

Medizinische Hochschule Hannover
Institut für Biometrie

RECPAM: RECURSIVE PARTITION AND AMALGAMATION

Kurzbeschreibung

H. Hecker

Contents

1	Anwendungsbereich und Voraussetzungen	1
2	Die Struktur des Zusammenhangs zwischen Prädiktor- und Responsevariablen.	1
2.1	Allgemeine Annahmen	1
2.2	Die Baumstruktur	1
2.3	Näheres zu den "Split Defining Statements".	2
3	Minimale Knoten	3
4	Kriterien für die Auswahl von Splits.	3
4.1	Ein globales Diskrepanzmaß	3
4.2	Das Split-Kriterium	4
4.3	Die verschiedenen Spezifikationen des Distanzmaßes	4
4.4	Die gesamte Baum-Erzeugung	5
4.4.1	Die Vorbereitung:	5
4.4.2	Der Algorithmus:	5
4.4.3	Behandlung fehlender Werte.	6
5	4. Kreuzvalidierung und AIC	6
5.1	Kreuzvalidierung	6
5.2	AIC	6
6	Zusammenlegen von Endknoten	7
6.1	AMALGAMATION	7
6.2	PRUNING	8
7	Literatur	8

1 Anwendungsbereich und Voraussetzungen

Es wird davon ausgegangen, daß der Einfluß von p ($p \leq 20$) *Prädiktorvariablen* X_1, X_2, \dots, X_p auf die Verteilung einer 1-dimensionalen Zielgröße (*Responsevariable*) Y untersucht werden soll.

Die *Prädiktorvariablen* X_1, X_2, \dots, X_p sind *ordinal* skaliert oder *qualitativ*. Auch in dem Fall, daß die X_i *quantitative* oder sogar *stetige Variable* sind, werden nur die Eigenschaften der Ordinalskala benutzt. Darüberhinaus muß

auch in diesem Fall der Wertebereich der Variablen in (≤ 20) *Kategorien* eingeteilt und diese Kategorien mit Labeln versehen werden.

Von der *Zielvariablen* wird angenommen, daß sie entweder *qualitativ* ist mit beliebiger Anzahl von Kategorien, oder *quantitativ*. Im ersten Fall ist die Verteilung von Y die *Multinomialverteilung* $M(p_1, p_2, \dots, p_k)$, wobei die Parameter p_1, p_2, \dots, p_k die Wahrscheinlichkeiten der Kategorien $1, 2, \dots, k$ kennzeichnen; im zweiten Fall wird angenommen, daß es sich um (rechts-)zensierte Daten handelt und daß die Verteilungsfunktionen (bzw. die Überlebenskurven) a) *exponentialverteilt* sind, oder b) *proportionale Hazardfunktionen* haben (COX-Modell) oder auch (c) gar keine speziellen Bedingungen erfüllen.

2 Die Struktur des Zusammenhangs zwischen Prädiktor- und Responsevariablen.

2.1 Allgemeine Annahmen

RECPAM geht (wie auch CART) davon aus, daß sich der Zusammenhang zwischen Prädiktor- und Responsevariablen (zumindest approximativ) dadurch beschreiben läßt, daß man *Untergruppen von Fällen* angibt, innerhalb derer die Verteilung von Y (angenähert) *konstant* ist. Dementsprechend wird nach Untergruppen gesucht, die in dieser Hinsicht (d.h. bezgl. der Verteilung von Y) *homogen* sind.

Diese Untergruppen müssen durch *Werte der Prädiktorvariablen charakterisierbar* sein. Darüber hinaus wird angenommen, daß diese Charakterisierung von homogenen Untergruppen mit Hilfe eines (*binären*) Baumes erfolgen kann.

2.2 Die Baumstruktur

Ein (binärer) Baum kann dadurch beschrieben werden, daß der Wertebereich \mathcal{X} des p -dimensionalen Vektors $X = (X_1, X_2, \dots, X_p)$, d.h. die Gesamtheit aller möglichen Wertekombinationen der Prädiktorvariablen, zunächst in zwei Teilräume aufgespalten wird. Jeder so entstandene Teilraum und alle weiteren Teilräume können sukzessive weiter aufgespalten werden, so daß insgesamt eine hierarchisch strukturierte Aufteilung entsteht.

In der Baumdarstellung erscheint zunächst der Gesamttraum \mathcal{X} als Kreis. Die Aufspaltung des Gesamttraumes und weiterer Teilräume wird dann dadurch symbolisiert, daß links und rechts darunter jeweils ein weiterer Kreis erscheint, der den entsprechenden Teilraum symbolisiert und dessen "Herkunft" durch entsprechende Verbindungslinien markiert ist.

Die Kreise, die also die Teilräume symbolisieren, werden die *Knoten* ("*nodes*") des Baumes genannt; der gesamte Raum, also der *erste* Kreis, heißt die *Wurzel* des Baumes (man beachte: der Baum wird i.d. Regel als *hängender* Baum dargestellt). Knoten, die sich nicht weiter verzweigen, heißen *Endknoten* ("*terminal nodes*"). Endknoten werden üblicherweise als Rechteck dargestellt.

Jede Verzweigung, also jeder *split*, kann durch eine "Frage" beschrieben werden, die *eine* (in besonderen Fällen (s.u.) auch *mehrere*) Prädiktorvariable betrifft, und die jeweils nur mit "Ja" oder "Nein" beantwortet werden kann. Analog kann man von "Aussagen" sprechen, die entweder

”richtig“ oder ”falsch“ sind. Diese Fragen (bzw. Aussagen) werden als ”SDS“ (= *Split Defining Statements*) bezeichnet.

In der Baumdarstellung bedeutet die Antwort ”Ja“ (bzw. ”richtig“) eine Verzweigung nach *links* und die Antwort ”Nein“ entsprechend eine Verzweigung nach *rechts*.

Die gleiche Beschreibung eines Baumes kann sich –statt abstrakt auf den Raum \mathcal{X} , den Wertebereich des Prädiktorvektors X – auch konkret auf die *Fälle einer Lernstichprobe* beziehen, die zur Konstruktion eines Baumes benutzt wird. Zu jedem *Knoten* gehören genau diejenigen *Fälle*, deren Prädiktorwerte x_1, x_2, \dots, x_p bei den einzelnen Fragen diejenigen Antworten erzeugen, die in diesen Knoten führen .

In dem folgenden Beispiel (Abb. 1) sind in den einzelnen Knoten des Baumes noch jeweils die Anzahl der Fälle aus der (Lern-)Stichprobe angegeben, die diese enthalten.

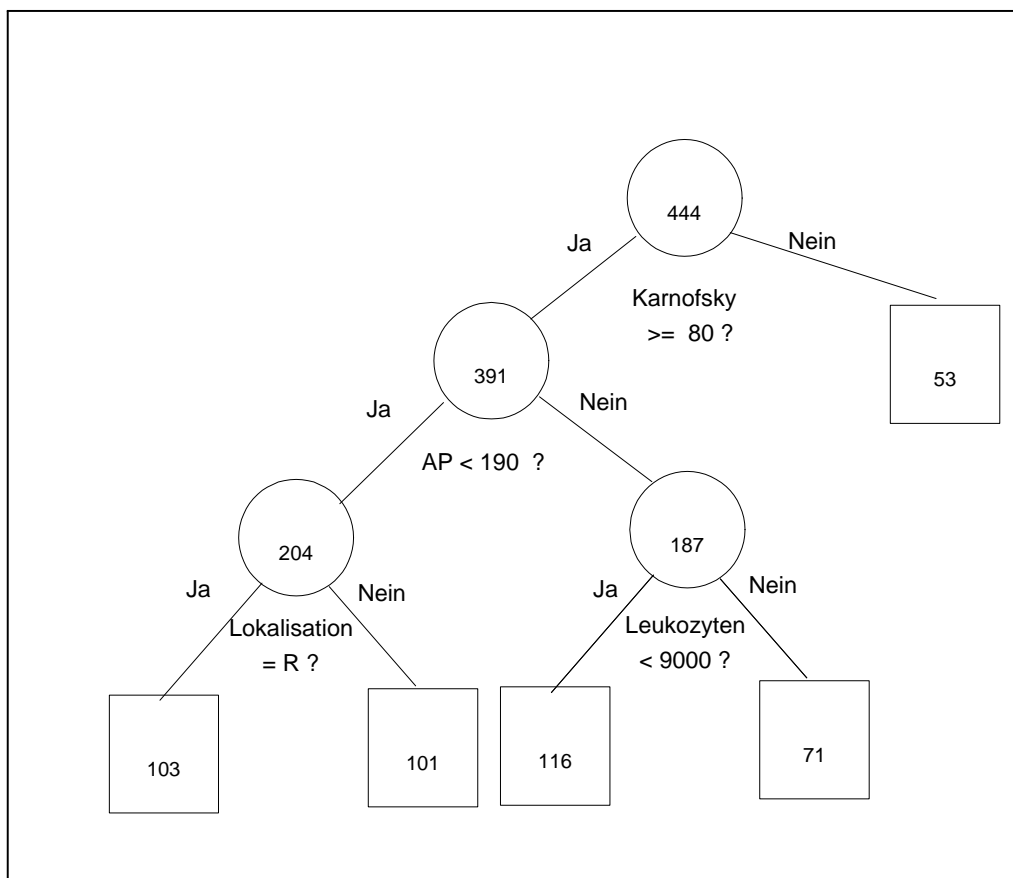


Abb. 1

2.3 Näheres zu den ”Split Defining Statements”.

Ist eine Prädiktorvariable X_j *ordinal*, so werden mit ihrer Hilfe ”Split Defining Statements” (SDS) der Form

$$X_j \leq x \quad (-\infty < x < \infty)$$

erzeugt. Da bei RECPAM maximal 20 Kategorien je Variable zugelassen sind, gibt es daher zu jeder ordinalen Prädiktorvariablen maximal 19 SDS’s, die zu einem Split führen können. (CART kennt diese Begrenzung nicht; die Anzahl der SDS’s je Variable ist dort aber durch die Anzahl der

Fälle in der Lernstichprobe begrenzt: Es gibt maximal $n-1$ *verschiedene* Aufteilungen der Lernstichprobe mit n Fällen aufgrund des Trennwertes einer ordinalen Prädiktorvariablen.)

Bei *qualitativen* Prädiktorvariablen X_j mit den Kategorien l_1, l_2, \dots, l_{m_j} lauten die SDS :

$$X_j \in A_j, \text{ wobei } A_j \text{ irgendeine Teilmenge aus } \{l_1, l_2, \dots, l_{m_j}\}.$$

In beiden genannten Fällen handelt es sich um *einfache SDS*, da sie durch nur *eine Prädiktorvariable* erzeugt werden. Unter der Bezeichnung *RP* (= Recursive Partition) werden bei RECPAM nur solche *einfachen* Splits zugelassen. Mit *RPB* werden dagegen auch *Boolesche Kombinationen* solcher einfachen Splits ermöglicht (dies ist eine Option im Programm RECPAM). Dabei werden die einfachen Statements mit "und" verbunden. Die Prädiktorvariablen, auf die sich die einfachen Statements beziehen, müssen dabei nicht vom selben Typ (ordinal bzw. qualitativ) sein. Beispielsweise ist das Statement "Alter ≤ 60 Jahre, Stadium III a" eine solche zugelassene Boolesche Kombination.

3 Minimale Knoten

Die weitere Aufteilung des Knotens eines Baumes darf nicht erfolgen, wenn einer oder beide der dabei entstehenden Knoten "zu klein" wird. Ein Knoten gilt dabei als "zu klein", wenn in der Lernstichprobe die *Anzahl der Fälle* in diesem Knoten einen bestimmten *Mindestwert unterschreitet*. Dieser Wert kann vom Anwender selber definiert werden.

Handelt es sich um die Analyse von *Überlebensdaten*, so ist die Anzahl der *nicht zensierten Fälle* (also beispielsweise die Anzahl der Patienten mit Rezidiv, wenn es das rezidivfreie Überleben untersucht wird) in einem Knoten maßgeblich dafür, ob dieser in dem Baum *zugelassen* ist. Die Mindestanzahl hierfür kann ebenfalls vom Anwender definiert werden.

(RECPAM verfährt hier anders als CART: Dort ist die Anzahl der Fälle in einem Knoten maßgebend dafür ob der Knoten *weiter gesplittet werden darf*. Daher ist es bei CART durchaus möglich, daß von einem großen Knoten im nächsten Split sehr kleine Knoten mit z.B. nur zwei oder drei Fällen abgespalten werden.)

4 Kriterien für die Auswahl von Splits.

4.1 Ein globales Diskrepanzmaß

Das Ziel der Baumanalyse besteht ja darin, Untergruppen (nach den bisherigen Annahmen speziell: "Knoten") zu finden, die bezüglich der Verteilung der Zielvariablen Y möglichst *homogen* sind, sich aber *untereinander* möglichst stark unterscheiden. Sowohl für die Auswahl von Splits in den einzelnen Knoten als auch für die gesamte Bewertung eines Baumes in Hinblick auf das genannte Ziel ist es daher notwendig, ein Kriterium zu definieren, mit dem die *Homogenität innerhalb* der Knoten im Vergleich zu den *Diskrepanzen zwischen* den einzelnen Knoten (jeweils in Bezug auf die Verteilung der Zielvariablen) gemessen wird.

RECPAM benutzt hierzu den Likelihood-Ratio-Quotienten: Es seien C_1, C_2, \dots, C_k die Endknoten eines Baumes. Ist die Verteilung von Y durch einen (mehrdimensionalen) Parameter γ charakterisiert, so bezeichne $\hat{\gamma}$ die Schätzung von γ unter der Nullhypothese, daß die Verteilung von Y in

allen Endknoten identisch ist, und $\hat{\gamma}_j$ die Schätzung von γ im Knoten C_j unter der Alternative, daß die Verteilungen in allen Knoten unterschiedlich sind. L_i bezeichne den Beitrag des Falles i zur Likelihood, und $K(i)$ den Endknoten, in dem der Fall i ist. Die Likelihood-Ratio-Statistik ρ für den Test "Alle Endknoten C_j haben dieselbe Verteilung" versus "Die Verteilungen sind verschieden" ist dann

$$\rho = 2 \log \frac{\prod_{i=1}^N L_i(\hat{\gamma}_{K(i)})}{\prod_{i=1}^N L_i(\hat{\gamma})} .$$

Diese Likelihood-Ratio-Statistik (*LRS*) dient zunächst als globales Diskrepanzmaß für den Baum mit den gegebenen Endknoten. Es wird später bei der Auswahl eines gesamten Baumes noch modifiziert, da es *zu weit* gehende Verzweigungen begünstigt, die eine "Überanpassung" an die Daten der Lernstichprobe bedeuten.

4.2 Das Split-Kriterium

Die Likelihood-Ratio-Statistik ist asymptotisch chiquadrat-verteilt mit $df = k-1$ Freiheitsgraden (k =Anzahl der Knoten). Es ist daher naheliegend, bei der Entscheidung darüber, *ob* bzw. *wie* ein Knoten aufgespalten werden soll, die *LRS* bzw. den P-Wert des entsprechenden Signifikanztests als Kriterium zu benutzen:

Unter allen *zugelassenen* Splits eines Knotens wird derjenige gewählt, bei dem der zugehörige Signifikanztest den kleinsten P-Wert liefert. Ist dieser größer als ein vorher festgelegtes Niveau α , so wird die Verzweigung nicht durchgeführt.

Will man die dabei berechneten P-Werte nicht nur als nominelle Werte interpretieren und stattdessen Wahrscheinlichkeitsaussagen ermöglichen, kann man in drei verschiedenen Versionen eine *Bonferroni-Korrektur* durchführen: Das vorgegebene Niveau α wird

1. durch die *Anzahl der Prädiktorvariablen*,
2. durch die *Anzahl der Tests im jeweiligen Knoten*, oder
3. durch die *Anzahl der bis dahin insgesamt durchgeführten Vergleiche*

dividiert.

4.3 Die verschiedenen Spezifikationen des Distanzmaßes

Für die Anwendung der oben beschriebenen Kriterien ist es nötig, in Abhängigkeit vom Variablentyp der Zielvariablen und den Annahmen über deren Verteilung die *LRS* bzw. den Signifikanztest zu spezifizieren. Dementsprechen werden in RECPAM folgende "Distanzmaße" zur Auswahl angeboten:

1. *Exponential LRS*. Hier wird also angenommen, daß die Zielvariable Y möglicherweise zensierte Beobachtungen enthält, und daß Y exponentiell verteilt ist.
2. *Logrank statistic*. Im Gegensatz zu 1. werden hier keine Annahmen über die Verteilung von Y gemacht.

3. *Kolmogorov-Smirnov-Statistic*. Hier handelt es sich um die *modifizierte* Version des Kolmogorov-Smirnov-Tests, bei der auch (rechts-)zensierte Beobachtungen zugelassen sind (FLEMING et al. Biometrics 36 (1980), 607-625.)
4. *Wilcoxon-Statistic*. Das ist der Wilcoxon-Gehan-Test für zensierte Beobachtungen.
5. *Chain binomial model LRS*. Die Zeitachse für die (rechtszensierte) Variable Y wird in Intervalle eingeteilt, die nach den Perzentilen der gemeinsamen Verteilungsfunktion gebildet werden. In jedem Zeitintervall wird dann fallweise das Ereignis "Tod eingetreten" mit den Ergebnissen "Ja" und "Nein" betrachtet und die entsprechende Likelihood dazu berechnet.
6. *Cox Proportional Hazard with covariates*. Bei dieser Version werden *vordefinierte* Kovariablen grundsätzlich in die Analyse einbezogen. Diese sind von den *Prädiktorvariablen*, die zu den Verzweigungen des Baumes führen sollen, zu unterscheiden. Der Signifikanztest, der über die Verzweigungen entscheidet, testet dann, ob die Indikatorvariable, die zwischen "Ja" und "Nein" eines Statements über eine odere mehrere Prädiktorvariablen unterscheidet, im (linearen) Cox-Modell zusätzlich zu den Kovariablen noch einen "signifikanten" Beitrag liefert.
7. *Multinomial LRS*. Die Zielgröße Y ist hier qualitativ und als Test im zugehörigen 2xK-Felder-Test wird der Likelihood-Quotienten-Test benutzt.
8. *Cox Proportional Hazard without covariates*. Wie 6, aber ohne Kovariablen.

4.4 Die gesamte Baum-Erzeugung

Die Erzeugung eines Baumes mit Hilfe einer Lernstichprobe kann nun zusammenhängend beschrieben werden:

4.4.1 Die Vorbereitung:

1. Die Wahl des *Distanzmaßes* (3.3) ist festzulegen.
2. Die Menge der *Split Defining Statements* wird angegeben. Man entscheidet also, ob nur *einfache statements* (RP) oder auch *Boolesche Kombinationen* (RPB) zugelassen sind.
3. Die *Mindestgröße* der Knoten (in die gesplittet wird), ist zu definieren: Bei nicht-zensierten Daten die Anzahl der Fälle, bei zensierten Daten die Anzahl der *nicht-zensierten* Fälle.
4. Das Niveau α ist anzugeben, welches unterschritten werden muß, damit eine in Frage stehende Verzweigung stattfinden kann. Dabei ist auch festzulegen, *ob* und ggf. *welche* Art der Bonferroni-Korrektur durchzuführen ist.

4.4.2 Der Algorithmus:

1. Ausgangspunkt ist die Gesamtheit aller Fälle der Lernstichprobe: die *Wurzel* des Baumes.
2. In *jedem Schritt* der Prozedur wird *jeder Knoten* des (gegenwärtigen) Baumes mit *allen zugelassenen Split Defining Statements* belegt.
3. Jede dabei erzeugte Verzweigung wird daraufhin überprüft, ob sie *zugelassen* ist, d.h. ob die entstehenden Knoten die vorgegebene Mindestgröße haben, und ob der zugehörige Signifikanztest einen P-Wert liefert, der das vorgegebene Minimum unterschreitet.

4. Unter allen danach zugelassenen Verzweigungen wird diejenige (und der entsprechende zugehörige Knoten) ausgewählt, die in dem durchgeführten Signifikanztest den kleinsten P-Wert liefert.
5. Das Verfahren endet, wenn es in dem gesamten Baum keine zugelassene Verzweigung mehr gibt.

4.4.3 Behandlung fehlender Werte.

Jede Beobachtungseinheit geht in die Analyse ein, auch wenn sie nicht für alle Prädiktorvariablen gültige Werte hat. Wird eine Verzweigung aufgrund einer Prädiktorvariablen X_j durchgeführt, so wird bei einer Beobachtungseinheit, bei der dieser Wert fehlt, nach derjenigen Variablen X_k gesucht, welche die betrachtete Verzweigung *am besten "imitiert"* (näheres bei CART). Diese Ersatzvariablen werden *Surrogate* genannt und auf Wunsch mit ausgedruckt.

5 4. Kreuzvalidierung und AIC

5.1 Kreuzvalidierung

Ein Verfahren zur Modellsuche wird "schlecht", wenn es sich "zu sehr" an den Daten der Lernstichprobe orientiert. Nimmt man die oben beschriebene Likelihood-Ratio-Statistik als globales Diskrepanzmaß für den gesamten Baum, so erhält man die größte Diskrepanz, wenn man die Verzweigung so weit wie möglich treibt (im Rahmen der Grenzen, die durch die Knotengröße und durch α vorgegeben sind): Die LRS wird mit jedem Schritt größer.

Eine realistischere Einschätzung der Diskrepanz erhält man durch eine besondere Version der Kreuzvalidierung. Sie beruht darauf, daß in der Likelihoodfunktion

$$\rho = 2 \log \frac{\prod_{i=1}^N L_i(\hat{\gamma}_K(i))}{\prod_{i=1}^N L_i(\hat{\gamma})}$$

der Beitrag jeder Beobachtungseinheit i dadurch ersetzt wird, daß die Parameter $\hat{\gamma}_K(i)$ und $\hat{\gamma}$ jeweils *ohne Berücksichtigung der Beobachtungseinheit i* berechnet werden.

Eine auf dieser Kreuzvalidierung beruhende Modellsuche geht so vor:

1. Lege das minimale Signifikanzniveau α fest, erzeuge -wie oben angegeben- den Baum und berechne durch Kreuzvalidierung die zugehörige *LRS* $\rho^{CV}(\alpha)$.
2. Variiere α und berechne so die Funktion $-\rho^{CV}(\alpha)$.
3. Suche das Minimum dieser Funktion. Wähle den zugehörigen Wert für α und den zugehörigen Baum als optimales Modell.
4. Wähle nach Bedarf einen einfacheren Baum mit kleinerem α nach der "Ellbogenregel": Wähle den Punkt in "Ellbogenentfernung" vom Minimum.

5.2 AIC

Die beschriebene Prozedur der Kreuzvalidierung ist sehr rechenintensiv und im Programm RECPAM nicht realisiert. Stattdessen wird dort das asymptotisch äquivalente "Akaike Information Criterion" verwendet:

$$AIC(T_\alpha) = -\rho_\alpha + 2p_\alpha$$

wobei T den Baum und p die Anzahl der Endknoten des Baumes bezeichnet; der Index α gibt dabei an, daß er unter Vorgabe des Signifikanzniveaus α erzeugt wurde.

Die Auswahl von α und damit des Baumes erfolgt dann in analoger Weise.

6 Zusammenlegen von Endknoten

6.1 AMALGAMATION

Die Konstruktion eines Baumes nach dem angegebenen Verfahren stellt sicher, daß sich jeweils die zwei aus demselben split entstehenden Knoten bezüglich der Verteilung der Zielvariablen Y "deutlich" (nach dem Kriterium "P-Wert $\leq \alpha$ ") voneinander unterscheiden. Das bedeutet aber nicht, daß dies auch für Knoten gilt, die von verschiedenen "Eltern" stammen.

Im oben dargestellten Beispiel (Abb.1) sind die Verteilungen der Zielvariablen Y im zweiten und dritten Knoten der vierten Zeile statistisch nicht unterscheidbar ($P > 0.05$). Sie wurden daher nachträglich zusammengelegt. Insgesamt wird daher die Lernstichprobe statt in 5 nur in 4 Untergruppen eingeteilt (Abb. 2).

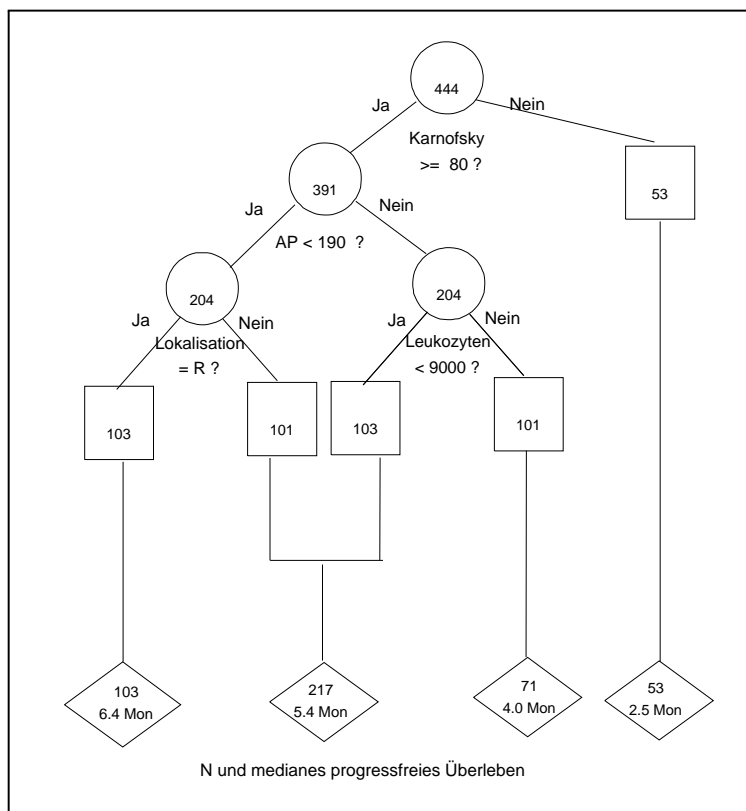


Abb. 2

In den Anwendungen der Baumanalyse sind solche Effekte, wonach sich die Knoten verschiedener "Eltern" wieder ähnlich werden, die Regel: Beim paarweisen Vergleich der Endknoten eines Baumes durch das P-Wert-Kriterium des ausgewählten Tests wird das vorgegebene minimale Signifikanzniveau α häufig überschritten. Wenn das Ziel der Analyse primär darin besteht, bezüglich der Zielvariablen *homogene Teilgruppen* zu finden, die sich *untereinander deutlich unterscheiden*, so ist also die nachfolgende Zusammenlegung von "ähnlichen" Endknoten konsequent. Dies bedeutet allerdings, daß dabei die Annahme aufgegeben wird, daß den Daten eine *Baumstruktur* zugrundeliegt: Der Baumansatz wird nur als *primärer Algorithmus* benutzt, welcher einen möglichst großen Teil der vorhandenen Unterschiede in der Verteilung von Y hervorbringt. Was an überflüssiger Differenzierung dabei mitgeliefert wurde, wird im nachfolgenden Prozeß wieder "glattgebügelt". Dieser Vorgang wird "*Amalgamation*" genannt und besteht darin, daß von den vorher erzeugten Untergruppen sukzessiv jeweils zwei zu einer neuen Untergruppe zusammengelegt werden:

1. Wähle als Ausgangsset von Untergruppen die Endknoten eines Baumes.
2. Lege ein "Signifikanzniveau" α_A für die Amalgamation-Prozedur fest.
3. Berechne in jedem Schritt der Prozedur für jedes Paar der vorhandenen Untergruppen den P-Wert des vorher ausgewählten Tests.
4. Wähle dasjenige Paar, das den *größten* P-Wert liefert.
5. Ist dieser P-Wert größer als α_A , lege die zugehörigen Untergruppen zusammen und gehe zum nächsten Schritt über.
6. Das Verfahren endet, wenn in einem Schritt alle paarweisen Vergleiche von Untergruppen einen P-Wert $\leq \alpha_A$ liefern.

Die beiden Prozeduren des Verfahrens -Recursive Partition (Erzeugung des Baumes) und Amalgamation- können in beliebiger Weise miteinander gekoppelt werden. Beispielsweise kann man durch die Wahl $\alpha = 1.00$ in der ersten Prozedur die Forderung, daß je zwei Knoten einer Verzweigung sich deutlich unterscheiden sollen, ganz aufgeben. Die Amalgamation-Prozedur ist dann der Clusteranalyse vergleichbar, wobei die "Fälle" durch die Untergruppen gebildet werden, als Ähnlichkeitsmaß "*1 - P - Wert*" gewählt wird und die Clusterbildung der Vereinigung von Untergruppen entspricht. Die Charakterisierung der Untergruppen durch die Prädiktorvariablen wird für die Amalgamation-Prozedur nicht mehr benötigt. Daher sind auch beliebige Kombinationen von Endknoten zu "Clustern" möglich. Dieses kann in einzelnen Anwendungen zu schwer interpretierbaren Ergebnissen führen.

6.2 PRUNING

Die Pruning-Prozedur ("Pruning" = "Baumschnitt") verfährt wie das Amalgamation-Programm, berücksichtigt beim Zusammenlagern aber immer nur solche Endknoten, die aus demselben Split stammen. Dies ist nur dann sinnvoll, wenn die P-Wert-Grenze α_P *niedriger* gesetzt wird als in der vorangehenden Rekursiven Partition, da das Pruning sonst keine Veränderung des initialen Baumes erzeugen kann.

7 Literatur

CIAMPI,A.,J.THIFFAULT: *Recursive partition and amalgamation (RECPAM) for censored survival data: criteria for tree selection*. Statistical Software Newsletter 14,2 (1988), 78-81

CIAMPI,A., S.H.HOGG, S.McKINNEY, J.THIFFAULT: *RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features.* Computer Methods and Programs in Biomedicine 26 (1988) 239-256.

CIAMPI,A.,J.THIFFAULT, U.SAGMAN: *RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. II. Applications to data on small cell carcinoma of the lung (SCCL).* Computer Methods and Programs in Biomedicine 30 (1989) 283-296.