
Data Warehousing und Data Mining

Eine Einführung in entscheidungsunterstützende Systeme

Interaktive Folien zu Kapitel 7
Regelinduktion

Expertensysteme



- ① Expertenbefragung
- ② Query the User
- ③ ev. Regelinduktion



Wissenserwerb



- ① Regelinduktion
- ② Query the User



Data Mining

Einordnung

Nutzwertanalyse am Beispiel von AHP

✓ Kioskstandort  , Personalauswahl 





Was Wenn-Analyse

✓ Erfolgsrechnung 
✓ Anzeigenplanung , Produktionsplanung 

Regelbasierte Systeme

✓ Spesen , Betriebskredit 
✓ Regelverkettung  

Data Warehouses

✓ Anlageberatung  
✓ Lieferfrist, Handel, Verkauf  

Data Mining - Ein Überblick

✓ Zeitschriften , Bank 

⇒ **Regelinduktion**

⇒ **Spesen** , **Bonitätsklassifikation**  

Neuronale Netze

- Bonitätsklassifikation , Bonitätsvorhersage 
- EindimPerzeptron  , ZweidimPerzeptron  
- MehrklassPerzeptron  
- MehrstufPerzeptron  

Unterrichtsmaterial

Software

- Demonstrationsversion von *XpertRule Profiler*
(Klassifikation mit Entscheidungsbäumen¹)
Neuere Version unter dem Namen XpertRule Miner
([Evaluation Copy](#))
- *Visual Basic* für Applikationen unter *MS Excel*

Beispiele und Übungen

- [Spesenabrechnung](#) mit *XpertRule* 📌
- [Bonitätsklassifikation](#) mit *XpertRule* 📌

Produktinformation

- <http://www.attar.com>

¹ Zum Begriff des Entscheidungsbaums vgl. “Regelbasierte Systeme”

Grundlagen¹

Grundlagen

- ⇒ Regelinduktion und regelbasierte Systeme 2
- ⇒ Regelinduktion und Data Mining 18

Entwicklung mit *XpertRule Profiler*

- Bonitätsklassifikation   30

Theorie

- Induktionsalgorithmus ID3 48

¹ Ein ausführliches Foliverzeichnis finden Sie am Ende des Kapitels

Motivationen für die Regelinduktion

Regeln von Experten sind manchmal mangelhaft



① Regeln für ›Expertensysteme



Ergänzung von Expertenregeln am Fallbeispiel Spesen

Ad hoc-Klassifikationen sind meist unzuverlässig



② ›Entscheidungsbäume für die ›Klassifikation



Data Mining am Fallbeispiel ›Bonitätsklassifikation

SPESEN - Redundante Regeln

Wenn nur die Abteilungen *Einkauf* und *Verkauf* existieren, gilt ...

① **Redundante** Wissensbasis

WENN

Funktion = Direktor UND

Abteilung = *Einkauf*

DANN

Antrag akzeptieren

WENN

Funktion = Direktor UND

Abteilung = *Verkauf*

DANN

Antrag akzeptieren

② **Redundanzfreie** Wissensbasis

WENN

Funktion = Direktor

DANN

Antrag akzeptieren

Eine Wissensbasis heisst **redundant**, wenn auch weniger Regeln oder Bedingungen die *gleichen* Folgerungen ergeben

Widersprüchliche Regeln

① **Widersprüchliche** Wissensbasis

WENN

Funktion = Direktor

DANN

Antrag *akzeptieren*

WENN

Funktion = Direktor

DANN

Antrag *ablehnen*

② **Widerspruchsfreie** Wissensbasis

nur eine der beiden Regeln.

Eine Wissensbasis heisst **widersprüchlich**, wenn Regeln mit gleichen oder äquivalenten Bedingungs-
teilen *unterschiedliche* Folgerungen ergeben

Lückenhafte Regeln

Wenn die Abteilungen *Rechnungswesen*, *Einkauf* und *Verkauf* existieren, gilt ...

① **Lückenhafte** Wissensbasis

WENN

Abteilung = *Rechnungswesen*

DANN

Hotelklasse erfragen

WENN

Abteilung = *Verkauf*

DANN

Hotelklasse irrelevant

② **Vollständige** Wissensbasis

Zusätzlich:

WENN

Abteilung = *Einkauf*

DANN

Hotelklasse

Eine Wissensbasis heisst **lückenhaft**, wenn die Bedingungen nur einen Teil der interessierenden Domäne abdecken

Ineffiziente Regeln

① Ineffiziente Wissensbasis

WENN

Funktion = Vizedirektor UND
Hotelklasse = Economy UND
Abteilung = Rechnungswesen

DANN Antrag akzeptieren

WENN

Funktion = Vizedirektor UND
Hotelklasse = Economy UND
Abteilung = Verkauf

DANN Antrag akzeptieren

② Effiziente Wissensbasis

WENN

Funktion = Vizedirektor UND
Hotelklasse = Economy UND
(Abteilung = Rechnungswesen ODER Abteilung = Verkauf)

DANN Antrag akzeptieren

Diese Formulierung ist effizienter, weil die Bedingungen
..... nur einmal ausgewertet werden müssen

Eine Wissensbasis heisst **ineffizient**, wenn syntaktisch geänderte Regeln - zum Beispiel anders geordnete Regeln - schneller ausgewertet werden können

Regeln automatisch extrahieren

Regeln von Experten sind manchmal ...

- redundant
- widersprüchlich
- lückenhaft oder
- ineffizient



Induziere Regeln auch aus automatisch erzeugten Attributewerttupeln, statt die Regeln nur von Experten zu erfragen

7.1 Elementarregeln

<i>Funktion</i>	<i>Abteilung</i>	<i>Hotelklasse</i>	<i>Spesenbescheid</i>
Direktor	Rechnungswesen	Komfort	?
Direktor	Rechnungswesen	Standard	?
Direktor	Rechnungswesen	Economy	?
...

Generierte **Attributwerttupel** bilden die Regelbedingungen



Experten ergänzen die Regelbedingungen durch **Folgerungen**



Ergänzte Attributwerttupel heissen **Elementarregeln**, z.B. ...

WENN

Funktion = Direktor UND

Abteilung = Rechnungswesen UND

Hotelklasse = Komfort

DANN

Antrag akzeptieren

Elementarregeln bestehen also aus ...
einem *generierten* **Bedingungsteil** und
einem *manuellen* **Folgerungsteil**

7.2 Elementarregeln sind vollständig ...

<i>Funktion</i>	<i>Abteilung</i>	<i>Hotelklasse</i>	<i>Spesenbescheid</i>
Direktor	Rechnungswesen	Komfort	Antrag akzeptieren
Direktor	Rechnungswesen	Standard	Antrag akzeptieren
Direktor	Rechnungswesen	Economy	Antrag akzeptieren
Direktor	Verkauf	Komfort	Antrag akzeptieren
Direktor	Verkauf	Standard	Antrag akzeptieren
Direktor	Verkauf	Economy	Antrag akzeptieren
Vizedirektor	Rechnungswesen	Komfort	...
Vizedirektor	Rechnungswesen	Standard	...
Vizedirektor	Rechnungswesen	Economy	...
Vizedirektor	Verkauf	Komfort	...
Vizedirektor	Verkauf	Standard	...
Vizedirektor	Verkauf	Economy	...
Prokurist	Rechnungswesen	Komfort	...
Prokurist	Rechnungswesen	Standard	...
Prokurist	Rechnungswesen	Economy	...
Prokurist	Verkauf	Komfort	...
Prokurist	Verkauf	Standard	...
Prokurist	Verkauf	Economy	...

... weil sie alle möglichen **Regelbedingungen** enthalten

... aber oft redundant und ineffizient

<i>Funktion</i>	<i>Abteilung</i>	<i>Hotelklasse</i>	<i>Spesenbescheid</i>
Direktor	Rechnungswesen	Komfort	Antrag akzeptieren
Direktor	Rechnungswesen	Standard	Antrag akzeptieren
Direktor	Rechnungswesen	Economy	Antrag akzeptieren
Direktor	Verkauf	Komfort	Antrag akzeptieren
Direktor	Verkauf	Standard	Antrag akzeptieren
Direktor	Verkauf	Economy	Antrag akzeptieren
...

Die generierten Elementarregeln sind **redundant**, weil sie zusammen dasselbe aussagen wie die folgenden Regel :

WENN

Funktion = Direktor

DANN

Antrag akzeptieren

Beurteilung von Elementarregeln

Vorteile

- vollständig

Nachteile

- redundant
- ineffizient



Die **Regelinduktion** fasst viele **spezielle** Elementarregeln zu einem **allgemeinen** Entscheidungsbaum bzw. wenigen allgemeinen Regeln zusammen

7.3 Regelinduktion am Beispiel SPESEN

Das Modul *Spesenbescheid* wird durch Regelinduktion **redundanzfrei** und **effizienter** als die Elementarregeln



WENN **Funktion** = Direktor
DANN Spesenantrag akzeptieren

WENN **Funktion** = Vizedirektor
DANN

WENN **Hotelklasse** = Komfort % 1. Schachtelung
DANN Antrag ablehnen

WENN **Hotelklasse** = Standard
DANN

WENN **Abteilung** = Rechnungswesen % 2. Schachtelung
DANN Spesenantrag akzeptieren

WENN **Abteilung** = Verkauf
DANN Spesenantrag ablehnen

WENN **Hotelklasse** = Economy
DANN Spesenantrag akzeptieren

WENN

Funktion = Prokurist UND
Hotelklasse = Economy ODER
Hotelklasse = Standard

DANN Spesenantrag ablehnen

WENN

Funktion = Prokurist UND
Hotelklasse = Economy

DANN Spesenantrag akzeptieren

Beurteilung der Regelinduktion

Regelinduktion := automatische Ableitung eines Entscheidungsbaums oder von Regeln aus vollständigen oder unvollständigen Elementarregeln

Vorteil

- weniger redundante
- weniger widersprüchliche
- weniger ineffiziente Regeln

Nachteil

- Syntax oft beschränkt
(Zahl der Bedingungen, Werte, Folgerungen)



Direkte Eingabe eines Entscheidungsbaums, wenn der Experte seine Regeln explizit formulieren kann.
Induktion, wenn das Wissen unstrukturiert vorliegt

Motivation

- ✓ **Expertensysteme** durch Regeln ergänzen, die sich nicht einfach durch Expertenbefragung erwerben lassen
- ⇒ Aus **Datenbanken** Klassifikations- und Vorhersagemuster gewinnen



Data Mining durch Regelinduktion

① *Datenbankabfragen*

Welche **Daten** passen zum Abfragemuster ?

≠

② *Data Mining*

Welches **Muster** passt zu den **Daten** ?



Sonderfall: Welcher *Entscheidungsbaum* passt zu den gegebenen Daten?

7.18 Methode im Vergleich

Kriterium	AHP	Optimierung	OLAP	Regelbasierte Systeme	Induktion	Neuronale Netze	Regression
Methode breit anwendbar	∅	–	+	∅	∅ ¹	∅	–
Automatisierungsgrad	–	+	–	∅	+	+	+
Ergebnis genau	–	+	+	∅	+	+	+
Unabhäng. Var. gewichtbar	–	–	–	–	∅ ²	–	+
Lösungsweg begründbar	∅	–	∅	+	+ ³	–	∅
Methode plausibel	+	∅	+	+	∅	–	∅
Ergebnis einbettbar	–	+	∅	∅	+	∅	+
Entwicklungsaufwand	+	+	–	–	+	∅	∅
Rechnerbelastung	+	+	–	∅	∅	–	+

Kriterien

- 1 Klassifikation und Vorhersage
- 2 Ordnung, aber keine Gewichtung
- 3 Entscheidungsbäume können durch ihre Grösse allerdings schwer verständlich werden

Gegeben

Eine Stichprobe *Bonitätsklassifikation* beschreibt eine Lernmenge bereits bearbeiteter Kreditgesuche

Ziel

Induzieren Sie einen Entscheidungsbaum, der jedes neue Kreditgesuch annimmt oder ablehnt

Eingaben

a) Unabhängige Variablen

Geschlecht

Zivilstand

Kinderzahl

Beschäftigung

Wohneigentum

Einkommen

Ausgaben

Ersparnisse

b) Abhängige Variable

Annahme bzw. Ablehnung eines Kreditgesuchs ("Annahme": Kreditnehmer, die alle Raten bezahlt haben, "Ablehnung": übrige)

Aufgabe

a) Lösen Sie die Aufgabe intuitiv.

b) Verwenden Sie ein Data Mining-Werkzeug.

Klassifikation und Regression

	<i>Abhängige Variable</i>	<i>Beispiel</i>
Klassifikationsbaum	diskret	[“Raten bezahlt”, “Raten nicht bezahlt”]
Regressionsbaum (Vorhersage)	kontinuierlich	Zahl der bezahlten Raten

7.4 Entscheidungsbauminduktion - Überblick

Problem



Lernmenge

<i>Unabhängige</i>	<i>Abhängige</i>
...	...
...	...

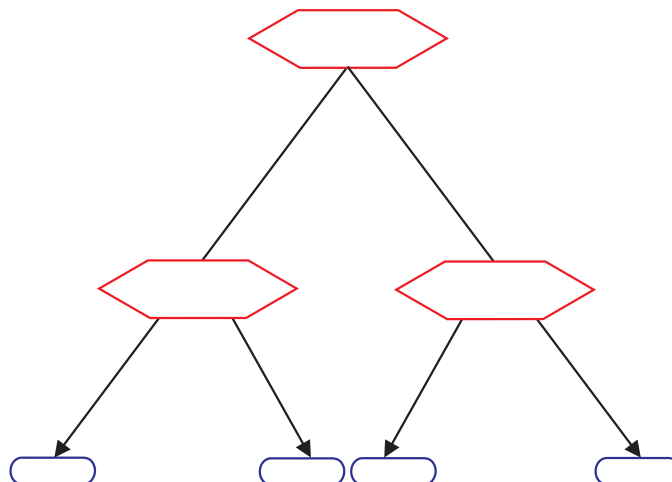


Induktion

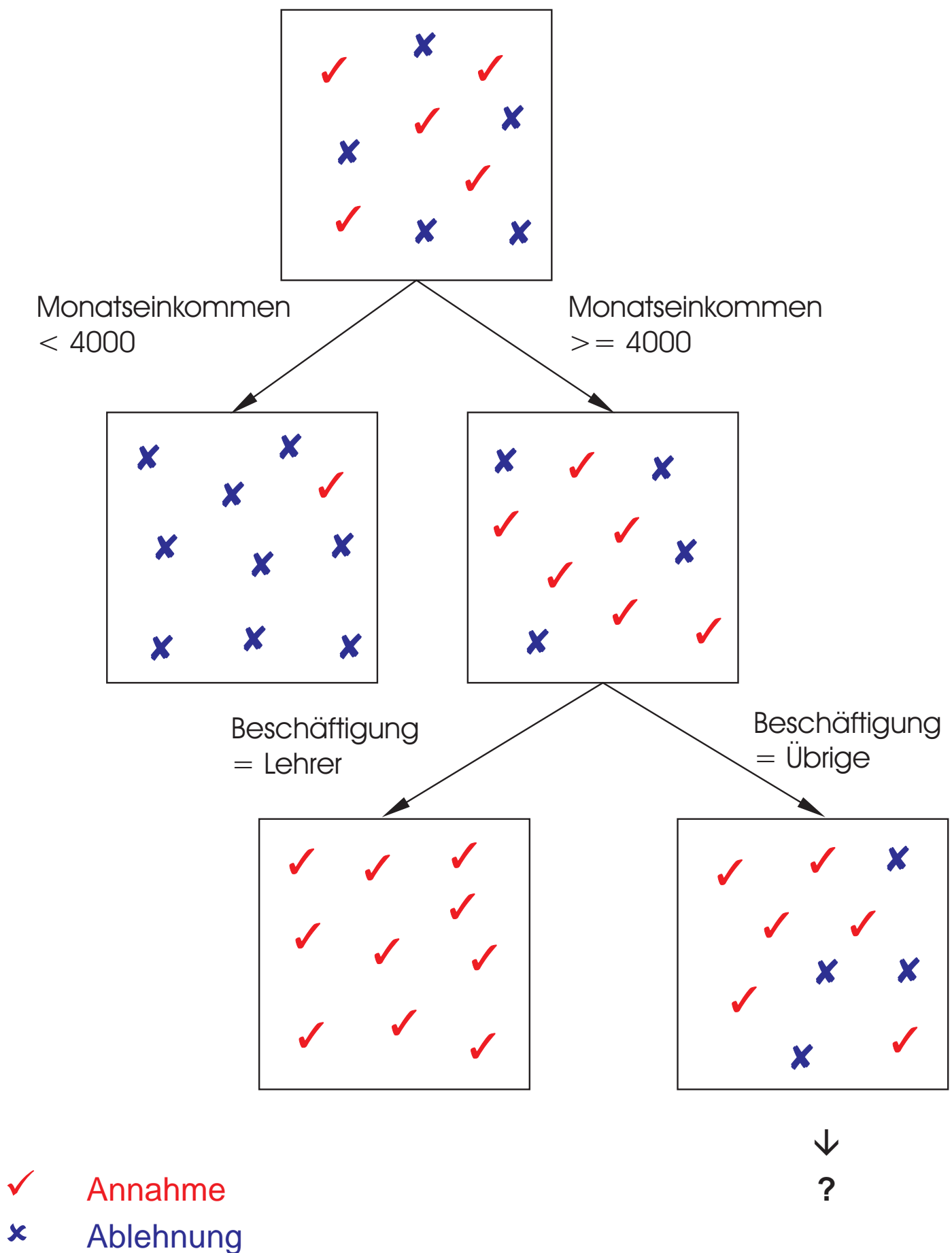
(aus Stichprobe
oder Grund-
gesamtheit)



*Entscheidungs-
baum*



7.5 Von heterogenen zu homogenen Klassen



Klassifikation mit Entscheidungsbäumen

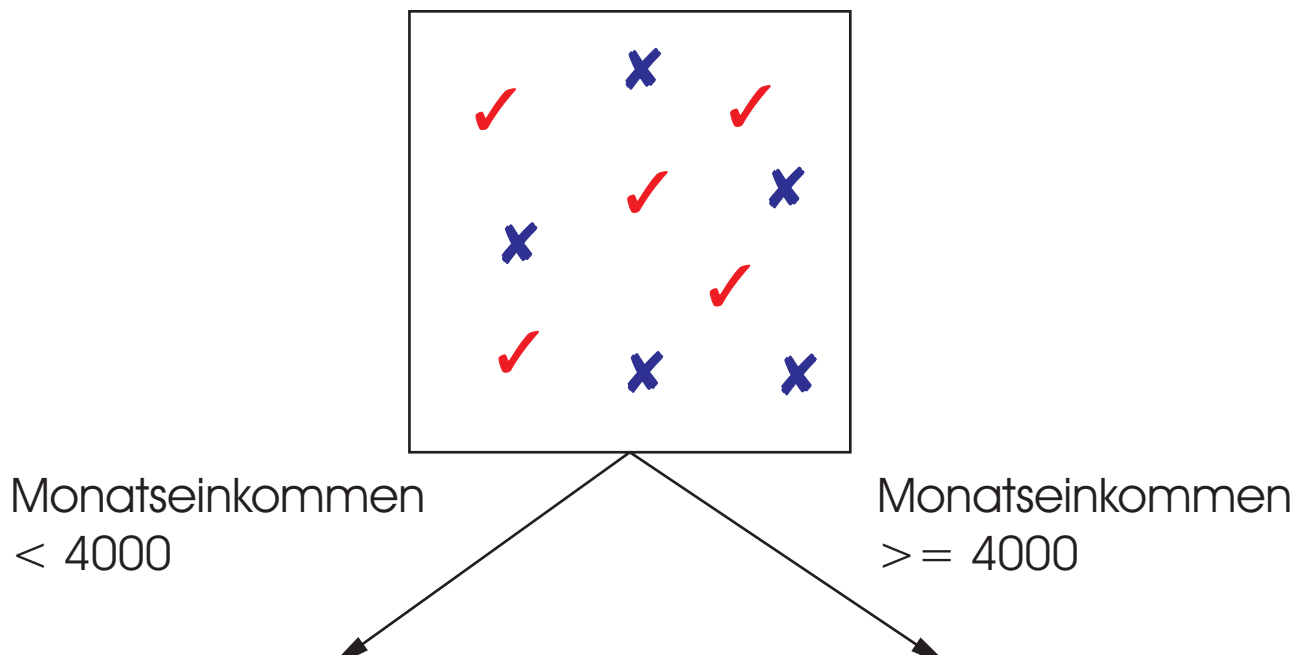
Klassifiziere baumtraversierend **rekursiv**

Beantworte die Frage des **Wurzel**knotens

BIS Du an einem **Blatt**knoten angelangt bist
Folge dem **Ast**, der auf Deine Antwort zutrifft
Beantworte die Frage des **nächsten** Knotens



Fragen eines Entscheidungsbaums bestimmen



- Weshalb *Monatseinkommen* statt *Geschlecht*, etc. ?
⇒ Welche **Attribute** trennen in welcher **Reihenfolge** am besten?
- Warum *Monatseinkommen < 4000* statt *< 3000* ?
⇒ Welche **Werte** trennen am besten ?



Wie messe ich **Homogenität** bzw. **Heterogenität**?

Reihenfolge der Entscheidungsknoten

Ein Entscheidungsbaum-Algorithmus ...

- ✓ fragt zuerst nach dem ›Attribut, das am **meisten** zur Klassifikation beiträgt und ...
- ✓ stellt dann in der **Reihenfolge** des Klassifikationsbeitrags weitere Fragen, ...
- ✓ bis die Klassifikationsgüte **genügt**

Enumerativer Ansatz

- Erzeuge alle möglichen Entscheidungsbäume
- Wähle den Entscheidungsbaum mit den wenigsten Fragen



Dieser Ansatz findet den am besten klassifizierenden Baum, ist aber ineffizient

►Heuristischer Ansatz

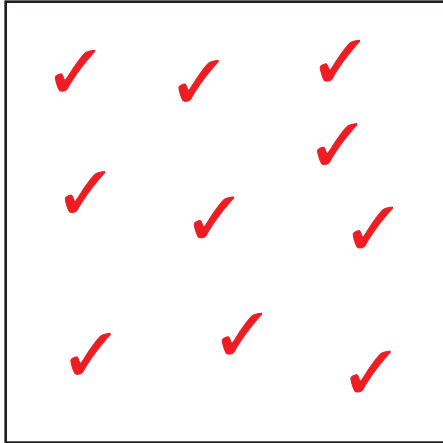
- Beginne mit einer zufälligen Stichprobe von Elementarregeln
- Generiere den Entscheidungsbaum rekursiv
- Wähle die Testattribute und -werte jedes rekursiven Teilbaums informationstheoretisch



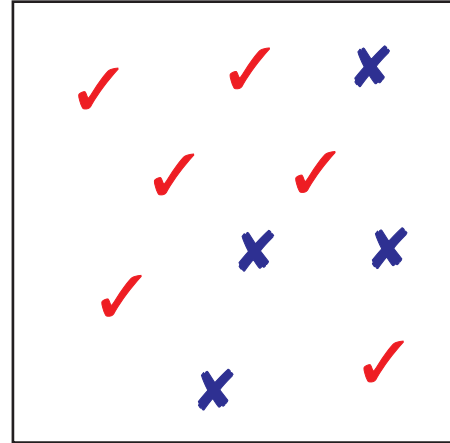
Dieser Ansatz ist zwar effizient, findet aber nicht immer den am besten klassifizierenden Baum

Testattribute verringern die Heterogenität

Homogene Klasse



Heterogene Klasse



Induktionsziel

- Attribute und Werte, welche die *Heterogenität* minimieren
- Höhe des Entscheidungsbaums minimieren



Ein mögliches Mass der Heterogenität (Unordnung) eines Entscheidungsbaums ist die informationstheoretische Entropie



Operationalisierung des Entropiebegriffs

Anwendungsentwicklung

Grundlagen

- ✓ Regelinduktion und Regelbasierte Systeme 2
- ✓ Regelinduktion und Data Mining 18

Entwicklung mit *XpertRule Profiler*

- ⇒ Bonitätsklassifikation   30

Theorie

- Induktionsalgorithmus ID3 48

Endbenutzer können ...

- die Daten in einem DBMS oder Tabellenkalkulationspakets **aufbereiten**
- eine **ODBC-Verbindung** zwischen der Datenquelle und *XpertRule Profiler* einrichten
- eine Lernmenge **definieren**
- einen binären Entscheidungsbaum aus der Lernstichprobe **induzieren** lassen

Programmierer können ...

- den induzierten Entscheidungsbaum in **C- oder SQL-Code** transformieren lassen
- diesen Code in ein **eigenes Programm** integrieren

BONITÄTSKLASSIFIKATION - Entwicklungsphasen

① Lernmenge **definieren**

Lernmenge *Bonitätsklassifikation* als Datenbanktabelle definieren

② **ODBC**-Verbindung einrichten

Induktionssoftware *XpertRule Profiler* über die ODBC-Schnittstelle mit der Datenquelle *Bonitätsklassifikation* verbinden

③ Daten nach einer ersten **Analyse** anpassen

- Unabhängige und Abhängige wählen (nicht zu viele Attribute!)
- Datentypen anpassen (z.B. ›kontinuierlich → ›diskret)

④ **Induktionsparameter** bestimmen

z.B. Signifikanzniveau

⑤ **Induzieren**

Profiler kommuniziert mit der ODBC-Datenquelle über ›SQL und erstellt einen Entscheidungsbaum, der das abhängige Attribut mit den unabhängigen Attributen klassifiziert

⑥ Entscheidungsbaum ev. an **Testdaten** validieren

⑦ Entscheidungsbaum auf **Produktionsdaten** anwenden ev. Entscheidungsbaum als C- oder SQL-Code integrieren

7.6 📌 ① Lernmenge

Geschlecht	Zivilstand	Kinder	Beschäftigung	Wohneigentum	Monatliches Einkommen	Ersparnisse	Kreditgesuch angenommen
W	V	1	ungelernt	N	2530	N	nein
M	V	0	ungelernt	N	2301	J	nein
M	V	1	angelernt	N	3484	N	nein
W	V	0	unbekannt	N	2188	N	nein
M	W	1	FacharbeiterIn	J	2071	J	nein
M	V	1	angelernt	J	2650	J	nein
M	V	0	unbekannt	N	2418	J	nein
M	V	2	FacharbeiterIn	J	2508	J	nein
W	W	0	FacharbeiterIn	N	2788	N	nein
W	V	2	angelernt	N	2778	N	nein
W	V	0	ungelernt	N	2599	N	nein
M	V	0	unbekannt	N	2166	N	nein
M	W	1	ungelernt	N	2092	J	nein
W	V	1	FacharbeiterIn	N	2444	N	nein
M	V	1	unbekannt	J	2844	N	nein
W	V	2	unbekannt	J	2448	N	nein
W	V	1	Verwaltungsang	N	4458	J	nein
W	V	0	Führungskraft	J	5896	N	nein
W	V	1	Führungskraft	N	3961	N	nein
W	V	1	ungelernt	N	2391	N	nein
M	V	0	unbekannt	J	2343	N	nein
M	A	1	angelernt	J	2501	N	nein

...

7.7 ③ Vorläufige Datenanalyse

Name	Type	Values / Range	Usage
Kinder	Continuous	0 to 3	Attribute
Monatliches_Einkommen	Continuous	2004 to 8975	Attribute
Geschlecht	Discrete	2	Attribute
Zivilstand	Discrete	3	Attribute
Beschaeftigung	Discrete	7	Attribute
Wohneigentum	Discrete	2	Attribute
Ersparnisse	Discrete	2	Attribute
Kreditgesuch_angenommen	Discrete	2	Outcome

Field Information ✕

Name: abhängige Variable

DB Column Name:

Description:

Field Type:

Number values: 4

Value	Frequency
1	193
2	148
0	52
3	11

Decimal places:

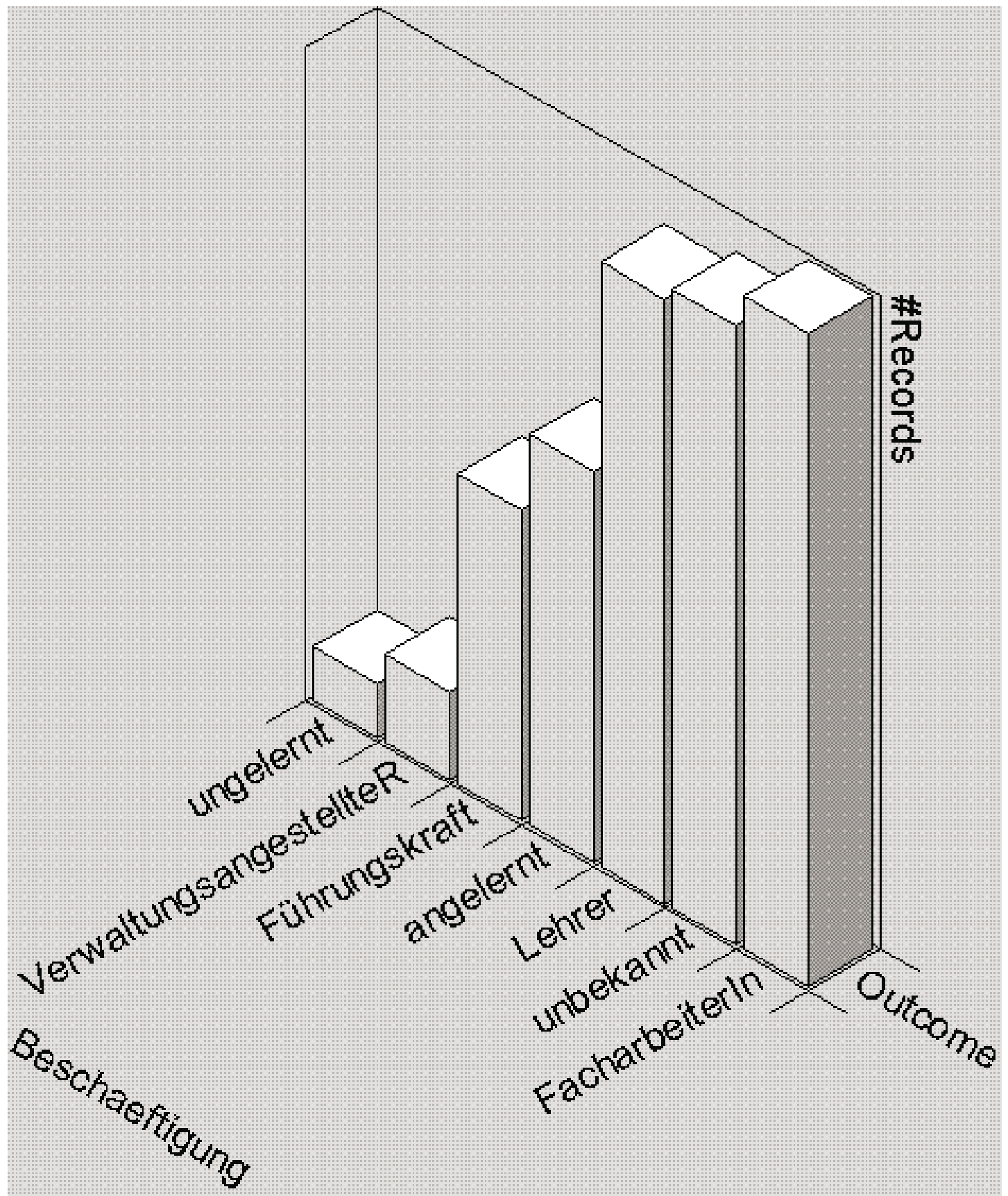
Minimum: 0

Maximum: 3

Average: 1.3

Standard Deviation: 0.7

③ Vorläufige Datenanalyse



④ Induktionsparameter

Interactive Induction X

Attribute	Entropy	Chi Sqr	S/
Monatsinkommen	191.62	155.96	▲
Beschaeftigung	211.12	115.74	
Ersparnisse	242.26	53.63	
Wohneigentum	256.95	40.35	
Kinder	262.31	29.78	▼

Split criteria: Chi Square ▼ Edit Split...

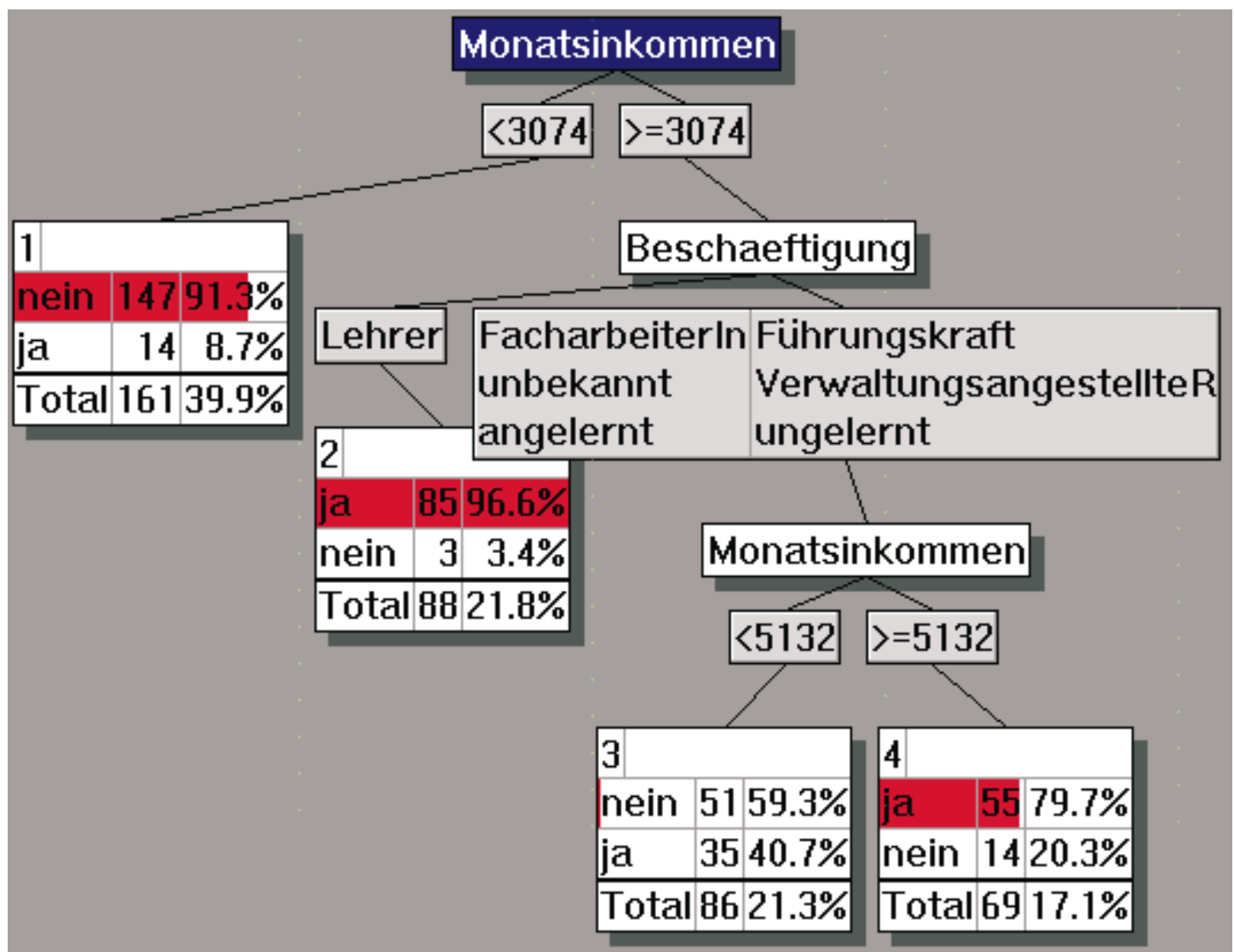
Min examples in branch:

Die ›Entropie oder der ›Chi Quadrat-Unabhängigkeitstest wählen das nächste Attribut des Entscheidungsbaums (split criterion). Je kleiner die Entropie oder je grösser der sogenannte Chi Quadrat-Abstand, desto grösser die Klassifikationsgüte des Attributs.

Weniger Entscheidungsfragen durch höheres ...

- Beispielminimum pro Ast
- Signifikanzniveau
(›Chi Quadrat - oder ›F-Test)

7.8 ⑤ Einfacher binärer Entscheidungsbaum



Bereits ein *Teil* der möglichen Testattribute klassifiziert gut !

Die Länge des roten Balkens ist proportional zur Überlegenheit der Entscheidungsbaum-Vorhersage über eine naive Vorhersage

Knotennummer

#

Grössere Klasse

ja (nein) p

$p / (p + q) \cdot 100 \%$

Kleinere Klasse

nein (ja) q

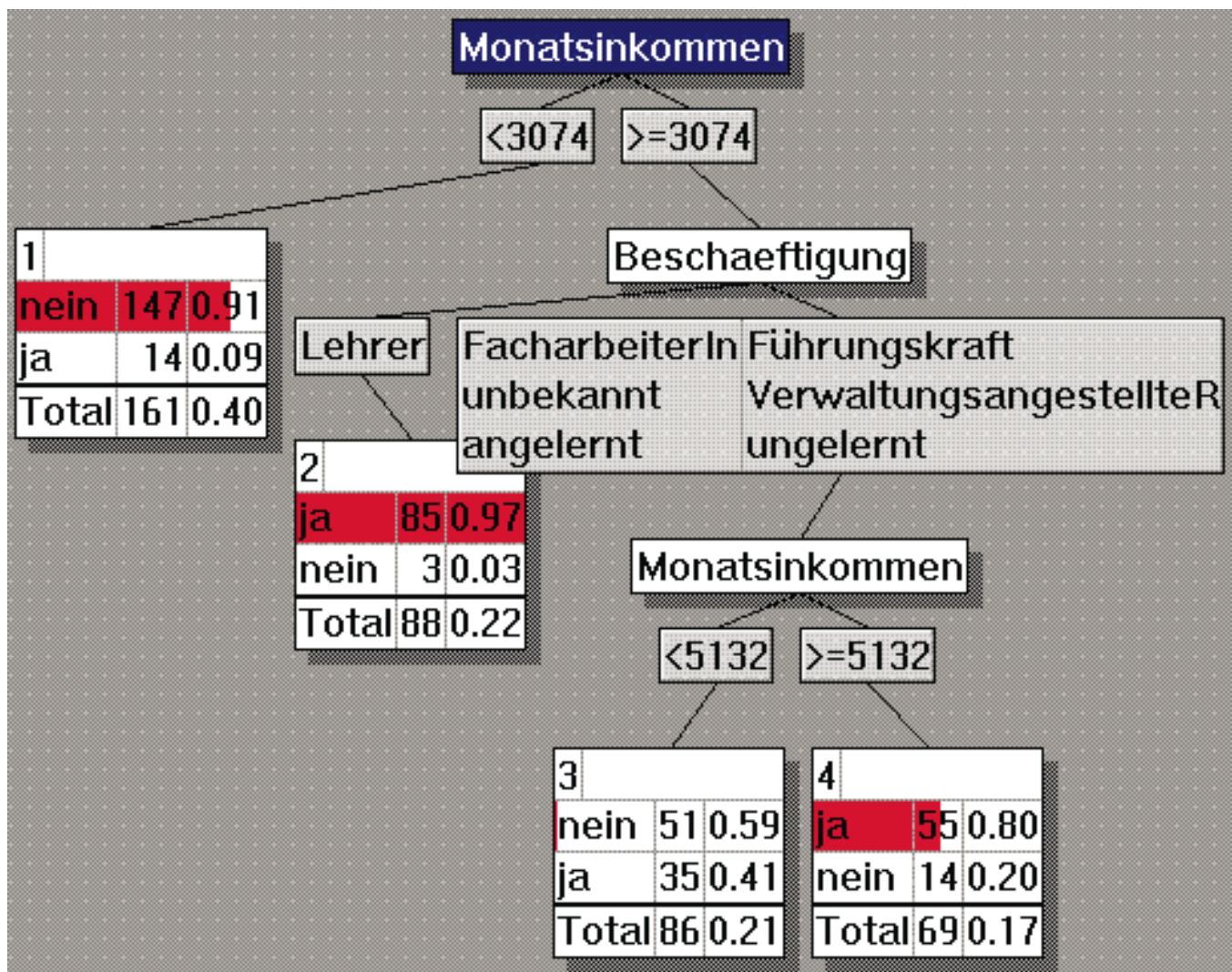
$q / (p + q) \cdot 100 \%$

Total

Total p + q

Anteil an der Stichprobe %

📌 Von der Baum- zur Textnotation



WENN *Monatseinkommen* < 3074

DANN *Ablehnung* mit einer Wahrscheinlichkeit von 0.91

SONST

WENN *Beschaeftigung* = Lehrer

DANN *Annahme* mit einer Wahrscheinlichkeit von 0.97

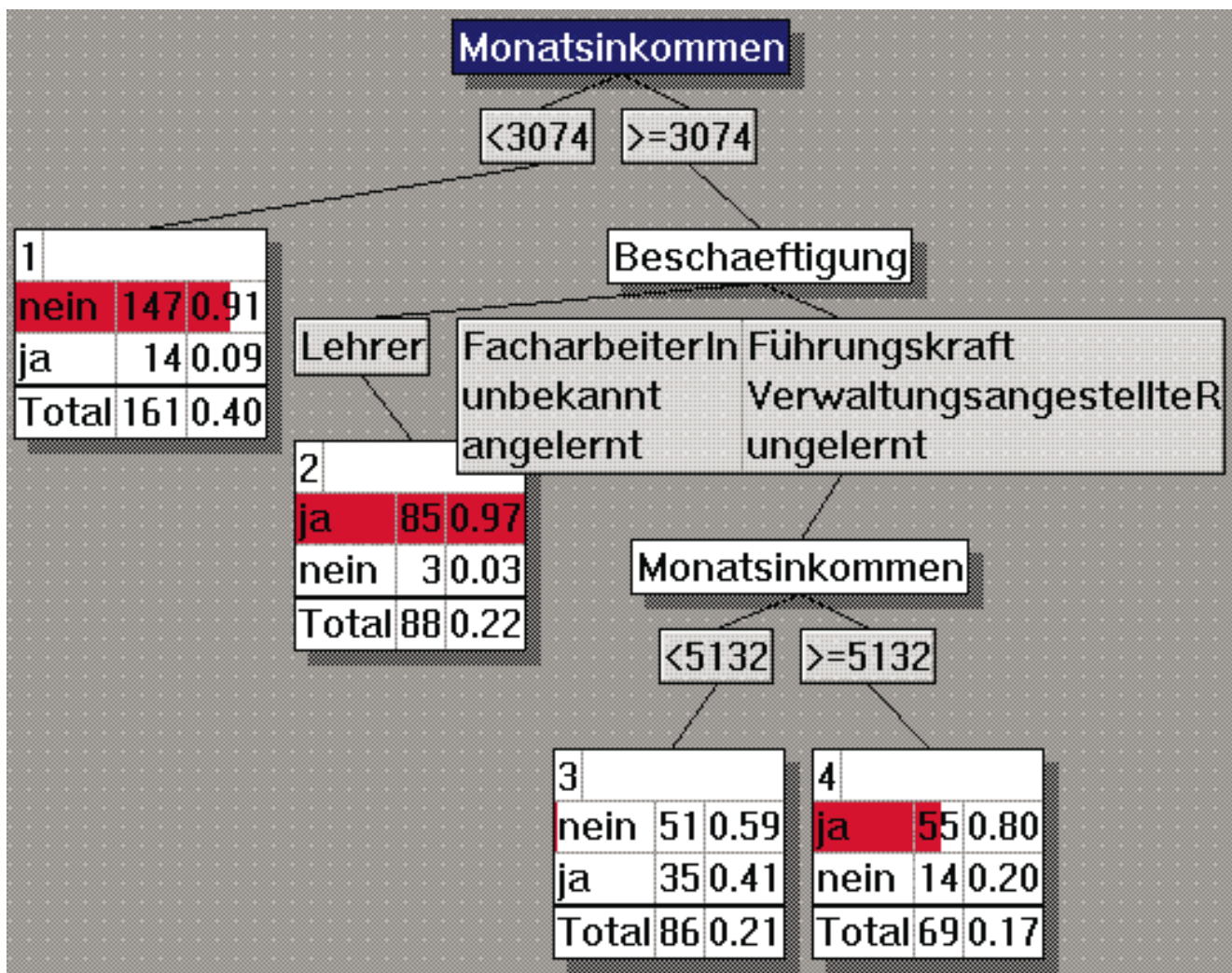
SONST

WENN *Monatseinkommen* < 5132

DANN *Ablehnung* mit einer Wahrscheinlichkeit von 0.59

SONST *Annahme* mit einer Wahrscheinlichkeit von 0.80

† Lässt sich der induzierte Baum verallgemeinern?



Wahrscheinlichkeiten (z.B. 0.91)
statt Prozentzahlen (z.B. 91.3%)



Lässt sich der Entscheidungsbaum auf
neue Stichproben **verallgemeinern** ?

Ordnen Sie die Fragen (Attribute) nach Ihrer Wichtigkeit !

Klassifikationsgüte

Trefferquote¹
(accuracy)



Beeinflusst von **Modellwahl** und **Induktionsparametern**, d.h.

- Zahl und Wahl der Attribute
- Minimaler Beispielszahl pro Ast
- Signifikanzniveau
- ...


¹ Nennen Sie Validitätsmasse Ihnen bekannter statistischer Verfahren, zum Beispiel der linearen Regression

Sequentielle und parallele Verarbeitung



<i>Regelinduktion</i>	<i>Eingabe/Ausgabe</i>	<i>Verarbeitung</i>
<i>Sequentiell</i>	Client	Client
<i>Parallel</i>	Client	Server

Zuerst werden Sie die bereits bekannte Lösung des Bonitätsproblems *nachvollziehen*. Dann erhalten Sie die Gelegenheit, Profiler *selbständig* kennen zu lernen.

1. XpertRule Profiler geleitet kennen lernen

Jede Zeile der Tabelle von  Bonitätsdaten.mdb enthält die Daten des Antragsformulars eines Kreditnehmers mit bereits bekannter Bonität. Kreditwürdig ist (ex post), wer alle Raten bezahlt hat. Ein Induktionsmodell soll die Spaltenattribute so verarbeiten, dass sie die Kreditwürdigkeit künftiger Antragsteller möglichst gut klassifizieren.

1.1 Datensicht

- Laden Sie  [Bonitätsdaten.mdb](#) in MS Access.
- Öffnen Sie die Tabelle KONSUMKREDIT.
 - Aus welchen Attributen besteht die Tabelle?
 - Welches sind die Datentypen der Attribute?
 - Welches Skalenniveau verwenden die Attribute?
- Starten Sie  XpertRule Profiler. Nach dem Start verlangt Profiler eine Legitimation. Klicken Sie stattdessen “Demo”. Sie können dann Profiler mit dem Beispiel und der Aufgabe dieses Kapitels benutzen. Allerdings können Sie ihre Ergebnisse vor dem Verlassen von Profiler nicht speichern. Sie erhalten auf drei Arten Hilfe:
 - Wenn Sie den Cursor auf ein Toolbar-Symbol positionieren, erscheint eine Kurzbeschreibung.
 - Ausführliche Hilfe zu Symbolleiste und Menü finden Sie unter dem Menüpunkt *Help/Inhalt/Reference*.
 - *Help/Contents* enthält unter dem Titel “Tutorial” eine ausführliche Beispielsitzung. Konsultieren Sie diese Hilfe, wenn der folgende Aufgabentext nicht genügt.

a) File/New führt zur Definition der Datenquelle (engl. data set).

b) Profiler greift über ODBC auf die Tabelle KONSUMKREDIT zu. Open Data Base Connectivity ist ein Industriestandard, der die Schnittstelle zwischen einem Anwendersystem (hier Profiler) und einem SQL-Datenbanksystem (hier MS Access) normiert. Im nächsten Schritt werden Sie deshalb eine ODBC-Verbindung (engl. data connection) zur Datenquelle KONSUMKREDIT einrichten. Klicken Sie dazu *Add* und füllen Sie das Formular aus:

- Name, zum Beispiel “Konsumkredit-Daten”
- Description, zum Beispiel “Lernmenge für eine Bonitätsklassifikation”
- Data Type: Case based
- Driver: 32bit ODBC CAF Driver (für Windows 95 und NT)
- User name: Admin
- Password: (auslassen)
- Edit Connection: Ein Klick führt zu einem Unterformular
- Data Source: Wählen Sie den Menüpunkt “MS Access 97-Datenbank” bzw. “Microsoft Access-Datenbank”
- Table Name: Wählen Sie zuerst Bonitätsdaten.mdb und dann KONSUMKREDIT aus dem Menü “Table Name”.

c) New Analysis: OK (alle Felder in die Analyse aufnehmen)

d) Profiler erstellt eine Liste der Attribute mit Name, Typ (discrete oder continuous, für unser Beispiel nicht von Belang), Wertebereich und Variablenart (“Attribute“ ist eine unabhängige und “Outcome“ eine abhängige Variable).

Klicken Sie doppelt auf einen Attributnamen und beschreiben Sie die Bedeutungen der Attribute und Werte, falls der Name nicht genügt. Beachten Sie ausserdem die statistischen Angaben am Ende des Dialogs.

e) Klicken Sie doppelt auf den Namen der abhängigen Variable und notieren Sie im Feld “Description”, dass es sich um die abhängige Variable handelt. Lassen Sie das gewählte Attribut markiert und

definieren Sie die abhängige Variable mit *Options/Field Usage/Outcome*. Beachten Sie die Änderung in der Spalte “Usage”.

- f) Verringern Sie die Zahl der Kommastellen der kontinuierlichen Attribute.
- g) Der Menüpunkt *Options/Ranking* ordnet die unabhängigen Variablen nach ihrem Klassifikationsbeitrag. Wenn die Zahl der Attribute gross ist, könnte man hier aus Effizienzgründen eine Vorauswahl treffen.

1.2 Entscheidungsbaum-Sicht

- h) Wechseln Sie mit *View/Decision Tree* in die Entscheidungsbaum-Sicht. Welche Bedeutung haben die gezeigten Knotenhäufigkeiten?
- i) Erzeugen Sie mit *Options/Induce Tree* einen Klassifikationsbaum. Ein Popup-Menü verlangt die folgenden Parameter:
 - Mindestzahl von Beispielen pro Klassifikationsknoten
 - Signifikanzniveau für jeden Ast
 - Klassifikationskriterium (Chi-Quadrat oder Entropie).

Belassen Sie die Voreinstellungen.

- j) Unterscheiden Sie Wurzel, Zwischenknoten, Blätter und Kanten.
- k) Interpretieren Sie die Knoten und Kanten inhaltlich und konvertieren Sie das Induktionsergebnis manuell in eine Regelfolge.

2. XpertRule Profiler selbständig erkunden

Beantworten Sie die folgenden Fragen (Konsultieren Sie wenn nötig die Hilfefunktion):

- l) Experimentieren Sie mit den unter *Options/Display* angebotenen Baumformaten.
- m) Experimentieren Sie mit verschiedenen Werten für die Induktionsparameter “Mindestzahl von Beispielen pro Klassifikationsknoten”

und “Signifikanzniveau” (vgl. i). Der Menüpunkt *Datei/Print* erlaubt die Ausgabe eines Baums auf den Drucker oder in eine Datei.

- Was bewirken geringere Anforderungen an die Beispielszahl und das Signifikanzniveau?
- Wie verhalten sich Anforderungen und Verallgemeinerungsfähigkeit?

n) Inspizieren Sie mit *Options/Inspect* die Entscheidung des Induktionsalgorithmus für ausgewählte Knoten. Vergessen Sie nicht, vorher einen Knoten zu markieren.

- Welches Klassifikationsattribut wird jeweils gewählt?
- Welche Beziehung besteht zwischen den Testkriterien Chi-Quadrat und Entropie?
- Was bedeutet die vierte Spalte? Experimentieren Sie zum Beispiel mit unterschiedlichen Besetzungszahlen.

o) Welche anwendungspraktischen Schlüsse ziehen Sie aus den in m) und n) angesprochenen Freiheitsgraden bei der Modellspezifikation?

Funktionalität von *XpertRule Profiler*

Eingabe

- ✓ Daten aus einer SQL-Datenbank importieren
- ✓ Variablen transformieren
- ✓ Daten vorläufig analysieren

Verarbeitung

- ✓ Entscheidungsbaum induzieren
- ✓ Klassifikationsgüte berechnen

Ausgabe

- ✓ Datenanalyse präsentieren
- ✓ Entscheidungsbaum präsentieren
- ✓ C- oder SQL-Code erzeugen

Theorie

Grundlagen

- ✓ Regelinduktion und Regelbasiertes Systeme 2
- ✓ Regelinduktion und Data Mining 18

Entwicklung mit *XpertRule Profiler*

- ✓ Bonitätsklassifikation 📌 🖱️ 30

Theorie

- ⇒ Induktionsalgorithmus ID3 48

Theoretischer Hintergrund

Entscheidungsbäume lassen sich **manuell** oder **automatisch** (induktiv) erstellen



Verbreitete Induktionsalgorithmen

CART (1984)¹ erzeugt binäre Entscheidungsbäume

CHAID (1976)² erzeugt mehrfache Entscheidungsbäume

ID3 (1986) und C4.5 (1988)

...



ID3 Iterative **D**ichotomiser **3**

von J. Ross Quinlan (1986)

¹ Classification And Regression Trees für kontinuierliche Unabhängige

² Chi Square Automatic Interaction Detection für diskrete Unabhängige

Gegeben

- **Lernmenge** von Wetterbeobachtungen, von denen jede genau einer von zwei **Klassen** angehört. Die erste Klasse enthält Tage mit “positivem” Wetter (+), die andere mit “negativem” Wetter (-).
- **Attribute**, welche die Klassenzugehörigkeit beeinflussen:
Wettercharakter, Temperatur, Feuchtigkeit, Wind

Gesucht

Vorschrift, die einen Entscheidungsbaum ableitet, der neue Wetterbeobachtungen aufgrund der vier **Attribute** einer der zwei Klassen zuordnet

¹ nach Quinlan 1986

7.10 ID3 - Lernmenge WETTER

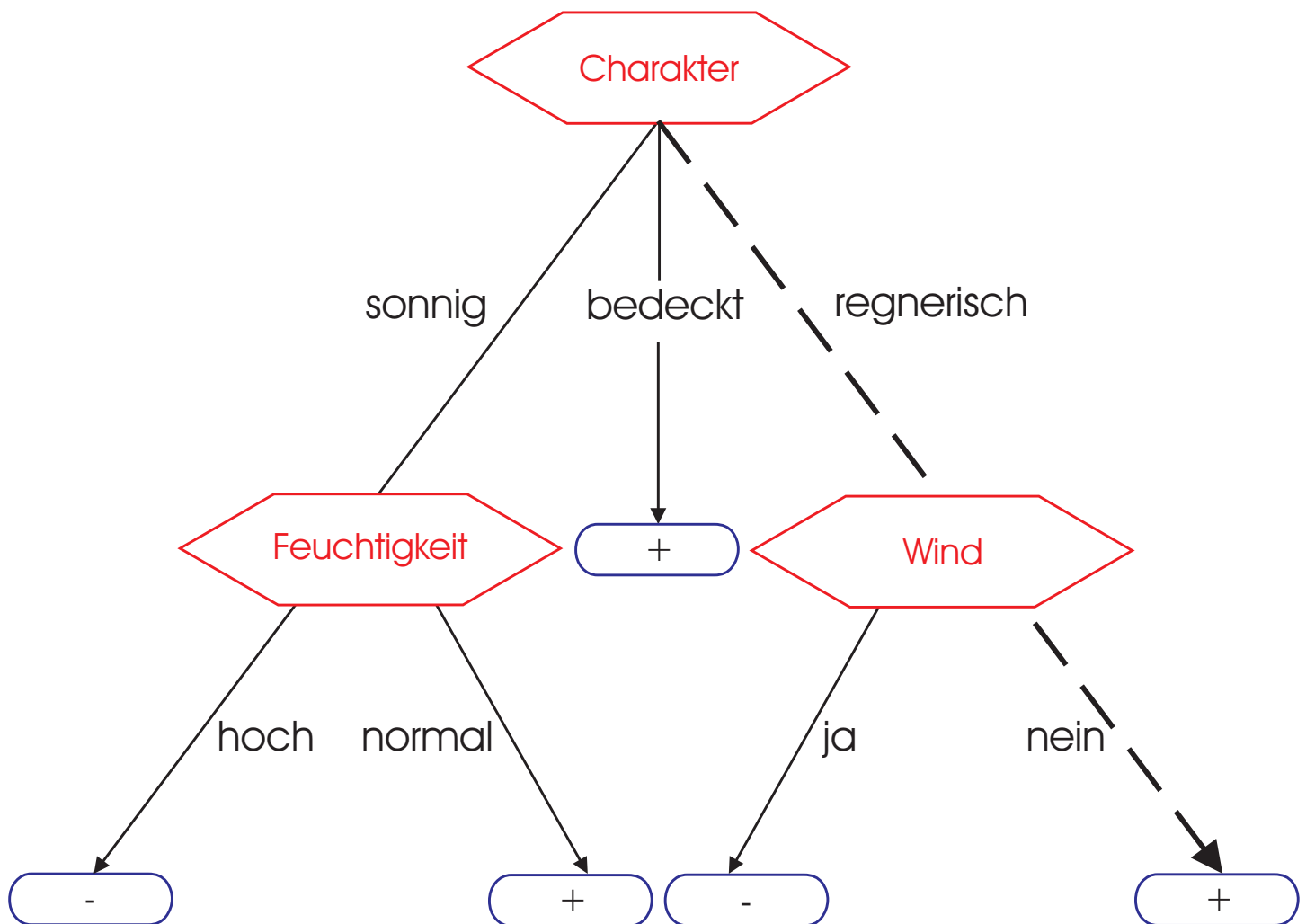
Lernmenge von Wetterbeobachtungen

<i>Element</i>	<i>Charakter</i>	<i>Temperatur</i>	<i>Feuchtigkeit</i>	<i>Wind</i>	<i>Wetter</i>
1	sonnig	heiss	hoch	nein	-
2	sonnig	heiss	hoch	ja	-
3	bedeckt	heiss	hoch	nein	+
4	regnerisch	mild	hoch	nein	+
5	regnerisch	kühl	normal	nein	+
6	regnerisch	kühl	normal	ja	-
7	bedeckt	kühl	normal	ja	+
8	sonnig	mild	hoch	nein	-
9	sonnig	kühl	normal	nein	+
10	regnerisch	mild	normal	nein	+
11	sonnig	mild	normal	ja	+
12	bedeckt	mild	hoch	ja	+
13	bedeckt	heiss	normal	nein	+
14	regnerisch	mild	hoch	ja	-

unabhängige Variablen

abhängige Variable

7.11 ID3 - Induzierter Entscheidungsbaum WETTER



Das Ergebnis des Induktionsalgorithmus teilt die Stichprobenelemente mit nur drei der vier **Attribute** den gegebenen Klassen + und - zu.

Klassifiziere mit dem Entscheidungsbaum den vierten Datensatz :

Element	Charakter	Temperatur	Feuchtigkeit	Wind	Klasse
4	regnerisch	mild	hoch	nein	?



⇒ Wie induziere ich einen Baum mit möglichst *wenig* Testattributen ?

⇒ Kann ich von der Stichprobe auf die Grundgesamtheit *schliessen* ?

Lerne einen Entscheidungsbaum

- Berechne für jedes Attribut, wie *gut* es allein die Elemente der Lernmenge klassifiziert
- Klassifiziere mit dem *besten* Attribut
- *Wiederhole* für jeden so entstandenen Teilbaum die ersten beiden Schritte
- Brich diesen rekursiven Prozess ab, sobald er ein bestimmtes *Abbruchkriterium* erfüllt



- ① Entscheidungsgehalt
- ② Entropie
- ③ Klassifikationsgewinn
- ④ Erwartungswert



Induktionsalgorithmus
mit Abbruchkriterium

Frage

Wie kann ich Information quantifizieren ?

Antwort

Durch ihren *Informationsgehalt* (*Entscheidungsgehalt*)

Motivation

Je mehr Binärentscheidungen die Darstellung einer Information benötigt, desto grösser der Gehalt dieser Information

Berechnung

4 Bit zur Darstellung von 10: "1010" ($10 = 2^3 + 2^1$),
genauer: 3.32 Bit ($10 = 2^{3.32}$)

$$3.32 = \text{Logarithmus von 10 zur Basis 2 } (\log_2 10)^1$$



Informationsgehalt :=

\log_2 (Zahl der Binärentscheidungen zur Informationsdarstellung)

Der Informationsgehalt misst
Information domänenunabhängig

¹ Berechne $\log_2(1/2)$ und $\log_2(1)$ ²

Entropie I eines binären Entscheidungsbaums

$$I(p, n) =$$

$$\begin{aligned} & - \frac{p}{p+n} \cdot \log_2 \left(\frac{p}{p+n} \right) \\ & - \frac{n}{p+n} \cdot \log_2 \left(\frac{n}{p+n} \right) \end{aligned}$$

p positive Klassenhäufigkeit (+)

n negative Klassenhäufigkeit (-)

Die Entropie ("Unordnung") quantifiziert die Information, die eine Klassifikation *ohne* Testattribute benötigen würde

7.12 ID3 - ② Entropie

p	n	$p/(p+n)$	$\text{ld}(p/(p+n))$	$n/(p+n)$	$\text{ld}(n/(p+n))$	$I(p, n)$
7	7	0.5	-1	0.5	-1	1
6	8	0.43	-1.22	0.57	-0.81	0.99
5	9	0.36	-1.49	0.64	-0.64	0.94
4	10	0.29	-1.81	0.71	-0.49	0.86
3	11	0.21	-2.22	0.79	-0.35	0.75
2	12	0.14	-2.81	0.86	-0.22	0.59
1	13	0.07	-3.81	0.93	-0.11	0.37

Je ungleich verteilter p und n,
desto kleiner die Entropie

Je grösser der Informationsgehalt,
desto kleiner die Entropie

Je höher ein Teilbaum,
desto kleiner seine Entropie

7.13 ID3 - Entropie des Gesamtbaums *Wetter*

positive Klassenhäufigkeit (+)

negative Klassenhäufigkeit (-)

$$I(9, 5) =$$

$$- 9 / (9+5) \cdot \log_2 (9 / (9+5))$$

$$- 5 / (9+5) \cdot \log_2 (5 / (9+5))$$

$$= 0.94$$

Element	Charakter	Temperatur	Feuchtigkeit	Wind	Klasse
1	sonnig	heiss	hoch	nein	-
2	sonnig	heiss	hoch	ja	-
3	bedeckt	heiss	hoch	nein	+
4	regnerisch	mild	hoch	nein	+
5	regnerisch	kühl	normal	nein	+
6	regnerisch	kühl	normal	ja	-
7	bedeckt	kühl	normal	ja	+
8	sonnig	mild	hoch	nein	-
9	sonnig	kühl	normal	nein	+
10	regnerisch	mild	normal	nein	+
11	sonnig	mild	normal	ja	+
12	bedeckt	mild	hoch	ja	+
13	bedeckt	heiss	normal	nein	+
14	regnerisch	mild	hoch	ja	-

ID3 - Entropie $I(p_i, n_i)$ des Teilbaums mit Wurzel i

$I_{\text{sonnig}}(2, 3)$ des Teilbaums mit der Wurzel *sonnig*

$$\begin{aligned}
 & - 2 / 5 \quad \log_2 (2 / 5) \\
 & - 3 / 5 \quad \log_2 (3 / 5) \\
 & = 0.971
 \end{aligned}$$

Element	Charakter	Temperatur	Feuchtigkeit	Wind	Klasse
1	sonnig	heiss	hoch	nein	-
2	sonnig	heiss	hoch	ja	-
3	bedeckt	heiss	hoch	nein	+
4	regnerisch	mild	hoch	nein	+
5	regnerisch	kühl	normal	nein	+
6	regnerisch	kühl	normal	ja	-
7	bedeckt	kühl	normal	ja	+
8	sonnig	mild	hoch	nein	-
9	sonnig	kühl	normal	nein	+
10	regnerisch	mild	normal	nein	+
11	sonnig	mild	normal	ja	+
12	bedeckt	mild	hoch	ja	+
13	bedeckt	heiss	normal	nein	+
14	regnerisch	mild	hoch	ja	-

Je weiter **oben** im Entscheidungsbaum, desto kleiner die Entropie des aufgespannten Teilbaums

ID3 - ③ Erwartungswert

Attribut A mit den Werten a_1, a_2, \dots, a_v teilt die Knotenelemente in v Partitionen

Erwartungswert E_A der für die Klassifikation mit Wurzelattribut A erforderlichen Information

:= gewichtetes Mittel der Entropien der durch a_i aufgespannten Teilbäume

$$E_A := \sum_{i=1}^v \frac{p_i + n_i}{p + n} \cdot I(p_i, n_i)$$

p positive (+) Klassenhäufigkeit

n negative (-) Klassenhäufigkeit

$I(p_i, n_i)$ Entropie des mit a_i erzeugten Teilbaums i

Der Erwartungswert misst die Information, die eine Klassifizierung mit Testattribut A benötigt

ID3 - Erwartungswert des Attributs Charakter

$$E_{\text{Charakter}} =$$

$$\sum_{i=1}^3 \frac{p_i + n_i}{p + n} \cdot I(p_i, n_i) =$$

$$\frac{5}{14} I(p_1, n_1) + \frac{4}{14} I(p_2, n_2) + \frac{5}{14} I(p_3, n_3) = 0.694$$

✓
?
?

Element	Charakter	Temperatur	Feuchtigkeit	Wind	Klasse
1	sonnig	heiss	hoch	nein	-
2	sonnig	heiss	hoch	ja	-
3	bedeckt	heiss	hoch	nein	+
4	regnerisch	mild	hoch	nein	+
5	regnerisch	kühl	normal	nein	+
6	regnerisch	kühl	normal	ja	-
7	bedeckt	kühl	normal	ja	+
8	sonnig	mild	hoch	nein	-
9	sonnig	kühl	normal	nein	+
10	regnerisch	mild	normal	nein	+
11	sonnig	mild	normal	ja	+
12	bedeckt	mild	hoch	ja	+
13	bedeckt	heiss	normal	nein	+
14	regnerisch	mild	hoch	ja	-

Klassifikationsgewinn des Attributs A :=

Entropie des Entscheidungsbaums mit Wurzel A –

Erwartungswert der durch A aufgespannten Teilbäume

:=

$$I_A(p, n) - E_A$$

Der Klassifikationsgewinn misst den Gewinn, den eine Klassifikation *mit* Testattribut A gegenüber einer Klassifikation *ohne* Testattribut erzielt

7.14 ID3 - Klassifikationsgewinn eines Attributs

Klassifikationsgewinn des Attributs *Wettercharakter* =

$$G_{\text{Charakter}} = I_{\text{Charakter}}(9, 14) - E_{\text{Charakter}} = 0.94 - 0.694 = 0.246$$

Teilbäume i	$I(p_i, n_i)$	Berechnung
sonnig ✓	0.971	- $2/5 \log_2(2/5)$ - $3/5 \log_2(3/5)$
bedeckt ?	0	
regnerisch ?	0.971	- $3/5 \log_2(3/5)$ - $2/5 \log_2(2/5)$

Berechnungselemente		Berechnung
② $I(p, n)$ ✓	0.94	- $9/14 \log_2(9/14)$ - $5/14 \log_2(5/14)$
③ $E_{\text{Charakter}}$ ✓	0.694	$5/14 I(p_1, n_1)$ (= $5/14 \cdot 0.971$) + $4/14 I(p_2, n_2)$ (= $4/14 \cdot 0$) + $5/14 I(p_3, n_3)$ (= $5/14 \cdot 0.971$)
④ $G_{\text{Charakter}}?$	0.246	0.94 - 0.694



Klassifikationsbeiträge der übrigen Attribute ?

7.16 ID3 - Höchster Klassifikationsgewinn

Wähle auf jeder Baumstufe das Wurzelattribut mit dem **höchsten** Klassifikationsgewinn¹



$G_{\text{Charakter}}$	= 0.246 ✓
$G_{\text{Temperatur}}$	= 0.029
$G_{\text{Feuchtigkeit}}$	= 0.151
G_{Wind}	= 0.048



Charakter spannt die drei Teilbäume “sonnig”, “bedeckt” und “regnerisch” auf. Für sie berechnet ID3 *wiederum* das Attribut mit dem höchsten Klassifikationsgewinn

Diese *rekursive* Berechnung von Testattributen dauert solange, bis alle Partitionen nur noch entweder positive oder negative Instanzen enthalten oder ein benutzerdefinierter Schwellenwert erreicht ist.

¹ Weil die Entropie des Entscheidungsbaums für alle Attribute gleich gross ist, genügt statt der Maximierung der Klassifikationsbeiträge auch eine Minimierung der Erwartungswerte

7.17 ID3 - Grober Induktionsalgorithmus

Untersuche aus Effizienzgründen jeweils nur eine zufällige Lernmenge der Datengesamtheit



Erstelle aus der Datengesamtheit einen Entscheidungsbaum

Wähle eine zufällige Lernmenge aus der Datengesamtheit

Erstelle daraus einen Entscheidungsbaum

BIS der Entscheidungsbaum die Datengesamtheit korrekt klassifiziert

Füge ausgewählte falsch klassifizierte Elemente zur Lernmenge

Erstelle aus der neuen Lernmenge einen Entscheidungsbaum

Erstelle einen Entscheidungsbaum

FÜR JEDES Attribut

Berechne den Klassifikationsgewinn ←

Erstelle aus der Partition einen Entscheidungsbaum

FALLS die Partition nur positive Instanzen enthält

Markiere den Knoten als Folgerung +

FALLS die Partition nur negative Instanzen enthält

Markiere den Knoten als Folgerung -



Erweiterungen ?

- Zufällige Klassifikationsfehler

Wie gross ist die Wahrscheinlichkeit, dass weitere Testfragen nichts mehr zur Klassifikation beitragen →

- Probabilistische Entscheidungsbäume

Mit welcher Wahrscheinlichkeit kann ein Element einer bestimmten Klasse zugeordnet werden?

- . . .

▸ Chi Quadrat-Wert und ▸ Entropie sind

- ✓ Masse für die Wahrscheinlichkeit, mit der ein
- ✓ Attribut noch zur
- ✓ Klassifikation mit einem
- ✓ Entscheidungsbaum beiträgt

Das Signifikanzniveau setzt einen ...

- ✓ Mindestwert für den
- ✓ ▸ Chi Quadrat-Wert bzw. die ▸ Entropie eines
- ✓ Attributkandidaten

ID3 - Vermeide Test eines irrelevanten Attributs A

Falls A nichts zur Klassifikation beiträgt, gilt:

$$p'_i = p \cdot \frac{p_i + n_i}{p + n} \quad \text{und} \quad n'_i = n \cdot \frac{p_i + n_i}{p + n}$$

d.h. A ergibt keinen Klassifikationsgewinn



Es lässt sich dann zeigen, dass die folgende Statistik annähernd **Chi Quadrat - verteilt** ist

$$\sum_{i=1}^v \frac{(p_i - p'_i)^2}{p'_i} + \frac{(n_i - n'_i)^2}{n'_i}$$

- A Attribut mit den Werten a_1, a_2, \dots, a_v , das die Knotenelemente in v Partitionen teilt
- p positive (+) Klassenhäufigkeit
 n negative (-) Klassenhäufigkeit
- p_i positive Häufigkeit des mit a_i erzeugten Teilbaums i
 n_i negative Häufigkeit des mit a_i erzeugten Teilbaums i
- p'_i erwarteter Wert von p_i
 n'_i erwarteter Wert von n_i

Folienverzeichnis (Ein Klick führt zur gewünschten Folie)

<u>Anwendungen der Regelinduktion</u>	<u>2</u>
<u>Einordnung</u>	<u>3</u>
<u>Unterrichtsmaterial</u>	<u>4</u>
<u>Grundlagen</u>	<u>5</u>
<u>Motivationen für die Regelinduktion</u>	<u>6</u>
<u>† SPESEN - Redundante Regeln</u>	<u>7</u>
<u>Widersprüchliche Regeln</u>	<u>8</u>
<u>Lückenhafte Regeln</u>	<u>9</u>
<u>Ineffiziente Regeln</u>	<u>10</u>
<u>Regeln automatisch extrahieren</u>	<u>11</u>
<u>7.1 Elementarregeln</u>	<u>12</u>
<u>7.2 Elementarregeln sind vollständig ...</u>	<u>13</u>
<u>... aber oft redundant und ineffizient</u>	<u>14</u>
<u>Beurteilung von Elementarregeln</u>	<u>15</u>
<u>7.3 Regelinduktion am Beispiel SPESEN</u>	<u>16</u>
<u>Beurteilung der Regelinduktion</u>	<u>17</u>
<u>Regelinduktion und Data Mining</u>	<u>18</u>
<u>Daten und Muster</u>	<u>19</u>
<u>7.18 Methode im Vergleich</u>	<u>20</u>
<u>† BONITÄTSKLASSIFIKATION - Fallbeispiel</u>	<u>21</u>
<u>Klassifikation und Regression</u>	<u>22</u>
<u>7.4 Entscheidungsbauminduktion - Überblick</u>	<u>23</u>

7.5 Von heterogenen zu homogenen Klassen	24
Klassifikation mit Entscheidungsbäumen	25
Fragen eines Entscheidungsbaums bestimmen	26
Reihenfolge der Entscheidungsknoten	27
Lösungsansätze	28
Testattribute verringern die Heterogenität	29
Anwendungsentwicklung	30
Entwicklung mit <i>XpertRule Profiler</i>	31
📌 BONITÄTSKLASSIFIKATION - Entwicklungsphasen	32
7.6 📌 ① Lernmenge	33
7.7 📌 ③ Vorläufige Datenanalyse	34
📌 ③ Vorläufige Datenanalyse	35
📌 ④ Induktionsparameter	36
7.8 📌 ⑤ Einfacher binärer Entscheidungsbaum	37
📌 Von der Baum- zur Textnotation	38
📌 Lässt sich der induzierte Baum verallgemeinern?	39
Klassifikationsgüte	40
Sequentielle und parallele Verarbeitung	41
🖱️ BONITÄT mit <i>XpertRule Profiler</i> (A 7.1)	42
Funktionalität von <i>XpertRule Profiler</i>	46
Theorie	47
Theoretischer Hintergrund	48
🌀 ID3 - Fallbeispiel WETTER	49
7.10 ID3 - Lernmenge WETTER	50

7.11 ID3 - Induzierter Entscheidungsbaum WETTER	51
ID3 - Skizze eines Induktionsalgorithmus	52
ID3 - ① Informationsmass	53
ID3 - ② Entropie	54
7.12 ID3 - ② Entropie	55
ID3 - ② Entropie	56
7.13 ID3 - Entropie des Gesamtbaums <i>Wetter</i>	57
ID3 - Entropie $I(p_i, n_i)$ des Teilbaums mit Wurzel i	58
ID3 - ③ Erwartungswert	59
ID3 - Erwartungswert des Attributs Charakter	60
ID3 - ④ Klassifikationsgewinn	61
7.14 ID3 - Klassifikationsgewinn eines Attributs	62
7.16 ID3 - Höchster Klassifikationsgewinn	63
7.17 ID3 - Grober Induktionsalgorithmus	64
ID3 - Erweiterungen	65
ID3 - Entropie und Chi Quadrat	66
ID3 - Vermeide Test eines irrelevanten Attributs A	67