Universitäts-Rechenzentrum Trier

AWS.AT.2

Trier, den 13.8.2004

Bernhard Baltes-Götz

Entscheidungsbaumanalyse mit AnswerTree 3.1

1 EINLEITUNG 4

1.1 Anwendungsbeispiel 4

1.2 Unterstützte Algorithmen 6

1.3 Hinweise zum Manuskript 6

2 DIE CHAID-VERFAHREN 7

2.1 Daten importieren 7

2.2 Modellspezifikation 8
2.2.1 Der Baumassistent 8
2.2.2 Skalenniveau und andere Variablenattribute 12

2.3 Baumaufbau 14
2.3.1 Automatische Segmentierung 14
2.3.2 Manuelle Steuerung der Segmentierung 16

2.4 Beurteilung von Entscheidungsbäumen 17
2.4.1 Treffer 17
2.4.2 Profit 21
2.4.3 Fehlklassifikationen 24
2.4.4 Profit und Fehlklassifikationskosten als Entscheidungsgrundlage 27
2.4.5 Knotendefinitionen exportieren 27

2.5 Methodische Details zu den CHAID-Verfahren 28
2.5.1 Baumaufbau 28
2.5.2 Bonferroni-adjustierte Überschreitungswahrscheinlichkeiten 29

3 DAS C&RT-VERFAHREN 34

3.1 Aufbau eines binären Baumes durch Reduktion von Inhomogenität 34
3.1.1 Nominalskalierte Kriterien 34
3.1.2 Metrische Kriterien 40
3.1.3 Ersatzvariablen bei fehlenden Werten 42

3.2 Klassifikationsbaum zu Fishers Iris-Daten 45

3.3 Kürzen zur Reduktion der Komplexität 48

3.4 Kreuzvalidierung 50

3.5 Regressionsbaum als Modellierungshilfe 53

4 DAS QUEST-VERFAHREN 57

4.1 Wahl eines Prädiktors 57

4.2 Metrisierung nominalskalierter Variablen 59

4.3 Trennwerte bestimmen 59
4.3.1 Bildung von Superklassen 59
4.3.2 Trennung am Punkt mit identischer Wahrscheinlichkeitsdichte für beide Superklassen 59
4.3.3 A-priori – Wahrscheinlichkeiten 61
4.3.4 Fehlklassifikationskosten 62

4.4 QUEST-Analyse der Iris-Daten 64

5 LITERATUR 65

6 ANHANG 66

6.1 Einsatzmöglichkeiten der einzelnen Algorithmen 66

6.2 Variablen im Marktforschungs-Beispiel 67

7 STICHWORTVERZEICHNIS 68

Herausgeber: Universitäts-Rechenzentrum Trier
Universitätsring 15
D-54286 Trier
Tel.: (0651) 201-3417, Fax.: (0651) 3921
Leiter: Prof. Dr.-Ing. Manfred Paul
Autor: Bernhard Baltes-Götz, E-Mail: baltes@uni-trier.de

1 Einleitung

AnswerTree unterstützt mehrere attraktive Algorithmen (CHAID, C&RT, QUEST) zum Aufbau von *baumartig strukturierten Klassifikationssystemen* aufgrund einer Entwicklungsstichprobe, die später auf neue Fälle angewendet werden können. Wesentlich bei diesen explorativen Analysemethoden ist die Suche nach Populationssegmenten, definiert durch Ausprägungskombinationen der beteiligten Prädiktorvariablen, die bezüglich eines vorgegebenen Kriteriums intern möglichst homogen und untereinander möglichst verschieden sind. Diese Optimalitätseigenschaft bezüglich eines Kriteriums unterscheidet die AnswerTree-Segmente z.B. von den Ergebnissen einer Clusteranalyse.

Man kann die Entscheidungsbaum-Verfahren in AnswerTree wie auch die Clusteranalyse oder die neuronalen Netze zu den **Data Mining** – Werkzeugen rechnen, mit denen bislang unbekannte Strukturen in komplexen Datensätzen (semi)automatisch aufgedeckt werden sollen. Im Vergleich zu den neuronalen Netzen liefert ein Entscheidungsbaum (wie auch die Clusteranalyse) transparente Klassifikationsregeln.

1.1 Anwendungsbeispiel

In folgendem Beispiel aus der Marktforschung soll eine postalische Zeitschriften-Werbeaktion geplant werden, wobei aus einer früheren Werbekampagne Erfahrungsdaten für 81040 Haushalte vorliegen (Beispiel aus Magidson & SPSS, 1993, S. 3ff). Man sucht Haushalte, die mit besonders hoher Wahrscheinlichkeit positiv auf eine Werbezusendung reagieren. Für jeden Haushalt in der Entwicklungsstichprobe[1] sind folgende Merkmale bekannt:

Merkmal	**Rolle**	**Skalenniveau**	**SPSS-Variable**
Antwort auf die Werbeaktion (2)	Kriterium	nominal	ANTW2
Alter des Haushaltsvorstandes	Prädiktor	ordinal	ALTER
Geschlecht des Haushaltsvorstandes	Prädiktor	nominal	GESCHL
Kinder im Haushalt	Prädiktor	nominal	KINDER
Haushaltseinkommen	Prädiktor	ordinal	HEINKOMM
Kreditkarte vorhanden	Prädiktor	nominal	KARTE
Anzahl der Personen im Haushalt	Prädiktor	ordinal	HHGRÖßE
Beruf des Haushaltsvorstandes	Prädiktor	nominal	BERUF

Ausführlichere Informationen zu den Variablen (insbesondere zu den einzelnen Ausprägungen) finden Sie im Anhang.

Es soll ein baumartiges Klassifikationssystem entwickelt werden, an dessen Verzweigungspunkten jeweils *beliebig viele* Äste entspringen dürfen. Da außerdem alle Prädiktoren maximal ordinales Skalenniveau besitzen, bietet sich der CHAID-Algorithmus (**C**hi-squared **A**utomatic **I**nteraction **D**etection) in seiner klassischen oder exhaustiven Variante (siehe unten) an[2]. Er besteht aus einer Sequenz von Zusammenfassungen und Zerlegungen, gesteuert durch die Ergebnisse von diversen Assoziationsanalysen, in denen jeweils der Zusammenhang des Kriteriums mit *einem* Prädiktor beurteilt wird. Leicht vereinfachend kann der CHAID-Algorithmus so beschrieben werden[3]:

---

1 Aliasnamen: Lernstichprobe, Trainingsstichprobe

2 Während der CHAID-Algorithmus bei der Zerlegung eines Knotens beliebig viele Unterknoten erlaubt, erzeugen C&RT und QUEST stets *binäre* Bäume. Von den zuletzt genannten Algorithmen werden *metrische* Prädiktoren in vielen Situationen besser unterstützt.

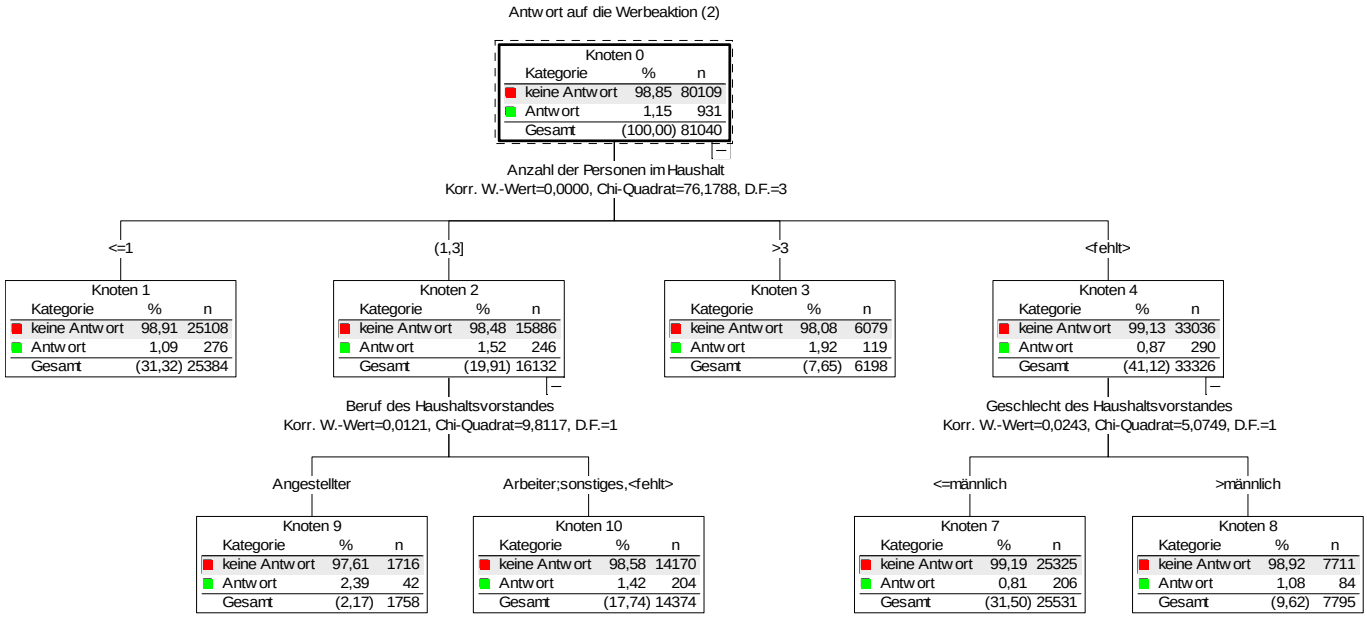
3 In Abschnitt 2 folgt eine detaillierte Beschreibung.

- Zunächst wird die Stichprobe in die Kategorien des besten Prädiktors aufgeteilt. *Gut* ist ein Prädiktor dann, wenn er eine signifikante Assoziation mit dem Kriterium nachweist (p-Level kleiner als 0,05). Unter den guten Prädiktoren wird derjenige mit dem kleinsten p-Level im Assoziationstest zum Besten gekürt. Dabei werden Prädiktor-Kategorien ohne bedeutsame Unterschiede hinsichtlich der Kriteriumsverteilung assoziations-optimierend zusammengefasst.
- Dann wird für jede der erhaltenen Gruppen eine weitere Zerlegung mit Hilfe der restlichen Prädiktoren versucht. Unter den Prädiktoren mit signifikanter Beziehung zum Kriterium wird wiederum der Beste bestimmt.
- Der Algorithmus läuft in jedem Zweig des entstehenden Baumes so lange, bis sich kein signifikanter Prädiktor mehr findet.

Allerdings verfügen Anwender des CHAID-Verfahrens durchaus über Mitspracherechte:

- Der Baumaufbau ist über etliche **Einstellungen** zu beeinflussen, z.B. durch die Wahl der Signifikanzgrenzen für das Zusammenfassen von Prädiktorkategorien bzw. für das Zerlegen von Knoten.
- Alternativ zum *vollautomatischen* Baumaufbau erlaubt AnswerTree auch **manuelle Eingriffe** bei der Festlegung von Verzweigungen (z.B. durch Wahl des Prädiktors).

Unter gewissen, noch zu besprechenden, Einstellungen resultiert für unser Beispiel unter Verwendung des (klassischen oder exhaustiven) CHAID-Verfahrens das folgende **Baumdiagramm**:



AnswerTree hat hier (teilweise mit Unterstützung des Anwenders, siehe unten) sechs Untergruppen (Lösungsknoten) ermittelt, die sich stark im Hinblick auf ihre Antwortquote unterscheiden:

- Im Knoten 9 (mit **Angestellter** überschrieben) beträgt die Quote immerhin 2,39%.
- Im Knoten 7 (mit **männlich** überschrieben) liegt sie bei 0,81%.

Die Rangordnung der identifizierten Segmente in Bezug auf ein Kriterium ist der wesentliche Unterschied zu den Ergebnissen einer Clusteranalyse, die ja ebenfalls aufgrund beschreibender Informationen die Stichprobe zerlegt.

1.2 Unterstützte Algorithmen

AnswerTree hat in der SPSS-Produktpalette das Programm **CHAID** abgelöst, das nur den eben skizzierten **CHAID**-Algorithmus (nach Kass, 1980) beherrscht. Demgegenüber bietet AnswerTree zusätzlich folgende Algorithmen (siehe SPSS 2002, S. 200ff):

- **Exhaustive CHAID**
  Während der klassische CHAID-Algorithmus bei der assoziations-steigernden Prädiktorvorbehandlung stoppt, wenn keine Kategorienpaare mehr aufgrund insignifikanter Kriteriumsunterschiede zu vereinigen sind, untersucht die von Biggs, De Ville & Suen (1991) vorgeschlagene exhaustive CHAID-Variante *alle* von einem Prädiktor ermöglichten Zerlegungen auf eine möglichst signifikante Kriteriums-Assoziation.
- **C&RT** (Classification and Regression Trees)[1]
  Bei dieser von Breiman, Friedman Olshen & Stone (1984) entwickelten Methode wird die Segmentierung *nicht* durch Assoziations-Signifikanztests, sondern durch die Minimierung von Inhomogenitätsmaßen gesteuert. Sie ist der CHAID-Methode oft überlegen, wenn auch *metrische* Prädiktoren vorliegen.
  Ein Knoten wird jeweils in *zwei* Unterknoten zerlegt, so dass ein binärer Baum resultiert.
- **QUEST** (Quick, Unbiased, Efficient, Statistical Tree)
  Beim Entwurf des QUEST-Verfahrens (siehe Loh & Shih 1997) standen folgende Ziele im Vordergrund:
  - Möglichst geringer Rechenaufwand
  - Vermeidung der (beim C&RT-Verfahren festzustellenden) Bevorzugung von Prädiktoren, die zahlreiche verschiedene Zerlegungen der Stichprobe erlauben.

  Beide Ziele profitieren von einer Entkopplung der folgenden Entscheidungen beim Zerlegen eines Knotens:
  - Wahl eines Prädiktors
  - Wahl eines Trennwertes zur Aufteilung der Stichprobe

  Wie der Q&RT- Algorithmus erstellt auch das QUEST-Verfahren *binäre* Bäume.

1.3 Hinweise zum Manuskript

Das Manuskript ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Segmentierung und Klassifikation mit AnswerTree 3.1

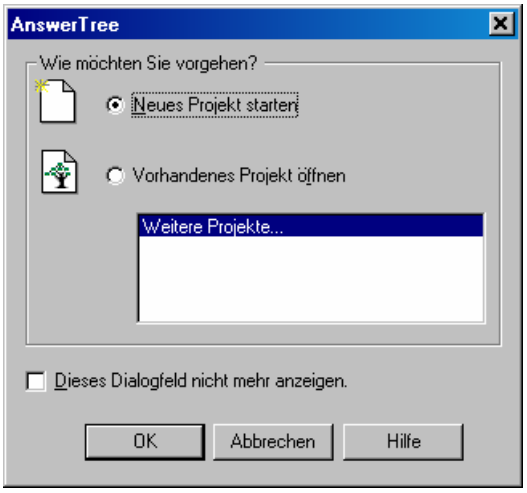
[1] Häufig wird auch das Kürzel *CART* verwendet.

2 Die CHAID-Verfahren

In diesem Abschnitt wird anhand eines ausführlich diskutierten Beispiels die klassische Variante des CHAID-Verfahrens vorgestellt. Im Zusammenhang mit einigen methodischen Details (siehe Abschnitt 2.5) kommen dann die Neuerungen der exhaustiven Variante zur Sprache.[1]
Neben den CHAID-Eigenschaften werden aber auch generelle AnswerTree-Bedienungsmerkmale behandelt (z.B. der Baumassistent).

2.1 Daten importieren

Starten Sie AnswerTree, und wählen Sie im Begrüßungsfenster (Startassistenten) die Option **Neues Projekt starten**:



Entscheiden Sie sich in der **Datenquellen**-Dialogbox für eine **SPSS-Datendatei**:

Wählen Sie anschließend aus dem **Samples**-Unterordner zum AnswerTree-Programmverzeichnis die Datei **subs.sav**.

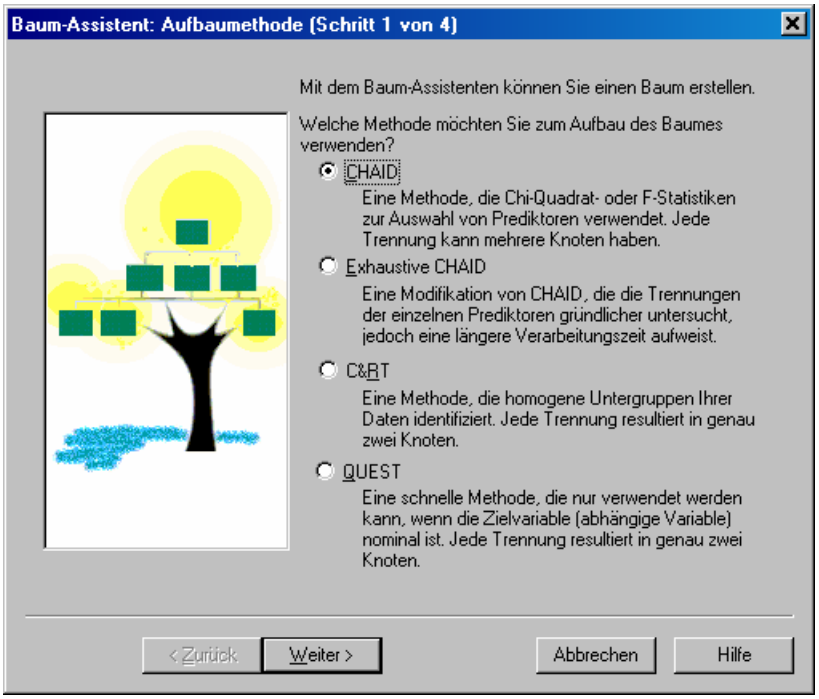
[1] Im Beispiel wirken sich die Unterschiede zwischen den beiden CHAID-Varianten allerdings *nicht* aus.

2.2 Modellspezifikation

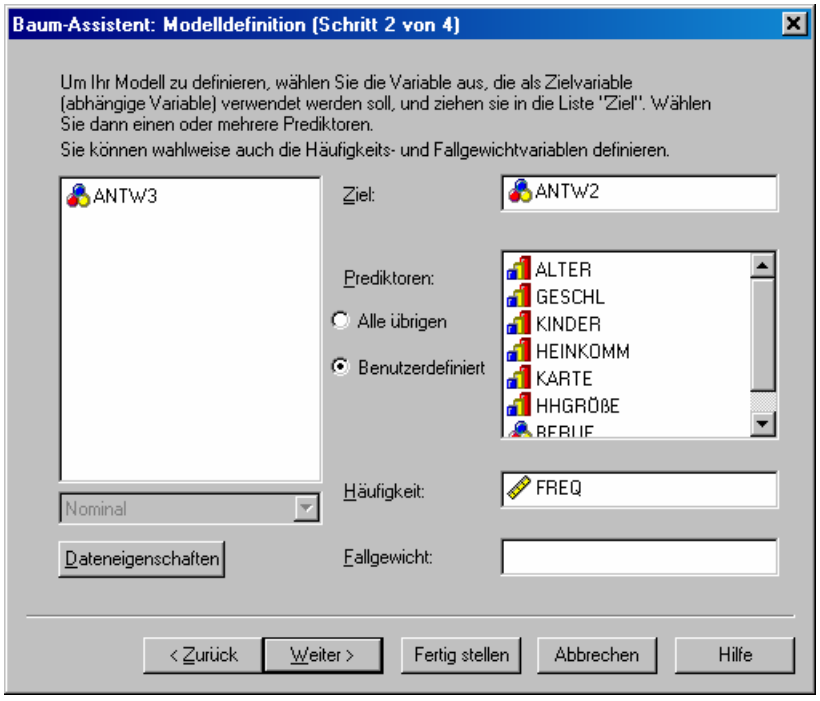
2.2.1 Der Baumassistent

Wir wollen nun einen Klassifikationsbaum erstellen. Falls der zugehörige Assistent nicht spontan auftritt, startet man ihn über den Schalter  oder den Menübefehl **Datei > Neuer Baum**.

Im ersten Schritt übernehmen wir als Segmentierungsmethode den voreingestellten CHAID-Algorithmus:



Im zweiten Dialogfenster des Assistenten lassen sich die Variablen per Maus durch Ziehen und Fallenlassen in Position bringen. In unserem Beispiel ist folgende Spezifikation sinnvoll:



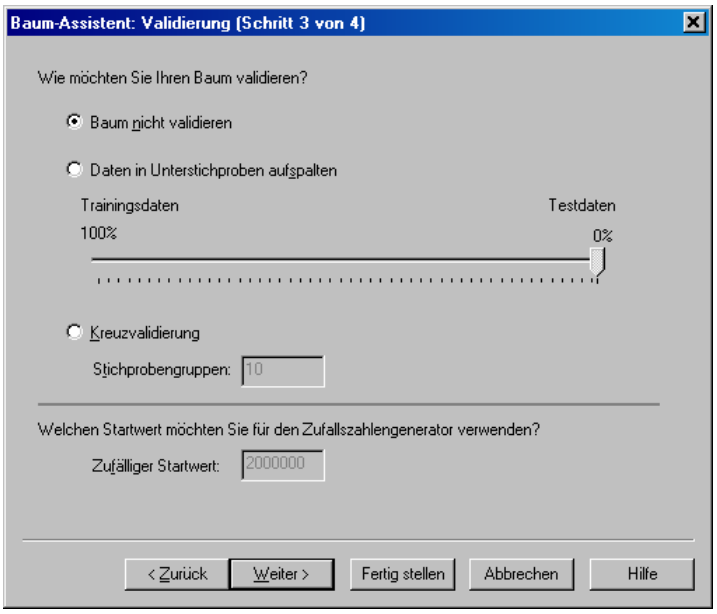
Um die teilweise abgeschnittenen und unübersichtlichen Variablenetiketten durch die prägnanteren Variablen*namen* zu ersetzen, wurde nach einem Rechtsklick auf eine Variablenliste die Option **Variablennamen anzeigen** gewählt.

Häufigkeitsvariablen haben in AnswerTree die selbe Funktion wie *Gewichtungsvariablen* in SPSS: Erhält ein Fall per Häufigkeitsvariable das Gewicht 12, läuft die gesamte Analyse so ab, als wäre der Fall 12 mal vorhanden. Dies ist sinnvoll und erforderlich bei *aggregierten* Dateien, deren Fälle jeweils für mehrere Beobachtungen mit identischen Variablenausprägungen stehen.
In unserem Beispiel ist die Häufigkeitsvariable **FREQ** zu verwenden.

Gewichtungsvariablen dienen im Unterschied zu den Häufigkeitsvariablen *nicht* dazu, die Erfassung *vorhandener* Fälle zu vereinfachen, sondern sie sind dann zu verwenden, wenn für die Populationen zu den einzelnen Wertekombinationen (Zellen) unterschiedliche Erhebungsraten realisiert worden sind. Dann spiegeln die relativen Häufigkeiten der Wertekombinationen in der Stichprobe die zugehörigen Teilpopulations-Proportionen *nicht* wieder, so dass die Fälle mit unterschiedlichen Gewichten in die Analyse eingehen sollten.
Sind Gewichtungsvariablen aktiv, wird beim CHAID-Verfahren der WLM-Algorithmus (Weighted Loglinear Modeling) verwendet (siehe Magidson & SPSS 1993, S. 127ff).
In unserem Beispiel ist keine Gewichtungsvariable vorhanden.

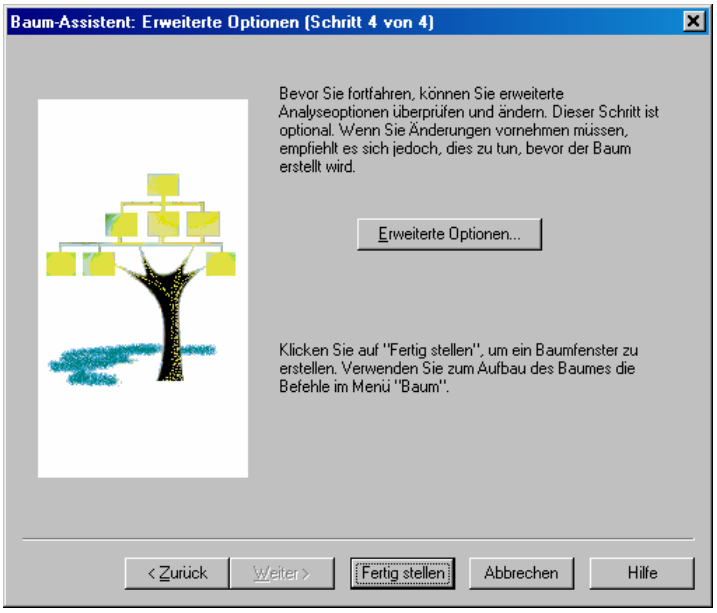
Das aus der SPSS-Datendatei übernommene und symbolisch angezeigte Skalenniveau der Variablen (siehe Abschnitt 2.2.2) lässt sich im aktuellen Assistentenschritt per Kontextmenü ändern.

Im dritten Assistentenschritt verzichten wir auf eine Validierung, d.h. auf eine Prüfung der Modell-Generalisierbarkeit:



Mit dem wichtigen Thema *Kreuzvalidierung* werden wir uns in Abschnitt 3.4 beschäftigen.

In der vierten Assistenten-Dialogbox nutzen wir die Möglichkeit, **erweiterte Optionen** festzulegen:



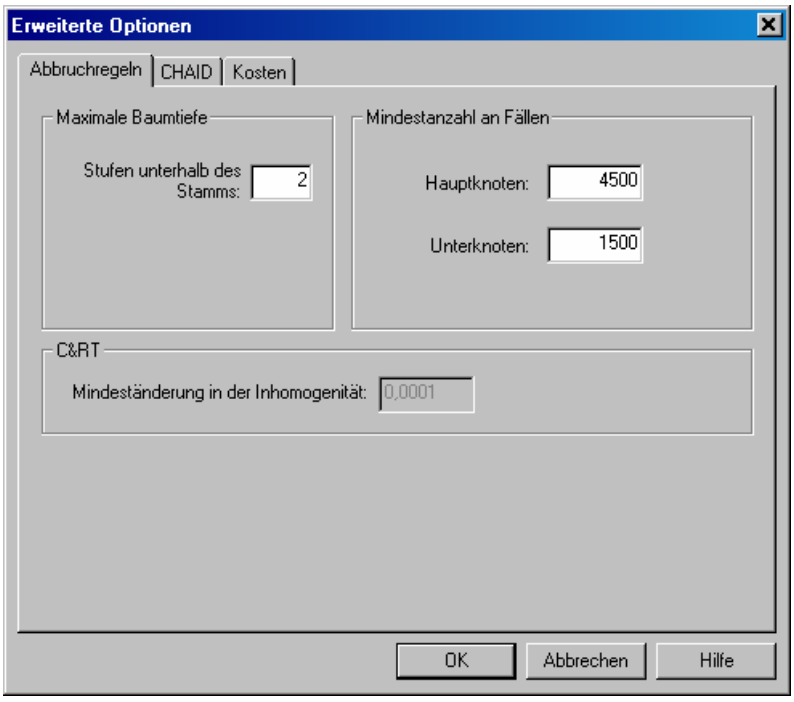
Auf dem Registerblatt **Abbruchregeln** vereinbaren wir:

Maximale Baumtiefe: 2
Damit wird der Segmentierungsbaum auf zwei Ebenen beschränkt.

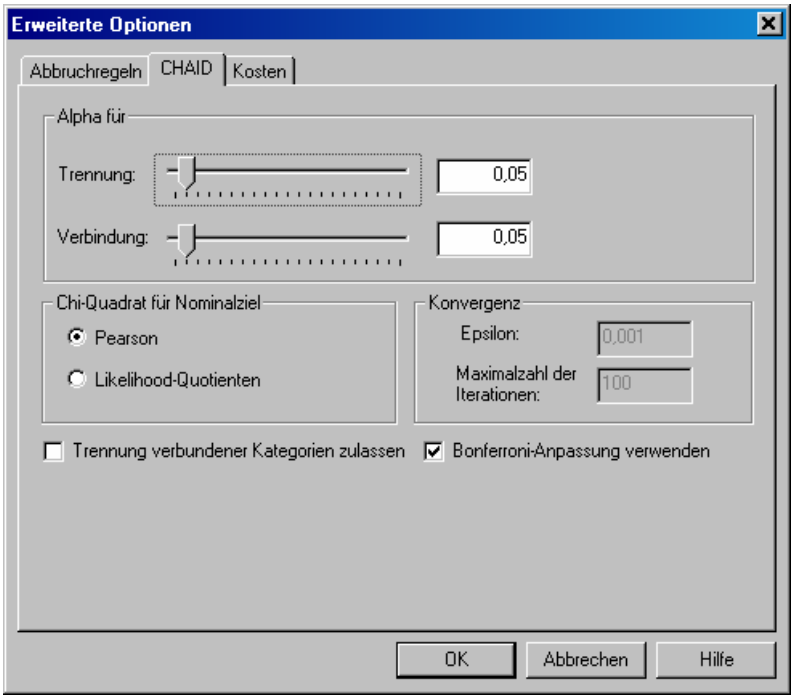
Mindestanzahl an Fällen

Hauptknoten: 4500
Ein Knoten darf nur dann aufgeteilt werden, wenn er mindestens 4500 Fälle enthält.

Unterknoten: 1500
Per Segmentierung darf kein Knoten entstehen, der weniger als 1500 Fälle enthält. Damit ist sichergestellt, dass in der endgültigen Lösung jeder Knoten mindestens 1500 Fälle enthält.



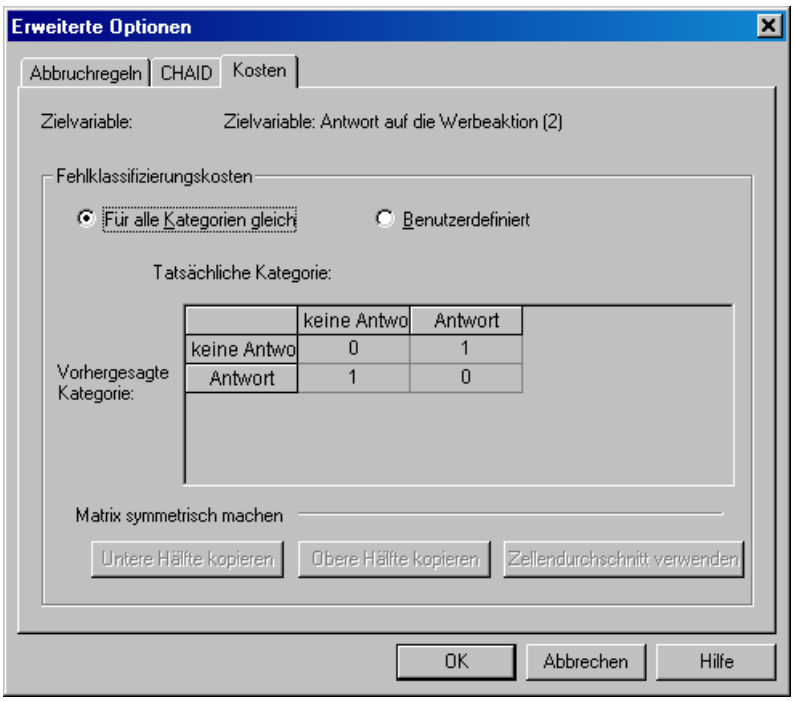
Auf dem Registerblatt **CHAID** übernehmen wir alle Voreinstellungen:



Damit haben wir uns u.a. für die Signifikanzgrenze von 0,05 beim Aufteilen von Knoten und beim Vereinigen von Prädiktorkategorien, für χ^2-Tests mit der Prüfgröße von Pearson und für eine Bonferroni-Anpassung der durch optimierende Klassenzusammenfassung gewonnenen Überschreitungswahrscheinlichkeiten entschieden (s.u.). Diese Einstellungen können die resultierende Segmentierung erheblich beeinflussen.

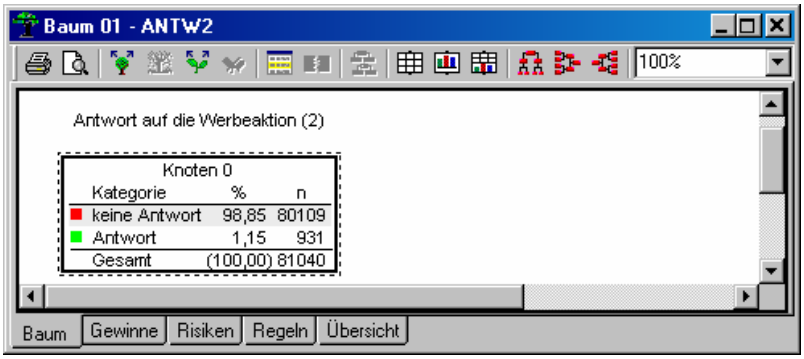
Nur bei Verwendung von Gewichtungsvariablen sind die **Konvergenz**-Einstellungen relevant, weil dann der iterative WLM-Algorithmus (Weighted Loglinear Modeling) zum Einsatz kommt (siehe Magidson & SPSS 1993, S. 128).

Auf die Festlegung **benutzerdefinierter Fehlklassifizierungskosten** wollen wir vorläufig verzichten:



Mit dieser Option lässt sich das Risiko für besonders teure Fehlentscheidungen (z.B. unerkannter Tumor) weiter reduzieren, wobei allerdings andere Fehlentscheidungen wahrscheinlicher werden. Bei der CHAID-Analyse wirken sich die Fehlklassifikationskosten *nicht* auf die Segmentierung aus, sondern nur auf die vom Programm für die Endknoten prognostizierte Zielkategorie (siehe Abschnitt 2.4.3). Bei den anderen Segmentierungsverfahren (C&RT, QUEST) können Fehlklassifikationskosten auch den Baumaufbau beeinflussen.

Zurück im vierten Schritt des Baum-Assistenten klicken wir auf **Fertig stellen**. Daraufhin präsentiert AnswerTree das **Baumfenster Baum 1**, das den Startknoten mit der gesamten Entwicklungsstichprobe zeigt und zahlreiche Analyseoptionen bietet:



Es wird allmählich Zeit, das Projekt mit

Datei > Projekt speichern unter

zu sichern, wobei eine Datei mit der voreingestellten Namenserweiterung **atp** (*AnswerTree Projekt*) entsteht.

Bevor wir den Entscheidungsbaum aufbauen (lassen), sind noch einige Details zu den Skalenniveaus und anderen Attributen der Variablen zu klären.

2.2.2 Skalenniveau und andere Variablenattribute

AnswerTree ordnet jeder Variablen ein Skalenniveau zu, wobei folgende Alternativen zur Verfügung stehen:

- **nominal**
- **ordinal**
- **kontinuierlich**

Beim Lesen einer SPSS-Datendatei (Namenserweiterung **sav**) übernimmt AnswerTree die dort vereinbarten Skalenniveaus und erlaubt eine Änderung im 2. Schritt des Baum-Assistenten über das Kontextmenü zu den Variablennamen.
In unserem Beispiel werden aus der SPSS-Datendatei **subs.sav** folgende Skalenniveaus übernommen:

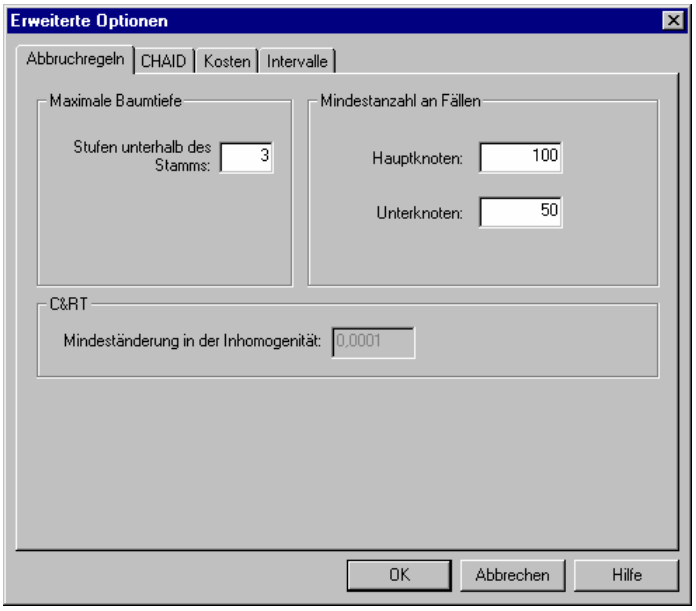
Variablenname	Variablenetikett	Skalenniveau	Symbol
ANTW2	Antwort auf die Werbeaktion (2)	nominal	
ALTER	Alter des Haushaltsvorstandes	ordinal	
GESCHL	Geschlecht des Haushaltsvorstandes	ordinal	
KINDER	Kinder im Haushalt	ordinal	
HEINKOMM	Haushaltseinkommen	ordinal	
KARTE	Kreditkarte vorhanden	ordinal	
HHGRÖßE	Anzahl der Personen im Haushalt	ordinal	
BERUF	Beruf des Haushaltsvorstandes	nominal	
FREQ	*Häufigkeitsvariable*	kontinuierlich	

Dass den Variablen **GESCHL**, **KINDER** und **KARTE** vom Urheber der Beispieldatei ein *ordinales* Skalenniveau zugeschrieben wurde, scheint nicht sehr plausibel. Weil es sich um dichotome Prädiktoren handelt, hat das zugeordnete Skalenniveau allerdings keinen Effekt auf den CHAID-Algorithmus, so dass wir auf eine Nachbesserung verzichten.

Soll *nach* dem Anlegen eines Baumes das Skalenniveau eines Prädiktors besichtigt oder geändert werden, ist die folgende Dialogbox zuständig:

Sie bezieht sich auf den im aktuell markierten Baumknoten zur Zerlegung gewählten Prädiktor und ist über das Kontextmenü dieses Knotens oder über den Menübefehl **Analyse > Skalenniveau** aufzurufen. Im aktuellen Beispiel steht sie mangels Baumaufbau noch nicht zur Verfügung.

Bei einer CHAID-Analyse werden kontinuierliche (metrische) Prädiktoren durch Intervallbildung auf ordinales Niveau gebracht. An Stelle der automatisch im Sinne einer Gleichverteilung gebildeten Intervalle können über das Registerblatt **Intervalle** in der Dialogbox **Erweiterte Optionen** auch *benutzerdefinierte* Intervalle vereinbart werden:



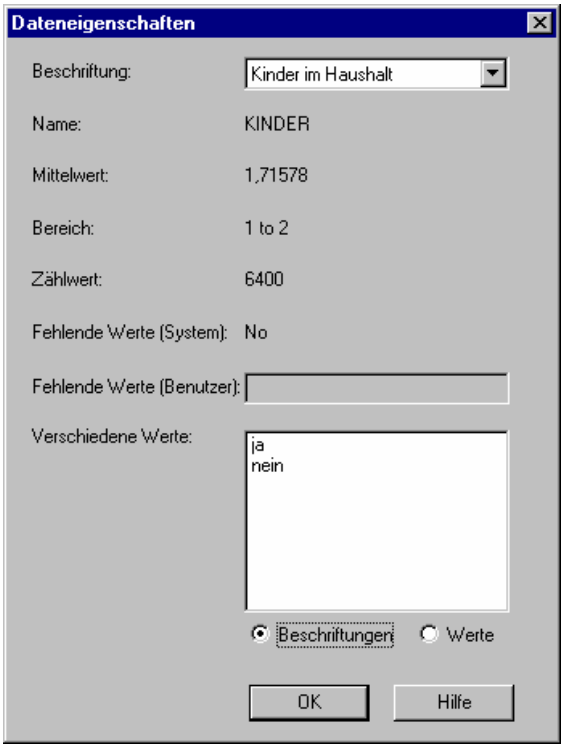
Diese Dialogbox stammt *nicht* aus dem aktuellen Beispiel, das keine metrischen Prädiktoren enthält.

Bei metrischen *Kriterien* nimmt der CHAID-Algorithmus *keine* Reduktion des Skalenniveaus vor, sondern verwendet zur Beurteilung der Assoziation mit den Prädiktoren an Stelle des χ^2-Tests der Kreuztabellenanalyse den F-Test der Varianzanalyse (siehe Abschnitt 2.5).

Die Ordinalität einer Prädiktorvariablen kommt im CHAID-Algorithmus beim Zusammenlegen nicht signifikant verschiedener Kategorien zum Tragen. Hier dürfen nur *benachbarte* Kategorien fusionieren, während bei nominalskalierten Variablen beliebige Kombinationen erlaubt sind. Ansonsten werden nominale und ordinale Prädiktoren identisch behandelt.

Fehlende Werte eines Prädiktors bilden in den CHAID-Algorithmen eine eigene Kategorie, was im Marktforschungs-Beispiel bei der Variablen **HHGRÖßE** zu beobachten ist. Bei nominalskalierten Prädiktoren darf diese Kategorie mit anderen fusionieren, bei ordinalskalierten Prädiktoren hingegen *nicht*.

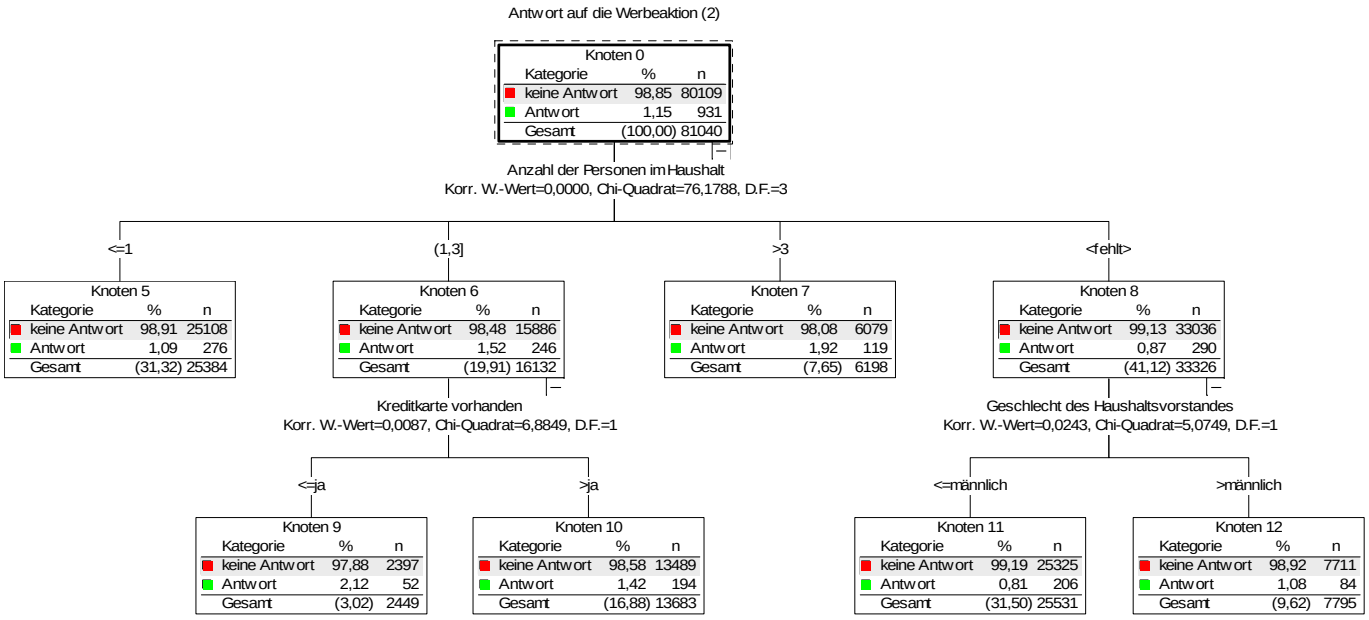
In der Dialogbox **Dateneigenschaften**, die im 2. Schritt des Baumassistenten per Schaltfläche und bei geöffnetem Baumfenster über den Menübefehl **Analyse > Dateneigenschaften** aufzurufen ist, werden etliche Variablenattribute angezeigt (allerdings nicht das Skalenniveau), z.B.:



2.3 Baumaufbau

2.3.1 Automatische Segmentierung

Im Baumfenster mit dem Stammknoten wollen wir AnswerTree nun per Mausklick auf das Symbol  , mit dem Menübefehl **Baum > Baumaufbau** oder via Stammknoten-Kontextmenü veranlassen, den auf Seite 5 beschriebenen CHAID-Algorithmus unter Verwendung der in Abschnitt 2.2.1 besprochenen Einstellungen vollautomatisch durchzuführen.
Das dabei erzielte Baumdiagramm weicht von der eingangs präsentierten Variante ab und zeigt insbesondere *nicht* mehr das besonders attraktive Marktsegment mit der höchsten Rücklaufquote von 2,39%:



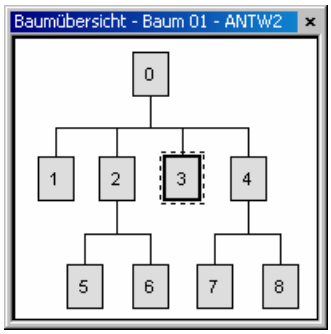
AnswerTree hat im ersten Schritt die Gesamtstichprobe (*n* = 81040, Antwortrate von 1,15 %, siehe Wurzelknoten) anhand der Variablen **HHGRÖßE** aufgeteilt, wobei die Kategorien 2 und 3 sowie die Kategorien 4 und 5 jeweils zusammengelegt wurden. Auf der zweiten Analyseebene wurde die neu gebildete Gruppe mit den 2- und 3-Personenhaushalten anhand der Variablen **KARTE** weiter zerlegt. Auch für die Gruppe mit Haushalten unbekannter Größe fand sich noch ein signifikanter Prädiktor (**GESCHL**) für eine weitere Unterteilung.

Ein Baumdiagramm kann übrigens aus dem AnswerTree-Baumfenster via Zwischenablage bequem in andere Anwendungen übernommen werden. Mit dem Menübefehl

Datei > Export > Baum

lässt es sich auch in eine Datei sichern (z.B. im **Enhenced Metafile** – Format).

Als Orientierungshilfe bei großen Baumdiagrammen bietet AnswerTree über den Menübefehl **Ansicht > Baumübersicht** oder den Symbol-Schalter eine Übersicht:



Per Mausklick auf einen Knoten bringt man ihn ins Blickfeld des Baumfensters.

Für Übersicht kann auch die Zoom-Einstellung im Baumfenster sorgen, die per Symbolleiste oder **Ansicht**-Menü zu ändern ist, z.B.:

2.3.2 Manuelle Steuerung der Segmentierung

Wie das am Anfang des Manuskriptes wiedergegebene Baumdiagramm beweist, ist die Zerlegung des Knotens mit den 2- und 3-Personenhaushalten anhand der Variablen **KARTE** nicht optimal, weil dabei der Unterknoten mit der besten Rücklaufquote übersehen wird (Kategorie **Angestellte** der Variablen **BERUF**). Wer genügend Spürsinn, Wissen oder Experimentierdrang besitzt, kann auf Zerlegungsentscheidungen Einfluss nehmen.

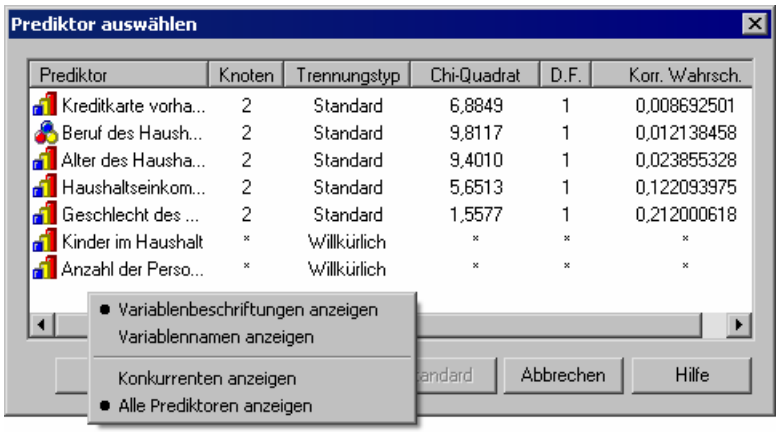
Um einen Ast aus einem AnswerTree-Entscheidungsbaum zu *entfernen* (eine Trennung aufzuheben), stehen nach dem Markieren des betroffenen Knotens drei alternative Bedienungselemente zur Verfügung:

- Menübefehl **Baum > Ast entfernen**
- Symbolschalter
- Option **Ast entfernen** im Kontextmenü des betroffenen Knotens

Um die Definition einer alternativen Trennung einzuleiten, stehen nach dem Markieren des betroffenen Knotens folgende Bedienungselemente zur Verfügung:

- Menübefehl **Baum > Prediktor auswählen**
- Symbolschalter
- Option **Prediktor auswählen** im Kontextmenü des betroffenen Knotens

Es erscheint ein Fenster mit allen Prädiktoren, das aber nur für „ernst zu nehmende“ **Konkurrenten** statistische Beurteilungen enthält:



Nach einem rechten Mausklick auf die Variablenliste kann man per Kontextmenü dafür sorgen, dass nur die von AnswerTree empfohlenen Konkurrenten angezeigt werden. Relevanter ist eventuell, die Anzeige von Variablen*namen* an Stelle der oft unübersichtlichen Beschriftungen anzufordern.

In der letzten Tabellenspalte findet sich das im CHAID-Algorithmus bei nominalskalierten Kriterien bevorzugte Maß zur Beurteilung eines Prädiktors: Das Bonferroni-adjustierte p-Level zum χ^2-Wert aus der Kreuztabelle mit dem Kriterium dem (ggf. optimal rekodierten) Prädiktor. Alle Berechnungen basieren natürlich auf der Teilstichprobe zum aktuellen Knoten.
Dass der Prädiktor **KARTE** eine bessere Bewertung erzielt als der Prädiktor **BERUF**, dessen Einsatz schließlich zu einem Knoten mit besonders günstiger Rückmeldequote führt, liegt an der relativ geringen Besetzung des optimalen Knotens (N = 1758).

Bei der Variablen **BERUF** schlägt AnswerTree vor, die Gruppe der Angestellten der Vereinigung aller anderen Kategorien gegenüber zu stellen. Für die Assoziation des derart rekodierten Prädiktors mit dem Kriterium **ANTW2** erhalten wir bei einer Kreuztabellenanalyse mit SPSS für Windows die folgenden Testergebnisse:

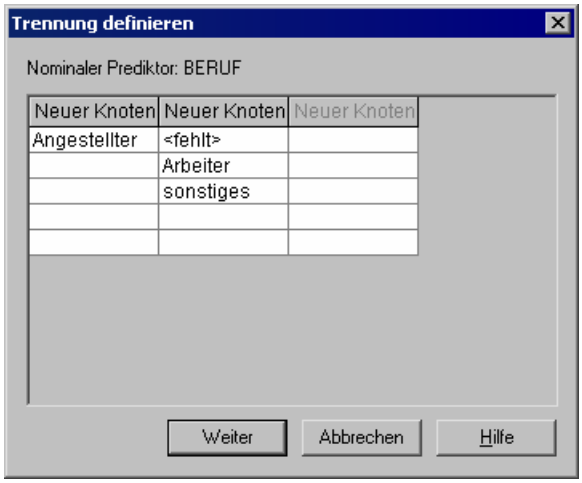
BERUF (rekodiert) vs. ANTW2 bei 2-3-Personen-Haushalten

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	9,8117	1	,00173
Anzahl der gültigen Fälle	16132		

Weil ausgehend von 4 Altkategorien eine von 7 möglichen (siehe Abschnitt 2.5.2) Zusammenfassungen zu 2 Neukategorien assoziations-maximierend gewählt wurde, berechnet AnswerTree die korrigierte Wahrscheinlichkeit, indem es das p-Level aus der optimalen Kontingenztafel mit 7 multipliziert:

$$7 \cdot 0{,}00173 = 0{,}01211$$

Eine von AnswerTree vorgenommene Gruppeneinteilung muss nicht klaglos hingenommen werden, sondern kann für den markierten Prädiktor nach **Trennung definieren** eingesehen und ggf. korrigiert werden, z.B.:



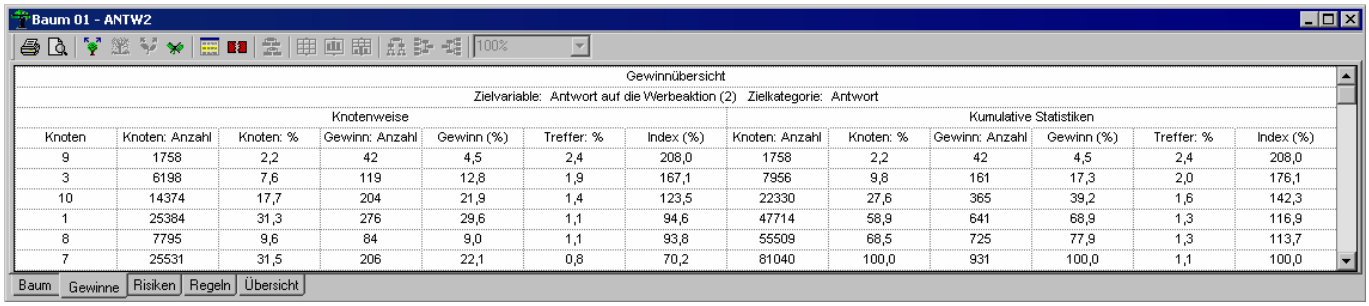
Wenn wir die Variable **BERUF** mit der von AnswerTree vorgeschlagenen Neugruppierung wählen und die Dialogbox **Prediktor auswählen** mit **Aufbau** quittieren, dann resultiert das am Anfang des Manuskriptes angegebenen Baumdiagramm.

2.4 Beurteilung von Entscheidungsbäumen

2.4.1 Treffer

2.4.1.1 Knotentabelle

Die von AnswerTree mit unserer Hilfe ermittelte Lösung besteht aus 6 Gruppen (Lösungsknoten), über die auf dem **Gewinne**-Registerblatt des Baumfensters wichtige Informationen zur Planung der Zeitschriften-Werbeaktion bereit stehen:



Knoten	Knoten: Anzahl	Knoten: %	Gewinn: Anzahl	Gewinn (%)	Treffer: %	Index (%)	Knoten: Anzahl	Knoten: %	Gewinn: Anzahl	Gewinn (%)	Treffer: %	Index (%)
9	1758	2,2	42	4,5	2,4	208,0	1758	2,2	42	4,5	2,4	208,0
3	6198	7,6	119	12,8	1,9	167,1	7956	9,8	161	17,3	2,0	176,1
10	14374	17,7	204	21,9	1,4	123,5	22330	27,6	365	39,2	1,6	142,3
1	25384	31,3	276	29,6	1,1	94,6	47714	58,9	641	68,9	1,3	116,9
8	7795	9,6	84	9,0	1,1	93,8	55509	68,5	725	77,9	1,3	113,7
7	25531	31,5	206	22,1	0,8	70,2	81040	100,0	931	100,0	1,1	100,0

Um die angegebenen Knoten-Nummern schnell im Baumdiagramm zu lokalisieren, nimmt man am besten die Baumübersicht zur Hilfe (siehe oben).
Die Haushalte im Knoten 9 (2- oder 3-Personen-Haushalte mit einem Angestellten als Haushaltsvorstand) versprechen den besten Erfolg der Werbekampagne. Wir erhalten folgende Informationen:

- **Knoten: Anzahl** Knoten 9 enthält 1758 Haushalte.
- **Knoten: %** Dies sind gerade 2,2% der Gesamtstichprobe.
- **Gewinn: Anzahl** Von den 1758 Haushalten in Knoten 9 haben 42 geantwortet.
- **Gewinn %** Die 42 Antworten in Knoten 9 sind ca. 4,5% von allen Antworten.
- **Treffer: %** Die Rücklaufquote im Knoten 9 beträgt ca. 2,4%.
- **Index (%)** Die Rücklaufquote im Knoten 9 ist ca. 2,08 mal größer als die Gesamtrücklaufquote von 1,15%.

Zusätzlich werden noch *kumulative* Informationen angeboten, z.B. über die Zusammenfassung der 3 besten Knoten (9, 3 und 10), nachzulesen in der Zeile zum Knoten 10:

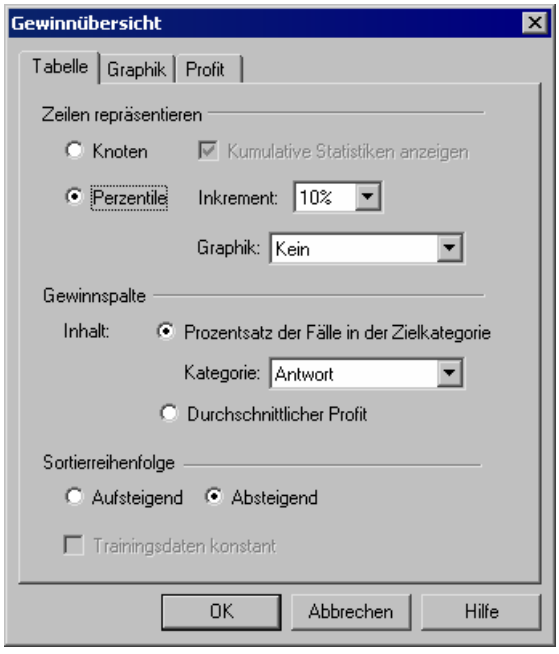
- **Knoten: Anzahl** Die besten drei Segmente enthalten zusammen 22330 Haushalte.
- **Knoten: %** Dies sind ungefähr 27,6% der Gesamtstichprobe.
- **Gewinn: Anzahl** Von den 22330 Haushalten in den Knoten 9, 3 und 10 haben 365 geantwortet.
- **Gewinn %** Die 365 Antworten aus den 3 besten Knoten sind ca. 39,2% von allen Antworten.
- **Gewinn (%)** Die Rücklaufquote für die Vereinigung der 3 besten Segmente beträgt ca. 1,6%.
- **Index (%)** Die Rücklaufquote für die Vereinigung der 3 besten Segmente ist ca. 1,42 mal größer als die Gesamtrücklaufquote von 1,15%.

2.4.1.2 Perzentiltabelle

Neben der eben beschriebenen knoten-orientierten Gewinntabelle bietet AnswerTree auch eine perzentil-orientierte Variante mit den geschätzten Antwortquoten beim Versandt an die ($k \cdot 10$) Prozent aussichtsreichsten Haushalte (k = 1, 2, ..., 10). Um diese Tabelle zu erhalten, muss man nach:

Format > Gewinne

in der Dialogbox **Gewinnübersicht** unter **Zeilen repräsentieren** die Option **Perzentile** wählen:



Es resultiert die folgende Tabelle[1]:

Gewinnübersicht						
Zielvariable: Antwort auf die Werbeaktion (2) Zielkategorie: Antwort						
Statistiken						
Knoten	Perzentil	Perz.: Anzahl	Gewinn: Anz.	Gewinn (%)	Treffer: %	Index (%)
9;3;10	10	8104	163	17,5	2,0	175,2
10	20	16208	278	29,9	1,7	149,4
10;1	30	24312	387	41,5	1,6	138,4
1	40	32416	475	51,0	1,5	127,5
1	50	40520	563	60,4	1,4	120,9
1;8	60	48624	651	69,9	1,3	116,5
8;7	70	56728	735	78,9	1,3	112,8
7	80	64832	800	86,0	1,2	107,4
7	90	72936	866	93,0	1,2	103,3
7	100	81040	931	100,0	1,1	100,0

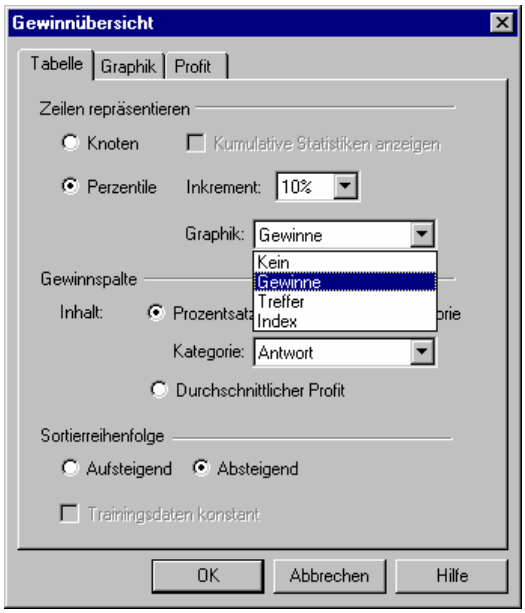
Für den Versandt an die 30% aussichtsreichsten Haushalte liefert uns AnswerTree z.B. folgende Schätzwerte:

- **Gewinn (%)** Es werden 41,5% der interessierten Haushalte erfasst.
- **Treffer: %** Die Rücklaufquote beträgt ca. 1,6%.

Um die 30% aussichtsreichsten Haushalte zusammen zu stellen, wurden die Knoten 9, 3, 10 (zusammen ca. 27,6 % der Gesamtstichprobe) durch zufällig gewählte Haushalte aus Knoten 1 ergänzt.

2.4.1.3 Perzentildiagramme

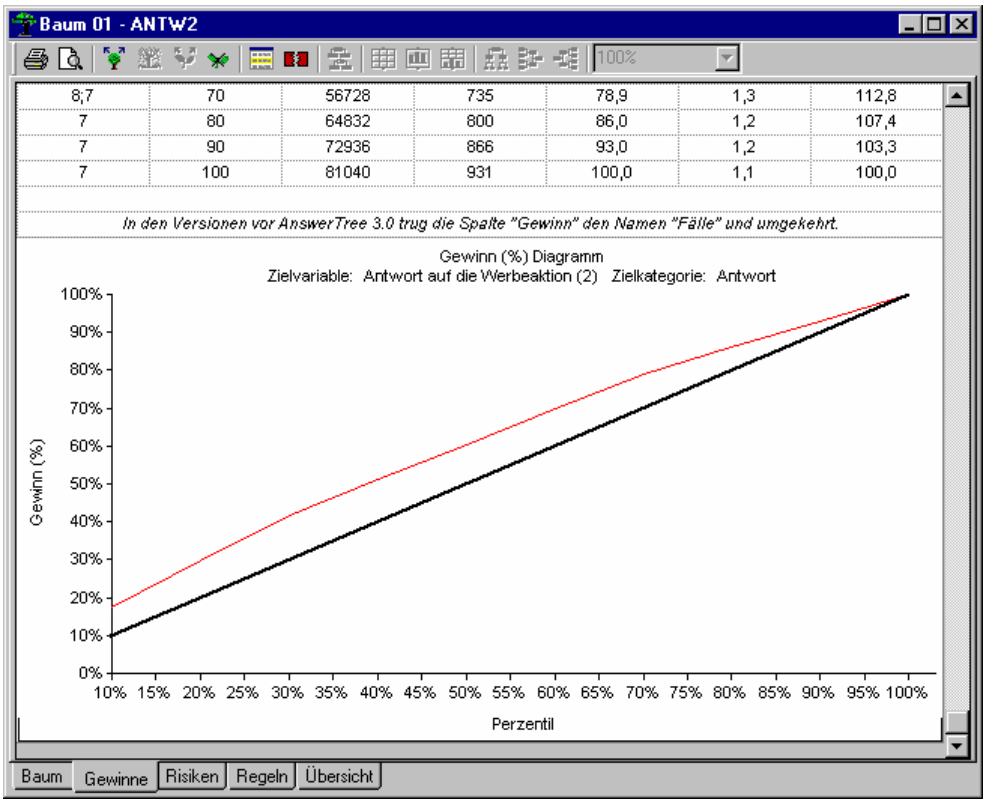
Zu einer perzentil-bezogenen Tabelle können für die Spalten **Gewinn (%)**, **Treffer: %** und **Index (%)** in der Dialogbox **Gewinnübersicht** graphische Darstellungen angefordert werden:



[1] Diese Tabelle wurde folgendermaßen von AnswerTree in das mit Microsoft Word erstellte Manuskript übertragen:

- Auf dem **Gewinne**-Registerblatt des AnswerTree-Baumfensters alle Datenzellen markieren.
- Menübefehl in AnswerTree: **Bearbeiten > Kopieren**
- Menübefehl in Word: **Bearbeiten > Einfügen**
- In Word den eingefügten Text markieren und Menübefehl: **Tabelle > Umwandeln > Text in Tabelle**

Das **Gewinne**-Diagramm zeigt, wie der Anteil der erreichten Interessenten mit dem Anteil der einbezogenen Haushalte wächst, wobei eine Basislinie mit Steigung 1 die Orientierung erleichtert:

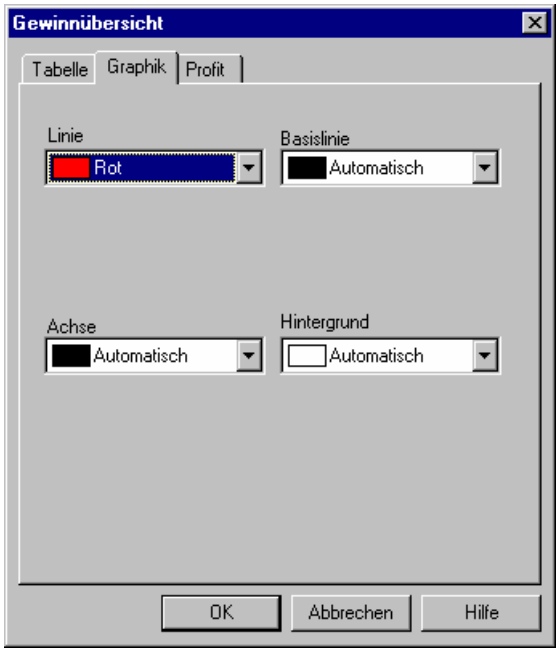


Günstig wäre eine *schnelle* Annäherung an das theoretische Maximum, was eine gute Ausschöpfung des Marktpotentials bei geringem Aufwand ermöglichen würde.

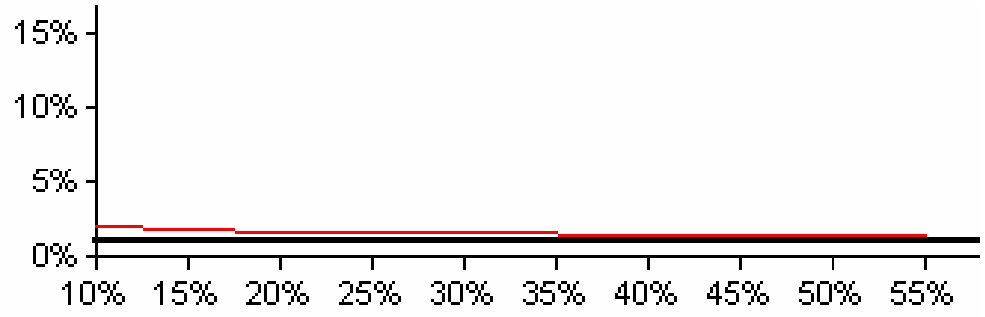
Die Graphiken erscheinen im Baumfenster *unter* der zugehörigen Tabelle und können ...

- im markierten Zustand mit **Bearbeiten > Kopieren** in die Zwischenablage befördert werden,
- mit **Datei > Export** in eine Datei gesichert werden (im BMP- oder EMF-Format).

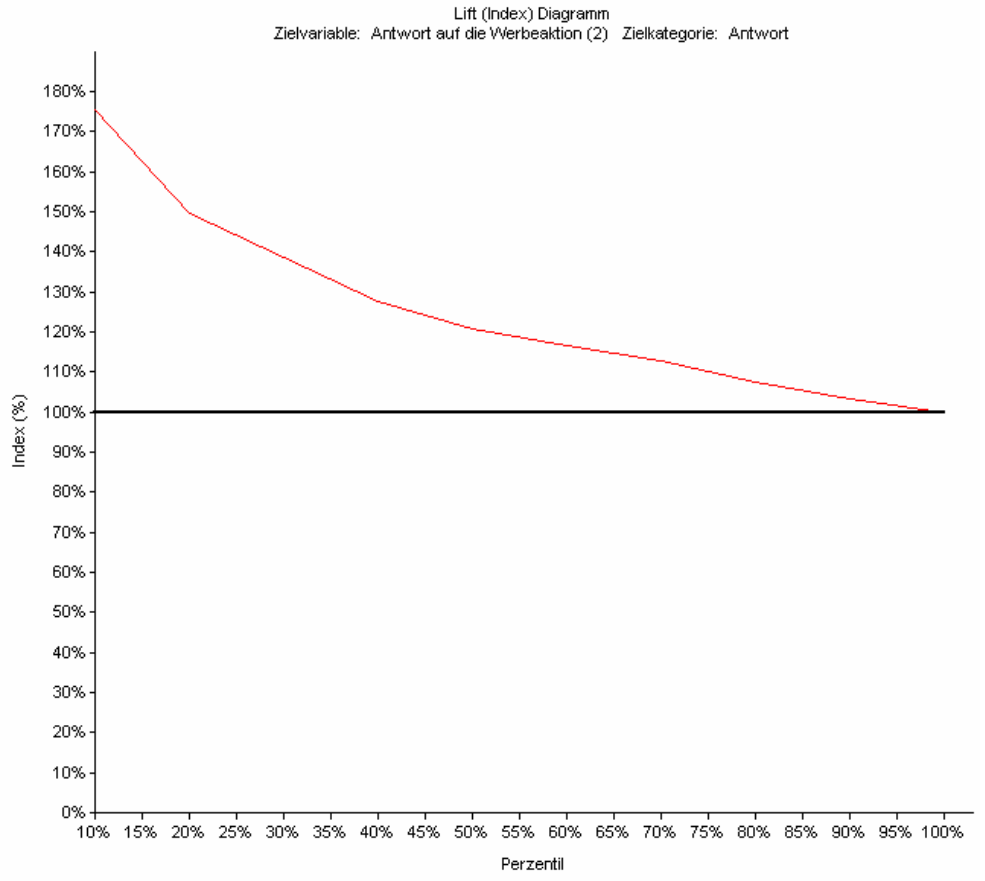
Auf dem **Graphik**-Registerblatt der Dialogbox **Gewinnübersicht** finden sich einige Möglichkeiten zur Gestaltung der Diagramme:



Das **Treffer**-Diagramm zeigt die (in unserem Beispiel sehr kleinen) Trefferraten zu den Perzentilen:



Etwas beeindruckender ist im Beispiel das **Index**-Diagramm, welches die Rücklaufquoten der Perzentile ins Verhältnis zur Gesamtquote setzt:

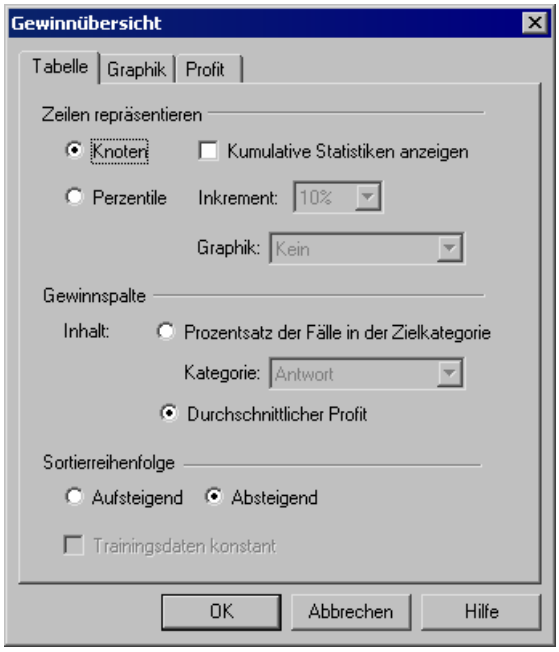


2.4.2 Profit

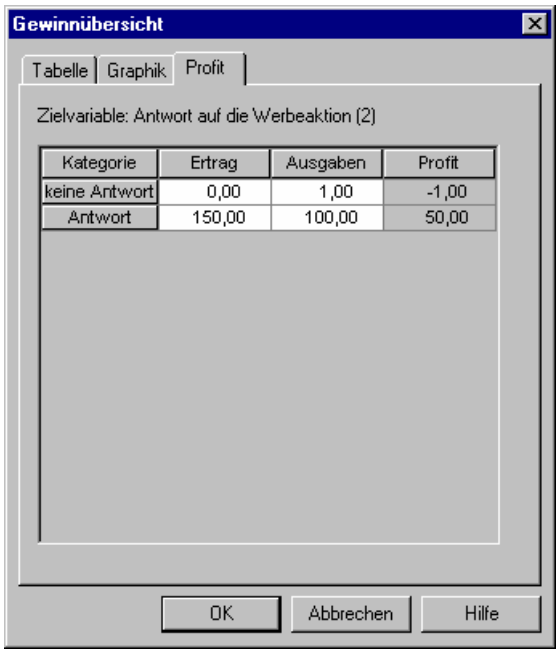
In den bisher diskutierten Gewinnübersichten wurde für Lösungsknoten oder Perzentile mitgeteilt, welcher Anteil ihrer Fälle in der günstigen Zielkategorie liegt. Alternativ kann AnswerTree den **durchschnittlichen Profit** berechnen, der sich aus den Wahrscheinlichkeiten der Zielkategorien sowie aus den jeweiligen Erträgen und Ausgaben ergibt.
Dazu öffnet man mit

Format > Gewinne

die Dialogbox **Gewinnübersicht** und markiert zunächst auf dem Registerblatt **Tabelle** unter **Gewinnspalte** den **durchschnittlichen Profit**:



Dann legt man auf dem Registerblatt **Profit** die Erträge und Ausgaben fest, z.B.:



Für unser Beispiel soll angenommen werden:

- Im Fall einer *Antwort* entsteht ein Ertrag von 150 € (z.B. durch Zahlungen des Abonnenten während der Vertragsdauer), dem Kosten von 100 € (z.B. für gelieferte Zeitschriften und Verwaltung) gegenüberstehen.
- Wer nicht antwortet, bringt keinen Ertrag, verursacht aber Portokosten von 1 €.

Unter dem *Profit* einer Kriteriumskategorie versteht AnswerTree die Differenz

Ertrag - Ausgaben

Bezeichnet man die Profitwerte der m Kriteriumskategorien mit u_i ($i = 1, ..., m$) und die m bedingten Wahrscheinlichkeiten der Kriteriumskategorien im Lösungsknoten a mit $P(i \mid a)$, dann ergibt sich für den Lösungsknoten a der folgende durchschnittliche Profit $U(a)$:

$$U(a) = \sum_{i=1}^{m} P(i \mid a) u_i$$

Im Beispiel erhalten wir für den Knoten 9:

$$U(9) = 0{,}0238908 \cdot (150 - 100) + 0{,}9761092 \cdot (0 - 1) = 0{,}2184308$$

Deaktiviert man auf dem Registerblatt **Tabelle** der Dialogbox **Gewinnübersicht** die **kumulativen Statistiken**, dann resultiert im Beispiel die folgende Tabelle mit den Profitwerten der Lösungsknoten:

Gewinnübersicht					
Zielvariable: Antwort auf die Werbeaktion (2)					
Statistiken					
Knoten	Knoten: Anz.	Knoten: %	Profit	ROI	Index (%)
9	1758	2,2	0,22	6,49	-52,7
3	6198	7,6	-0,02	-0,72	5,0
10	14374	17,7	-0,28	-11,48	66,7
1	25384	31,3	-0,45	-21,45	107,6
8	7795	9,6	-0,45	-21,79	108,8
7	25531	31,5	-0,59	-32,72	142,1

In der **ROI**-Spalte (*Return On Investment*) erscheint die folgendermaßen definierte **Investitionsrentabilität**:

$$\text{ROI} = \frac{\text{Profit}}{\text{Investition}} \cdot 100$$

In unserem Beispiel ergibt sich für den Knoten 9 die mittlere Investition (Ausgabe):

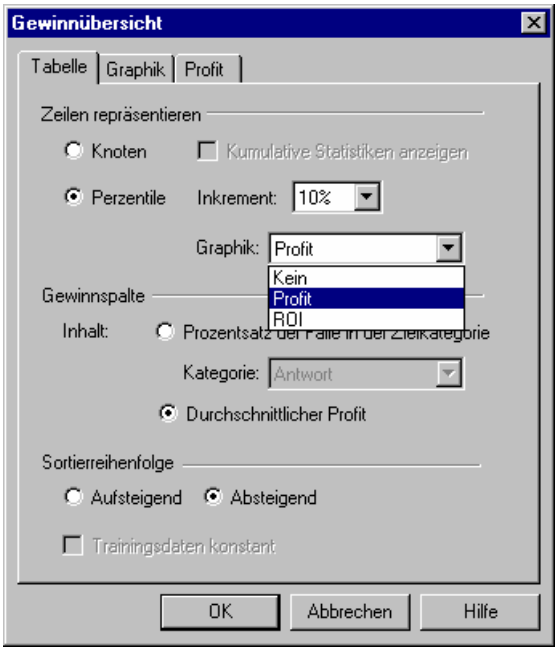
$$0{,}0238908 \cdot 100 + 0{,}9761092 \cdot 1 = 3{,}3651892$$

Damit resultiert die Investitionsrentabilität:

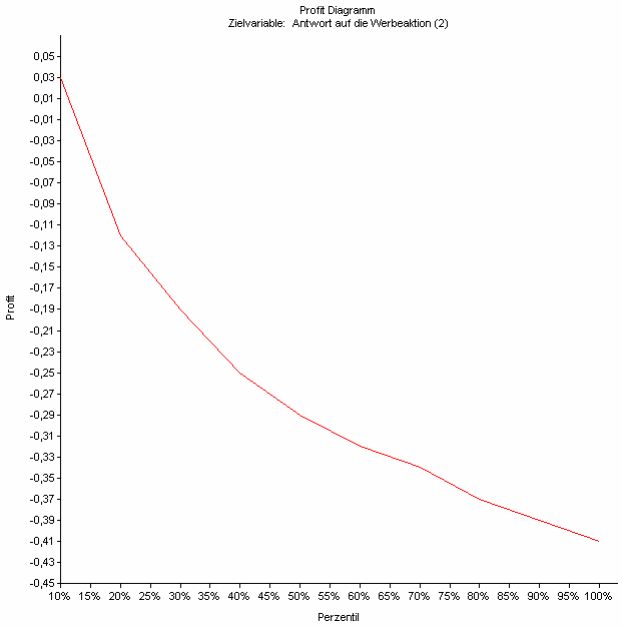
$$\frac{0{,}2184308}{3{,}3651892} \cdot 100 \approx 6{,}49$$

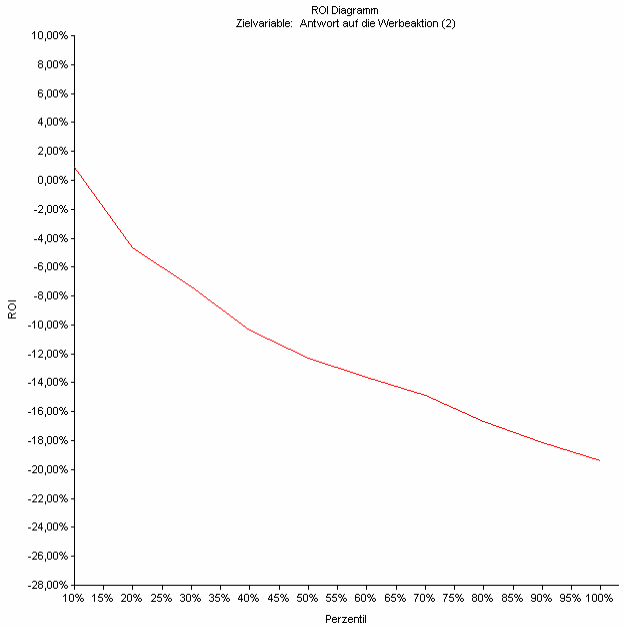
Der letzten Tabellenspalte ist z.B. zu entnehmen, dass im Knoten 7 ca. 1,42 mal mehr Verlust (negativer Profit) zu erwarten ist als in der Gesamtstichprobe. Für den Knoten 9 erhält man die wenig anschauliche Information, dass der hier zu erwartende Profit betragsmäßig ungefähr halb so groß ist wie der für einen beliebigen Fall der Gesamtstichprobe zu erwartende Verlust.

Zu einer perzentil-orientierten Tabelle bietet AnswerTree in der Dialogbox **Gewinnübersicht** auch *graphische* Darstellungen für die Profit- und ROI-Werte an:



Im Beispiel bestätigen **Profit**- und **ROI**-Diagramm, dass sich eine Werbekampagne auf die attraktivsten Knoten bzw. Perzentile beschränken sollte:





Aus den tabellarischen und graphischen Profit- und Rentabilitätsinformationen zu einem Lösungsbaum können kaufmännische Schlussfolgerungen gezogen werden, auf Verlauf und Ergebnis der Segmentierung haben sie jedoch *keinen* Einfluss.
Ähnlich verhält es sich beim CHAID-Algorithmus mit den Klassifikationsfehlern eines Modells und den damit verbundenen Kosten:

2.4.3 Fehlklassifikationen

In unserem Beispiel liegen die Rücklaufquoten aller Lösungsknoten unter 50 %, so dass unter den voreingestellten symmetrischen Fehlklassifikationskosten (siehe Abschnitt 2.2.1) für alle Lösungsknoten ein negatives Ergebnis (keine Antwort) prognostiziert wird. Für jeden Lösungsknoten wird die Entscheidung für eine Kriteriumskategorie nämlich so getroffen, dass die geringsten Fehlklassifikationskosten zu er-

warten sind. Auch beim günstigsten Lösungsknoten (Trefferrate: 2,389 %) resultiert somit eine negative Prognose:

Fehlklassifikationskosten bei positiver Prognose:	0,97611 · 1
Fehlklassifikationskosten bei negativer Prognose:	0,02389 · 1

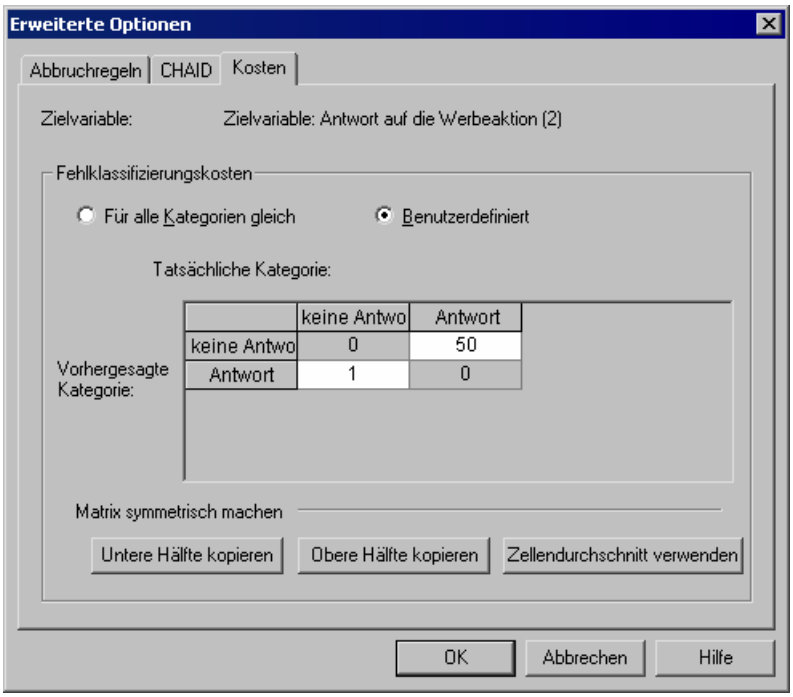
Das **Risiken**-Registerblatt des Baumfensters zeigt demzufolge eine recht langweilige **Fehlklassifikationsmatrix**:

Fehlklassifizierungsmatrix				
		Tatsächliche Kategorie		
		keine Antwort	Antwort	Gesamt
Vorhergesagte Kategorie	keine Antwort	80109	931	81040
	Antwort	0	0	0
	Gesamt	80109	931	81040
		Risikostatistiken		
Risikoschätzung		0,0114882		
Std.f. der Risikoschätzung		0,00037434		

Um eine andere Matrix zu provozieren, wählen wir bei geöffnetem Baumfenster den Menübefehl

Analyse > Erweiterte Optionen

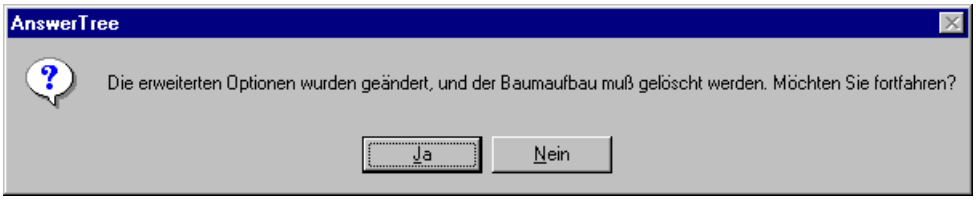
und legen dann neue Fehlklassifikationskosten fest:



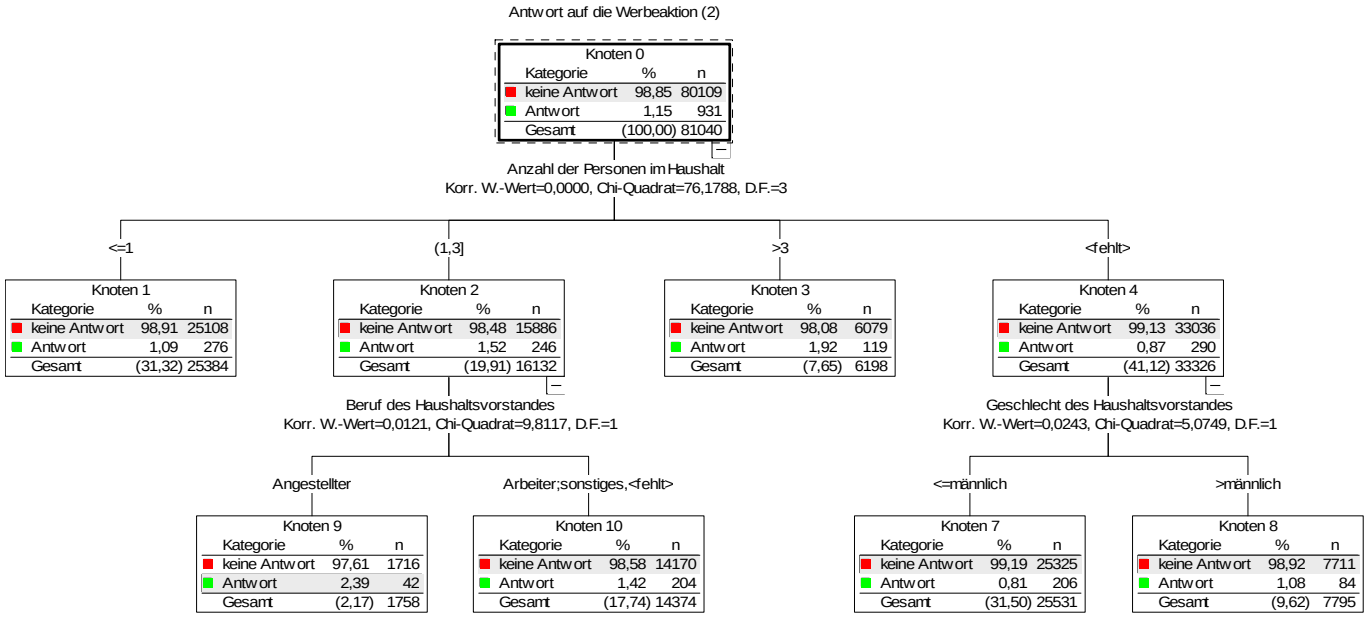
Hier werden falsch-negative Klassifikationen erheblich stärker bestraft als falsch-positive, so dass beim Lösungsknoten 9 eine Entscheidung für die Kategorie *Antwort* nunmehr kostengünstiger ist:

Fehlklassifikationskosten bei positiver Prognose:	0,97611 · 1
Fehlklassifikationskosten bei negativer Prognose:	0,02389 · 50 = 1,1945

Analog zu den Fehlklassifikationskosten lassen sich auch andere Analyse-Optionen jederzeit ändern, wobei AnswerTree nötigenfalls den Entscheidungsbaum anschließend neu aufbaut:



Das wirkt in unserem konkreten Fall zunächst überflüssig und lästig, weil die Segmentierungen einer CHAID-Analyse von den Fehlklassifikationskosten *nicht* beeinflusst werden. Allerdings werden Sie im neu aufgebauten Baum bei genauerem Hinsehen *doch* eine Änderung bemerken. AnswerTree markiert nämlich generell pro Knoten die beste (weil kostengünstigste) Kriteriums-Prognose durch einen grauen Hintergrund. Für den Knoten 9 prognostiziert AnswerTree aufgrund der geänderten Kosten jetzt die Kriteriumskategorie *Antwort*:



Dies zeigt sich auch in der neuen Fehlklassifizierungsmatrix:

Fehlklassifizierungsmatrix				
		Tatsächliche Kategorie		
		keine Antwort	Antwort	Gesamt
Vorhergesagte Kategorie	keine Antwort	78393	889	79282
	Antwort	1716	42	1758
	Gesamt	80109	931	81040
		Risikostatistiken		
Risikoschätzung		0,569669		
Std.f. der Risikoschätzung		0,0182939		

Ferner wird klar, dass mit der *Risikoschätzung* die *erwarteten Fehlklassifikationskosten* pro Entscheidung gemeint sind. AnswerTree summiert über alle *Fehlerzellen* der Klassifikationsmatrix die Produkte aus der geschätzten Wahrscheinlichkeit und den festgelegten Kosten, so dass sich im Beispiel ergibt:

$$\frac{1716}{81040}\cdot 1+\frac{889}{81040}\cdot 50 \approx 0{,}021175\cdot 1+0{,}01097\cdot 50 = 0{,}021175+0{,}5485 = 0{,}569675$$

2.4.4 Profit und Fehlklassifikationskosten als Entscheidungsgrundlage

Die im letzten Abschnitt beschriebene *kostenorientierte* Entscheidungsregel soll nun für den Fall eines *zweistufigen* Kriteriums mit der *profitorientierten* Entscheidungsregel auf Basis der in Abschnitt 2.4.2 beschriebenen Informationen verglichen werden.

Für einen Lösungsknoten *a* die *erste* Kategorie des Kriteriums zu prognostizieren, kann mit einer *Investitionsentscheidung* identifiziert werden (im Beispiel: der Lösungsknoten wird in die Werbekampagne einbezogen). Dabei orientiert man sich am Profit (=Ertrag – Ausgaben) der beiden Kriteriumskategorien (Bezeichnung: u_1, u_2) sowie an ihren bedingten Wahrscheinlichkeiten im Lösungsknoten *a* (Bezeichnung: $P(1 \mid a)$, $P(2 \mid a)$). Man entscheidet sich genau dann für die Investition (also für die Kategorie 1), wenn das Gesamtergebnis $P(1 \mid a)\, u_1 + P(2 \mid a)\, u_2$ wirtschaftlich sinnvoll ist (eine hinreichende Rendite verspricht). Wir wollen der Einfachheit halber annehmen, dass jedes positive Ergebnis akzeptabel ist, so dass für einen Lösungsknoten *a* eine Investitionsentscheidung genau dann erfolgt, wenn gilt:

$$P(1 \mid a)\, u_1 + P(2 \mid a)\, u_2 > 0$$

Bezeichnet man mit

- e_1 die Kosten für eine Fehlinvestition (falsche Entscheidung für Kategorie 1)
- e_2 die Kosten für entgangenen Gewinn (falsche Entscheidung für Kategorie 2)

dann fällt bei Orientierung an den Fehlklassifikationskosten die Entscheidung genau dann zu Gunsten der ersten Kategorie aus, wenn gilt

$$P(2 \mid a)\, e_1 < P(1 \mid a)\, e_2$$
$$\Leftrightarrow$$
$$P(1 \mid a)\, e_2 - P(2 \mid a)\, e_1 > 0$$

Offenbar enden die beiden Entscheidungsregeln z.B. dann mit demselben Ergebnis, wenn gilt:

$$e_1 = -u_2 \text{ und } e_2 = u_1$$

Auf Kriteriumsvariablen mit mehr als zwei Kategorien lassen sich obige Äquivalenzüberlegungen allerdings nicht verallgemeinern: Bei 3 Kategorien enthält die Matrix mit den Fehlklassifikationskosten bereits 6 unabhängige Einträge, denen lediglich 3 Nutzenwerte gegenüber stehen. In dieser Situation können mit Hilfe der Fehlklassifikationskosten differenziertere Entscheidungen erzielt werden.

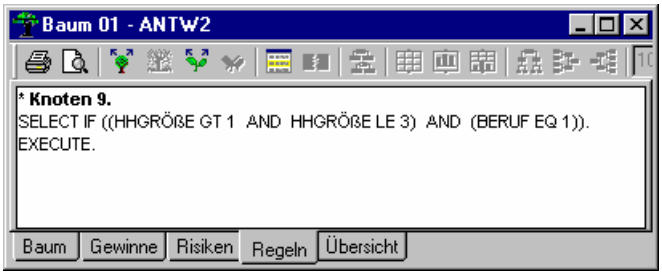
2.4.5 Knotendefinitionen exportieren

Um die identifizierten Segmente mit anderen Programmen näher zu analysieren, kann man im Baumfenster die Definition eines markierten Knotens auf dem Registerblatt **Regeln** anzeigen lassen und dann (z.B. per Zwischenablage) in die Zielanwendung übernehmen. Nach dem Menübefehl

Format > Regeln

lässt sich u.a. der gewünschte Syntax-**Typ** festlegen.

Im Anwendungsbeispiel erhalten wir für den Knoten Nr. 9 folgende SPSS-Syntax:



2.5 Methodische Details zu den CHAID-Verfahren

Anschließend wird der CHAID-Algorithmus (klassische und exhaustiv) für Kriterien mit nominalem, ordinalem oder metrischem Skalenniveau beschrieben (vgl. AnswerTree-Handbuch, SPSS 2002, S. 203ff). Neben dem generellen Vorgehen beim Baumaufbau kommen Details zur Bonferroni-Adjustierung der p-Levels zu optimal rekodierten Prädiktoren zur Sprache.

2.5.1 Baumaufbau

2.5.1.1 Klassische Variante

Ein noch nicht als *final* erkannter Knoten wird folgendermaßen untersucht und ggf. aufgeteilt:

Teilschritt 1: Zusammenlegen von Prädiktorkategorien (Merging)

Jeder verfügbare, d.h. im betrachteten Knoten nicht konstante, Prädiktor wird zunächst optimiert:

- Unter den Kategorien-Paaren, deren Vereinigung nicht verboten ist, wird dasjenige mit dem geringsten Unterschied hinsichtlich der Kriteriumsverteilung bestimmt, beurteilt über das p-Level des zugehörigen Tests. Bei ordinalen oder metrischen Prädiktoren dürfen nur *benachbarte* Kategorien fusionieren, während bei nominalskalierten Prädiktoren beliebige Paare erlaubt sind.
  In Abhängigkeit vom Skaleniveau des Kriteriums werden die p-Levels zur Beurteilung der Kategorien-Unterschiede folgendermaßen berechnet:
  - Metrisches Kriterium
    Es kommt der F-Test zum Einsatz.
  - Ordinales Kriterium
    Es wird ein so genanntes Y-Verknüpfungsmodell (siehe z.B. Magidson 1992) angepasst und per Likelihood-Quotienten-Test beurteilt.
  - Nominales Kriterium
    Es wird der χ^2-Test (mit Pearson- oder Likelihood-Quotienten-Prüfgröße) zur Homogenitäts-Nullhypothese für die zweidimensionale Kontingenztabelle mit dem Kriterium und dem betrachteten Prädiktor gerechnet.

- Ist das maximale p-Level *größer* als der Grenzwert α_{merge} für die Verbindung von Kategorien (siehe **Analyse > Erweiterte Optionen > CHAID**), dann werden die beiden Kategorien zusammengefasst.
- Die Suche nach Fusionskandidaten wird fortgesetzt, bis das maximale p-Level den kritischen Wert α_{merge} nicht mehr übersteigt.

So resultiert für jeden Prädiktor ein überarbeiteter Satz von Kategorien.

Teilschritt 2: Aufteilen des Knotens (Splitting)

- Für alle verfügbaren Prädiktoren wird die Assoziation mit dem Kriterium getestet, wobei die Auswahl des Verfahrens vom Skaleniveau des Kriteriums abhängt (siehe Teilschritt 1). Bei Analysen mit rekodierten Prädiktoren wird der p-Wert Bonferroni-adjustiert, um die künstliche Deflationierung durch die optimale Zusammenlegung von Kategorien auszugleichen. Dieses Thema wird in Abschnitt 2.5.2 noch ausführlich behandelt.
- Liegen p-Werte unterhalb des Grenzwerts α_{split} für die Aufteilung von Knoten vor (siehe **Analyse > Erweiterte Optionen > CHAID**), dann wird der Prädiktor mit dem kleinsten p-Level zur Segmentierung herangezogen.

Stop-Kriterien

Die Segmentierung stoppt, wenn entweder kein signifikanter Prädiktor gefunden wird, oder eine Abbruchregel in Kraft tritt (siehe **Analyse > Erweiterte Optionen > Abbruchregeln**):

- Die maximale Baumtiefe ist erreicht.
- Die minimale Größe für teilbare Knoten ist unterschritten.
- Beim Aufteilen würde die minimale Größe für Unterknoten unterschritten.

Dann liegt ein Endknoten vor.

2.5.1.2 Besonderheiten der exhaustiven Variante

Vom gerade beschriebenen *klassischen* Verfahren unterscheidet sich der **exhaustive CHAID-Algorithmus** (Biggs et al. 1991) nur durch einen erhöhten Aufwand bei der Suche nach optimalen Rekodierungen für die Prädiktoren in Teilschritt 1. Der klassische CHAID-Algorithmus stellt seine Vereinigungsbemühungen ein, sobald die verbliebenen Kategorien sich paarweise signifikant hinsichtlich der Kriteriumsverteilung unterscheiden (alle p-Level kleiner als α_{merge}). Demgegenüber setzt der exhaustive CHAID-Algorithmus die Vereinigung der beiden jeweils ähnlichsten Kategorien so lange fort, bis nur noch zwei Kategorien übrig bleiben. Nach jedem Vereinigungsschritt wird außerdem mit dem aktuellen Kategoriensatz das p-Level zur Assoziation mit dem Kriterium ermittelt. Schließlich wählt der exhaustive CHAID-Algorithmus denjenigen Kategoriensatz aus, der zu einem minimalen p-Level in der Assoziationsanalyse führt.
Außerdem beseitigt die exhaustive Variante einige Ungereimtheiten in der vom klassischen Verfahren eingesetzten Bonferroni-Adjustierung (siehe Seite 32).

2.5.2 Bonferroni-adjustierte Überschreitungswahrscheinlichkeiten

Der CHAID-Algorithmus beginnt in jedem Schritt mit einer assoziations-steigernden Kategorienverschmelzung für jeden Prädiktor: Die I vorhandenen Kategorien eines Prädiktors werden so zu $I' \leq I$ Kategorien zusammengelegt, dass letztlich für die rekodierte Variable ein möglichst kleines p-Level im Assoziationstest bzgl. des Kriteriums resultiert. Bei der Beurteilung dieses p-Levels muss berücksichtigt werden, dass es als vermutliches Minimum aus einer Menge möglicher p-Werte ermittelt wurde. Wir wollen

die Logik und das Vorgehen an einem künstlichen Beispiel mit dem dichotomen Kriterium **AV** sowie den beiden nominalskalierten Prädiktoren **UV1A** und **UV2** untersuchen[1]:

		UV2				Gesamt	
		1		2			
		p	N	p	N	p	N
UV1A	1	,471	170	,471	170	,471	340
	2	,471	170	,471	170	,471	340
	3	,471	170	,471	170	,471	340
	4	,500	160	,600	200	,556	360
	5	,500	160	,600	200	,556	360
Gesamt		,482	830	,527	910	,506	1740

In den mit **p** bezeichneten Spalten sind die relativen Häufigkeiten angegeben, mit denen die Untersuchungsteilnehmer in den Zellen, Zeilen oder Spalten positiv auf eine Anfrage reagiert haben (**AV** = 1). Wir stellen sofort fest, dass sich die ersten drei Kategorien des Prädiktors **UVA1** hinsichtlich der Zustimmungsrate nicht unterscheiden. Dasselbe gilt für die beiden letzten Kategorien von **UVA1**, so dass dieser Prädiktor vermutlich vom CHAID-Algorithmus in der Merging-Phase auf zwei Kategorien reduziert wird.
SPSS liefert für die Kreuztabellenanalyse **UV1A** × **AV** folgende Resultate:

UV1A * AV Kreuztabelle

			AV		
			0	1	Gesamt
UV1A	1	Anzahl	180	160	340
		% von UV1A	52,9%	47,1%	100,0%
	2	Anzahl	180	160	340
		% von UV1A	52,9%	47,1%	100,0%
	3	Anzahl	180	160	340
		% von UV1A	52,9%	47,1%	100,0%
	4	Anzahl	160	200	360
		% von UV1A	44,4%	55,6%	100,0%
	5	Anzahl	160	200	360
		% von UV1A	44,4%	55,6%	100,0%
Gesamt		Anzahl	860	880	1740
		% von UV1A	49,4%	50,6%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	12,190[a]	4	,016
Likelihood-Quotient	12,209	4	,016
Zusammenhang linear-mit-linear	9,188	1	,002
Anzahl der gültigen Fälle	1740		

a. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 168,05.

[1] Die Daten und das AnswerTree-Projekt zum Beispiel finden Sie in den Dateien **Bonferroni.sav** bzw. **Bonferroni.atp** (genaue Bezugsquelle: siehe Einleitung).

Wir erhalten einen Pearson-χ^2-Wert von 12,19 und bei 4 Freiheitsgraden die Überschreitungswahrscheinlichkeit p = 0,016. Diesen Wert können wir erheblich verbessern, indem wir alle Kategorien mit identischer Zustimmungsrate zusammenfassen und so die folgende Variable **UV1B** als rekodierte Variante von **UV1A** bilden:

UV1A		**UV1B**
1	→	1
2	→	1
3	→	1
4	→	2
5	→	2

Die Optimierung wirkt sich natürlich günstig auf die Kreuztabellenanalyse aus:

UV1B * AV Kreuztabelle

			AV		
			0	1	Gesamt
UV1B	1	Anzahl	540	480	1020
		% von UV1B	52,9%	47,1%	100,0%
	2	Anzahl	320	400	720
		% von UV1B	44,4%	55,6%	100,0%
Gesamt		Anzahl	860	880	1740
		% von UV1B	49,4%	50,6%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	12,190[b]	1	,000		
Kontinuitätskorrektur[a]	11,852	1	,001		
Likelihood-Quotient	12,209	1	,000		
Exakter Test nach Fisher				,001	,000
Zusammenhang linear-mit-linear	12,183	1	,000		
Anzahl der gültigen Fälle	1740				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 355,86.

Nun resultiert ein p-Level von 0,000480, dessen Interpretierbarkeit aber durch unsere „Manipulation“ belastet ist. Man kann sich nämlich leicht vorstellen, dass auch bei einem puren Zufalls-„Prädiktor“ mit ebenfalls 5 Stufen durch optimale Rekodierung mit relativ hoher Wahrscheinlichkeit eine „Signifikanz“ erzielt werden kann. Der artifiziellen p-Level-Schrumpfung können wir mit einer Bonferroni-Adjustierung zu Leibe rücken, die in Rechnung stellt, aus wie vielen potentiellen zweidimensionalen Tabellen bzw. **UV1A**-Rekodierungen die minimale Überschreitungswahrscheinlichkeit gewählt wurde.

Werden *M* unabhängige* Tests zum Niveau α' bei Gültigkeit aller Nullhypothesen durchgeführt, dann resultiert folgende Wahrscheinlichkeit, mindestens einen Fehler erster Art zu begehen:

$$1-(1-\alpha')^{M^*}$$

Bei *abhängigen* Tests kann die globale Fehlerwahrscheinlichkeit nicht exakt ausgerechnet, aber auf jeden Fall nach oben abgeschätzt werden durch

$$M^* \cdot \alpha'$$

Das simple Prinzip einer Bonferroni-Adjustierung besteht darin, das Globalrisiko durch α zu begrenzen, indem jeder Einzeltest zum Niveau

$$\alpha' := \frac{\alpha}{M^*}$$

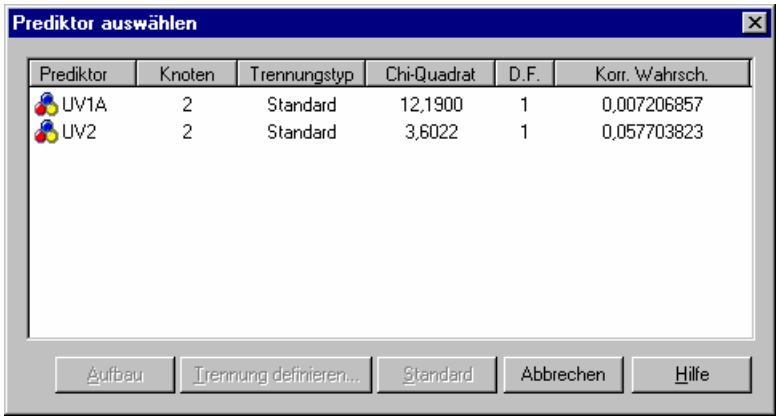
durchgeführt wird. Bei einer auf analoge Weise adjustierten Prädiktor-Beurteilung im CHAID-Algorithmus muss die für einen optimal rekodierten Prädiktor ermittelte Überschreitungswahrscheinlichkeit p' mit der Anzahl M^* der möglichen Zerlegungen multipliziert werden, um ein vergleichbares p-Level zu erhalten:

$$p := M^* p'$$

Zur Frage, wie das M^* zur Merging-Phase des CHAID-Algorithmus zu bestimmen ist, bestehen unterschiedliche Auffassungen, die im Aufsatz von Biggs et al. (1991) ausführlich diskutiert werden. Unbestritten ist zunächst, dass bei einer nominalskalierten Variablen mit I Kategorien die Anzahl der möglichen Rekodierungen zu I' Kategorien nach folgender Formel bestimmt werden kann (siehe z.B. Magidson & SPSS 1993, S. 133)[1]:

$$\mathrm{M}(I, I') = \sum_{i=0}^{I'-1} (-1)^i \frac{(I'-i)^I}{i!(I'-i)!} \qquad (1)$$

Für unser künstliches Beispiel mit $I = 5$ und $I' = 2$ ergibt sich z.B. M(5, 2) = 15.
Im klassischen CHAID-Algorithmus nach Kass (1980) wird die in der Merging-Phase (datenabhängig!) ermittelte Kategorienzahl I' als einzig relevant betrachtet und M^* mit dem zugehörigen M(I, I') gleich gesetzt. Alle Rekodierungen mit abweichender Kategorienzahl bleiben also bei der Ermittlung des Bonferroni-Faktors unberücksichtigt, obwohl sie ebenfalls zur Wahl standen.
Dementsprechend liefert AnswerTree bei einem CHAID-Entscheidungsbaum zu den oben beschriebenen künstlichen Daten (Kriterium **AV**, Prädiktoren **UV1A** und **UV2**) für die Aufteilung des Wurzelknotens folgende Entscheidungshilfen:



Zu dem dichotomisierten Prädiktor UV1A wird als korrigierte Wahrscheinlichkeit das 15-fache des p-Levels aus der **UV1B**×**AV** – Kontingenztabelle geliefert:

$$0{,}0072 = 15 \cdot 0{,}00048$$

Dieses Vorgehen ist *nicht* korrekt, weil dem Merging-Algorithmus auch die Lösungen mit 3, 4 oder 5 Gruppen zur Verfügung standen, nach obiger Formel (1) in folgenden Häufigkeiten:

[1] Die Formel (1) zur Berechnung der Anzahl möglicher Rekodierungen bei vorgegebener Gruppenanzahl gilt nur für nominalskalierte Prädiktoren. Bei ordinalskalierten Prädiktoren sind bei der Gruppenverschmelzung Restriktionen zu beachten, so dass die Anzahl zulässiger Rekodierungen kleiner ausfällt (siehe Magidson & SPSS 1993, S. 133).

I' = 3: 24
I' = 4: 10
I' = 5: 1

Die im klassischen CHAID-Algorithmus verwendete Methode zur Bestimmung des Bonferroni-Faktors hat nicht nur eine Rechtfertigungs-Lücke, sondern auch eine sachlich kaum begründbare Abneigung gegen Prädiktoren, die aus der Merging-Phase mit einer mittleren Kategorienzahl hervorgegangen sind. Wie im Beispiel zu sehen war, führt bei einer ehemals fünfstufigen Variablen die Rekodierung auf 3 Stufen zu einem besonders großen Bonferroni-Faktor.

Im *exhaustiven* CHAID-Algorithmus berechnet AnswerTree den Bonferroni-Faktor nach einem Vorschlag von Biggs et al. (1991), wobei die Nachteile der Kass-Methode vermieden werden:

- Der Bonferroni-Faktor stellt *alle* im Merging-Prozeß realisierbaren Kategorienzahlen in Rechnung.
- Weil der Bonferroni-Faktor zu einem Prädiktor nur von seiner *ursprünglichen* Kategorienzahl abhängt, wird keine rekodierte Kategorienzahl benachteiligt oder bevorzugt.

Außerdem begründen die Autoren mit wahrscheinlichkeitstheoretischen Argumenten und Simulationsdaten, dass der nahe liegende Bonferroni-Faktor

$$\sum_{i=2}^{I} \mathrm{M}(I,i)$$

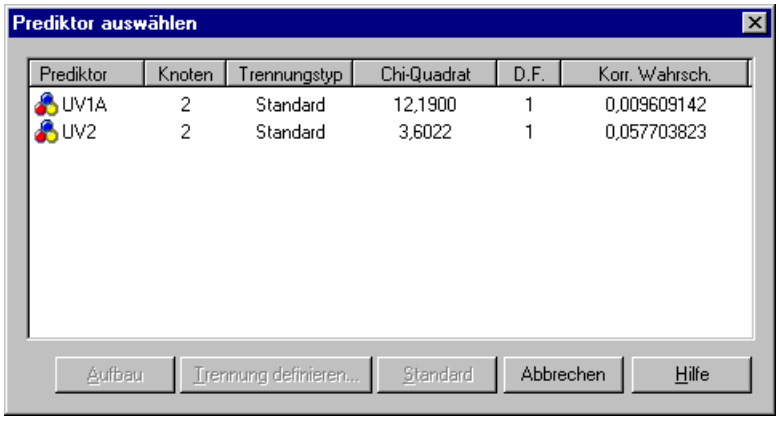
zu konservativ ist. Ihr im exhaustiven CHAID-Algorithmus realisierter Alternativ-Vorschlag lautet:

$$\mathrm{M}_{\mathrm{EC}}(I) = 1 + \sum_{i=1}^{I-2} \mathrm{M}(I-i+1, I-i)$$

Für den fünfstufigen Prädiktor UV1A in unserem künstlichen Datensatz ergibt sich:

$$\begin{aligned} \mathrm{M}_{\mathrm{EC}}(5) &= 1 + \mathrm{M}(5,4) + \mathrm{M}(4,3) + \mathrm{M}(3,2) \\ &= 1 + 10 + 6 + 3 \\ &= 20 \end{aligned}$$

Dementsprechend liefert AnswerTree bei einem Exhaustive-CHAID-Entscheidungsbaum zu den oben beschriebenen künstlichen Daten (Kriterium **AV**, Prädiktoren **UV1A** und **UV2**) für die Aufteilung des Wurzelknotens folgende Entscheidungshilfen:



Zum dichotomisierten Prädiktor **UV1A** wird als korrigierte Wahrscheinlichkeit das 20-fache des p-Wertes aus der **UV1B**×**AV** – Kontingenztabelle geliefert:

$$0{,}0096 = 20 \cdot 0{,}00048$$

3 Das C&RT-Verfahren

Das von Breiman et al. (1984) entwickelte C&RT-Verfahren[1] (**C**lassification & **R**egression **T**rees) unterstützt nominale (Classification) und metrische (Regression) Kriterien.

3.1 Aufbau eines binären Baumes durch Reduktion von Inhomogenität

Pro Analyseschritt wird jeweils ein Knoten mit Hilfe eines Prädiktors in genau *zwei* homogenere Unterknoten aufgeteilt. Die Aufteilung wird nicht (wie bei den CHAID-Verfahren) durch die empirischen Überschreitungswahrscheinlichkeiten von Assoziationstests gesteuert, sondern über ein zu minimierendes Inhomogenitätsmaß. In Abhängigkeit von Skalenniveau des Kriteriums sind folgende Maße verfügbar:

Inhomogenitätsmaß	**vorausgesetztes Skalenniveau des Kriteriums**
Gini	nominal
Twoing	nominal
ordinales Twoing	ordinal (geordnete Kategorien)
LSD (Least Squared Deviation)	metrisch

Während die gleich präsentierten Beschreibungen zum Gini- und zum LSD-Inhomogenitätsmaß einige technische Details enthalten, ist die Grundlogik des C&RT-Verfahrens sehr einfach:

- Zum Wurzelknoten wird diejenige Zerlegung in *zwei* Unterknoten gesucht, welche den stärksten Abfall der Inhomogenität ergibt. Dazu müssen die Inhomogenitäten der entstehenden Unterknoten für *alle möglichen Dichotomisierungen jedes verfügbaren Prädiktors* bestimmt werden.
- Mit demselben Verfahren wird sukzessive auch für die jeweils entstandenen Unterknoten eine optimale Zerlegung gesucht.
- Eine C&RT-Analyse endet, wenn bei keiner weiteren Knotenzerlegung die festgelegte Mindeständerung der Inhomogenität erreicht wird (siehe **Analyse > Erweiterte Optionen > Mindeständerung**) oder eine andere Abbruchregel erfüllt ist (vgl. Abschnitt 2.2.1).

3.1.1 Nominalskalierte Kriterien

3.1.1.1 Gini-Index

Der **Gini-Index** g(a) für einen Knoten a wird folgendermaßen definiert:

$$g(a) := \sum_{i \neq j} P(i \mid a)\, P(j \mid a)$$

Dabei ist $P(i \mid a)$ die bedingte Wahrscheinlichkeit der Kriteriumskategorie i im Knoten a.
Durch elementare Umformungen erhält man einen alternativen Ausdruck für g(a):

$$\begin{aligned}\sum_{i \neq j} P(i \mid a)\, P(j \mid a) &= \sum_i \sum_{j \neq i} P(i \mid a)\, P(j \mid a) = \sum_i P(i \mid a) \sum_{j \neq i} P(j \mid a) \\ &= \sum_i P(i \mid a)\,(1 - P(i \mid a)) = \sum_i P(i \mid a) - \sum_i P(i \mid a)^2 \\ &= 1 - \sum_i P(i \mid a)^2\end{aligned}$$

[1] Oft wird auch das Kürzel *CART* verwendet.

Der Gini-Index erreicht seinen minimalen Wert 0, wenn alle Fälle im Knoten *a* zur *selben* Kategorie der Zielvariablen gehören. Wenn sich die Fälle eines Knotens hingegen gleichmäßig auf die *m* Kategorien der Zielvariablen verteilen, resultiert der Maximalwert $1-\frac{1}{m}$.

Zur Bestimmung der optimalen Aufteilung eines Knotens *a* in zwei Unterknoten a_1 und a_2 wird im C&RT-Verfahren *für jede mögliche Dichotomisierung t jedes verfügbaren Prädiktors* folgende **Gini-Zielfunktion** $\Phi(t, a)$ ausgewertet:

$$\Phi(t,a) := g(a) - P(a_1 \mid a)\,g(a_1) - P(a_2 \mid a)\,g(a_2)$$

Dabei ist $P(a_1 \mid a)$ bzw. $P(a_2 \mid a)$ die bedingte Wahrscheinlichkeit des ersten bzw. zweiten Unterknotens, und

$$P(a_1 \mid a)\; g(a_1) + P(a_2 \mid a)\; g(a_2)$$

ist dementsprechend das gewichtete Mittel der Inhomogenitäten in den beiden Unterknoten. Der Algorithmus wählt die Trennung *t* mit dem maximalen $\Phi(t, a)$-Wert.

Gewichtet mit der relativen Häufigkeit von Knoten *a* gibt AnswerTree dieses Maximum als **Verbesserung** im Baumdiagramm aus.

Zur Demonstration der Berechnungen betrachten wir ein Beispiel[1] mit dreistufigem, gleichverteiltem Kriterium **Y**, dessen Wurzelknoten die maximale Inhomogenität

$$1-\frac{1}{3}=\frac{2}{3}$$

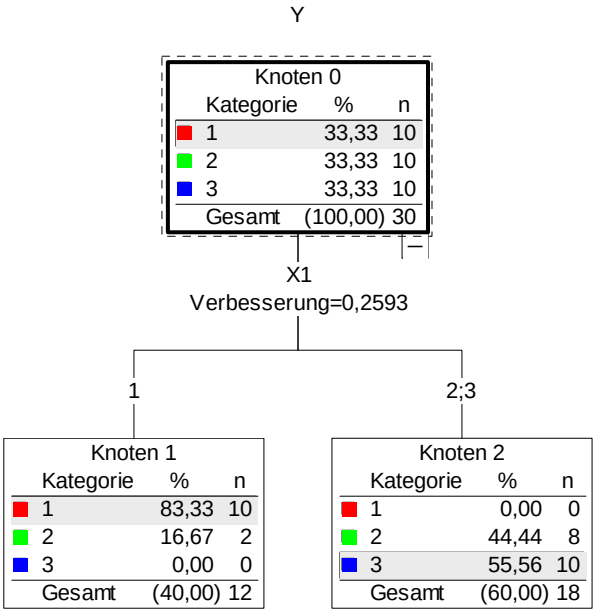
besitzt. In der per Wurzelknoten-Kontexmenü oder

Baum > Prediktor auswählen

verfügbaren Dialogbox **Prediktor auswählen** werden für die beiden Prädiktoren **X1** und **X2** folgende Verbesserungen angegeben:

Ein einstufiger Baumaufbau unter Verwendung des besten Prädiktors führt zu folgenden Unterknoten:

[1] Die Daten und das AnswerTree-Projekt zum Beispiel finden Sie in den Dateien **Gini.sav** bzw. **Gini.atp** (genaue Bezugsquelle: siehe Einleitung).



Knoten 1 besitzt die Gini-Inhomogenität:

$$1-\left(\frac{5}{6}\right)^2-\left(\frac{1}{6}\right)^2=1-\frac{25}{36}-\frac{1}{36}=\frac{10}{36}=0,2\overline{7}$$

Knoten 2 besitzt die Gini-Inhomogenität:

$$1-\left(\frac{4}{9}\right)^2-\left(\frac{5}{9}\right)^2=1-\frac{16}{81}-\frac{25}{81}=\frac{40}{81}\approx 0,494$$

Als gewichtetes Mittel erhalten wir:

$$0,4\cdot 0,2\overline{7}+0,6\cdot 0,494\approx\cdot 0,4074$$

Daraus ergibt sich die Verbesserung:

$$0,\overline{6}-0,4074\approx 0,2593$$

3.1.1.2 A-priori – Wahrscheinlichkeiten

Zur Schätzung der bedingten Wahrscheinlichkeiten $P(i\,|\,a)$, die den Gini-Index und somit den Baumaufbau bestimmen, können a-priori – Wahrscheinlichkeiten $P(i)$ der Kriteriumskategorien (meist als $\pi(i)$ notiert) herangezogen werden, die von den relativen Häufigkeiten der Trainingsstichprobe abweichen. So entsteht ein Entscheidungsbaum mit optimaler Eignung für Klassifikationsaufgaben in einer Population mit den vorgegebenen a-priori – Wahrscheinlichkeiten.
Für die bedingte Wahrscheinlichkeit der i-ten Kriteriumskategorie im Knoten a gilt definitionsgemäß:

$$P(i\,|\,a)=\frac{P(i,a)}{P(a)}$$

Dabei ist $P(a)$ die Wahrscheinlichkeit von Knoten a sowie $P(i,a)$ die gemeinsame Wahrscheinlichkeit von Kriteriumskategorie i und Knoten a. Letztere lässt sich über die a-priori – Wahrscheinlichkeit $\pi(i)$ sowie die bedingte Wahrscheinlichkeit $P(a\,|\,i)$ berechnen:

$$P(i,a)=\pi(i)\,P(a\,|\,i)$$

$P(a\,|\,i)$ schätzt man aus der Trainingsstichprobe:

$$\hat{P}(a \mid i) = \frac{N(i,a)}{N(i)}$$

Dabei ist N(i) die Häufigkeit der i-ten Kriteriumskategorie und N(i,a) die Häufigkeit von Knoten a in der der i-ten Kriteriumskategorie. Insgesamt schätzt man P(i,a) also durch:

$$\hat{P}(i,a) = \pi(i)\,\frac{N(i,a)}{N(i)}$$

Durch Summieren über alle Kriteriumskategorien gewinnt man eine Schätzung für P(a):

$$\hat{P}(a) = \sum_i \hat{P}(i,a)$$

Somit erhält man Schätzungen für die bedingten Wahrscheinlichkeiten $P(i \mid a)$ passend zur Population mit den vorgegebenen a-priori – Wahrscheinlichkeiten.

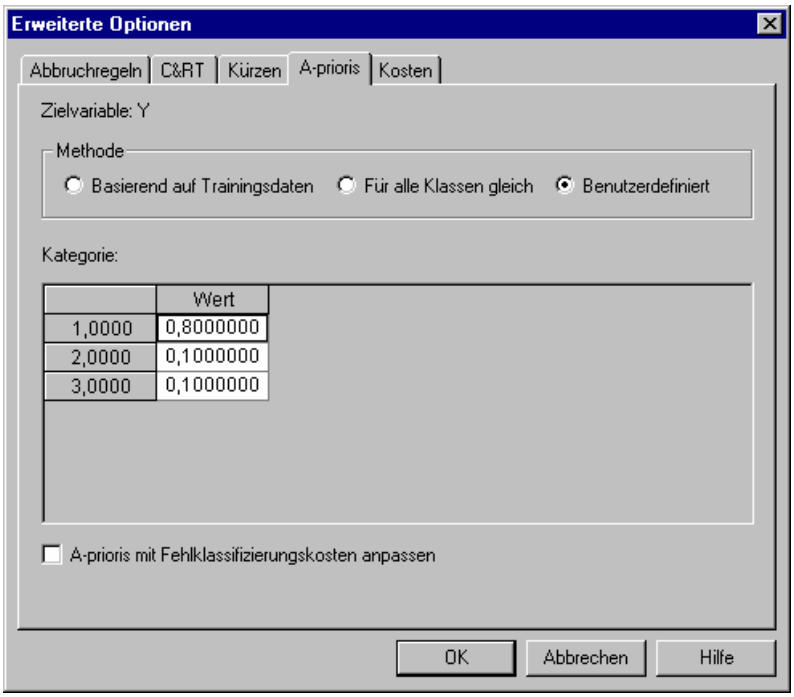
An einem Beispiel mit dreistufigem Kriterium **Y** soll demonstriert werden, wie sich a-priori – Wahrscheinlichkeiten auf die Verbesserungs-Bewertungen von Prädiktoren auswirken.[1] Bei der Aufteilung des Wurzelknotens stehen zwei Prädiktoren **X1** und **X2** zur Verfügung:

- Prädiktor **X1** trennt die Kriteriumskategorie 1 perfekt von den restlichen Kategorien, zwischen denen er aber nicht mehr differenzieren kann.
- Prädiktor **X2** trennt die Kriteriumskategorie 2 perfekt von den restlichen Kategorien, zwischen denen er aber nicht mehr differenzieren kann.

In einem AnswerTree-Projekt sollen nach

Analyse > Erweiterte Optionen

auf der Registerkarte **A-prioris** folgende benutzerdefinierte Wahrscheinlichkeiten festgelegt werden:



Über die Kontextmenü-Option **Prediktor auswählen** zum Wurzelknoten wird eine Dialogbox mit den Verbesserungs-Fähigkeiten der Prädiktoren angefordert:

[1] Die Daten und das AnswerTree-Projekt zum Beispiel finden Sie in den Dateien **GiniPriors.sav** bzw. **GiniPriors.atp** (genaue Bezugsquelle: siehe Einleitung).

Erwartungsgemäß wird derjenige Prädiktor deutlich besser bewertet (und infolgedessen als Trennvariable bevorzugt), welcher die Kategorie mit der höchsten a-priori – Wahrscheinlichkeit perfekt separieren kann.
Unter den alternativen a-priori – Wahrscheinlichkeiten

	Wert
1,0000	0,1000000
2,0000	0,8000000
3,0000	0,1000000

ergibt sich das umgekehrte Bild:

3.1.1.3 Fehlklassifikationskosten

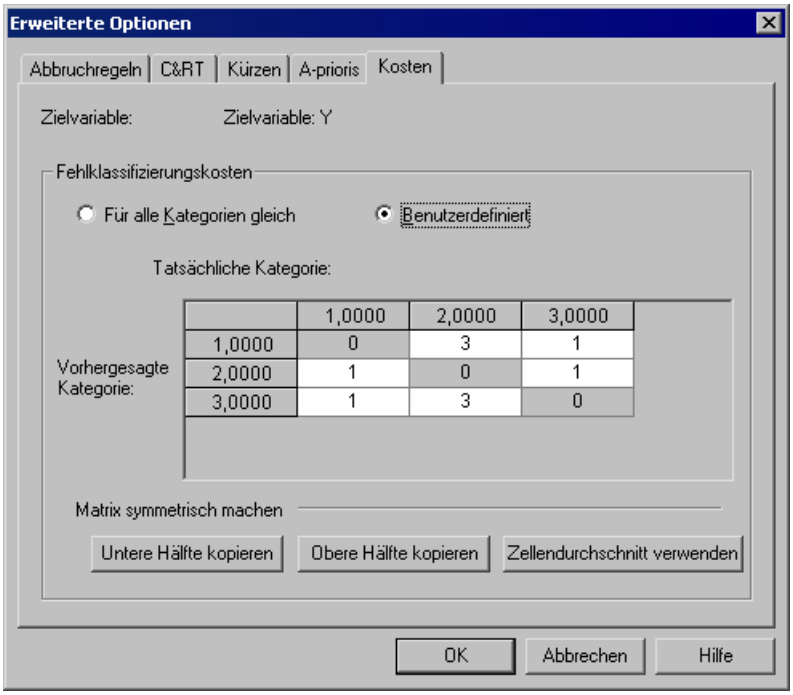
Während sich Fehlklassifikationskosten bei den CHAID-Verfahren nur auf dem **Risiken**-Registerblatt zur Beurteilung einer Lösung auswirken (siehe Abschnitt 2.4.3), steuern sie beim C&RT - (und beim QUEST -) Verfahren auch den Baumaufbau.
Im Gini-Index werden die Kosten der fehlerhaften Klassifikation eines Falles aus Kriteriumskategorie *j* in die Kategorie *i* folgendermaßen berücksichtigt:

$$g(a) := \sum_{i \neq j} C(i \mid j)\, P(i \mid a)\, P(j \mid a)$$

Wir greifen das Beispiel aus Abschnitt 3.1.1.2 wieder auf, stornieren aber die a-priori – Wahrscheinlichkeiten und vereinbaren stattdessen nach

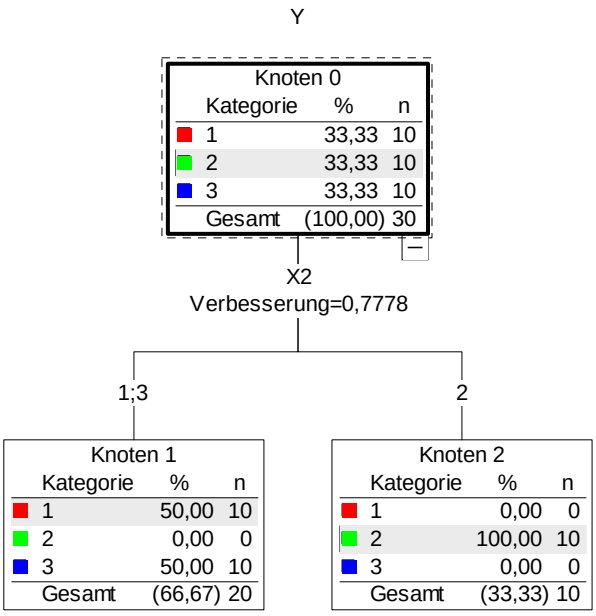
Analyse > Erweiterte Optionen

benutzerdefinierte Kosten:



Weil es besonders teuer ist, einen Fall aus Kategorie 2 falsch einzuordnen, wird der als perfekter Kat-2-Detektor bekannte Prädiktor **X2** deutlich besser bewertet als sein Konkurrent

und infolgedessen als Trennvariable bevorzugt:



Wie bei der CHAID-Analyse (vgl. Abschnitt 2.4.3) ist auf dem Registerblatt **Risiken** dokumentiert, welche Fehler sich aus den (kostenminimierenden) Klassifikationsentscheidungen des Baumes ergeben. Im Beispiel kann bereits der einstufige (noch unvollendete) Baum die teuren Fehlentscheidungen vermeiden:

Fehlklassifizierungsmatrix					
		Tatsächliche Kategorie			
		1	2	3	Gesamt
Vorhergesagte Kategorie	1	10	0	10	20
	2	0	10	0	10
	3	0	0	0	0
	Gesamt	10	10	10	30

	Risikostatistiken
Risikoschätzung	0,333333
Std.f. der Risikoschätzung	0,0860663

Die erwarteten Fehlklassifikationskosten pro Entscheidung (*Risikoschätzung*) betragen:

$$\frac{0\cdot 1+0\cdot 3+10\cdot 1}{30}=0,\overline{3}$$

3.1.2 Metrische Kriterien

Bei metrischen Kriterien wird der Gini-Index durch die Varianz im Knoten *a* ersetzt, die bei Berücksichtigung einer Häufigkeitsvariablen *f* und einer Gewichtungsvariablen *w* (vgl. Abschnitt 2.2.1) folgendermaßen definiert ist (mit Kriteriumsvariable *y*):

$$v(a) := \frac{1}{\sum_{i\in a} f_i} \sum_{i\in a} f_i (y_i - \bar{y}_w(a))^2 \quad \text{mit} \quad \bar{y}_w(a) := \frac{1}{\sum_{i\in a} f_i w_i} \sum_i f_i w_i y_i$$

Bei der Inhomogenitätsberechnung wird also nicht das Stichprobenmittel verwendet, sondern das unter Berücksichtigung der Ziehungsraten geschätzte Populationsmittel. Man bezeichnet v(*a*) auch als **LSD-Index** (Least Squared Deviation) zum Knoten *a*.

Zur Demonstration der Berechnung betrachten wir den Wurzelknoten eines einfachen Beispiel, dessen Daten von AnswerTree nach dem Menübefehl

Ansicht > Daten

in folgender Dialogbox angezeigt werden:

Datenanzeige

	y	x1	f	g
1	-,91	1,00	1,00	1,00
2	-,04	1,00	2,00	1,00
3	-,28	1,00	3,00	1,00
4	-,36	1,00	4,00	1,00
5	-1,86	1,00	5,00	1,00
6	,23	2,00	1,00	10,00
7	1,68	2,00	2,00	10,00
8	3,63	2,00	3,00	10,00
9	1,81	2,00	4,00	10,00
10	1,68	2,00	5,00	10,00

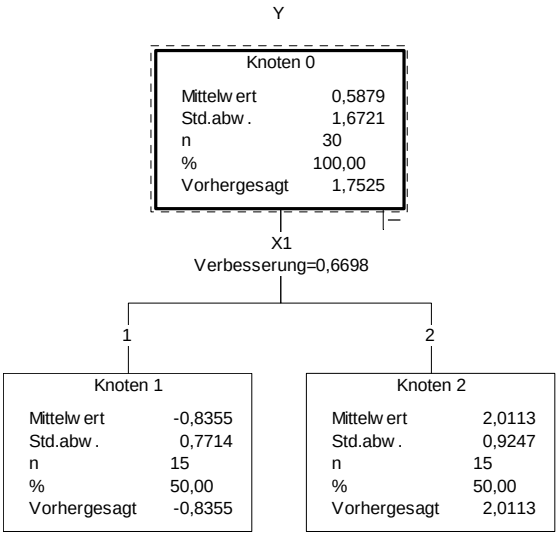
Als Kriteriums-Wurzelknotenmittel nach der Formel $\bar{y}(a) := \frac{1}{\sum_{i \in a} f_i} \sum_{i \in a} f_i \, y_i$ erhält man 0,5879. Bei der Kriteriumsprognose berücksichtigt AnswerTree aber die Ziehungsraten und berechnet $\bar{y}_w(a)$:

Über die Inhomogenität v(a) des Wurzelknotens kann man sich auf dem **Risiken**-Registerblatt informieren:

An die Stelle der Gini-Zielfunktion für nominale Kriterien tritt die **LSD-Zielfunktion**, die auf der Suche nach einer optimalen Aufteilung des Knotens a in zwei Unterknoten a_1 und a_2 mit den bedingten Wahrscheinlichkeiten P(a_1| a) und P(a_2| a) für jede mögliche Dichotomisierung t jedes verfügbaren Prädiktors auszurechnen ist:

$$\Phi(t,a) := \mathrm{v}(a) - \mathrm{P}(a_1 \mid a)\,\mathrm{v}(a_1) - \mathrm{P}(a_2 \mid a)\,\mathrm{v}(a_2)$$

Im Beispiel kommt nur *eine* Aufteilung in Frage:



Die Inhomogenitäten in den Unterknoten betragen 0,5553 bzw. 0,7982, so dass ein gewichteter Mittelwert von 0,676744 resultiert, den AnswerTree auch als neue Risikoschätzung für den Baum (bzw. seine Endknoten) angibt:

	Risikostatistiken
Risikoschätzung	0,676744
Std.f. der Risikoschätzung	0,162884

Weil die im Baumdiagramm angegebenen knoteninternen Standardabweichungen *erwartungstreu* berechnet sind, muss man wie im folgenden Beispiel vorgehen, um den LSD-Index gemäß obiger Definition daraus zu berechnen:

$$\frac{0{,}7714^2 \cdot (15-1)}{15} \approx 0{,}555$$

Das als *Risikoschätzung* mitgeteilte gewichtete Mittel der Endknoten-Inhomogenitäten kann übrigens auch als **gepoolte Varianz** der Endknoten aufgefasst werden (vgl. Abschnitt 3.5):

$$\frac{0{,}7714^2 \cdot (15-1) + 0{,}9247^2 \cdot (15-1)}{15+15} \approx \frac{8{,}3308 + 11{,}9710}{30} \approx 0{,}6767$$

Das mit dem Anteil aller Fälle im aufgeteilten Knoten gewichtete Maximum der LSD-Zielfunktion erscheint im Baumdiagramm als **Verbesserung**[1].

3.1.3 Ersatzvariablen bei fehlenden Werten

Beim C&RT- sowie beim QUEST-Algorithmus wird das Problem fehlender Prädiktorwerte auf recht überzeugende Weise gelöst. Dazu ermitteln die Algorithmen bei jeder Knotenaufteilung zum Sieger des Prädiktor-Auswahlverfahrens mehrere Ersatzvariablen mit möglichst hoher **prädiktiver Assoziation**. Mit diesem Begriff ist die Fähigkeit der Ersatzvariablen gemeint, das Trennverhalten des Siegers zu imitieren, also bei möglichst vielen Fällen dieselbe Unterknoten-Zuordnung vorzunehmen.

Ist der *Wurzelknoten* aufzutrennen, geht die prädiktive Assoziation einher mit der einfachen (bivariaten) Assoziation zwischen den Ersatzvariablen und dem Meister-Separator. Ansonsten *divergieren* beide Begriffe, weil für die prädiktive Assoziation nur das lokale Verhalten im aufzutrennenden Knoten relevant ist.
Besitzt ein Fall bei einer Aufteilung keinen Wert für den primären Prädiktor, kommt die beste Ersatzvariable zum Einsatz, deren Wert beim fraglichen Fall vorhanden ist.

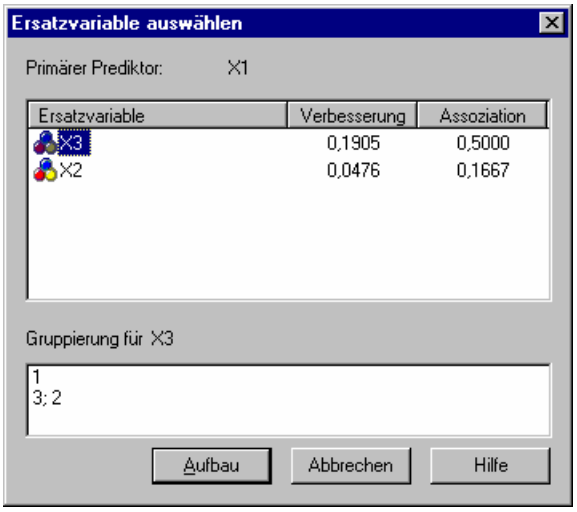
Über den Eintrag **Ersatzvariable auswählen** im Kontextmenü eines aufgeteilten Knotens bzw. über den Menübefehl

Baum > Ersatzvariable auswählen

ist in AnswerTree die folgende Dialogbox verfügbar, die für alle Surrogatvariablen u.a. die prädiktive Assoziation enthält, z.B.:[2]

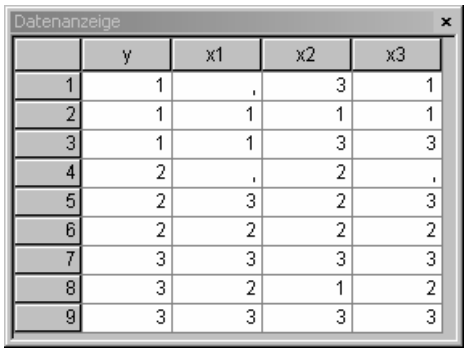
[1] Bei Abwesenheit einer Gewichtungsvariablen sind die von AnswerTree berichteten Verbesserungen unmittelbar nachvollziehbar. *Mit* Fallgewichten (wie im Beispiel) ist zumindest das Ergebnis nach einer Zerlegung des Wurzelknotens unplausibel.

[2] Die Daten und das AnswerTree-Projekt zum Beispiel finden Sie in den Dateien **Surrogat.sav** bzw. **Surrogat.atp** (genaue Bezugsquelle: siehe Einleitung).



Für den aufgrund seiner Trennungsleistung

trotz fehlender Werte

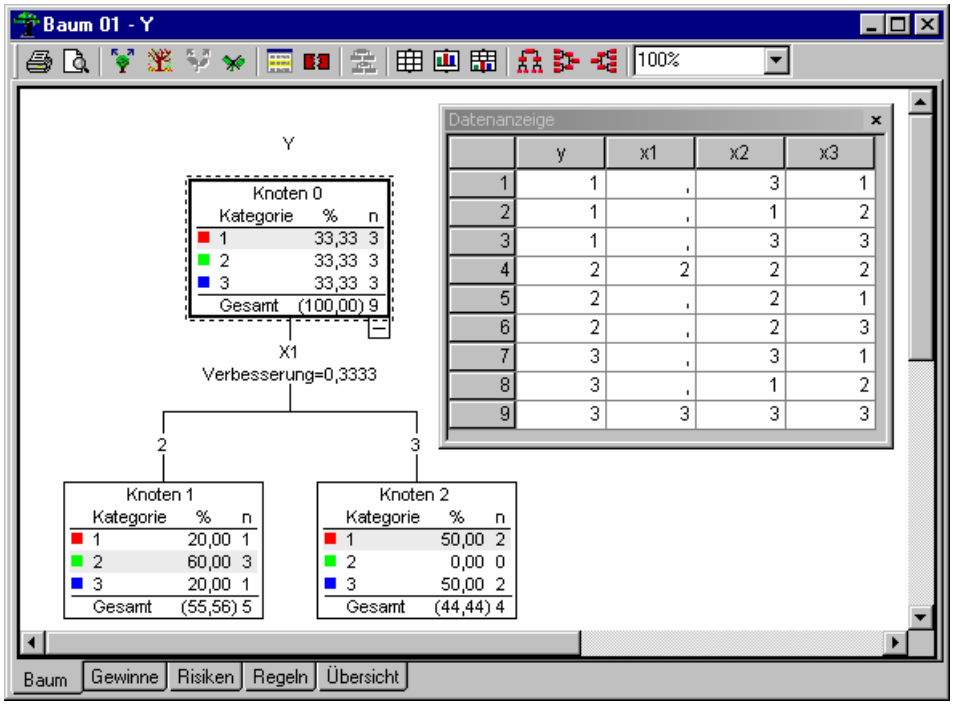


	y	x1	x2	x3
1	1	,	3	1
2	1	1	1	1
3	1	1	3	3
4	2	,	2	,
5	2	3	2	3
6	2	2	2	2
7	3	3	3	3
8	3	2	1	2
9	3	3	3	3

im Beispiel gewählten Prädiktor **X1** ist in erster Linie der Prädiktor **X3** als Ersatzvariable geeignet. Dementsprechend wird bei Fällen ohne **X1**-Wert nach Möglichkeit auf die **X3**-basierte Zuordnung zurück gegriffen (siehe Fall 1). Fehlt jedoch mit dem **X1**- auch der **X3**-Wert, so kommt die die nächst beste Ersatzvariable (im Beispiel: **X2**) zum Einsatz (siehe Fall 4).

In der obigen Dialogbox **Ersatzvariable wählen** wird übrigens die **Verbesserung** beim optimalen Imitieren des **X1**-Zuordnungsverhaltwens angegeben, wobei **X2** schlecht abschneidet, weil seine eigenen Qualitäten dabei nicht zum Tragen kommen. Benutzt man den **Aufbau**-Schalter dieser Dialogbox, kommt nicht etwa **X1** mit spezieller Ersetzung fehlender Werte zum Einsatz, sondern die gerade markierte Ersatzvariable darf sich als **X1**-Imitator bewähren.

Wie das folgende Beispiel zeigt, kann es bei Prädiktoren mit fehlenden Werten allerdings zu unplausiblen Verbesserungs-Berechnungen kommen:[1]



Aus den Knoten-Inhomogenitäten:

$$g(a_0) = 1-\frac{1}{3}=\frac{2}{3}, \quad g(a_1) = 1-\frac{2}{25}-\frac{9}{25}=\frac{14}{25}, \quad g(a_2) = 1-\frac{2}{4}=\frac{1}{2}$$

resultiert die Verbesserung:

$$g(a_0)-\frac{5}{9}g(a_1)-\frac{4}{9}g(a_2)=\frac{2}{3}-\frac{70}{225}-\frac{2}{9}=\frac{150}{225}-\frac{70}{225}-\frac{50}{225}=\frac{30}{225}=0,1\overline{3}$$

Bei einem Prädiktor mit vollständig vorhandenen Werten und obiger Aufteilungsleistung kommt AnswerTree zu genau dieser Beurteilung. Dass die fehlenden Werte zu einem deutlich besseren Urteil führen, ist nicht unbedingt sinnvoll.

[1] Die Daten und das AnswerTree-Projekt zum Beispiel finden Sie in den Dateien **Surrogat2.sav** bzw. **Surrogat2.atp** (genaue Bezugsquelle: siehe Einleitung).

3.2 Klassifikationsbaum zu Fishers Iris-Daten

In Anlehnung an das AnswerTree - Handbuch soll die C&RT-Methode mit der berühmten Iris-Aufgabe konfrontiert werden, die auf Fisher (1936) zurück geht. Die 3 Iris-Sorten

Setosa

Versicolor

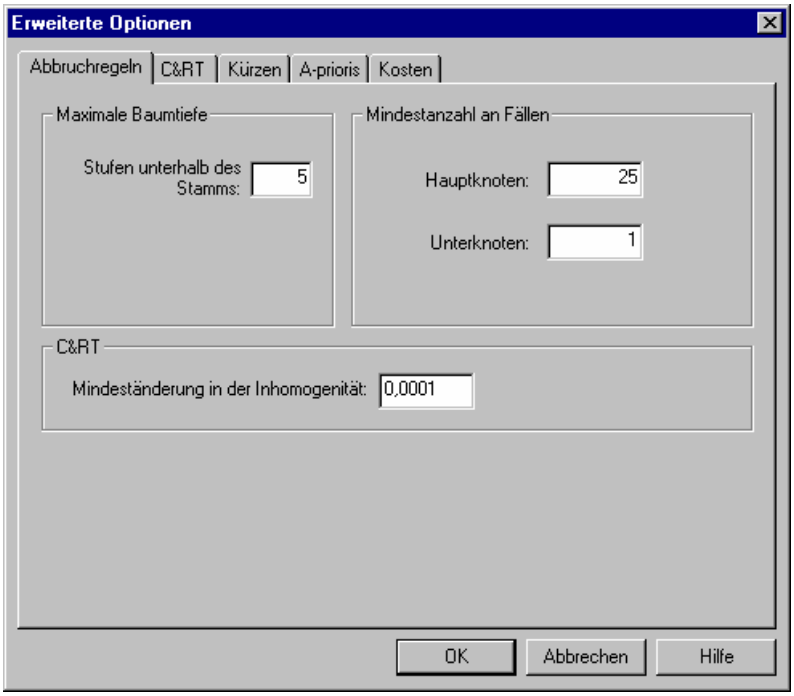
Virninica

sollen mit Hilfe von 4 metrischen Prädiktoren getrennt werden:

- Breite des Blütenblattes
- Länge des Blütenblattes
- Breite des Kelchblattes
- Länge des Kelchblattes

Im AnswerTree-Programmverzeichnis finden Sie Fishers Daten in der SPSS-Datei **Iris.sav**. Verwenden Sie diese Datei in einem neuen Projekt und fordern Sie einen C&RT-Baum an. Machen Sie im 2. Schritt des Baumassistenten die Variable **Spezies** zum **Ziel**, und lassen Sie **alle übrigen** als Prädiktoren zu:

Weil die drei Teilstichprobe jeweils nur aus 50 Fällen bestehen, sollten Sie im 4. Assistenten-Schritt die Dialogbox mit den **erweiterten Optionen** anfordern und die Mindeststärke für Haupt- und Unterknoten reduzieren:



Der Wurzelknoten zeigt eine Gleichverteilung der Irisarten:

Spezies

Knoten 0		
Kategorie	%	n
Iris-setosa	33,33	50
Iris-versicolor	33,33	50
Iris-virginica	33,33	50
Gesamt	(100,00)	150

Verlangt man in seinem Kontextmenü einen *einstufigen* Bauaufbau, präsentiert AnswerTree folgendes Zwischenergebnis:

Spezies

Knoten 0		
Kategorie	%	n
Iris-setosa	33,33	50
Iris-versicolor	33,33	50
Iris-virginica	33,33	50
Gesamt	(100,00)	150

Länge Blütenblatt
Verbesserung=0,3333

<=2,4500000000000002

Knoten 1		
Kategorie	%	n
Iris-setosa	100,00	50
Iris-versicolor	0,00	0
Iris-virginica	0,00	0
Gesamt	(33,33)	50

>2,4500000000000002

Knoten 2		
Kategorie	%	n
Iris-setosa	0,00	0
Iris-versicolor	50,00	50
Iris-virginica	50,00	50
Gesamt	(66,67)	100

Mit Hilfe des Prädiktors **Länge Blütenblatt**, dichotomisiert bei 2,45, kann die Spezies **Iris Setosa** perfekt von den restlichen Sorten getrennt werden.

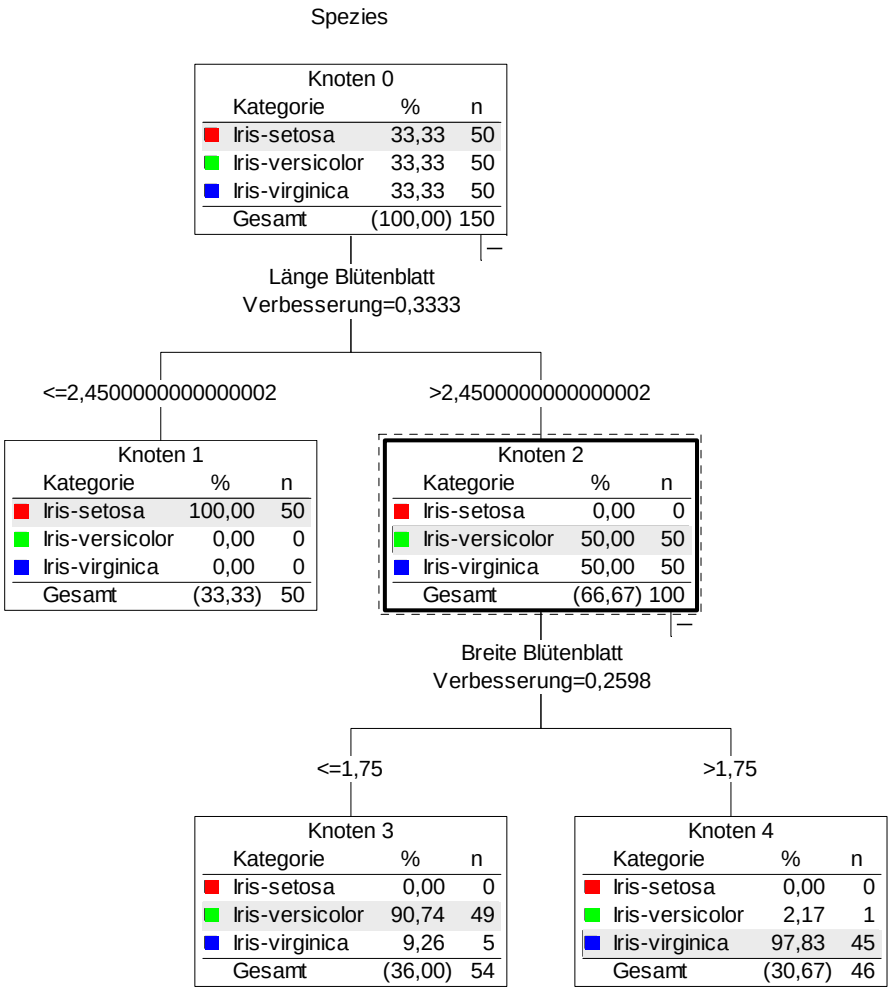
Die angegebene Verbesserung von $0,\overline{3}$ kommt nach den Formeln von Abschnitt 3.1 folgendermaßen zu Stande:

$$g(a_0) = 1 - \frac{1}{3} = \frac{2}{3}; \quad g(a_1) = 0; \quad g(a_2) = \frac{1}{2}$$

$$\Phi(t, a_0) = \frac{2}{3} - \frac{1}{3}0 - \frac{2}{3}\frac{1}{2} = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

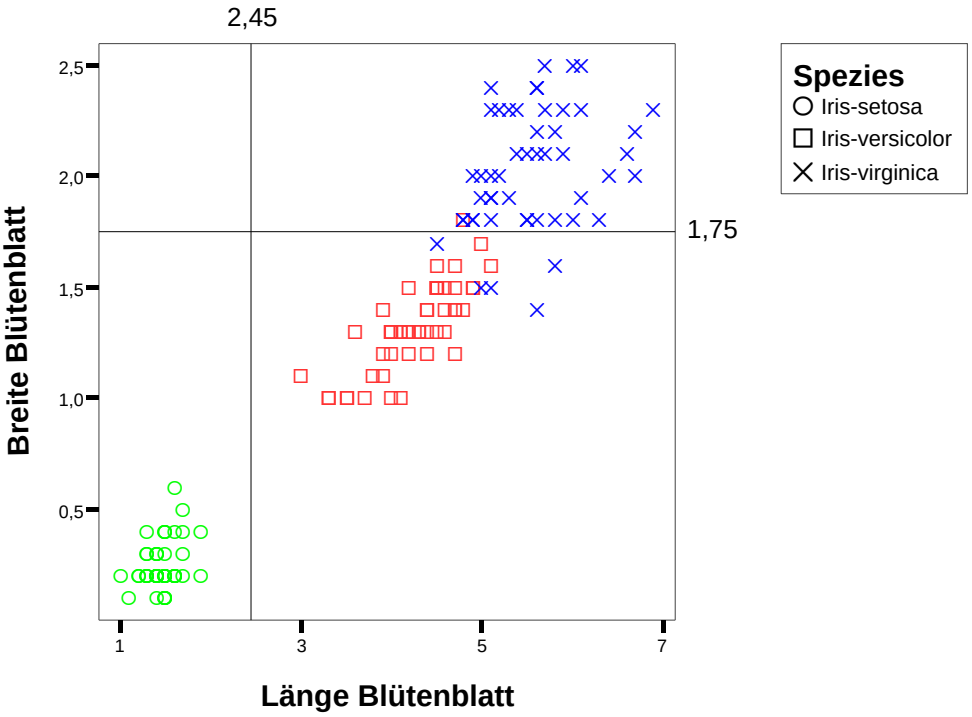
$$\text{Verbesserung} = 1\ \Phi(t, a_0) = \frac{1}{3}$$

Lässt man den gemischten Knoten durch die Kontextmenü-Option **Einstufiger Astaufbau** weiter zerlegen, kommt der Prädiktor **Breite Blütenblatt** zum Einsatz:



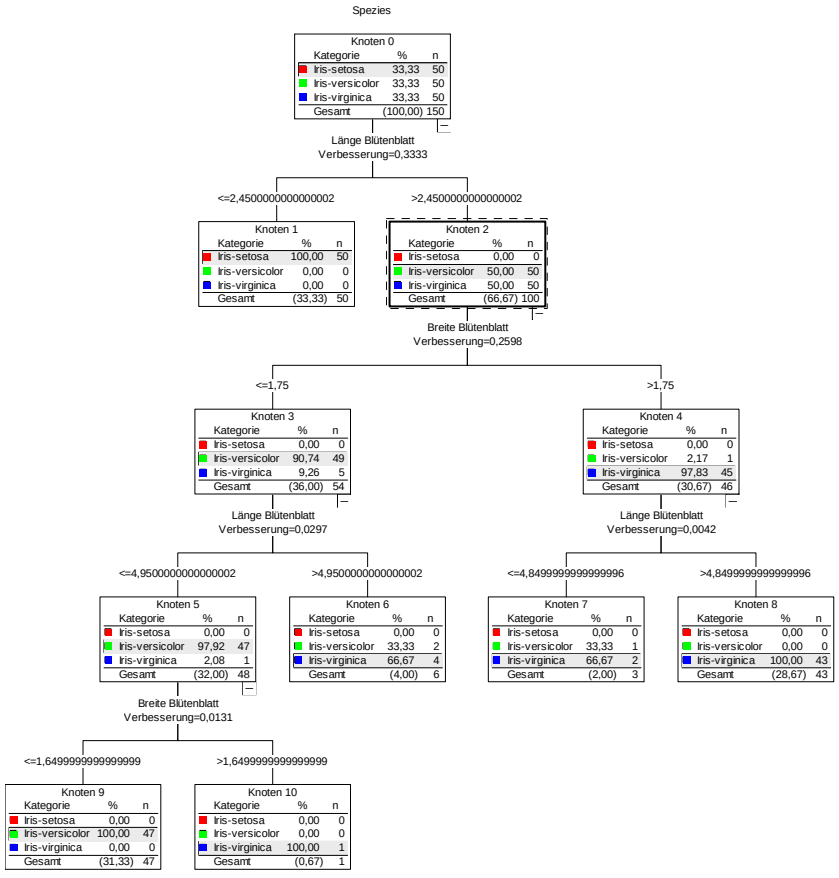
Nun sind auch die beiden Spezies **Iris Versicolor** und **Iris Virginia** recht gut getrennt.

Wie das folgende Streuungsdiagramm zeigt, stellt Fishers Segmentierungsaufgabe keine allzu große Herausforderung dar:



3.3 Kürzen zur Reduktion der Komplexität

Lässt man im Beispiel mit **Baum > Baumaufbau** den Algorithmus weiter laufen, resultiert ein recht unübersichtlicher Baum mit vielen schwach besetzten Knoten:



Es muss zudem befürchtet werden, dass viele Endknoten nur zufällige Stichproben-Verhältnisse reflektieren und sich bei einer Kreuzvalidierung nicht bewähren („Overfitting“).

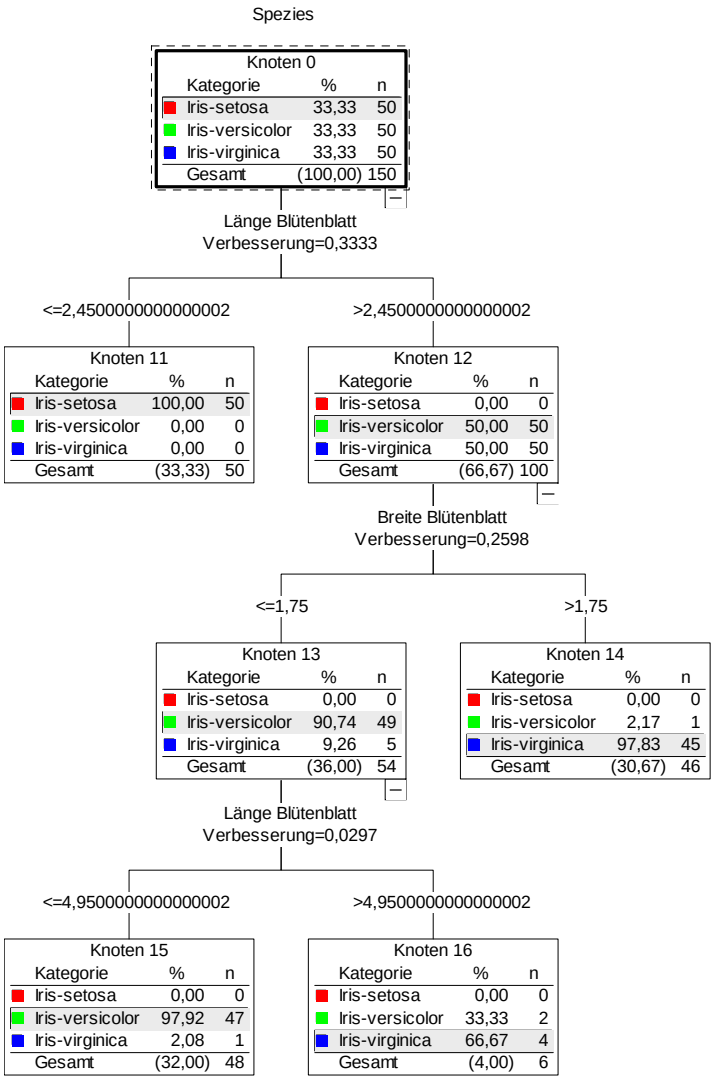
Dass eine Variable *wiederholt* zur Aufteilung herangezogen wird (auch auf verschiedenen Stufen), ist übrigens bei allen AnswerTree-Algorithmen möglich.

Inhomogenitäts-minimierende C&RT-Bäume tendieren zu übertriebener Komplexität. Schließlich besitzt eine Lösung mit einem einzigen Fall pro Knoten eine optimale Homogenität. Um solche „Lösungen“ zu vermeiden, schlagen Breiman et al. (1984) vor, einen Inhomogenitäts-minimierenden C&RT-Baum anschließend zu *kürzen*. Dabei wird versucht, die Anzahl der Endknoten möglichst stark zu reduzieren, ohne die auf dem **Risiken**-Registerblatt des Baumfensters angegebene Risikoschätzung wesentlich zu erhöhen.

Der Menübefehl

Baum > Baumaufbau und Kürzen

führt zu einer übersichtlichen Lösung mit 4 Endknoten:



Die Risikoschätzung steigt im Vergleich zur komplexen Lösung (mit 6 Endknoten) nur unwesentlich an (von 0,02 auf 0,027):

Fehlklassifizierungsmatrix					
		Tatsächliche Kategorie			
		Iris-setosa	Iris-versicolor	Iris-virginica	Gesamt
Vorhergesagte Kategorie	Iris-setosa	50	0	0	50
	Iris-versicolor	0	47	1	48
	Iris-virginica	0	3	49	52
	Gesamt	50	50	50	150
		Risikostatistiken			
Risikoschätzung		0,0266667			
Std.f. der Risikoschätzung		0,0131544			

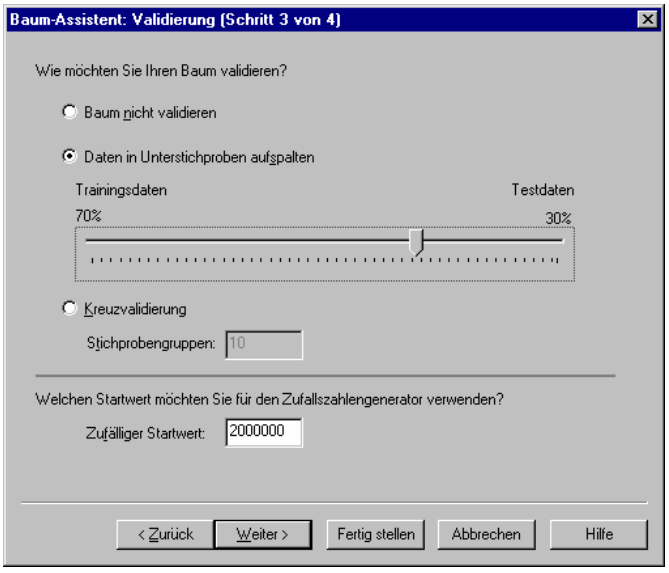
Das Klassifikationsergebnis des C&RT-Verfahren (50, 47 und 49 Treffer) wird von der **linearen Diskriminanzanalyse**, die bei metrischen Prädiktoren ebenfalls einsetzbar ist, nur unwesentlich übertroffen (50, 48 und 49 Treffer):

		Vorhergesagte Gruppenzugehörigkeit			
	Spezies	Iris-setosa	Iris-versicolor	Iris-virginica	Gesamt
Original	Iris-setosa	50	0	0	50
	Iris-versicolor	0	48	2	50
	Iris-virginica	0	1	49	50

Im Unterschied zur Diskriminanzanalyse ist die C&RT-Methode jedoch keinesfalls auf metrische Prädiktoren beschränkt.

3.4 Kreuzvalidierung

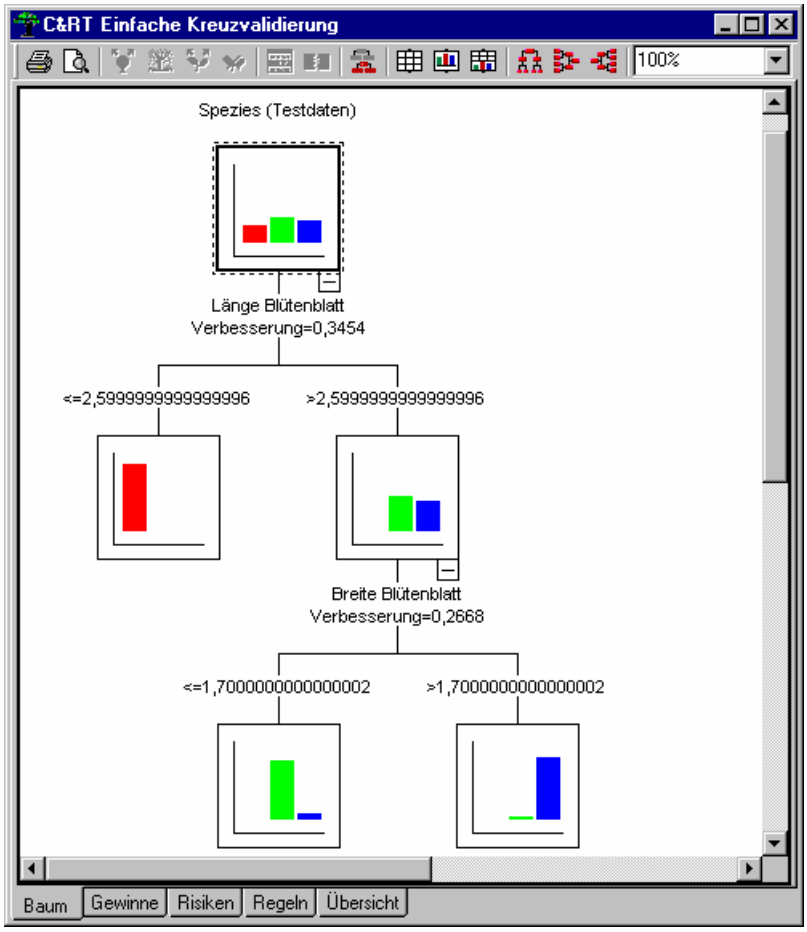
Bei den bisher vorgestellten Projekten wurden alle verfügbaren Fälle in die Lern- bzw. Trainingsstichprobe einbezogen. Um die Generalisierbarkeit bzw. die diagnostische Tauglichkeit eines Entscheidungsbaumes zu prüfen, wird zusätzlich eine unabhängige Kreuzvalidierungsstichprobe benötigt.
AnswerTree bietet dazu im dritten Schritt des Baumassistenten



folgende Optionen:

- **Daten in Unterstichproben aufspalten**
  Bei dieser einfachen Kreuzvalidierung werden die verfügbaren Fälle auf eine Trainings- und eine Teststichprobe aufgeteilt. Weil in der Entwicklungsphase ausschließlich die Fälle der Lernstichprobe mitwirken, steht die Teststichprobe für eine Kreuzvalidierung zur Verfügung.
- **Kreuzvalidierung**
  Bei der Jack Knife – Kreuzvalidierung findet eine besonders präzise Risikoschätzung durch den Mittelwert aus *mehreren* Kreuzvalidierungen statt. Man zerlegt die Stichprobe in *K* Gruppen und klassifiziert jede Gruppe durch einen Baum, der auf den jeweils restlichen Fällen basiert.
  Einschränkungen:
  - Nur einsetzbar bei *vollautomatischer* Baumentwicklung
  - Wegen des hohen Rechenzeitbedarfs nur bei relativ kleinen Stichproben möglich
- **Startwert für den Pseudozufallszahlengenerator** festlegen

Bei einer Q&RT-Analyse auf der Basis von 70% der Iris-Fälle resultierte folgender Klassifikationsbaum mit wenig Änderungen im Vergleich zum Produkt der Vollstrichprobe:



Mit dem Schalter wurde dafür gesorgt, dass zur Beschreibung der Knoten Balkengrafiken an Stelle der voreingestellten Tabellen erscheinen.

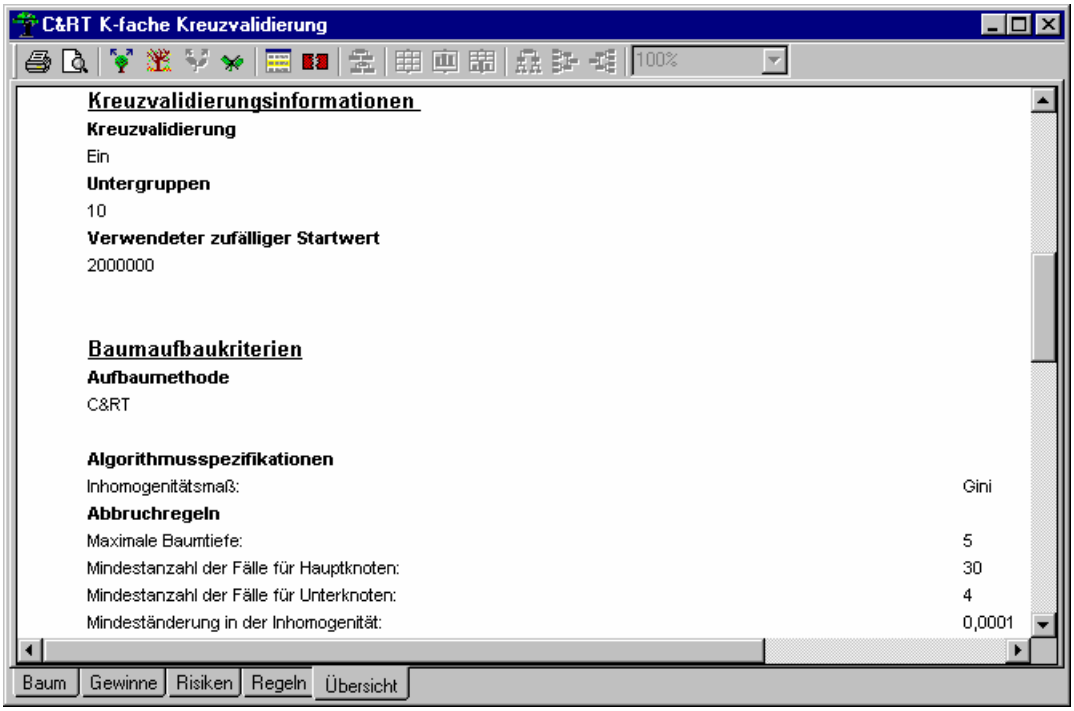
Über den Schalter im Baumfenster kann man auf allen Registerblättern zwischen Trainings- und Teststichprobe umschalten. So zeigt sich etwa bei der Risikoschätzung für die Testdaten mit 0,064 ein deutlich erhöhter Wert im Vergleich zur Trainingsstichprobe (0,029):

Testdaten					
Fehlklassifizierungsmatrix					
		Tatsächliche Kategorie			
		Iris-setosa	Iris-versicolor	Iris-virginica	Gesamt
Vorhergesagte Kategorie	Iris-setosa	13	0	0	13
	Iris-versicolor	0	17	2	19
	Iris-virginica	0	1	14	15
	Gesamt	13	18	16	47
		Risikostatistiken			
Risikoschätzung		0,0638298			
Std.f. der Risikoschätzung		0,0356566			

Eine Jack Knife – Prozedur mit 10 Gruppe ergibt, dass die wahre Fehlerrate in der Nähe von 0,04 liegt:

	Risikostatistiken	Kreuzvalidierung
Risikoschätzung	0,0266667	0,04
Std.f. der Risikoschätzung	0,0131544	0,016

An dieser Stelle soll ein nicht ganz unwichtiges Detail der AnswerTree-Baumfenster vorgestellt werden: die **Übersicht**-Registerkarte. Hier sind alle Baumparameter aufgelistet, so dass sich z.B. in Erfahrung bringen lässt, wie viele Untergruppen bei der Jack Knife – Kreuzvalidierung gebildet wurden:



3.5 Regressionsbaum als Modellierungshilfe

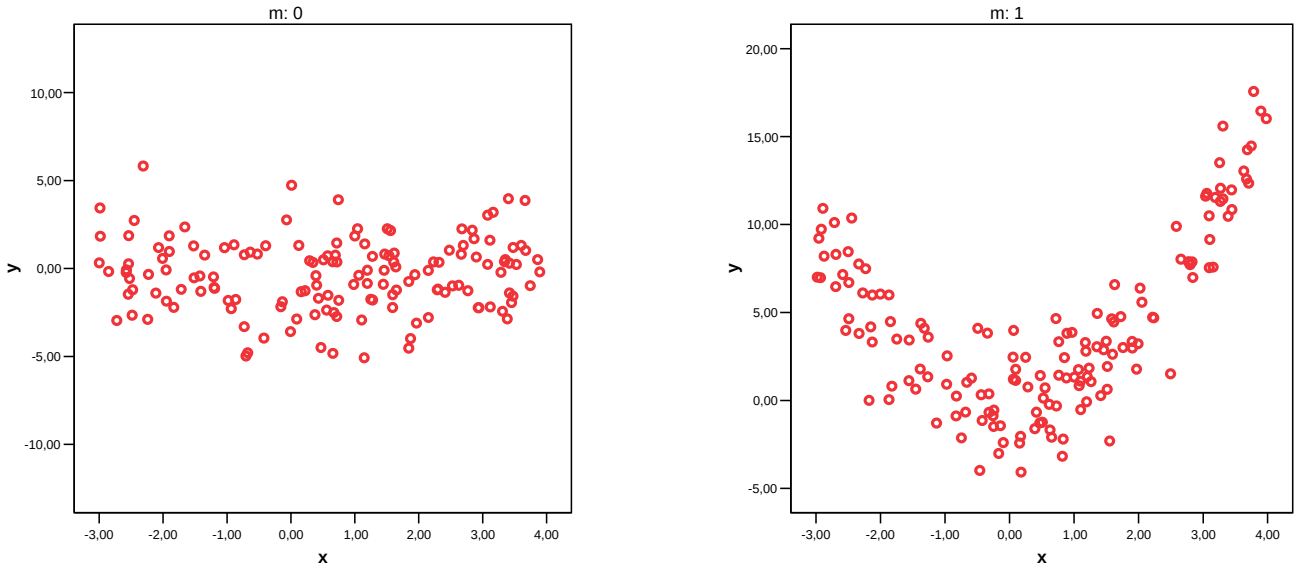
In diesem Abschnitt soll an einem Beispiel demonstriert werden, wie eine Entscheidungsbaumanalyse mit metrischem Kriterium (also eine Regressionsbaumanalyse) die Entwicklung eines parametrischen Regressionsmodells unterstützten kann, indem sie Hinweise auf relevante Regressoren und Interaktionen liefert. Es kommt ein künstlicher Datensatz[1] zum Einsatz, der für 300 Fälle folgende Variablen enthält:

- das Kriterium **Y**
- einen dichotomen Regressor **M** mit den gleichwahrscheinlichen Werten 0 und 1
- einen metrischen Regressor **X** (gleichverteilt auf dem Intervall [-3, 4])

Im wahren Modell hat **X** bei **M** = 0 *keinen* und bei **M** = 1 einen *quadratischen* Effekt auf das Kriterium, so dass beim Modelldesign sowohl eine Wechselwirkung als auch ein nichtlinearer Zusammenhang zu entdecken ist:

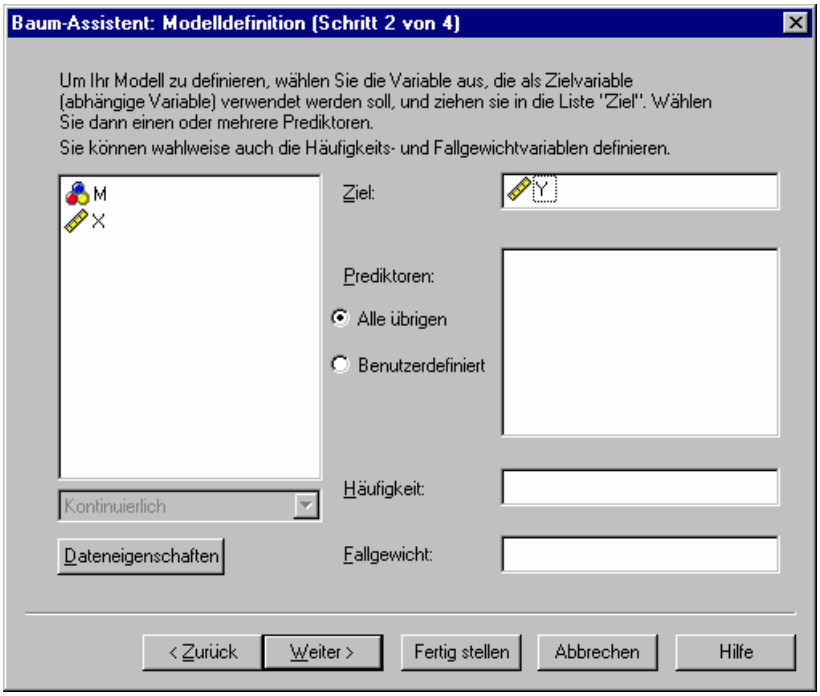
$$\mathbf{Y} = \mathbf{M} \cdot \mathbf{X}^2 + \varepsilon,\ \varepsilon \sim N(0, 4)$$

Für die beiden **M**-Kategorien ergeben sich in der Stichprobe prägnant verschiedene **X**-**Y**-Streudiagramme:

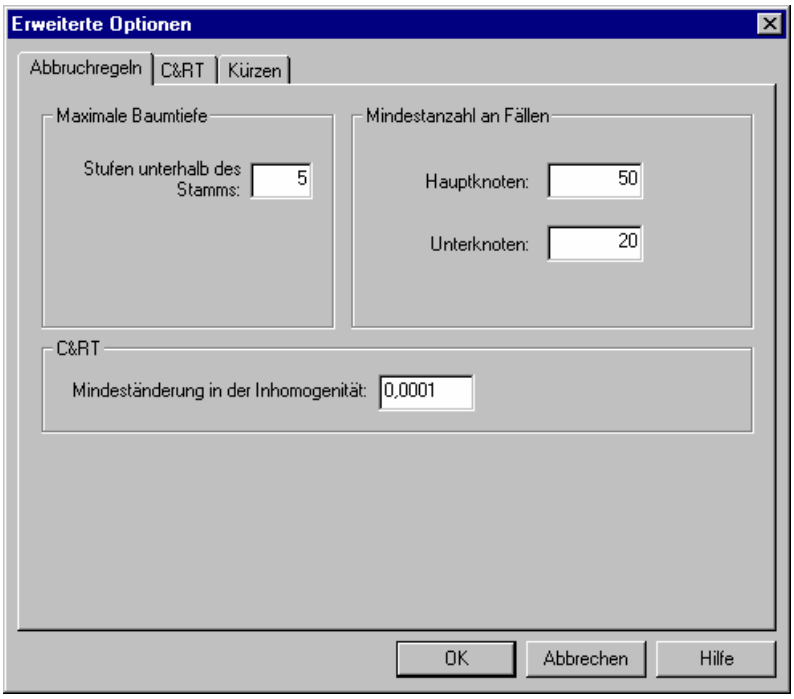


Wir wählen im AnswerTree-Baumassistenten die Aufbaumethode **C&RT** und weisen den Variablen ihre Rollen zu:

[1] In den Dateien **RegressionTree.sav** bzw. **RegressionTree.atp** finden Sie die Daten und das AnswerTree-Projekt zum Beispiel (genaue Bezugsquelle: siehe Einleitung).



Über den Schalter **Erweiterte Optionen** im vierten Assistentenschritt ändern wir die Abbruchregeln (vgl. Abschnitt 2.2.1):



Wir erhalten folgenden Wurzelknoten mit einer LSD-Inhomogenität von 18,583, die wir als *aufzuklärende Varianz* auffassen können:

Y

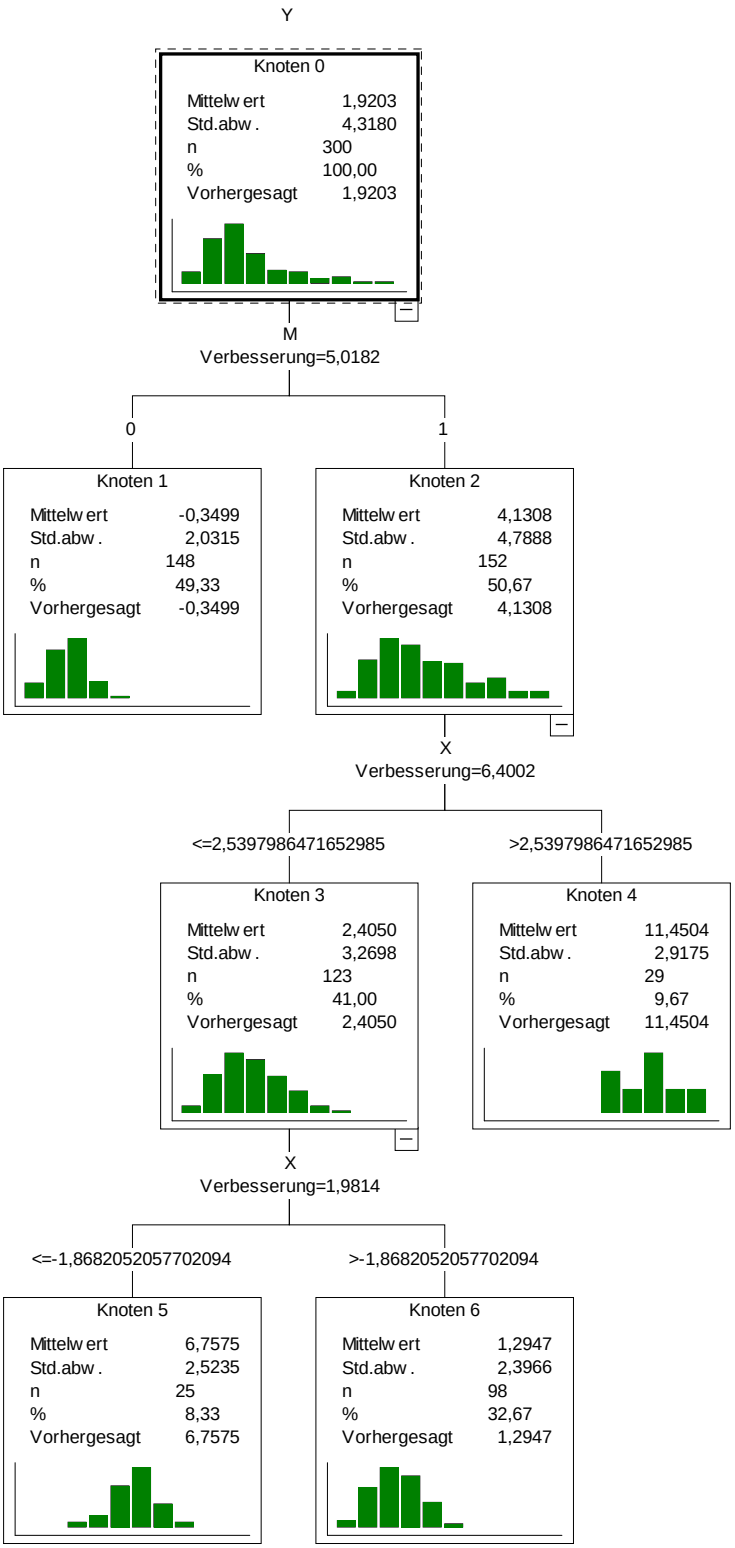
Knoten 0	
Mittelwert	1,9203
Std.abw.	4,3180
n	300
%	100,00
Vorhergesagt	1,9203

	Risikostatistiken
Risikoschätzung	18,583
Std.f. der Risikoschätzung	1,971

Die mit

Baum > Baumaufbau und Kürzen

angeforderte „Vollautomatik“ liefert ein brauchbares Ergebnis:



Mit dem Baumfenster-Schalter wurde für die Knoten neben der Ergebnistabelle auch ein Histogramm aktiviert. Wer ausschließlich das Histogramm sehen möchte, wählt den Schalter .

Beide Suchaufgaben sind gut gelöst:

- Der C&RT-Algorithmus „erkennt“ die Wechselwirkung und entwickelt für die beiden **M**-Kategorien verschiedene Modelle.
- Im Ast für **M** = 1 entstehen 3 Endknoten (nach den **X**-Werten geordnet mit den Nummern 5, 6 und 4), deren Kriteriumsverteilungen deutlich für einen quadratischen **X**-Effekt sprechen (siehe Mittelwerte oder Histogramme).

Wie ein Blick auf die **Risiken**-Registerkarte des Baumes zeigt, ist die gepoolte Varianz in den Endknoten deutlich geringer als die Varianz im Wurzelknoten:

	Risikostatistiken
Risikoschätzung	5,18324
Std.f. der Risikoschätzung	0,405434

Aus den beiden Fehlervarianzen lässt sich folgender Determinationskoeffizient berechnen:

$$R^2 = 1 - \frac{5{,}18324}{18{,}583} \approx 0{,}721$$

Bei einer Regressionsanalyse mit perfektem Modell, also mit dem Produkt aus **M** und **X** als Regressor, ergibt sich ein R^2-Wert von 0,786. In der Differenz spiegelt sich die reduzierte Präzision des Regressionsbaums aufgrund der 3-stufigen **X**-Variante im Vergleich zu den **X**-Individualwerten.

Auf analoge Weise kann ein Q&RT-*Klassifikationsbaum* (mit nominal- oder ordinalskaliertem Kriterium) oder auch ein per CHAID- bzw. QUEST-Algorithmus entwickelter Entscheidungsbaum zur Unterstützung einer *logistischen Regressionsanalyse* dienen.

Aus dem (durchaus realistischen) Beispiel dieses Abschnitts darf man allerdings nicht folgern, dass ein Q&RT-Regressionsbaum generell als vollautomatischer „Modelldetektor“ arbeitet. In vielen Fällen wird erst ein teilweise manuell gesteuerter Baumaufbau wichtige Strukturen sichtbar machen.

4 Das QUEST-Verfahren

Das von Loh & Shih (1997) vorgeschlagene QUEST–Verfahren (Quick Unbiased Efficient Statistical Tree) teilt viele Vor- und Nachteile mit dem Q&RT-Verfahren, z.B.:

- Aufbau eines *binären* Baums
- Tendenz zum Erzeugen komplexer Bäume, die aber durch Kürzungsverfahren kompensiert werden kann
- Gute Eignung für metrische Prädiktoren
- Behandlung fehlender Fälle durch Ersatzvariablen

Zwei potentielle Probleme des Q&RT-Verfahrens werden jedoch im QUEST-Algorithmus gezielt vermieden:

- Kein großer Rechenaufwand
- Keine Bevorzugung von Prädiktoren mit vielen Ausprägungen

Dies wird vor allem durch eine Entkopplung der folgenden Teilaufgaben beim Zerlegen eines Knotens erreicht:

- Wahl eines Prädiktors
- Aufteilung der Stichprobe mit Hilfe des gewählten Prädiktors

Weitere Unterschiede zum Q&RT-Verfahren sind:

- QUEST unterstützt nur *nominalskalierte* Kriterien.
- Eine Fallgewichtung ist *nicht* möglich.

Im gleich näher zu beschreibenden QUEST-Algorithmus zum schnellen und fairen Zerlegen eines Knotens kommen diverse statistische Methoden zum Einsatz, die aus anderen Kontexten wohlbekannt sind:

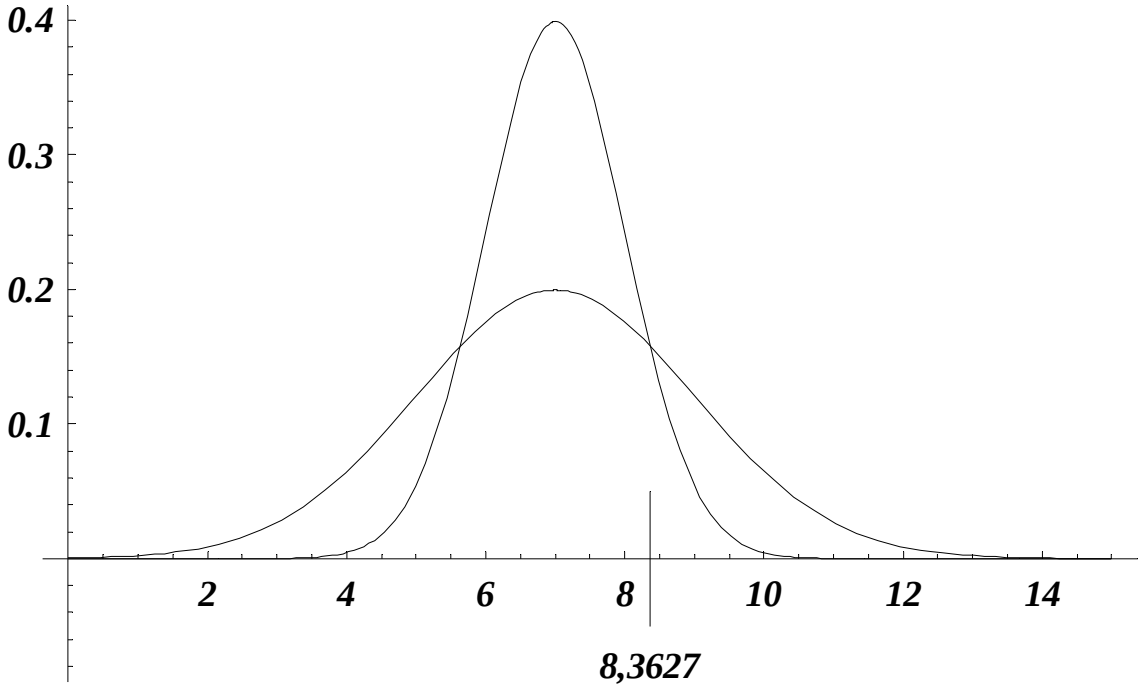
4.1 Wahl eines Prädiktors

Zunächst wird für jeden Prädiktor über einen passenden Signifikanztest geprüft, welchen Informationsgehalt er bzgl. des (prinzipiell nominalskalierten!) Kriteriums besitzt:

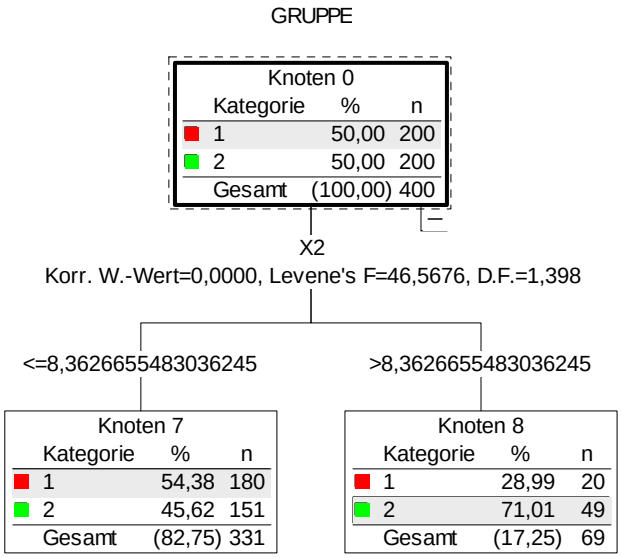
- Bei nominalskalierten Prädiktoren wird Pearsons χ^2-Test zur Homogenitätshypothese durchgeführt. Es findet *keine* χ^2-maximierende Überarbeitung des Kategoriensatzes statt (vgl. Abschnitt 2.5.1).
- Bei ordinalen oder metrischen Prädiktoren wird der F-Test der einfaktoriellen Varianzanalyse verwendet.

Unterschreitet das kleinste p-Level ein festgelegtes α-Niveau (Voreinstellung bei AnswerTree: 0,05), so wird der zugehörige Prädiktor gewählt.
Ist die Zahl K der Prädiktoren größer als 1, wird eine Bonferroni-Adjustierung vorgenommen und gegen α/K getestet.

Findet sich kein signifikanter Prädiktor, so wird bei den mindestens ordinalen Prädiktoren mit dem **Levene-Test** nach signifikanten *Varianzunterschieden* zwischen den Kriteriumskategorien gesucht. In der folgenden Abbildung wird für ein dichotomes Kriterium demonstriert, dass ein Prädiktor auch bei identischen Mittelwerten in beiden Kriteriumsgruppen für relevante Trennwerte sorgen kann, sofern seine Varianzen in den beiden Gruppen verschieden sind (im Beispiel: 1 bzw. 4):



Die X-Koordinate zu einem Schnittpunkt der Dichten als Trennwert zu verwenden, scheint unmittelbar plausibel. Wir werden diese aus der quadratischen Diskriminanzanalyse stammende Idee aber gleich noch diskutieren.
Beauftragt man AnswerTree bei der beschrieben Datenlage, mit dem QUEST-Verfahren einen Knoten zu zerlegen, resultiert abgesehen von Stichprobenschwankungen tatsächlich der aus obiger Abbildung mit den theoretischen Verteilungen zu erwartende Trennwert:[1]



Auch die p-Levels der Levene-Tests werden mit dem festgelegten α-Niveau verglichen. Ebenso findet eine Bonferroni-Adjustierung statt, wobei mittlerweile $(K + K_1)$ Tests zu berücksichtigen sind: K χ^2- bzw. F-Tests und K_1 Levene-Tests bei den ordinalen oder metrischen Prädiktoren.

Findet sich auch kein signifikanter Levene-Test, gewinnt der Prädiktor mit dem kleinsten p-Level im Pearson- bzw. F-Test.

[1] In den Dateien **Levene.sav** bzw. **Levene.atp** finden Sie die Daten und das AnswerTree-Projekt zum Beispiel (genaue Bezugsquelle: siehe Einleitung).

4.2 Metrisierung nominalskalierter Variablen

Nominalskalierte Prädiktoren werden nach dem bei Gnanadesikan (1977) beschriebenen CRIMCOORD-Verfahren „metrisiert", wobei die diskriminatorische Information des Prädiktors in Bezug auf das Kriterium genutzt wird, um Rangordnung und Abstände der transformierten Werte zu ermitteln. Bei Loh & Shih (1997, S. 823) wird folgendes Beispiel mit einem dichotomen Kriterium und einem 3-stufigen Prädiktor angegeben:[1]

a) Kontingenztabelle

Kriterium	Prädiktor		
	1	2	3
1	4	1	5
2	2	2	6

b) Metrisierung durch CRIMCOORD-Werte

Prädiktor-kategorie		CRIMCOORD-Wert
1	→	1
2	→	-1
3	→	-0,273

Metrisierte nominalskalierten Prädiktor werden bei der Knotenzerlegung genauso behandelt wie ordinale bzw. metrische Prädiktoren (siehe Abschnitt 4.3). Aus dem dabei ermittelten Trennwert (im Beispiel: 0,142) gewinnt man eine Aufteilungsregel für die Prädiktorkategorien (im Beispiel: Kategorie 1 gegen Rest).

4.3 Trennwerte bestimmen

Für den zuvor gewählten und nötigenfalls metrisierten Prädiktor wird nun ein Trennwert zur Aufteilung des Knotens *a* in zwei Unterknoten bestimmt.

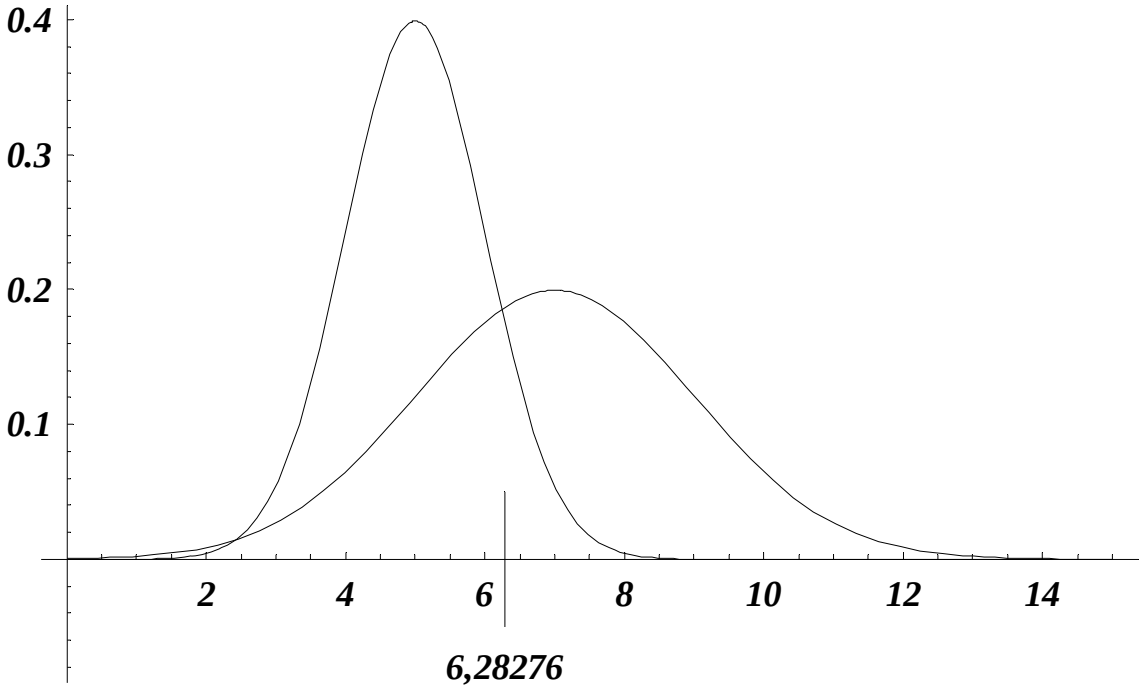
4.3.1 Bildung von Superklassen

Sind im aktuellen Knoten mehr als 2 Kriteriumskategorien realisiert, werden zunächst auf Basis der Prädiktor-Kategorienmittelwerte 2 Superklassen gebildet. Dies geschieht durch ein Cluster-Verfahren nach Hartigan and Wong (1979), ausgehend von den beiden extremen Mittelwerten.

4.3.2 Trennung am Punkt mit identischer Wahrscheinlichkeitsdichte für beide Superklassen

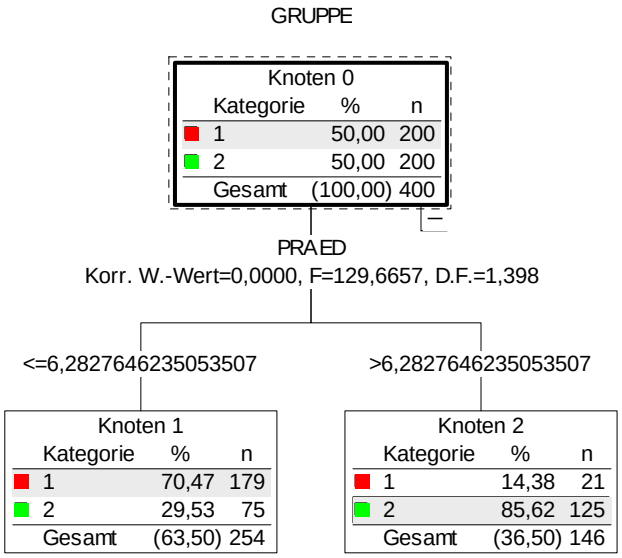
Nun wird die Grundidee der *quadratischen Diskriminanzanalyse* dazu genutzt, einen Trennwert für die beiden Superklassen in Bezug auf den als metrisch behandelten Prädiktor zu bestimmen. Man nimmt für beide Superklassen eine Normalverteilung des Prädiktors an, schätzt die Verteilungsparameter durch die Mittelwerte ($\bar{x}_1$, $\bar{x}_2$) und die Varianzen (s_1^2, s_2^2) der Stichproben und ermittelt die Schnittpunkte der beiden Dichtefunktionen. In der folgenden Abbildung sind die Dichten der N(5, 1) – und der N(7, 4) – Verteilung zu sehen:

[1] In den Dateien **crimcoord.sav** bzw. **crimcoord.atp** finden Sie die Daten und ein AnswerTree-Projekt zum Beispiel (genaue Bezugsquelle: siehe Einleitung).



Weil das QUEST-Verfahren eine *binäre* Aufteilung vornehmen möchte, beschränkt es sich auf den rechten (näher an beiden Mittelwerten liegenden) Schnittpunkt der Dichtefunktionen und verwendet dessen X-Koordinate (im Beispiel: 6,28276) als Trennwert.

Bei der QUEST-Zerlegung eines Knotens, dessen 400 Fälle jeweils zur Hälfte bzgl. des Prädiktors aus einer N(5, 1) – bzw. aus einer N(7, 4) – Normalverteilung stammen, ermittelt AnswerTree abgesehen von Stichprobenschwankungen tatsächlich den nach obiger Abbildung zu erwartenden Trennwert:[1]



In obiger Darstellung der Trennwert-Bestimmung wurde der Einfachheit halber die Rolle der (im Beispiel identischen) bedingten Wahrscheinlichkeiten P(1|*a*) und P(2|*a*) der beiden Kriteriums-Superklassen im aufzuteilenden Knoten *a* unterschlagen. Zur Bestimmung des Trennwertes ist die folgende Gleichung nach *x* aufzulösen (siehe Loh & Shih 1997, S. 5):

[1] Die zum Erzeugen der Daten benutzte SPSS-Syntax, die SPSS-Daten und das AnswerTree-Projekt zum Beispiel finden Sie in den Dateien **qdasep.sps**, **qdasep.sav** bzw. **qdasep.atp** (genaue Bezugsquelle: siehe Einleitung).

$$\hat{P}(1\,|\,a)\,\frac{1}{s_1}e^{-\frac{(x-\bar{x}_1)^2}{2s_1^2}} = \hat{P}(2\,|\,a)\,\frac{1}{s_2}e^{-\frac{(x-\bar{x}_2)^2}{2s_2^2}}$$

Weil sich der QUEST-Algorithmus bei der Trennwertbestimmung auf die *quadratische* Diskriminanzanalyse stützt, die im Unterschied zur linearen Variante unterschiedliche Varianzen berücksichtigt, taugt das Verfahren auch bei Prädiktoren, die aufgrund eines signifikanten Levene-Tests gewählt wurden (vgl. Abschnitt 4.1).

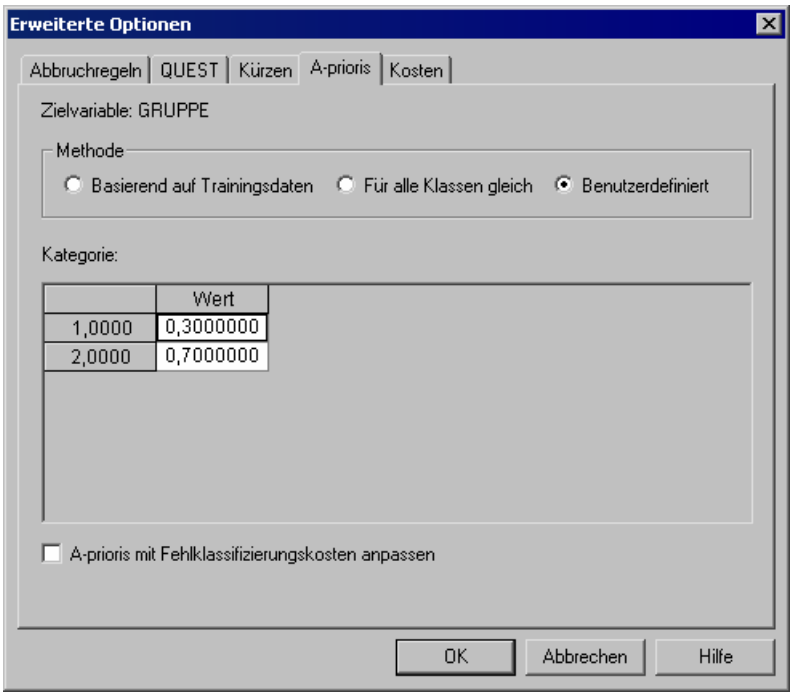
Bei halbwegs symmetrischen und eingipfligen Verteilungen ist die Normalverteilungsannahme unproblematisch, ansonsten aber eher fragwürdig.

4.3.3 A-priori – Wahrscheinlichkeiten

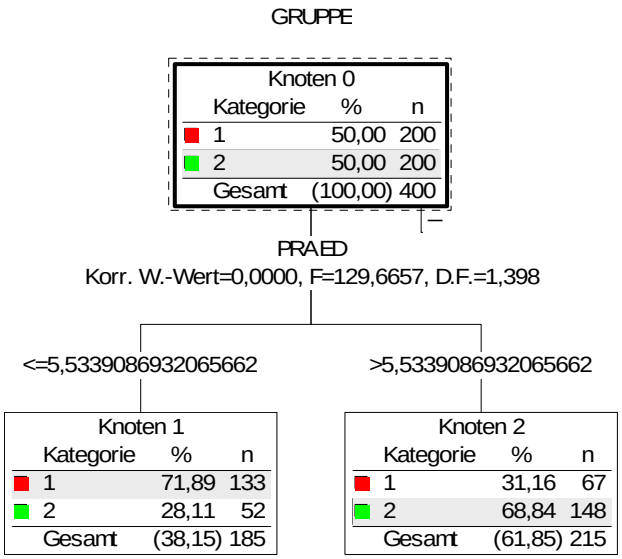
Wie das C&RT-Verfahren kann auch der QUEST-Algorithmus zur Schätzung der bedingten Superklassen-Wahrscheinlichkeiten $P(i\,|\,a)$ im aufzuteilenden Knoten *a*, welche die Trennwerte und damit den Baumaufbau beeinflussen, eine vorgegebene a-priori – Wahrscheinlichkeitsverteilung des Kriteriums berücksichtigen (siehe Abschnitt 3.1.1.2).
In obigem Beispiel sollen nach

Analyse > Erweiterte Optionen

auf der Registerkarte **A-prioris** folgende benutzerdefinierte Wahrscheinlichkeiten festgelegt werden:

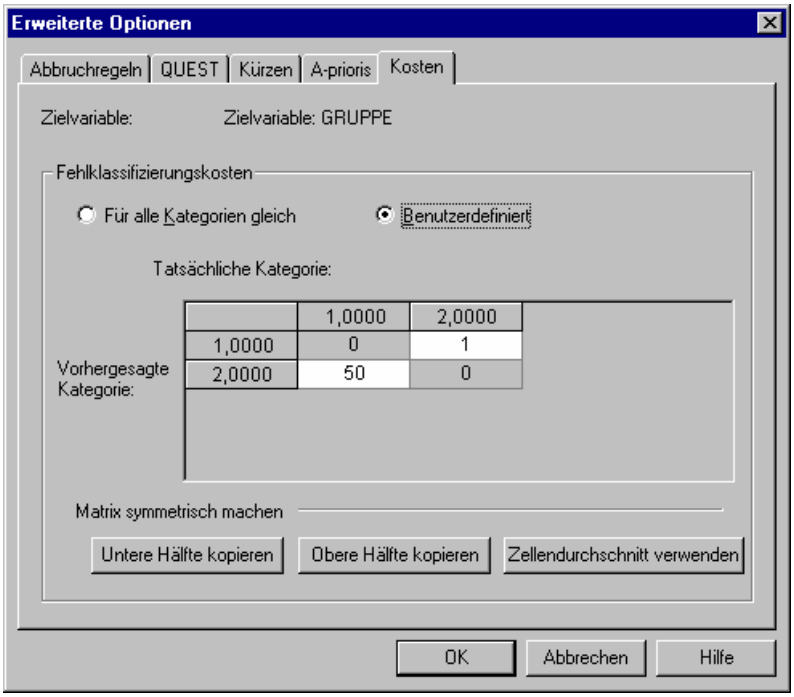


Durch das relative Anheben der rechten Dichte wandert der Dichtenschnittpunkt nach links:

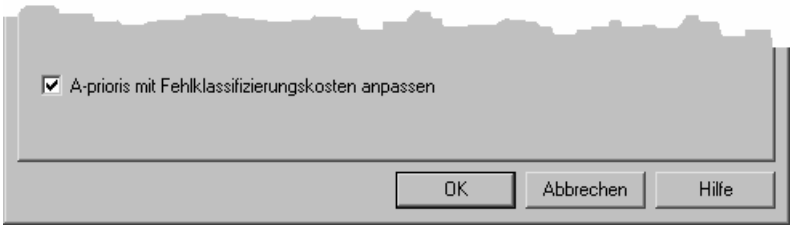


4.3.4 Fehlklassifikationskosten

Benutzerdefinierte **Kosten** wirken sich im QUEST-Algorithmus zunächst nur auf dem **Risiken**-Registerblatt aus (vgl. Abschnitt 2.4.3), jedoch *nicht* beim Baumaufbau.



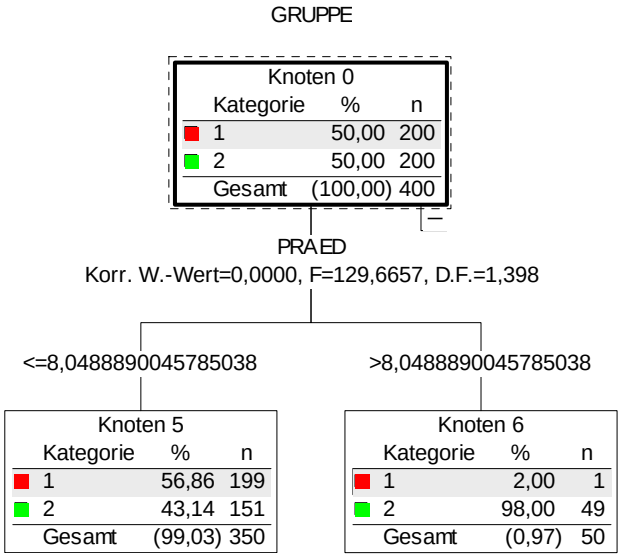
Über das Kontrollkästchen **A-prioris mit Fehlklassifikationskosten anpassen** auf dem **A-prioris** – Registerblatt der Dialogbox mit den erweiterten Optionen kann man aber dafür sorgen, dass die Kosten in die a-priori – Wahrscheinlichkeiten einfließen und damit den Baumaufbau beeinflussen (vgl. Abschnitt 4.3.3):



Der QUEST-Algorithmus adjustiert dabei die bisherigen a-priori – Wahrscheinlichkeiten $\pi(i)$ folgendermaßen:

$$\pi_a(j) := \frac{C(j)\,\pi(j)}{\sum_i C(i)\,\pi(i)} \quad \text{mit} \quad C(j) := \sum_i C(i \mid j)$$

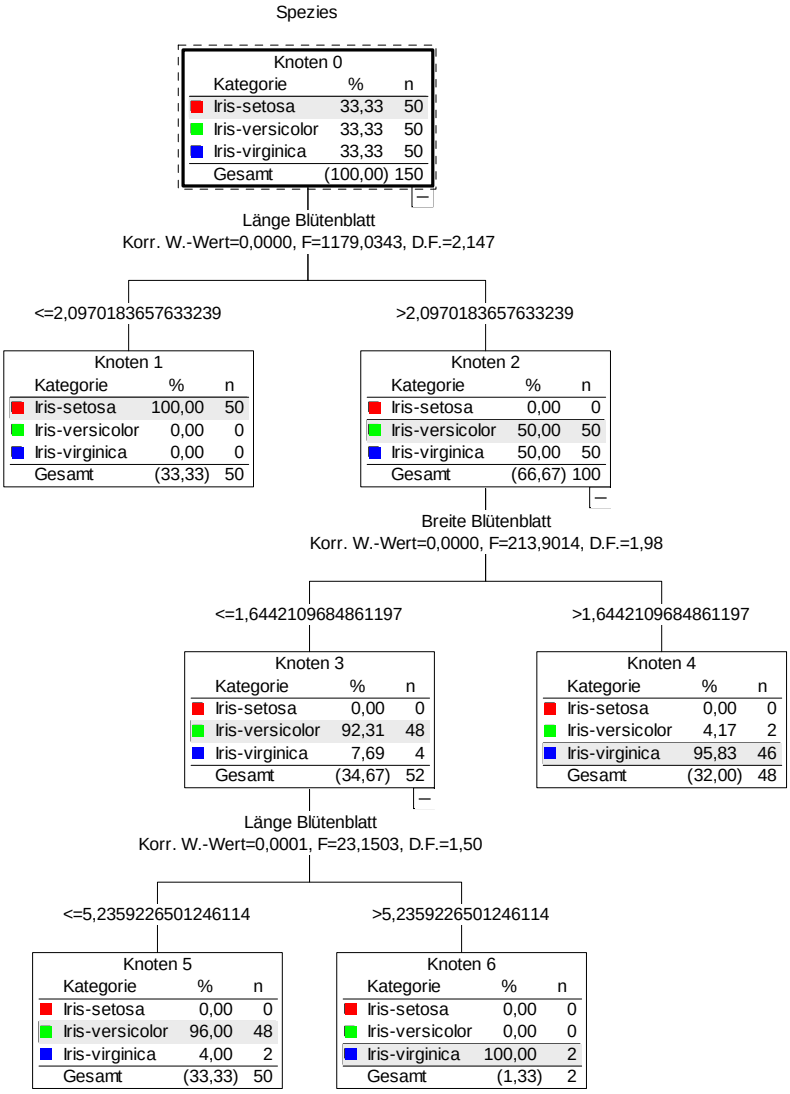
Im Beispiel sorgt die harte Bestrafung von Zuordnungsfehlern bei Objekten der ersten Kategorie dafür, dass der Trennweit deutlich nach rechts wandert:



4.4 QUEST-Analyse der Iris-Daten

Wie die Q&RT- soll auch die konkurrierende QUEST-Methode mit Fischers Iris-Daten konfrontiert werden (vgl. Abschnitt 3.2).

Der automatisch aufgebaute und gekürzte QUEST-Baum zeigt kaum Unterschiede zum Q&RT-Produkt:



Bei allen Zerlegungen kommen dieselben Prädiktoren zum Einsatz, und auch die Trennwerte fallen sehr ähnlich aus.

5 Literatur

Biggs, D., De Ville, B. & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18(1), 49-62.

Breimann, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179-188.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm 136. A k-means clustering algorithm, *Applied Statistics* 28 100.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119-127.

Loh, W. Y. & Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7, 815-840.

Magidson, J. (1992). Chi-squared analysis of a scalable dependent variable. In *Proceedings of the 1992 Annual Meeting of the American Statistical Association*, Educational Statistics Section.

Magidson, J. & SPSS, Inc. (1993). *SPSS for Windows CHAID 6.0*. Chicago, IL.

SPSS Inc. (2001). *AnswerTree 3.0. Benutzerhandbuch*. Chicago, IL: SPSS.

SPSS Inc. (2002). *AnswerTree 3.1. User's Guide*. Chicago, IL: SPSS.

6 Anhang

6.1 Einsatzmöglichkeiten der einzelnen Algorithmen

In der folgenden Tabelle aus dem AnswerTree-Handbuch (SPSS Inc., 2002, S. 215) werden die Einsatzmöglichkeiten der einzelnen Verfahren zusammengestellt:

Verfahren	**Typ des Kriteriums**	**Fall-gewichte**	**Häufig-keits-variablen**	**Werte[19] (ordinal)**	**Profit (ordinal/ nominal)**	**Kosten (ordinal/ nominal)**	**A-prioris (ordinal/ nominal)**
CHAID	metrisch, ordinal, nominal	ja	ja	ja	ja	ja	nein
Exhaustive CHAID	metrisch, ordinal, nominal	ja	ja	ja	ja	ja	nein
C&RT	metrisch, ordinal, nominal	ja	ja	nein	ja	ja	ja
QUEST	nominal		ja	nein	ja	ja	ja

[19] Mit dieser im Manuskript nicht diskutierten Option kann man bei ordinalen Kriterien die Anordnung der Kategorien verändern.

6.2 Variablen im Marktforschungs-Beispiel

Beschreibung der Variablen aus dem Beispiel zum CHAID-Algorithmus (Zeitschriften-Werbeaktion, siehe Abschnitt 1.1 bzw. 2):

Abhängige Variable:

Variable		**Kategorie**	
Name	**Label**	**Wert**	**Label**
ANTW2	Antwort auf die Werbeaktion (2)	1 2	Antwort Keine Antwort

Prädiktoren:

Variable		**Kategorie**	
Name	**Label**	**Wert**	**Label**
ALTER	Alter des Haushaltsvorstandes in Jahren	1 2 3 4 5 6 7	18-24 25-34 35-44 45-54 55-64 65+ unbekannt
GESCHL	Geschlecht des Haushaltvorstands	1 2	männlich weiblich
KINDER	Kinder im Haushalt	1 2	ja nein
HEINKOMM	Haushaltseinkommen	1 2 3 4 5 6 7 8	unter $8.000 $8.000-$9.999 $10.000-$14.999 $15.000-$19.999 $20.000-$24.999 $25.000-$34.999 $35.000-$49.999 $50.000 oder mehr
KARTE	Kreditkarte vorhanden	1 2	ja nein
HHGRÖßE	Anzahl der Personen im Haushalt	1 2 3 4 5 6	 Fünf oder mehr unbekannt
BERUF	Beruf des Haushaltsvorstandes	1 2 3 4	Angestellter Arbeiter Sonstiges unbekannt

7 Stichwortverzeichnis

A

Abbruchregel 29
Abbruchregeln 10
a-priori – Wahrscheinlichkeit
  C&RT 36
  QUEST 61
Assoziation 42

B

Baumassistent 8
Baumfenster 12
Bonferroni-Adjustierung 29, 57
Bonferroni-Anpassung 11, 16

C

C&RT-Methode 34
CART 6, 34
CHAID-
  Algorithmus 8
  Analyse 7
  Programm 6
CHAID-Algorithmus
  exhaustiver 29, 33
Classification & Regression Trees 34
Clusteranalyse 4, 5
CRIMCOORD 59

D

Data Mining 4
Datensätze 6
Determinationskoeffizient 56
Diagramme 19
Diskriminanzanalyse 50
  quadratische 59

E

Enhenced Metafile 15
Ersatzvariablen 42
Erwartete Kosten 26
Exhaustiver CHAID-Algorithmus 29, 33
Export der Knotendefinitionen 27
Exportieren
  Baumdiagramm 15

F

Fehlende Werte
  C&RT, QUEST 42
  CHAID 14
Fehlklassifikationskosten
  CHAID 25, 38
  QUEST 62
Fehlklassifikationsmatrix 24
Fehlklassifizierungskosten 11

G

Gewichtungsvariablen
  C&RT 40
  CHAID 9
Gewinne 17
Gewinne-Diagramm 20
Gini-Index 34
Gini-Zielfunktion 35

H

Häufigkeitsvariablen 9, 40

I

Importieren
  SPSS-Daten 7
Index-Diagramm 21
Inhomogenität 34
Intervalle 13
Investitionsrentabilität 23

J

Jack Knife 51

K

Knotendefinition
  exportieren 27
Konkurrenten 16
Kosten 11
  CHAID 25, 38
  QUEST 62
Kreuzvalidierung 50
Kürzen
  der Komplexität 48

L

Levene-Test 57
Logistische Regressionsanalyse 56
LSD-Index 40

M

Merging 28
Modellspezifikation 8
Monotonen Prädiktoren 32

O

Overfitting 49

P

Pearson-Chi-Quadrat-Test 11
Perzentildiagramme 19
Perzentiltabelle 18
Prädiktive Assoziation 42
Profit 21

Profit-Diagramm 24

Q

Quadratische Diskriminanzanalyse 59

R

Regressionsbaum 53
Return On Investment 23
Risikoschätzung 26
ROI 23
ROI-Diagramm 24

S

Segmentierung
  automatische 14
  manuelle 16
Skalenniveau 12
Splitting 29
Superklassen 59

T

Treffer-Diagramm 21

U

Überschreitungswahrscheinlichkeiten 29

V

Verbesserung 35, 42

W

WLM-Algorithmus 9

Z

Zusammenlegen
  von Kategorien 28