt.

Aber aas sind dann "richtige" Merkmale?

Es gibt viele Möglichkeiten das zu definieren.

- Merkmale mit hohem Informationsgewinn (Information-Gain) wie wir es von Entscheidungsbäumen kennen (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: Motiviert durch primäre Merkmale

Sei

- X_i ein Merkmal, $\mathbf{S_i}$ die Menge aller Merkmale ohne X_i und sei $\mathbf{s}_i$ eine Wertzuweisung an alle Merkmale in $\mathbf{S}_i$.

dann

- ist X_i stark relevant genau dann wenn es x_i und $\mathbf{s}_i$ gibt mit $P(X_i = x_i, \mathbf{S}_i = \mathbf{s}_i) > 0$, so dass

$$P(Y = y | X_i = x_i, \mathbf{S}_i = \mathbf{s}_i) \neq P(X_i = x_i, \mathbf{Y} = y | \mathbf{S}_i = \mathbf{s}_i)$$

Im Wesentlichen beschreibt das unsere Intuition, dass X_i einen Einfluss auf die Vorhersage hat!

Probleme mit der Definition

Aber ist das eine gute Definition?

- Wir betrachten binäre Merkmale. 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 Funktioni sei $Y = X_1\ XOR\ X_2 = X_1\ XOR\ \neg X_4$.
- Per constructionem sind die Merkmale X_3 und X_5 irrelevant.
- Gemäß der Definition sind X_2 und X_4 auch irrelevant, weil sie $\mathbf{S}_2$ und $\mathbf{S}_4$ keine Information hinzufügen.
- Aber: 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

Probleme mit der Definition

So etwas passiert häufiger als man denkt!

- Oft werden nominale Werte als Binärzahlen kodiert. Betrachten wir als ein Beispiel $X_1 = rot$, $X_2 = gruen$, $X_3 = blau$. Also wenn ein Object nur eine Farbe hat, dann 001, 010 bzw. 100.
- Jedes Merkmal ist aus den beiden übrigen abzuleiten.
- Die Merkmale x_1, x_2 und x_3 ergeben keine zusätzliche Information zu $\mathbf{S}_1$, $\mathbf{S}_2$ und $\mathbf{S}_3$.
- Damit werden alle Farbinformationen irrelevant!

Ein Ausweg: schwache Relevanz

- Ein Merkmal X_i ist schwach relevant genau dann wenn es nicht stark relevant ist und es gibt eine Menge von Merkmalen $\mathbf{s}' \subseteq \mathbf{S}_i$ gibt, für die es es x_i und $\mathbf{s}'_i$ gibt mit $P(X_i = x_i, \mathbf{S}'_i = \mathbf{s}\prime_i) > 0$, so dass

$$P(Y = y | X_i = x_i, \mathbf{S}'_i = \mathbf{s}'_i) \neq P(X_i = x_i, \mathbf{Y} = y | \mathbf{S}'_i = \mathbf{s}'_i)$$

Im Wesentlichen erlauben wir jetzt Teilmengen der verbleibenden Merkmale!

Aber Achtung: Relevanz ist nicht gleich Optimalität

- Es gilt nicht notwendigerweise: alle relevanten Merkmale sind in der optimalen Merkmalsmenge.
- Betrachten wir als Hypothesenraum die Menge aller Ausdrücke mit nur einer binären Variable. Die Trainingsbeispiele sind aber mittels $(X_1 \wedge X_2) \vee X_3$ generiert worden.
- Alle drei Eingangsvariablen sind relevant.
- Die optimale Merkmalsmenge ist X_3.

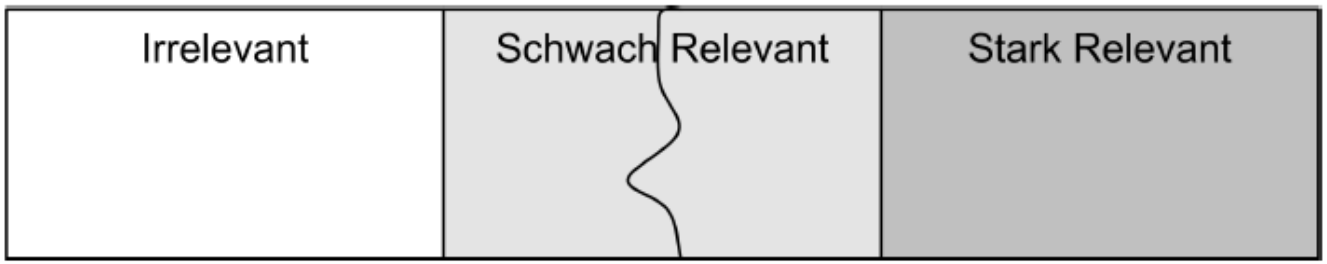
Aber Achtung: Relevanz ist nicht gleich Optimalität

- 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 festen Schwellwert $\theta = 0$

$$Y = 1 \text{ gdw. } \theta < \sum_{m \text{ Merkmal}} w \cdot m,$$

  und 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.

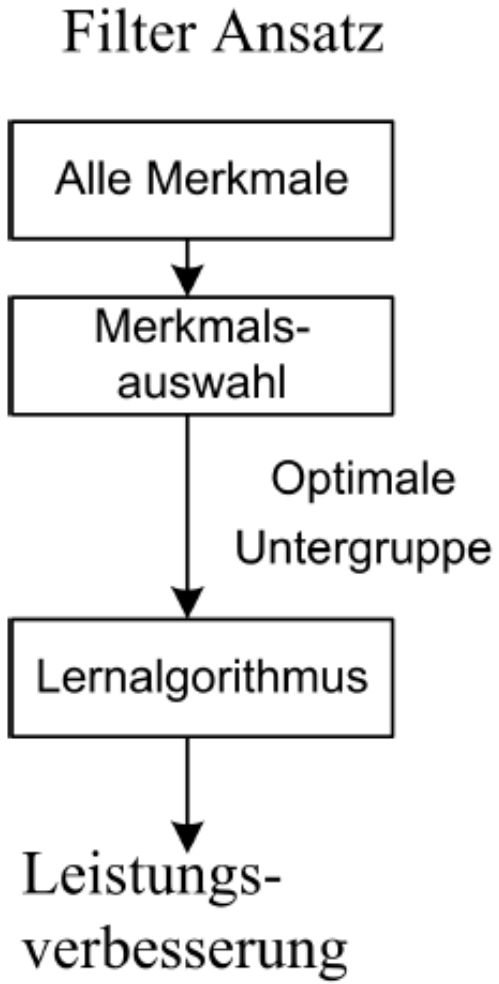
Aber wie wählen wir denn nun Merkmale aus?



- Wir wissen jetzt was "relevante" Merkmale prinzipiell sind.
- Aber wir wählen wir denn nun solche Merkmale aus?

Der Filter-Ansatz

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



Relief

- Ziel: Finde alle (schwach und stark) relevanten Merkmale!
- Vergehen ist heuristisch basierend auf der Euklidischen Distanz:
  - Ziehe Beispiele zufällig.
  - Für jedes gezogene Beispiel finde den nächsten Nachbarn derselben Klasse (near hit), als auch
  - den nächsten Nacharn der anderen Klasse finden (near miss)
  - Intuition: Ein Merkmal, das zwischen einer Instanz und einem near hit unterscheidet ist irrelevant! Ein Merkmal, das zwischen einer Instanz und einem near miss unterscheidet, ist relevant!
  - Daher, erhöhe eie Relevanzwerte für verschiedene Ausprägungen bei near miss.
  - Merkmale mit genügend hohem Relevanzgewicht werden ausgewählt.

Relief: Algorithmus

Algorithm 1 Relief

```
Relief(Attribute, Instanzen, Testanzahl, Toleranz)
    Lösung = ∅
    Initialisiere alle Gewichte, W_i = 0
    For i = 1 to Testanzahl
        Zufällige Wahl einer Instanz x aus Instanzen
        Finde NearHit(x) und NearMiss(x)
        For j= 1 to N
            W_j = W_j − diff(x_j, NearHit_j)^2 + diff(x_j, NearMiss_j)^2
        For j = 1 to N
            If W_j ≥ Toleranz
                Lösung += Attribut_j
    return Lösung
```

Relief: Update Instanz 2

	Flag 1	Flag 2	Flag 3	Attribut X	Zielklasse	Abstand
Instanz 0	0	0	0	a	1	3
Instanz 1	1	1	1	b	1	1
Instanz 2	1	0	1	b	0	-
Instanz 3	1	0	1	c	0	1
Instanz 4	0	0	0	a	1	3
Instanz 5	0	1	1	b	0	2
Gewichtung	0	1	0	-1		

Relief: Gesamtbewertung

	Flag 1	Flag 2	Flag 3	Attribut X	Zielklasse
Instanz 0	0	0	0	a	1
Instanz 1	0	1	1	b	1
Instanz 2	1	0	1	b	0
Instanz 3	1	1	1	c	0
Instanz 4	0	0	0	a	1
Instanz 5	0	1	1	b	0
Gewichtung	3	3	3	2	

Merkmalsauswahl mittels Suche

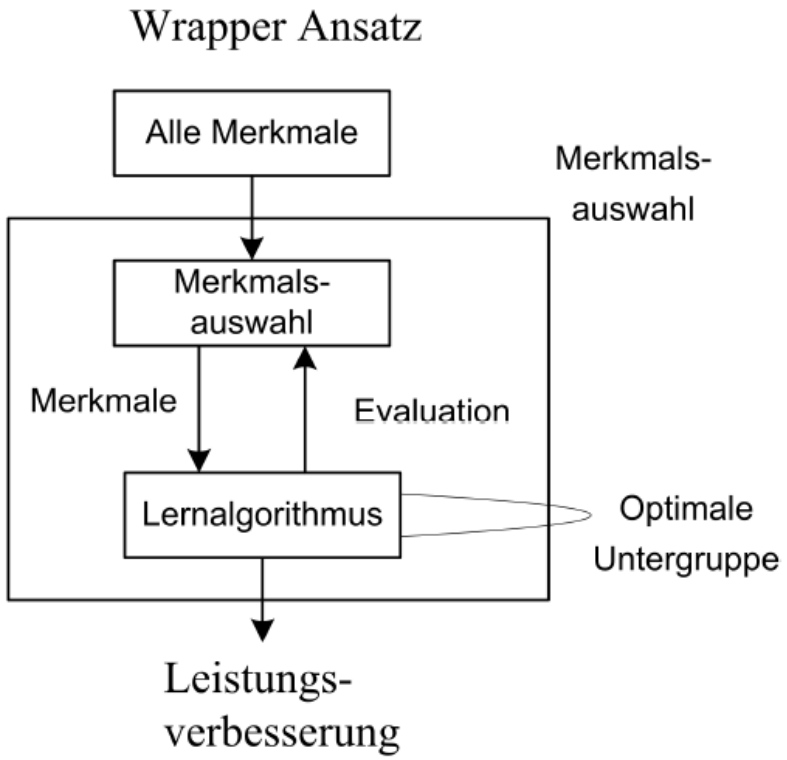
Ein Alternative sind Filter-Ansätze mittels Suche

- Suchraum: Verband der 2^m Teilmengen von m Merkmalen
- Suchoperatoren: Im Wesentlichen ein Merkmal löschen oder hinzufügen
- Suchalgorithmus: Bergsteigen, Bestensuche usw.
- Suchstrategie: Forwardselection (wir fangen mit der leeren Menge an) oder Backward elimination (wir fangen mit allen Merkmalen an)
- Bewertungsfunktion f wie z.B. Information-Gain

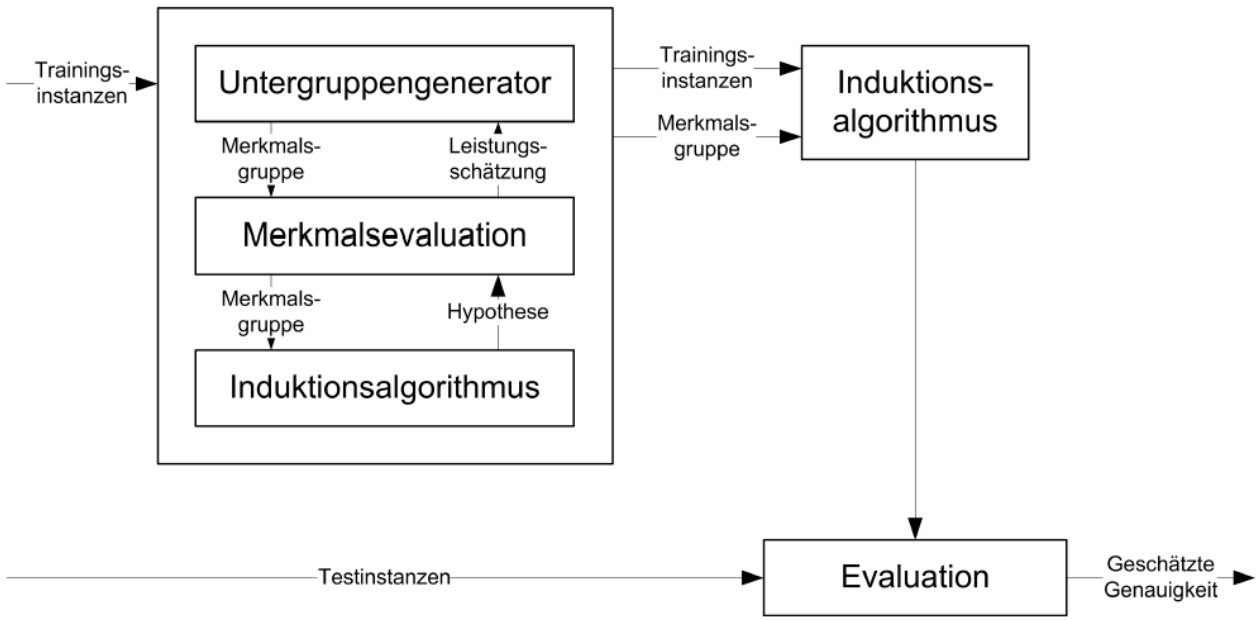
Sie führen auch zu den Wrapper-Ansätzen, wenn wir die Bergwertung mittels des Lernalgorithmus durchführen

Der Wrapper-Ansatz

- Beim Wrapper-Ansatz wird der Lernalgorithmus als Blackbox genutzt. Die Modelgüte einer Menge von ausgewählten Merkmalen wird als Bewertungsfunktion verwendet.
- Der Lernalgorithmus wird beachtet!
- Keine "aufwendige" Extraprogrammierung einer Bewertungsfunktion.
- Aber er ist sehr kostspielig/aufwendig !!



Der Wrapper-Ansatz



Daher werden die Beispiele aufgeteilt. Zum Beispiel:

- Kreuzvalidierung 1
  - Trainingsdaten Merkmalsauswahl
  - Testdaten Merkmalsauswahl
- Kreuzvalidierung 2
  - Trainingsdaten Lernen
  - Testdaten Lernen

Bergsteigen

1. Sei v der initiale Zustand (ausgewählte Menge an Merkmalen).
2. Expandiere v zu den Kindern von v
3. Bewerte jedes Kind w von v gemäß $f(w)$ (z.B. dem Lernalgorithmus)
4. Sei v' das Kind mit höchstem $f(v')$
5. Wenn $f(v') > f(v)$ dann $v := v'$; GOTO 2
6. Gib v aus.

Bergsteigen: Ergebnisse

- ID3 (Entscheidungsbaum) und 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 abstieg.
- Naive Bayes wird durch die Merkmalsauswahl nicht schlechter, die Ergebnisse aber verständlicher.
- Die ausgewählten Merkmalsmengen für ID3 und Naive Bayes überlappen sich, sind aber verschieden.

Bestensuche

1. Sei der initiale Zustand der einzige Zustand in der Liste OPEN und auch der initialer Wert für BEST. Die Liste CLOSED sei leer.
2. $v := \arg\max_{w \in OPEN} f(w)$.
3. Lösche v aus OPEN und trage v in CLOSED ein.
4. Wenn $f(v) - \epsilon > f(BEST)$, 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.

Bestensuche: 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.

Vergleich C4.5, Wrapper, Relief

- C4.5 "pur"
- RLF: Filter Relief vor C4.5 Anwendung
- BFS: C4.5 Wrapper mit backward elimination Bestensuche und kombinierten Suchoperatoren (wir wenden die *i* besten Operatoren gemeinsam an)?

- Obwohl C4.5 Merkmalsauswahl und pruning hat, wird es durch den Wrapper verbessert.

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

Vergleich: Reduktion der Zahl der Merkmale

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.