



Vorlesung
*Methodische Grundlagen des
Software-Engineering*
im Sommersemester 2014

Prof. Dr. Jan Jürjens

TU Dortmund, Fakultät Informatik, Lehrstuhl XIV

Teil 2.2: Data-Mining

v. 23.05.2014

1



2.2 Data-Mining

[mit freundlicher Genehmigung basierend
auf einem englischen Foliensatz von
Prof. Dr. Wil van der Aalst (TU Eindhoven)]

Literatur:

[vdA11] Wil van der Aalst: **Process Mining: Discovery, Conformance and Enhancement of Business Processes**, Springer-Verlag. 2011.

Unibibliothek (6 Exemplare): <http://www.ub.tu-dortmund.de/katalog/titel/1332248>

(Bei Engpässen kann eine **Kopiervorlage** der relevanten Ausschnitte zur Verfügung gestellt werden.)

- **Kapitel 3**

[R10] Thomas A. Runkler: **Data Mining: Methoden und Algorithmen Intelligenter Datenanalyse**, Vieweg+Teubner. 2010.

Unibibliothek: <http://www.ub.tu-dortmund.de/katalog/titel/1294605>

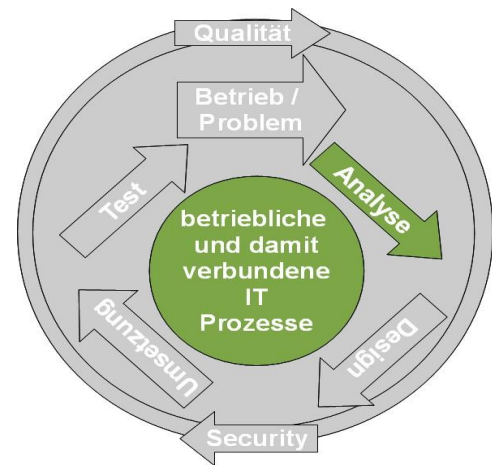
- **Kapitel 8**

- Geschäftsprozessmodellierung

- **Process-Mining**

- Einführung: Process-Mining
- Petrinetze
- **Data-Mining**
- Datenbeschaffung
- Prozessextraktion
- Konformanzanalyse
- Mining: Zusätzliche Perspektiven
- Betriebsunterstützung
- Werkzeugunterstützung
- Analysiere „Lasagne Prozesse“
- Analysiere „Spaghetti Prozesse“
- Kartographie und Navigation
- Epilog

- Modellbasierte Entwicklung sicherer Software





- **Vorheriger Abschnitt:** Verifikation von Petrinetzen und Einschränkungen modellbasierter Analyse.
- **Dieser Abschnitt:** „Data-Mining“:
 - Datenbasierte Modellanalyse



- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen
- Konfusionsmatrix
- Cluster-Analyse
- Assoziationsregel-Lernen
- Sequence- und Episode-Mining
- Hidden-Markov-Modell und Validierung



Process-Mining: Data-Mining auf Prozess-Daten.

→ Kurzer Hintergrund zu Data-Mining.

Vergleich Data-Mining vs. Process-Mining:

- **Process-Mining:** Ende-zu-Ende Prozesse.
- **Data-Mining:** **Datenbasiert** und nicht prozessbasiert.
- **Qualität** von Data-Mining und Process-Mining bewerten: viele Ähnlichkeiten, aber auch Unterschiede.
- Process-Mining-Techniken können Vorteile aus Erfahrungen im Bereich des Data-Mining ziehen.

6

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3: Seite 59 1. Abschnitt



drinker	smoker	weight	age
yes	yes	120	44
no	no	70	96
yes	no	72	88
yes	yes	55	52
no	yes	94	56
no	no	62	93
...	...		

Daten über 860 kürzlich verstorbene Personen.
Studie über Auswirkung von Alkoholkonsum, Rauchen
und Körpergewicht auf Lebenserwartung.

Fragen:

- Welchen Einfluss hat Rauchen und Trinken auf Körpergewicht?
- Trinken Leute, die rauchen, auch?
- Welche Faktoren beeinflussen Lebenserwartung am meisten?
- Gruppen mit ähnlichem Lebensstil identifizierbar?

7

Literatur:

Wil van der Aalst: Process Mining: Discovery,
Conformance and Enhancement of Business
Processes

- Kapitel 3.1: Seite 60



linear algebra	logic	program- ming	operations research	workflow systems	...	duration	result
9	8	8	9	9	...	36	cum laude
7	6	-	8	8	...	42	passed
-	-	5	4	6	...	54	failed
8	6	6	6	5	...	38	passed
6	7	6	-	8	...	39	passed
9	9	9	9	8	...	38	cum laude
5	5	-	6				
...				

Daten über 420 Studenten, um Verbindungen zwischen Kursnoten und Gesamtleistung im Bachelor zu untersuchen.

Fragen:

- Haben Noten bestimmter Kurse eine hohe Korrelation?
- Welche Wahl treffen Studenten mit Auszeichnung (cum laude)?
- Welche Kurse verzögern den Abschluss signifikant?
- Warum brechen Studenten ab?
- Gruppen mit ähnlichem Lernverhalten identifizierbar?

8

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3:1: Abbildung: Seite 60, Fragen Seite 61



cappuccino	latte	espresso	americano	ristretto	tea	muffin	bagel
1	0	0	0	0	0	1	0
0	2	0	0	0	0	1	1
0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	2	0
0	0	0	1	1	0	0	0
...

Daten über 240 Bestellungen in einem Café, aufgenommen von der Kasse.

Fragen:

- Welche Produkte werden häufig zusammen gekauft?
- Wann kaufen Kunden bestimmte Produkte?
- Gruppen typischer Kunden charakterisierbar?
- Wie fördert man Verkauf von Produkten mit hoher Gewinnmarge?

9

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.1: Abbildung & Fragen Seite 61



- Datensatz besteht aus **Instanzen** (Individuen, Entitäten, Fälle, Objekte oder Aufzeichnungen).
- **Variablen**: als Attribute, Features oder Datenelemente bezeichnet.

Zwei Typen:

- **Kategorielle Variablen**:
 - **Ordinal** (hoch – mittel – niedrig, mit Auszeichnung – bestanden – durchgefallen) oder
 - **Nominal** (true – false, rot – pink – grün).
- **Numerische Variablen** (geordnet, können nicht einfach aufgezählt werden).

10

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.1: Seite 61 letzter Abschnitt nach Fragen – Seite 62 1. Abschnitt



- **Klassifizierte Daten** (Labeled Data): Jede Instanz durch **Response-Variable** gekennzeichnet.
- **Ziel:** Erkläre **Response-Variable** (abhängige Variable) in Form von **Predictor-Variablen** (unabhängige Variable).
- **Klassifikationstechniken** (z.B.: Lernen mit Entscheidungsbäumen):
Setzen kategorielle Response-Variablen voraus.
Ziel: Instanzen anhand Predictor-Variablen klassifizieren.
- **Regressionstechniken:** Benötigen numerische Response-Variablen.
Ziel: Zu Daten passende Funktion mit wenigsten Fehlern finden.

11

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.1: Seite 61 Letzter Abschnitt - Seite 64
1. Abschnitt



- **Nicht überwachtetes Lernen** verwendet **unlabeled** data.
Variablen nicht in Response- und Predictor-Variablen unterteilt.

Beispiele:

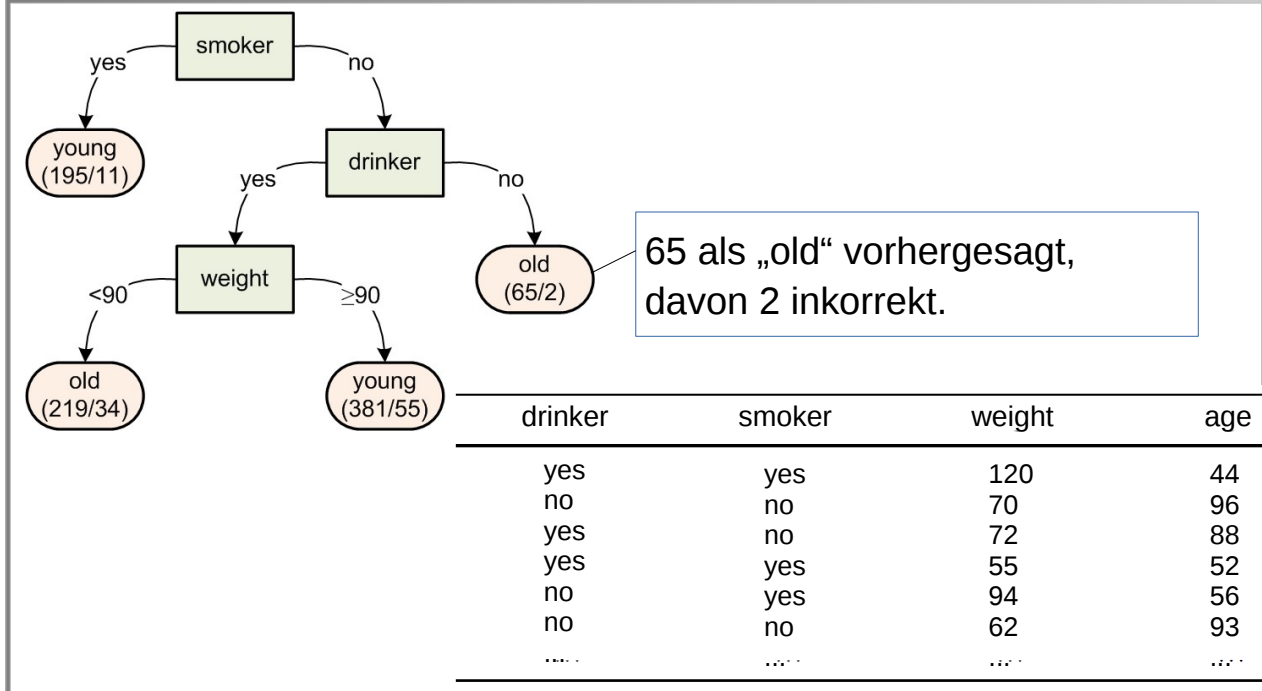
- **Clustering** (z.B.: k-means clustering und agglomerative hierarchical clustering).
- **Pattern discovery** (association rules).

12

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

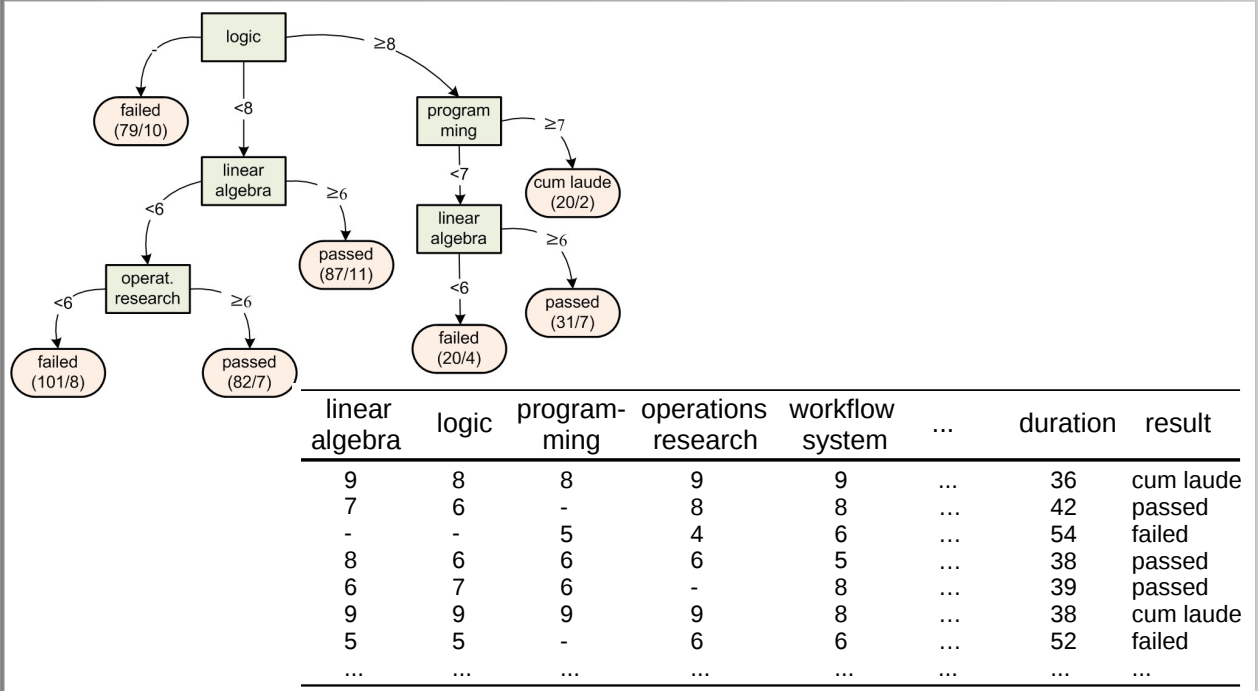
- Kapitel 3.1: Seite 64 Abschnitt „Unsupervised Learning“



Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

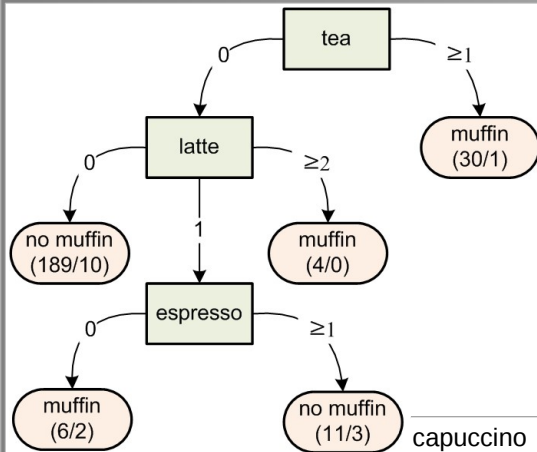
- Kapitel 3.1: Seite 65



Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.2: Obige Abbildung Seite 66, untere Abbildung Seite 60



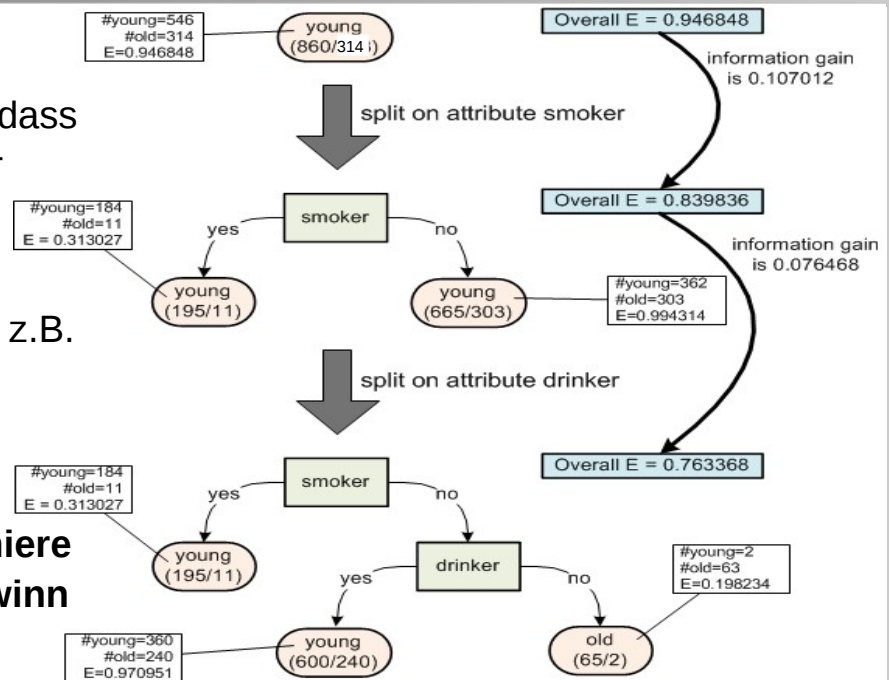
	capuccino	latte	espresso	americano	ristretto	tea	muffin	bagel
	1	0	0	0	0	0	1	0
	0	2	0	0	0	0	1	1
	0	0	1	0	0	0	0	0
	1	0	0	0	0	0	0	0
	0	0	0	0	0	1	2	0
	0	0	0	1	1	0	0	0

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.2: Obige Abbildung Seite 66, untere Abbildung Seite 61

- Teile Menge der Instanzen auf, sodass **Variation** in jeder Teilmenge **kleiner** wird.
- Variation messen z.B. durch **Entropie**.
- **Minimiere** durchschnittliche **Entropie**; **maximiere Informationsgewinn** pro Schritt.



16

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.2: Seite 68 letzter Abschnitt – Seite 70



Entropie E : **informationstheoretisches Maß** für „Chaos“ in einer Multimenge:

- Benötigte Anzahl Bits zur **Kodierung** der Multimenge ($E \geq 0$).
- Häufiges Ziel: Minimierung von E anstreben.

Annahme: Element v_i in Multimenge c_i -mal enthalten; Multimenge hat n Elemente.

Einzelentropie: $-\log_2(p_i)$, wobei $p_i = c_i/n$.

Gesamt-Entropie: gewichtetes Mittel der Einzel-Entropien:
$$E = -\sum_{i=1}^k (p_i \cdot \log_2(p_i))$$

Beispiel 1: Alle Elemente haben denselben Wert ($p_1=1$) $\rightarrow E = -\log_2 1 = 0$

Beispiel 2:

- 4 mögliche Elemente $V = \{a,b,c,d\}$
- Sequenz: *bacaabadabadcaba...*
- Relative Häufigkeiten: $p_1 = 0,5$; $p_2 = 0,25$; $p_3 = 0,125$; $p_4 = 0,125$
- $E = 1,75$

17

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.2: Seite 67-68



Anwendung:

- Bei großer Datenmenge viele verschiedene Attribute von Bedeutung.
→ Entscheidungsbaum ohne große Berechnungen generieren.

Algorithmus:

- Iterativ
- Benutzt **Entropie** zur Bestimmung von Baum-Knoten.
- Abbruch, falls jedem Blattknoten eine Klassifikation zugeordnet.

Eingabe:

- Menge zu klassifizierender Objekte, Wurzelknoten, Menge noch zu vergebener Merkmale

Ausgabe:

- Struktur mit Tupeln (Entscheidungsbaum)

18

Literatur:

Thomas A. Runkler: Data Mining: Methoden und
Algorithmen Intelligenter Datenanalyse

<http://www.ub.tu-dortmund.de/katalog/titel/1294605>

- Abschnitt 8.6 (Entscheidungsbäume, ab S.99)
- ID3 (S.102-104)



Aufruf von ID3 ausgehend von

- Menge $\{1, \dots, n\}$ der Indizes aller zu klassifizierender Objekte,
- Wurzelknoten des Entscheidungsbaums,
- Menge $\{1, \dots, p\}$ aller noch zu vergebenden Merkmale.

Am aktuellen Knoten gehören **alle Daten derselben Klasse** → **Abbruch!**

Sonst:

- Erwarteten **Informationsgewinn gi** für jedes zu vergebendes Merkmal berechnen.
- Merkmal j bestimmen, der höchsten Informationsgewinn liefert.
- Anhand Werte des **Gewinnermerkmals j** Datensatz in **disjunkte Teilmengen** zerlegen.
- Für jede **nichtleere Teilmenge** neuen Knoten anhängen.
- Für jeden angehängten Knoten rekursiv zugehörigen Teilbaum berechnen.

Ausgabe: Entscheidungsbaum bzw. Struktur mit Tupeln

19

Literatur:

Thomas A. Runkler: Data Mining: Methoden und
Algorithmen Intelligenter Datenanalyse

<http://www.ub.tu-dortmund.de/katalog/titel/1294605>

- Abschnitt 8.6 (Entscheidungsbäume, ab S.99)
- ID3 (S.102-104)



Gegeben: $X = \{x_1, \dots, x_n\} \subset \{1, \dots, v_1\} \times \dots \times \{1, \dots, v_p\}$, $y = \{y_1, \dots, y_n\} \subset \{1, \dots, c\}$

Aufrufen: $ID3(\{1, \dots, n\}, \text{Wurzel}, \{1, \dots, p\})$

Prozedur $ID3(I, N, K)$

1. Wenn alle $y(I)$ gleich dann abbrechen
2. Berechne Informationsgewinn $g_j((X(I), y(I))) = E(X) - \sum_{j \in y} \left(\frac{|X_j|}{|X|}\right) E(X_j) \forall j \in K$
X: Datensatz, E(X): Entropie im Datensatz, c: Klassifikation
 X_j : Untermenge von X, $|X|$: Mächtigkeit von X, $|X_j|$: Mächtigkeit von X_j
3. Bestimme Gewinnermerkmal $i = \text{argmax} \{g_j((X(I), y(I)))\}$
4. Zerlege I in v_i disjunkte Teilmengen
 $I_j = \{k \in I \mid x_k^{(i)} = k\} \quad j = 1, \dots, v_i$
5. für j mit $I_j \neq \{\}$
 - Generiere neuen Knoten N_j und hänge ihn an N
 - Aufrufen: $ID3(I_j, N_j, I \setminus \{i\})$

20

Literatur:

Thomas A. Runkler: Data Mining: Methoden und
Algorithmen Intelligenter Datenanalyse

<http://www.ub.tu-dortmund.de/katalog/titel/1294605>

- Abschnitt 8.6 (Entscheidungsbäume, ab S.99)
- ID3 (S.102-104)
- Besonders Abb. 8.7 (S.103)



Bedingte Entropie:

- Maß über den Wert einer Zufallsvariable, welche verbleibt, nachdem das Ergebnis einer anderen Zufallsvariable bekannt wird.

$$H(C|A_t) = - \sum_{i=1}^{k_t} p_i \sum_{j=1}^k \left(\frac{x_{i,j}}{x_i} \right) \log_2 \left(\frac{x_{i,j}}{x_i} \right)$$

Größe:

	giftig	essbar
klein	1	2
groß	0	2

$$x_{\text{klein}} = 3 \quad p_{\text{klein}} = \frac{3}{5} = 0.6$$

$$x_{\text{groß}} = 2 \quad p_{\text{groß}} = \frac{2}{5} = 0.4$$

$$\begin{aligned} H(C|A_{\text{Größe}}) &= - \left[0.6 \left(\frac{1}{3} \log_2 \left(\frac{1}{3} \right) + \frac{2}{3} \log_2 \left(\frac{2}{3} \right) \right) + 0.4 \left(\frac{0}{2} \log_2 \left(\frac{0}{2} \right) + \frac{2}{2} \log_2 \left(\frac{2}{2} \right) \right) \right] \\ &= - \left[0.4 \left(-\log_2(3) + \frac{2}{3} \right) + 0 \right] \approx \mathbf{0.4562} \end{aligned}$$

21

Literatur:

Thomas A. Runkler: Data Mining: Methoden und
Algorithmen Intelligenter Datenanalyse

<http://www.ub.tu-dortmund.de/katalog/titel/1294605>

- Abschnitt 8.6 (Entscheidungsbäume, ab S.99)
- ID3 (S.102-104)



- **Punkte:**

	giftig	essbar
ja	1	1
nein	0	3

$$x_{ja}=2 \quad p_{ja}=\frac{2}{5}=0.4$$

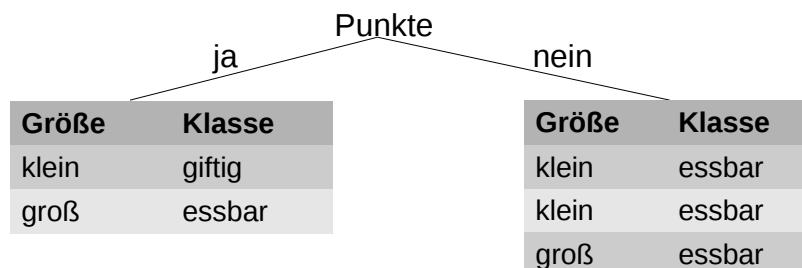
$$x_{nein}=3 \quad p_{nein}=\frac{3}{5}=0.6$$

$$H(C|A_{Punkte}) = -\left[0.4\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + 0.6\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right)\right]$$

$$- [0.4(-1) + 0] = 0.4$$

→ **Bedingte Entropie minimal** für das Attribut Punkte.

- Entscheidungsbaum nach **Rekursionsebene I:**



22

Literatur:

Thomas A. Runkler: Data Mining: Methoden und
Algorithmen Intelligenter Datenanalyse

<http://www.ub.tu-dortmund.de/katalog/titel/1294605>

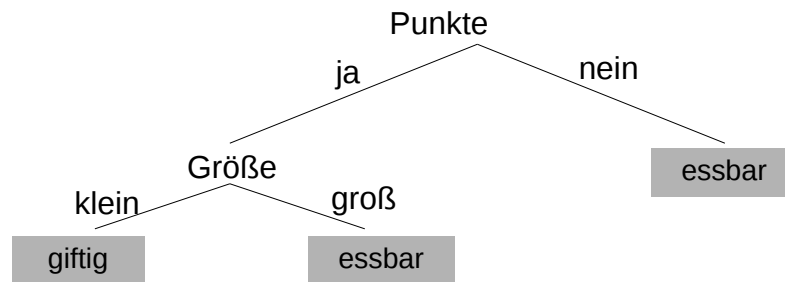
- Abschnitt 8.6 (Entscheidungsbäume, ab S.99)
- ID3 (S.102-104)



Rekursionsstufe II:

- Objektsammlung M_{nein} besteht aus 3 Beispielen, die alle in die Klasse der essbaren Pilze gehören. $\rightarrow M_{nein}$ trivial
- Objektsammlung M_{ja} trivial, der Entscheidungsbaum besteht nur aus Blättern mit Label giftig bzw. essbar.

Entscheidungsbaum nach Rekursionsebene II:



http://www2.informatik.uni-hamburg.de/wsv/teaching/praktika/GwvPRAK_WiSe10/aufg05_ID3_script.pdf

23

Literatur:

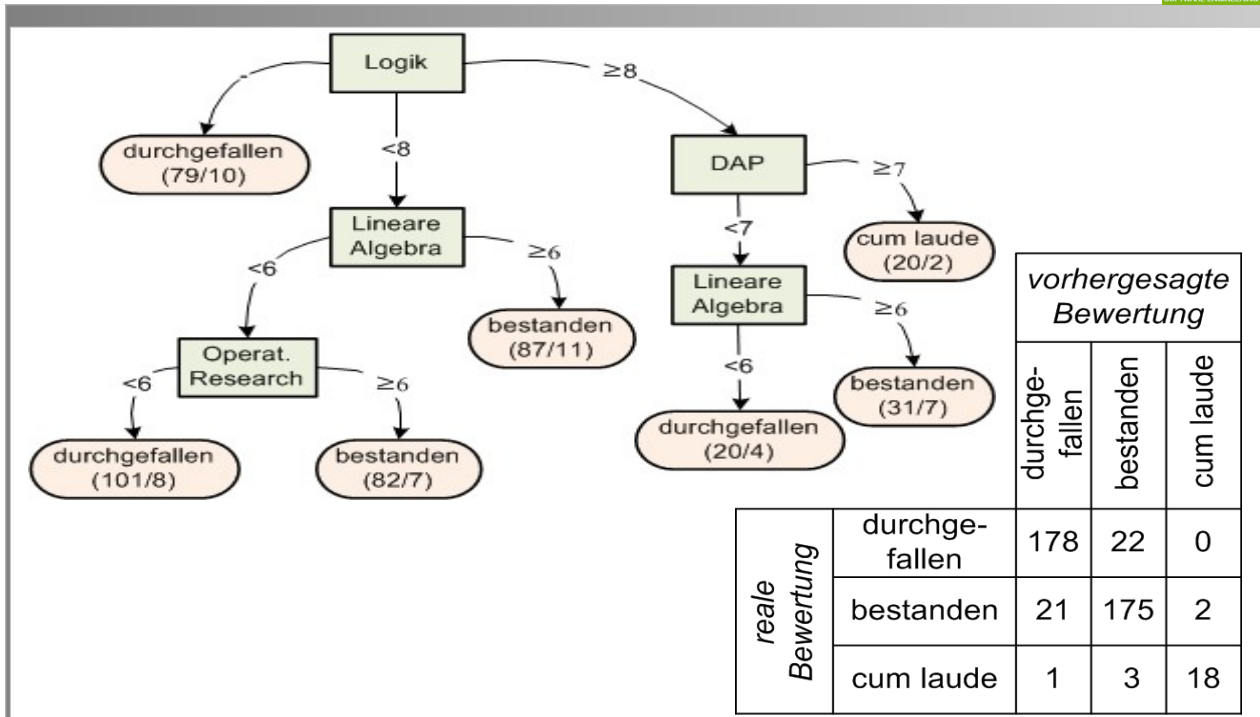
Thomas A. Runkler: Data Mining: Methoden und
Algorithmen Intelligenter Datenanalyse

<http://www.ub.tu-dortmund.de/katalog/titel/1294605>

- Abschnitt 8.6 (Entscheidungsbäume, ab S.99)
- ID3 (S.102-104)



- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen
- **Konfusionsmatrix**
- Cluster-Analyse
- Assoziationsregel-Lernen
- Sequence- und Episode-Mining
- Hidden-Markov-Modell und Validierung



25

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.2: Obige Abbildung Seite 66, untere Abbildung Seite 83 (Kap. 3.6)



Bsp. für 2 Klassen + und -		vorhergesagte Klasse		
		+	-	
reale Klasse	+	tp	fn	p
	-	fp	tn	n
		p'	n'	N

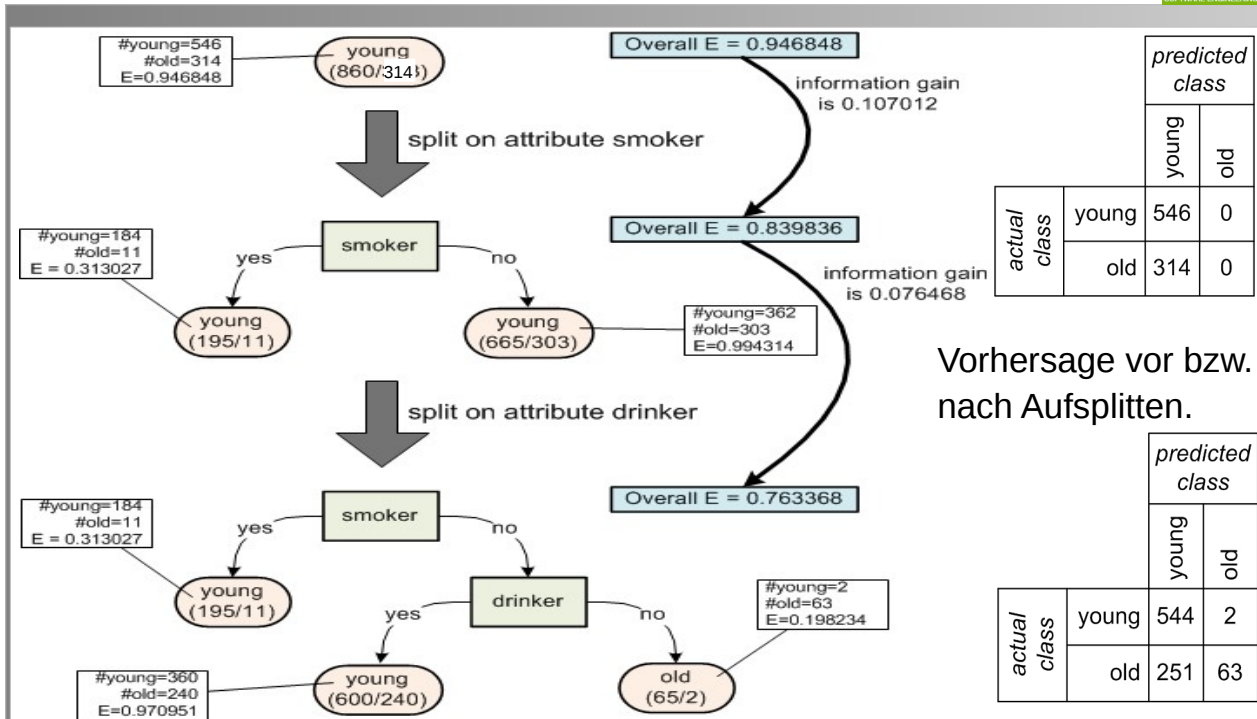
Name	Formel
Error	$(fp+fn)/N$
accuracy	$(tp+tn)/N$
tp-Rate	tp/p
fp-Rate	fp/n
precision	tp/p'
recall	tp/p (= tp-Rate!)

tp: Anzahl **true positives**; korrekterweise als positiv klassifiziert.
fn: Anzahl **false negatives**; als negativ klassifiziert, aber positiv.
fp: Anzahl **false positives**; als positiv klassifiziert, aber negativ.
tn: Anzahl **true negatives**; korrekterweise als negativ klassifiziert.

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.6: Seite 83-84



27

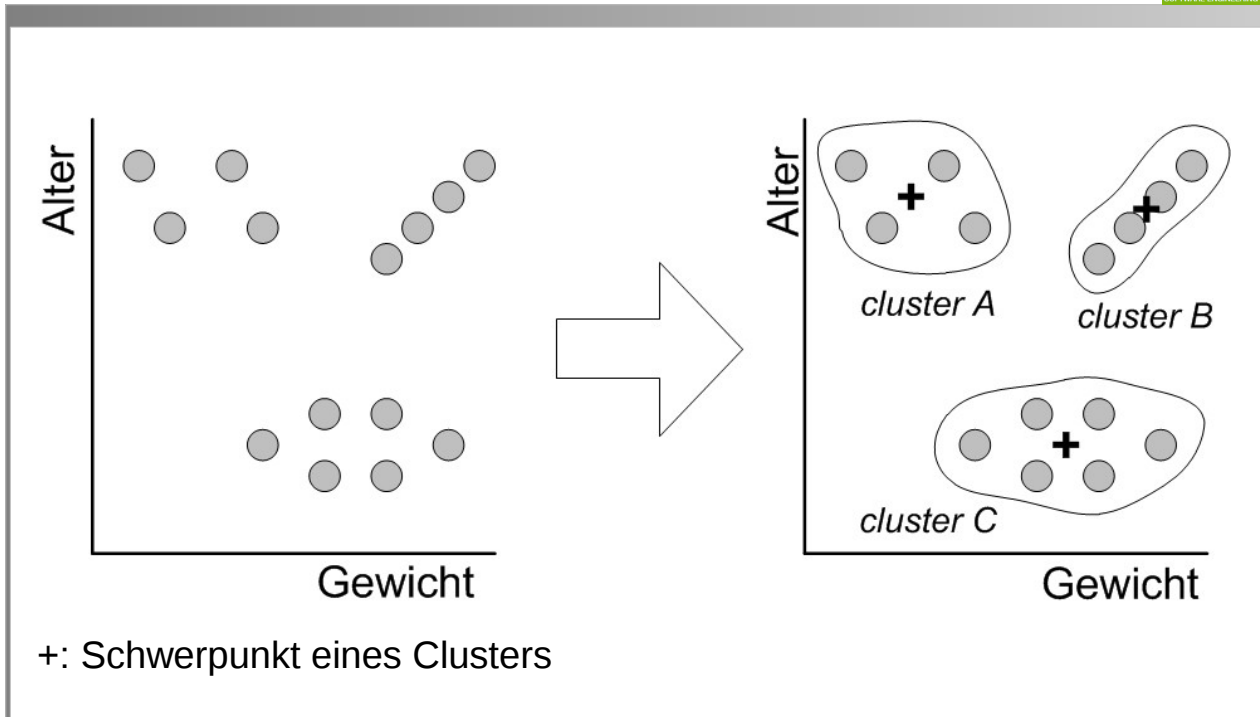
Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Linke Abbildung: Kap. 3.2 Seite 69
- Rechte Abbildung: Kap. 3.6 Seite 85



- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen
- Konfusionsmatrix
- Cluster-Analyse
- Assoziationsregel-Lernen
- Sequence- und Episode-Mining
- Hidden-Markov-Modell und Validierung



29

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kapitel 3.3: Seite 71



Von **Stuart P. Lloyd** 1982 vorgestellt (benannt: k-Means).

- Anzahl K zu ermittelnder Cluster vorher festlegen.
- Start ($i=0$): Positionen der **Clusterschwerpunkte** zufällig initialisieren.
- Objekte **nächstgelegenen Schwerpunkten** zuordnen. ($i=1$)
- Bei jeder Iteration i Schwerpunkt und nächstliegende Kandidaten neu berechnen.
- Dies Wiederholen bis **Summe quadratischer Distanz** einzelner Objekte zu ihrem jeweiligen Clusterschwerpunkt über alle Cluster ein **Minimum** erreicht.

→ Mathematische Darstellung:
$$J = \sum_{n=1}^N \sum_{k=1}^K \| \vec{x}_n - \vec{\mu}_k \|^2$$

- \vec{x}_n Datensätze und $\vec{\mu}_k$ Schwerpunkte der Cluster.
- **Entscheidungskriterium** für Cluster-Zugehörigkeit der Testobjekte: Abstände der Testvektoren von Clusterschwerpunkten.

30

Literatur:

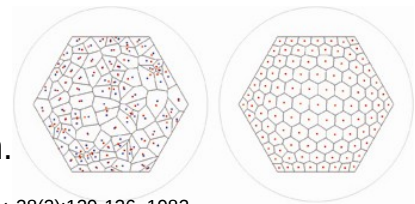
Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.3: Seite 71 letzter Abschnitt – Seite 72 letzter Abschnitt

k -MEANS(P, k):

- 1: Wähle Punkte $C = \{c_1, \dots, c_k\}$ zufällig gleichverteilt
- 2: **repeat**
- 3: Assoziiere jeden Punkt aus P mit dem Zentrumspunkt c_j mit dem geringsten Abstand, um eine Partitionierung P_1, P_2, \dots, P_k zu erhalten
- 4: Berechne für jede Teilmenge P_j den Zentroid und verwende diese Menge von Zentroiden als neue Menge von Zentrumspunkten
- 5: **until** Menge der Zentrumspunkte ändert sich nicht mehr

- Beginn mit **zufälligen** Zentrumspunkten.
- Zur **Verringerung der Gesamtfehler** der **Approximation** neue Punkte in 3. und 4. wählen.



Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129-136, 1982.

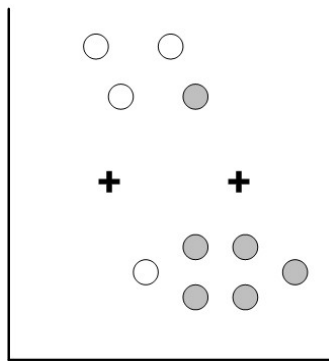
31

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.4: Seite 74

k-Means-Clustering: Anzahl k von Clustern a-priori festgelegt.



(a)

k-Means-Algorithmus veranschaulicht für $k=2$:

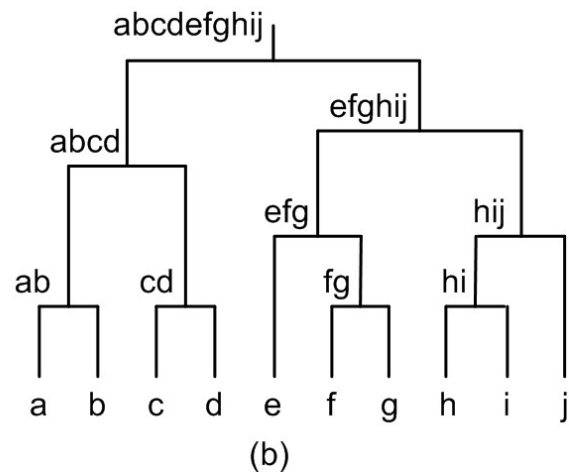
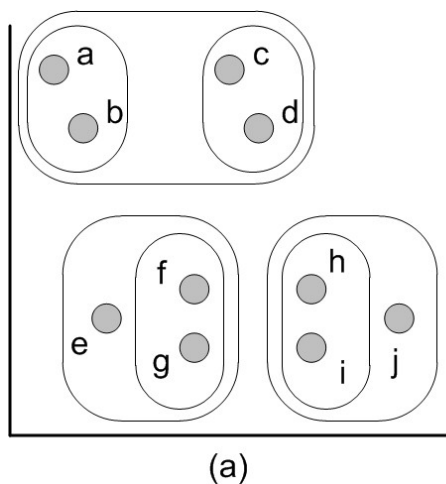
(a) „Schwerpunkte“ zufällig setzen; Punkte nächstgelegenen Schwerpunkten zuordnen.

32

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.3: Seite 71 letzter Abschnitt – Seite 72 letzter Abschnitt



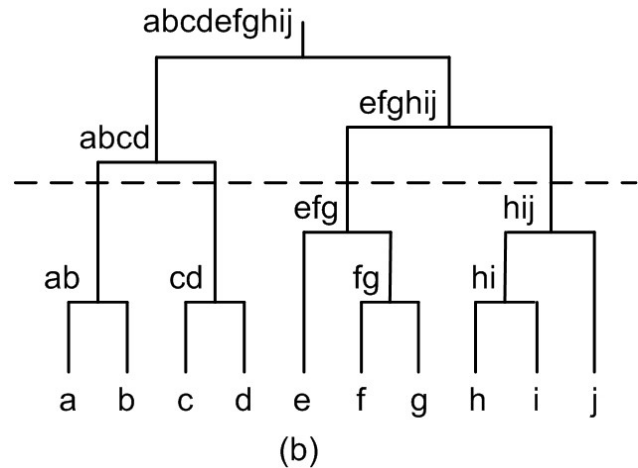
(a) Agglomerative hierarchische Cluster-Analyse: **Bottom-up**-Vorgehen.
(b) Visualisierung durch Dendrogramm.

35

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.3: Seite 72 letzter Abschnitt – Seite 73 erster Abschnitt



(b) **Horizontale Linie** in Dendrogramm korrespondiert zu konkretem Clustering auf bestimmter **Abstraktionsebene**.

36

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.3: Seite 73



- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen
- Konfusionsmatrix
- Cluster-Analyse
- Assoziationsregel-Lernen
- Sequence- und Episode-Mining
- Hidden-Markov-Modell und Validierung



Ziel: Regeln der Form "IF X THEN Y" lernen: $X \Rightarrow Y$.

Definiere Maße für **Relevanz / Gültigkeit / Aussagekraft** der Regel:

$$\text{support}(X \Rightarrow Y) = N_{X \wedge Y} / N$$

$$\text{confidence}(X \Rightarrow Y) = N_{X \wedge Y} / N_X$$

$$\text{lift}(X \Rightarrow Y) = \frac{N_{X \wedge Y} / N}{(N_X / N) (N_Y / N)} = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

(N_X : Anzahl Datensätze, die alle Eigenschaften in X erfüllen).

Welche Bedeutung könnten hohe / niedrige Werte haben ?

39

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.4: Seite 74

Beispiel: Einkaufswagen-Analyse



cappuccino	latte	espresso	americano	ristretto	tea	muffin	bagel
1	0	0	0	0	0	1	0
0	2	0	0	0	0	1	1
0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	2	0
0	0	0	1	1	0	0	0
...

Zwei Beispiel-Hypothesen:

$tea \wedge latte \Rightarrow muffin$ → Leute, die *tea* oder *latte* bestellen,
bestellen auch *muffins*.

$tea \Rightarrow muffin \wedge bagel$

41

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Obige Abbildung: Kap. 3.1 Seite 61
- Untere Abbildung: Kap. 3.4 Seite 74 letzter Abschnitt – Seite 75 1. Abschnitt



$tea \wedge latte \Rightarrow muffin$, d.h.: $X = tea \wedge latte$ und $Y = muffin$

$$support(X \Rightarrow Y) = N_{X \wedge Y} / N = N_{tea \wedge latte \wedge muffin} / N = 15 / 240 = 0.0625$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y} / N_X = N_{tea \wedge latte \wedge muffin} / N_{tea \wedge latte} = 15 / 20 = 0.75$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y} = \frac{N_{tea \wedge latte \wedge muffin} N}{N_{tea \wedge latte} N_{muffin}} = \frac{15 \times 240}{20 \times 40} = 4.5$$

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.4 Seite 75 ab 1. Abschnitt



Definition: Für gegebenen *support*-Schwellwert *minsup*: Menge *Z* heißt **frequent item-set**, wenn $N_Z / N \geq \text{minsup}$.

Assoziationsregeln wie folgt **generierbar**:

- 1) Generiere **frequent item-sets**: alle Mengen *Z* sodass N_Z / N größer als gegebener Schwellwert für *support* und $|Z| \geq 2$.
- 2) Für jedes *frequent item-set* *Z* betrachte Partition in nicht-leere Teilmengen *X*, *Y*.
Behalte Regeln $X \Rightarrow Y$, für die **confidence** gegebenen Schwellwert überschreitet oder gleich ist.

43

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.4: Seite 75 letzter Abschnitt



Beobachtung:

Unter welcher Voraussetzung ist **Teilmenge** einer *frequent item-set* ebenfalls eine *frequent item-set* ?

44

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.5: Seite 76



1. Create I_1 . This is the set of singleton frequent item-sets, i.e., item-sets with a support above the threshold *minsup* containing just one element.
2. $k := 1$
3. If $I_k = \emptyset$, then output $\bigcup_{i=1}^k I_i$ and end. If $I_k \neq \emptyset$, continue with the next step.
4. Create C_{k+1} from I_k . C_{k+1} is the candidate set containing item-sets of cardinality $k + 1$. Note that one only needs to consider elements that are the union of two item-sets A and B in I_k such that $|A \cap B| = k-1$ and $|A \cup B| = k + 1$. (wg. o.g. Beobachtung)
5. For each candidate frequent item-set $c \in C_{k+1}$: examine all subsets of c with k elements; delete c from C_{k+1} if any of the subsets is not a member of I_k . (wg. o.g. Beobachtung)
6. For each item-set c in the pruned candidate frequent item-set C_{k+1} , check whether c is indeed frequent. If so, add c to I_{k+1} . Otherwise, discard c .
7. $k := k + 1$ and return to Step 3.

46

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.5: Seite 76



- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen
- Konfusionsmatrix
- Cluster-Analyse
- Assoziationsregel-Lernen
- **Sequence- und Episode-Mining**
- Hidden-Markov-Modell und Validierung

Sequence-Mining

customer	seq. number	timestamp	items
Wil	1	02-01-2011:09.02	{cappuccino}
	2	03-01-2011:10.06	{espresso,muffin}
	3	05-01-2011:15.12	{americano,cappuccino}
	4	06-01-2011:11.18	{espresso,muffin}
	5	07-01-2011:14.24	{cappuccino}
	6	07-01-2011:14.24	{americano,cappuccino}
Mary	1	30-12-2010:11.32	{tea}
	2	30-12-2010:12.12	{cappuccino}
	3	30-12-2010:14.16	{espresso,muffin}
	4	05-01-2011:11.22	{bagel,tea}
Bill	1	30-12-2010:14.32	{cappuccino}
	2	30-12-2010:15.06	{cappuccino}
	3	30-12-2010:16.34	{bagel,espresso,muffin}
	4	06-01-2011:09.18	{ristretto}
	5	06-01-2011:12.18	{cappuccino}

... ..
 $X = \langle \{cappuccino\}, \{espresso\} \rangle$

... ..
 $Y = \langle \{cappuccino\}, \{espresso\}, \{latte, muffin\} \rangle$

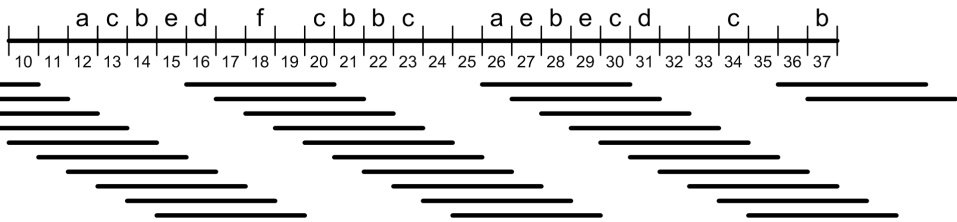
48

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

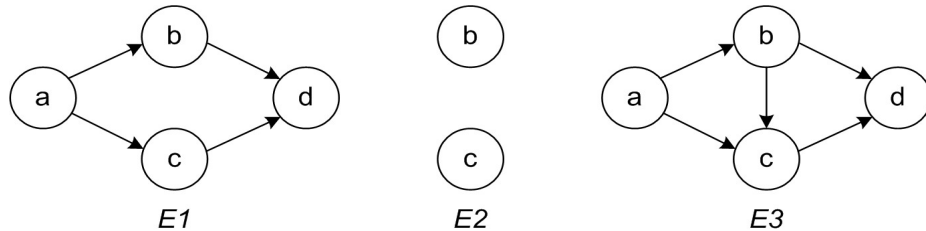
- Kap. 3.5: Seite 78

32 Zeitfenster
der
Länge 5:



„Episode“: Prozessmodell-Fragment.

3 Beispiele:



(weitere sind möglich, z.B. unter Verwendung von e,f).

Wie oft treten diese Episoden hier auf (d.h. wie oft gibt es vollständigen Ablauf eines Modells in Zeitfenster der Länge 5) ?

49

Literatur:

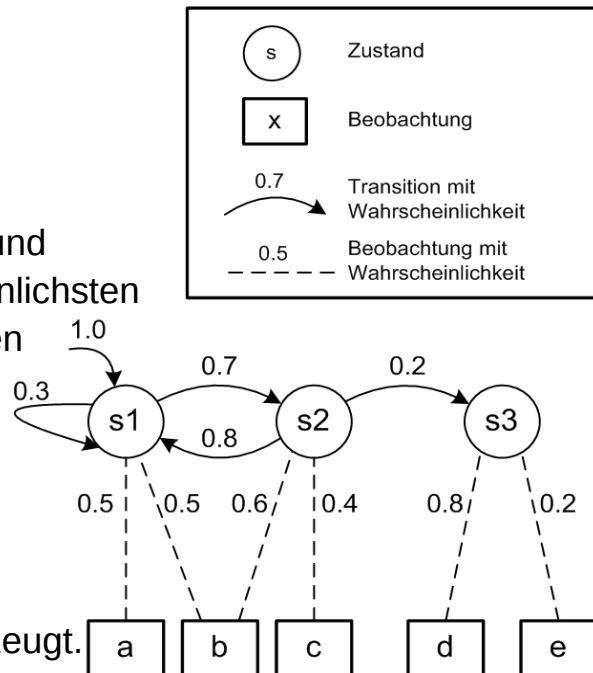
Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.5: Seite 79



- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen
- Konfusionsmatrix
- Cluster-Analyse
- Assoziationsregel-Lernen
- Sequence- und Episode-Mining
- Hidden-Markov-Modell und Validierung

- 1) **Wahrscheinlichkeit** der Sequenz berechnen (gegeben Beobachtungssequenz und Hidden-Markov-Modell).
- 2) Gegeben Beobachtungssequenz und Hidden-Markov-Modell, wahrscheinlichsten „**hidden path**“ im Modell berechnen (= interne Zustandsfolge; kann zur z.T. beobachtet werden).
- 3) Bei gegebener Beobachtungssequenz **Hidden-Markov-Modell** ableiten, welches mit max. Wahrscheinlichkeit Sequenzen erzeugt.

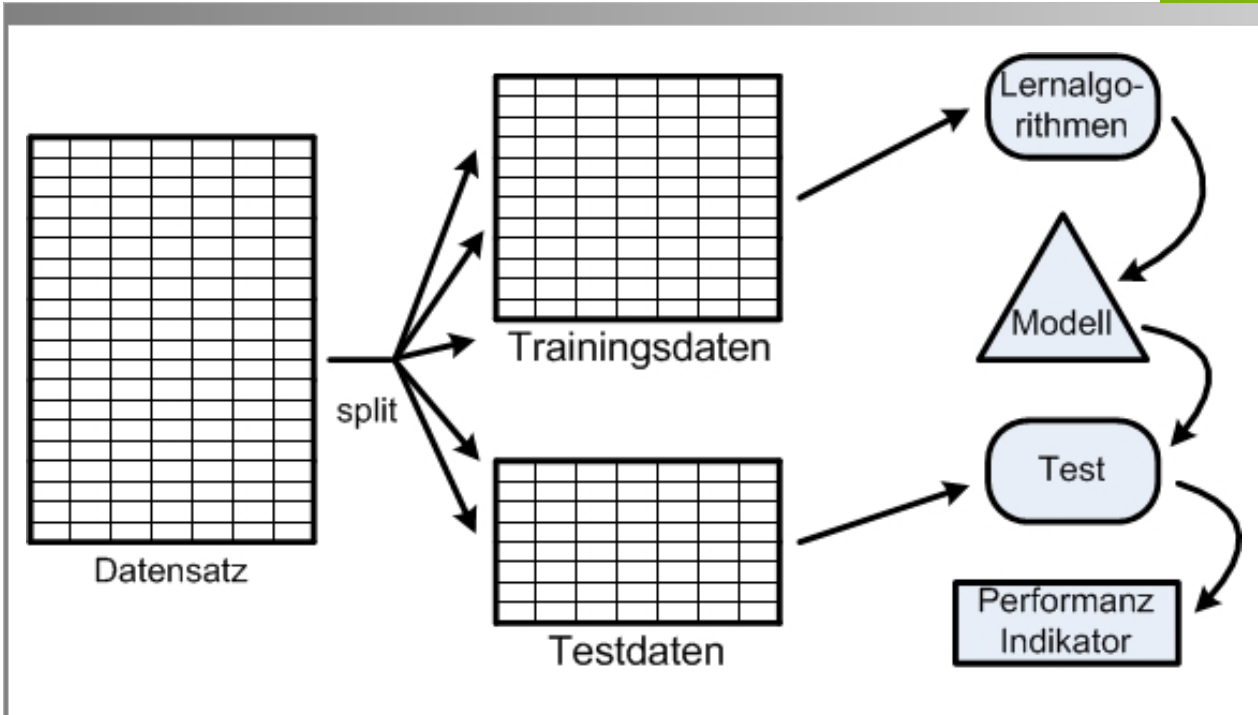


52

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.5: Seite 81 vorletzter Abschnitt; Abbildung Seite 82

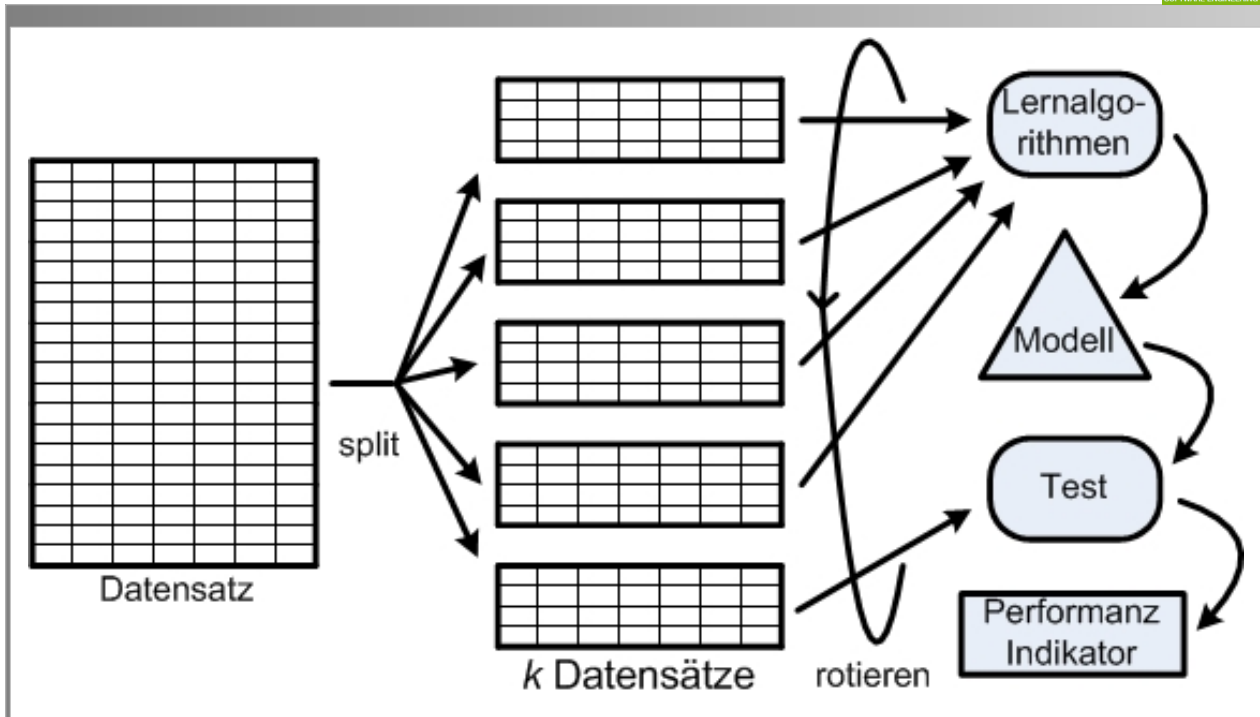


53

Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.6: Seite 86



Literatur:

Wil van der Aalst: Process Mining: Discovery, Conformance and Enhancement of Business Processes

- Kap. 3.6: Seite 87



In diesem Abschnitt:

- Von Datensätzen zu Entscheidungsbäumen durch überwachtes / nicht überwachtes Lernen.
- Konfusionsmatrix.
- Cluster-Analyse.
- Assoziationsregel-Lernen.
- Sequence- und Episode-Mining.
- Hidden-Markov-Modell und Validierung.

Im nächsten Abschnitt:

- Datenbeschaffung.